Applying GIS Methods to Public Health Research at Harvard University

Applying GIS Methods to Public Health Research at Harvard University

Shortened title: GIS and Public Health at Harvard University

Jeffrey C. Blossom[1], Julia L. Finkelstein[2], Weihe Wendy Guan[3], and Bonnie Burns[4]

Abstract

The Center for Geographic Analysis (CGA) at Harvard University supports research and teaching that relies on geographic information. This includes supporting geographic analysis for public health research at Harvard. This article reviews geographic concepts that apply to public health, pertinent data available in geographic format, and GIS analytical techniques. The workflow methodology the CGA has developed for conducting research with geographic data will be presented, highlighting successful practices to follow and pitfalls to avoid. Applications of this workflow are illustrated through an in-depth discussion of specific case studies in public health research at the University.

Introduction

The Center for Geographic Analysis (CGA) at Harvard University supports research and teaching that relies on geographic information. In this role, CGA provides support for geographic analysis in public health research and other disciplines across the University. Harvard University researchers relying on geographic information are abundant and cover dozens of different

[1] Corresponding author, Harvard University Center for Geographic Analysis. 1737 Cambridge St., Suite 350, Cambridge, MA 02138 USA. jblossom@cga.harvard.edu.

[2] Harvard School of Public Health; Harvard University Center for Geographic Analysis

[3] Harvard University Center for Geographic Analysis

[4] Harvard Map Collection

disciplines. Since its inception in 2006, the CGA has served members of the public health community at Harvard every single month, with several months containing more than 50 active projects for public health or medical related research or tasks. The majority of public health researchers aided by the CGA are from the Harvard School of Public Health and Harvard Medical School. On a less frequent basis the CGA aids public health research for other Harvard entities including the Harvard Humanitarian Initiative (HHI), and Harvard Initiative for Global Health (Guan et al. 2011). The CGA's close collaboration with the Harvard Map Collection (HMC) and Harvard Geospatial Library (HGL) are essential in order to provide all researchers with the best possible service. The HMC holds 400,000 maps, more than 6,000 atlases and thousands of reference books. HGL holds over 6,000 digital data layers that are ready for use in geographic information systems (GIS) (ibid).

Discussed within the context of this existing infrastructure at Harvard, this article will draw on CGA's experience applying geographic analysis to public health research. Geographic concepts and how they can be applied to public health research will be discussed, pertinent data available in geographic format will be reviewed, the full project workflow CGA employs to conduct projects will be articulated, and case studies in public health geography will be presented.

1.0 Applicable Geographic Concepts

Understanding the spatial and temporal distribution of disease and of public health issues is central in conducting public health research. Quantification, visualization, and analysis of information in a geographic context can often reveal trends in data that would otherwise go unnoticed. Applying the fundamental geographic concepts of proximity, travel time, correlation,

and normalization by population and area to public health datasets can also target the source, and reveal transmission patterns of specific health problems. Use of these techniques within a GIS enables the exploration of a broad range of determinants (e.g., demographic, socioeconomic, geographic, environmental) that influence disease risk and transmission and population health.

The successful application of GIS technology toward research or analysis nearly always requires the proper understanding and utilization of different geographic concepts. Having a priori knowledge of specific geographic analysis techniques is essential in order to effectively consult and advise those seeking help. In this section, the geographic concepts of proximity, travel time, normalization by area and population, and correlation will be discussed in terms of how they can be applied to public health research.

1.1 Proximity

Proximity is an integral component of geographic analysis and public health research. It refers to the location of features through measurement of distance between points in a given area. Location and proximity are an important step in linking risk factors and disease outcomes, as well as hypothesis generation to further understand disease etiology.

GIS techniques are ideally suited for public health surveillance and infectious disease control, as transmission of infectious diseases is closely related to geographic proximity. The use of spatial analysis in public health furthers knowledge of disease dynamics; understanding the location of vectors and cases is important for understanding transmission dynamics and mapping public health surveillance. In the fields of environmental health and chronic disease epidemiology, geographic analysis has also been applied to link proximity to exposures and incidence of diseases, such as air pollution and risks of cancers and cardiovascular disease.

Proximity is an important consideration in examining access to health care services and health care utilization, monitoring and evaluation of disease control programs, and program planning and resource allocation. For example, in public health research, distance to health services is an important proxy for access to services and health care utilization, and examining potential sources of disease (vectors), health metric outcomes, and targeted interventions.

1.2 Travel Time

Travel time is a geographic concept that is often applied in health related studies through the use of a GIS. The shortest distance between two points either directly or along a network (such as roads) can be used to derive travel time. Attributes such as length, speed, restrictions on travel direction, and level of congestion can be taken into account when calculating travel time (Longley et al. 2010). Human movement plays a significant role in the transmission of pathogens, spread of infectious diseases and drug resistance, and emergence of novel pathogens (Prothero 1977). Human locations, dispersal patterns, and redistribution areas can be modeled with a GIS. Including travel time in these models can reveal the rate with which an infectious disease can spread. Potentially affected areas can be mapped and classified in terms of when the pathogen or infectious disease will arrive. Travel time can be calculated between a person's place of residence and areas where pathogens originate (such as areas of standing water for mosquito-borne illnesses). This can be used as a factor to identify at risk populations.

Perhaps the most frequent use of travel time in public health research is in evaluating accessibility to health care. Travel distance to health care providers has been recognized as a significant barrier to health care access in the U.S. as far back as 1875 (Guagliardo 2004). Increased travel distance has been associated with decreases in the utilization of mental health

and alcoholic treatment services (Fortney et al. 1995), breast cancer treatment (Meden et al. 2002), and primary care treatment services (Guagliardo 2004). These are a few examples from many that utilize the concept of travel time. Travel time analysis within a modern GIS is currently a common and relatively straightforward operation to perform, enabling its use for many future studies in terms of when the pathogen or infectious disease is likely to arrive.

1.3 Normalization by Area and Population

Health-related research often relies on data summarized by political states, districts, or other areal enumeration units. Total counts of thematic information such as health care facilities, practicing health care professionals, and pharmacies is often summarized and published by governmental or other organizations. Many countries conduct census counts and surveys of population at certain time intervals (for example, the U.S. Census Bureau's decennial census). These census surveys report numerous variables that are helpful for public health research.

To properly interpret this total count census information, the geographic area and population setting must be considered. If the data are reported per areal divisions such as political state, the size of each area must be factored into the dataset to account for the variations in land area of the different states. This adjustment is calculated by dividing the total count values per state by the total area of each state, yielding a value expressed in terms of "count per square mile" or other areal unit. This process, known as "normalizing by area," effectively removes the factor of varying state size, allowing for statistical comparison between states on equal terms[1].

Likewise, "normalizing by population" is often necessary to compare data between urban and rural areas on equal terms. This is accomplished by dividing the total count statistic by the

total population of the reporting area, yielding a "count per 100,000 people" or other population amount. Normalizing by population allows for equal comparison of variables among areas with different population densities.

1.4 Correlation

In epidemiology and biostatistics research, correlation is a statistical procedure to examine the interdependence of two random variables, ranging in value from 0 (no correlation) to -1 (perfect negative correlation) or +1 (perfect positive correlation).

In the context of geographic analysis, spatial correlation reflects the degree to which two or more factors are associated in space. Spatial association or dependence is a fundamental concept in geographic analysis. Through examination of the spatial dependence of data observations, one can explore the associations between variables in a geographic context. Allowing for spatial dependence is essential in the analysis of geographically distributed data.

In public health research, correlations represent an initial analytic step to formulate and examine hypotheses regarding a risk factor and a health outcome. Proximity and correlation are important in examining associations between exposures and outcomes.

In any statistical analysis, a correlation between two geographic features does not demonstrate causation. Rather, this represents an initial point of hypothesis generation for further examination using quantitative geographic analysis and statistical techniques.

2.0 How Geographic Analysis can be Applied to Public Health Research

The application of geographic analytic techniques to understand the distribution of disease and determinants of health is at the core of public health research. During the past two decades,

technological advances have made it possible to examine spatial and temporal trends in large-scale epidemiological data (Rushton 2003; Paolino 2005; Jerrett 2010). The wide variety of GIS applications toward public health research have been documented in several publications including *GIS and Public Health* by Ellen K. Cromley and Sara L. McLafferty (2002), *Public health, GIS, and spatial analytic tools* by G. Rushton (2003), and *GIS—a proven tool for public health analysis*" by R. H. Jenks and J. M. Malecki (2004).

GIS techniques are ideally suited for public health surveillance and disease control, as transmission of infectious diseases is closely related to geographic proximity. This includes case location, identification of clustering, mapping of epidemic dynamics, and mapping disease burden and response. More recently, geographic analytic methods have been applied to the fields of chronic disease epidemiology and environmental health, to link proximity to exposures to incidence of non-communicable diseases. Application of GIS methods to public health research has also helped inform health care delivery and resource allocation (McLafferty 2003; Bazemore 2010; Gobalet 1996).

GIS methods can support the summation of large amounts of data for disease surveillance and health reporting, identify new cases and at-risk populations, stratify risk factors, and quantify risk and transmission patterns. Visualization of complex health data can inform health care policy and resource allocation and targeted interventions. Geographic analysis and examining medical research questions spatially have helped to integrate public health research across disciplines and departments at Harvard University (Guan et al. 2011).

Understanding spatial and temporal distribution of disease is integral to public health research (Rushton 2003; Ricketts 2003). Quantitative and statistical analysis methods within a GIS facilitate exploration of a broad range of determinants, including demographic,

socioeconomic, geographic, and environmental factors that influence disease transmission. This section will discuss how thematic mapping, visualization of spatial and temporal disease, and the usage of quantitative and statistical analysis can be applied to public health research.

2.1 Thematic Mapping

The discipline of cartography recognizes two major types of maps: general reference and thematic. General reference maps such as the U.S. topographic map series produces by the U.S. Geological Survey are intended to portray the locations of different spatial phenomena such as rivers, roads, buildings, etc. Thematic maps are used to emphasize the spatial patterns of one or more specific attributes, such as population density or temperature (Slocum et al. 2009). The use of thematic maps can be particularly helpful when trying to visualize disease, risk, and other health-related issues across space and time. Consider the disease polio: Polio is eradicated in most countries, but in India there are still occurrences (Bandyopadhyay 2010). Creating a thematic map depicting where these cases occur may reveal existing spatial patterns.

FIGURE 1 GOES HERE

The map in Figure 1 clearly indicates a much higher incidence of polio in Uttar Pradesh and Bihar states, along with localized clustering of cases. In addition, symbolizing the polio cases by year of occurrence increases the information revealed by this thematic map. A clear trend of polio cases can be observed originating from a cluster in Bihar and progressing into Uttar Pradesh. This applied use of GIS illustrates the utility of mapping disease and risk in order to better understand distribution and transmission patterns. This is just one example from a vast amount of literature regarding spatial modeling for use in public health surveillance (Sonneson 2003). Once spatial models are applied, the results can be visualized through the proper

application of thematic mapping.

2.2 Quantitative and Statistical Analysis

Three major approaches are commonly used to perform geographic analysis: *geographic information systems, remote sensing,* and *spatial statistics*. In the context of public health research, GIS methods and remote sensing techniques are integral to formulating hypothesis and examining research questions visually. Statistical analytic methods can be used to further explore and quantify the statistical significance of observed trends in location and spatial distribution.

Geographic information system methods are used to capture, store, analyze, manage, and present data that are linked to locations. GIS brings together fields of geography and statistical analysis, and has a wide range of applications in urban planning, land surveying, and geography, and, more recently, has been transformative in the field of public health research. For example, the application of GIS analysis to epidemiological methods and statistical analysis in medical research can be used to examine the statistical significance of clustering and spatial patterns through visualization (patterns), exploration (clustering), and modeling (predictive modeling, spatial diffusion). GIS tools can also be used to inform spatial statistics and statistical analysis. For example, one can use GIS methods to create covariates for inclusion in statistical models, and to visualize the output from statistical models.

Remote sensing is the process of acquiring geographic information through distant recording devices, such as on airplanes or satellites. Remote sensing techniques often represent an efficient and cost-effective method of collecting large amounts of data in regions where it is logistically challenging or unsafe to collect data directly (e.g., conflict zones, or areas of extreme environmental conditions). These techniques have been increasingly utilized in public health

research to map disease vectors (e.g., animal sources of disease transmission), and environmental characteristics that may play an important role in risk and transmission dynamics."

Spatial statistics, the application of geographic analytic methods to statistical analysis, refers to the use of formal statistical techniques to examine the topological, geometric, or geographic properties of features. In public health research, spatial statistics can help to explore and quantify the statistical significance of observed trends in location and spatial distribution. These techniques can be used to examine additional aspects of data which may not be visually apparent (e.g., spatial patterns or trends in data, how different attributes are distributed spatially), investigate the distribution of values, and identify potential outliers and errors in the geographic data.

Spatial statistics provide a method for statistical modeling, hypothesis testing and inference. They aid in forming conclusions regarding geographic data, and they provide a medium to fit and smooth spatial surfaces, as well as to integrate spatial and non-spatial data or attributes (i.e., directly accessed from a layer's feature attribute table, e.g., mean, standard deviation). Statistical techniques can be used to analyze data and examine hypotheses to confirm the orientation and strength of a spatial pattern in the data, and the extent to which the data exhibit grouping in close proximity, i.e. clusters.

Descriptive statistics (e.g., histograms, normal Q-Q plots) can be used for hypothesis generation and to identify and confirm spatial patterns in data, such as the center or direction of geographic features. If specific features form data clusters, descriptive statistics can also be used to examine non-spatial aspects of data (e.g., comparison of data distributions to normal distributions) (Rushton 2003; Chung 2004). *Analytical statistical methods* are used to test hypotheses and develop conclusions regarding the distribution or characterization of the

analyzed data. These statistical techniques include: 1) spatial autocorrelation (e.g., Moran's I and Geary's C), used to measure and analyze the degree of dependency in locations of observations, 2) spatial interpolation methods (e.g., inverse distance weighting, kriging), used to estimate the variables at unobserved locations, based on locations of observed locations, 3) spatial regression (e.g., geographically weighted regression), used to examine spatial dependency in regression models, and 4) inferential statistics, which use probability theory to predict the likely occurrence of values, or the likelihood that any pattern or observed trend is not due to chance alone.

A common research question in public health is the identification of clusters of cases to examine risk factors, transmission dynamics, and inform targeted interventions. To examine the phenomenon of clustering, a variety of statistical analysis functions and statistical modeling techniques are available through the Spatial Statistics, Spatial Analyst, and Geostatistical Analyst extensions in the ArcGIS software. Other statistical methods may be more optimally utilized in combination with other geographic analytic software programs such as GeoDa or SaTScan, or statistical software packages, such as R or S-Plus.

3.0 Review of Geographic Data for Use in Support of Public Health Research

Geographic data that support health research can be categorized as follows:

1) Data about health care capacities, such as health facilities, employment and administration.
2) Data about the population and their health conditions and health care needs.
3) Data about the environment, both natural and social, which affects people's health.
4) Data about transportation and address location, which is essential in understanding accessibility for care, dissemination of medical supply, transmission of infectious

diseases, distribution of patients, and many other relationships among data of the above three categories.

We will examine each of these data categories, introduce the major data sources and popular data formats, and explain the most common usage of such data in health related research.

3.1 Data about Health Care Capacities

The US Department of Health and Human Services Health Resources and Services Administration (HRSA) maintains a Geospatial Data Warehouse, and makes health care related data available for download, including:

1) The Health Centers and Look-alike Sites.
2) Currently-designated Health Professional Shortage Areas.
3) The National Sample Survey of Registered Nurses.
4) The Primary Care Service Areas project (a collaboration between HRSA and Dartmouth College)

Such data are in text or spreadsheet format, including postal address addresses of the entities, and can be geocoded into spatial points data (see section 3.4 below for geocoding).

A summary of local health departments in the US was published by Hua Lu and James Holt in 2009[2]. It was evident that local health services in the US is fragmented, managed at different levels of local government agencies. It has been a challenge to develop a standard nation-wide coverage data set to support effective collaboration.

Many states' GIS agencies or health agencies maintain their own spatial health resource data sets for public use. Examples include the State of Massachusetts[3] and New York[4]. In many states, such geospatial warehouses are often housed at state universities.

For data about health care providers, commercial data vendors are an alternative to public

agencies. The ArcGIS Business Analyst is a licensed package of spatial data and analytical tools. It contains Infogroup Business Locations which represents the locations of over 12 million private and public companies in United States, among them health care facilities. It includes address information, estimated sales or assets, number of employees, franchise/specialty information, and Standard Industrial Classification (SIC) and North American Industry Classification System (NAICS) codes. These codes are 4 digit numbers (standardized by the U.S. Government) that classify industries, and are very useful in selecting out specific business types.

The availability of international health care data varies greatly from country to country, often depending on the national or provincial government agencies' data management practice. Health care provider information is harder to find in areas outside of the U.S. The US National Geospatial-Intelligence Agency (NGA) maintains an international GeoNames service (GNS). It contains designations for Medical Centers, Hospitals, and clinics[5]. The database is updated on a weekly basis, but locations are approximate, and by no means complete for the globe.

3.2 Data about the Population and Health

Population data for health research can be further categorized into three categories. One is the global population distribution and density data, which is aggregated and processed from nation-based census and other sources[6]. The second category is demographic health survey (DHS) data, which are surveys in many countries describing disease outbreak rates and other health indicators of the population, often published with map coordinates of survey clusters[7]. The third category is the more widely-known national census data from many countries such as the U.S.[8], Great Britan[9], and Israel[10] just to mention a few.

Population data are usually derived from interpolation and projection models, which provide continuous coverage for the globe and an estimated value for the time periods when no census was conducted. Data are often organized in tables, graphs or grid formats, which require further processing to match administrative polygon units in a GIS. Five of these global population datasets are listed below:

1) Gridded Population of the World – GPW v3[11]
2) United Nations Population Information Network (POPIN) [12]
3) Home HYDE – the Netherlands Environmental Assessment Agency (PBL) [13]
4) LandScan[6]
5) International Data Base – U.S. Census Bureau[14]

Demographic health survey (DHS) data are plentiful. The World Health Organization (WHO) maintains a searchable database online[15]. It includes data on mortality and health status, diseases, coverage of services, risk factors, health systems, and world health statistics. The WHO Global InfoBase is a data warehouse that collects, stores, and displays information on chronic diseases and their risk factors for all WHO member states. Data is presented for download in graphs, maps and data tables formats.

The Demographic and Health Survey (DHS) program[16] collects, analyzes and disseminates data on population, health, HIV and nutrition through more than 200 surveys in over 75 countries. Data are organized in administrative units and can be mapped to the administrative polygons.

The United Nations Children's Fund (UNICEF) maintains a website which contains statistical information on the well beings women and children around the globe[17]. It also supports the Multiple Indicator Cluster Surveys (MICS), which are a major source of global development data downloadable for users with a login[18]. The data are organized by country and

year, in SPSS format. It requires further processing to match with the country boundaries for global statistical analysis.

In the United States, national level health survey data are available from many sources. The National Center for Health Statistics[19], part of the Center for Disease Control and Prevention (CDC), provides downloadable, tabular data. Some data are mappable by address geocoding or by matching table records to survey geographic units, such as administrative units or metropolitan/micropolitan statistical areas (MMSAs).

Behavioral Risk Factor Surveillance System (BRFSS), also sponsored by the CDC, provides downloadable GIS data[20]. These files contain data and documentation, and are available in Zip Archive File (ZIP) format. The zip files contain BRFSS data that is mapped for both the states and MMSAs. These data files are a subset of the BRFSS data intended for use with a GIS package. Complete data sets and documentation for these data years are available in the BRFSS and SMART sections of the site[21].

CDC's WONDER, Wide-ranging Online Data for Epidemiologic Research, is a menu-driven system that provides access to a wide array of public health information by the CDC[22]. It allows for access to statistical research data published by CDC, as well as reference materials, reports and guidelines on health-related topics. Public-use data sets about mortality (deaths), cancer incidence, HIV and AIDS, tuberculosis, vaccinations, natality (births), census data and many other topics are available for query, and the requested data are summarized and analyzed, with dynamically calculated statistics, charts and maps. The data are ready for use in desktop applications such as word processors, spreadsheet programs, or statistical and geographic analysis packages. File formats available include plain text (ASCII), web pages (HTML), and spreadsheet files (Tab Separated Values).

The Health and Medical Care Archive (HMCA)[23] is the data archive of the Robert Wood Johnson Foundation (RWJF), the largest philanthropy devoted exclusively to health and health care in the United States. Operated by the Inter-university Consortium for Political and Social Research (ICPSR) at the University of Michigan, HMCA preserves and disseminates data collected by selected research projects funded by the Foundation and facilitates secondary analyses of the data. The HMCA includes health care provider locations, household survey data, community disease information, and many more datasets used in numerous studies.

Local scale health surveys in the United States may be found at city or county government websites, such as the New York City Community Health Survey[24], and the Los Angeles County Health Assessment[25]. Some other countries publish their health survey results online as well. Examples are the China Health and Nutrition Survey[26] and the India National Family Health Survey[27].

Census/demographic data are usually published by the respective countries' census organizations, as well as by commercial data vendors, who process the raw census data from the government, and add value by applying statistical, geographic, or other analyses to produce derived demographic data that is more convenient for end users. In the United States, census data can be obtained from the US Census Bureau FactFinder[28], and from a number of publicly accessible or subscription-based data providers, such as Geolytics[29], Social Explorer[30] , ICPSR[31] , and ESRI Community Analyst.[32]

Many other countries offer their census data through a fee-based service online or upon request. Examples include the China census data[33] and the India census data[34]. Most census data have its own geographic base units, either the same as local administration units (such as counties, districts and townships in China), or a uniquely defined system (such as tracks, block

groups and blocks in the U.S.). The census data are usually organized in tabular format, which need to be matched to the geographic units for mapping or spatial analysis.

3.3 Data about the Environment

Many environmental conditions impact people's health. For example, the distribution of wetlands may affect the dispersion of malaria, while groundwater aquifer and the location of superfund sites may impact drinking water quality, which in turn affect residents' health. Because of the huge variety of data subjects that could belong to this category, we will not categorize data further by its content here, but by the major source format instead, which includes remotely sensed image data, geo-referenced survey data, and statistical data.

The US Geological Survey (USGS) maintains a map-centric data discovery tool Earth Explorer[35], which provides public image data of different scales, from global satellite images, national aerial photos, world-wide digital elevation models, land cover classifications, and weather monitoring images.

The commercial image data vendor, GeoEye, maintains an online image search tool GeoFUSE[36], which allows users to find images they need, and to examine the cloud cover or other quality issues, before submitting a purchase request. The Food and Agriculture Organization of the United Nations (FAO) maintains several statistical and spatial databases on agriculture, nutrition, fisheries, forestry, food aid, land use and population[37].

The US Environmental Protection Agency (EPA) maintains many public databases and tools for analysis environmental data[38]. Its Geospatial Data Access Project provides downloadable files of environmental facilities or sites in various GIS formats, including extensible markup language (XML) file, keyhole markup language (KML) file, ESRI Shapefile

and ESRI Feature Class[39]. The EnviroMapper is an online mapping system for browsing environmental quality data interactively[40].

The ESRI Data and Maps and ArcGIS Business Analyst packages offer licensed users access to US parks and recreational land use data, as well as food and beverage business locations, and retail locations of tobacco and alcoholic beverages, etc.

More specific local environmental data are often collected first-hand by researchers equipped with GPS receivers on the streets or in the fields.

3.4 Data about Transportation and Address Location

Accurate and detailed transportation networks are critical for two purposes. One is for traffic routing, which provides information on best route and travel time between any two points. The other is for address geocoding, which is the process of converting postal addresses into longitude and latitude coordinates. Data sets used for traffic routing and for address geocoding may look identical when displayed on a map – both appear to be linear networks. However they require different preparation and contain different attributes on the street or road segments.

The most critical property of a traffic routing network dataset is connectivity among street/road segments. Other attributes in support of traffic routine include speed limits, one way directions, turn restrictions, and vehicle height or weight limits. On the other hand, street network in support of address geocoding require each street segment to contain street names, house number ranges, as well as other postal address components such as city and state names, or zip codes. The geocoding software looks for matches between the input and reference data, and when matches are found the corresponding location is interpolated from the street segment whose value range contains the matching address number. Many of the commercial street

network datasets are prepared for both traffic routing and address geocoding functions. The two combined allow for instant routing between a pair of user input addresses.

It is worth noting that both traffic routing and address geocoding datasets are readily available for North America and Europe, but not for many other parts of the world. Without a street level geocoder, addresses may be matched to zip code zones, town or city centers, or other geographic features. The locations obtained from such matching process is less precise than that of address geocoding, but nevertheless useful for some research purposes.

Information about foreign geographic feature names can be obtained from the GEOnet Names Server (GNS), developed and maintained by the National Geospatial-Intelligence Agency (NGA). The GNS database is the official repository of foreign place-name decisions approved by the U.S. Board on Geographic Names[41].

The Geographic Names Information System (GNIS) is the U.S. federal and national standard for geographic nomenclature[42], containing information about physical and cultural geographic features of all types in the United States, associated areas, and Antarctica. It includes current and historical place names, but not roads and highways. GNIS was developed by the USGS in support of the U.S. Board on Geographic Names as the official repository of domestic geographic names. Data was collected through a broad program of partnerships with Federal, State, and local government agencies and other authorized contributors. “Hospital” is a category among the GNIS data types[43].

Both address geocoding (using a street network), and location identification (matching place names with entries) in the GNIS or GNS, achieve the purpose of mapping tabular records or descriptive data to a geographic location. Given that the vast majority of traditional health data reside in tabular files and other non-GIS file formats, many of which containing addresses - such

as patient records, clinic records, or hospital records - geocoding is an essential feature. Once turned into locations, they can be studied through spatial analysis, revealing new patterns, relationships, trends, and meanings. The potential to identify address locations of patients raises a privacy issue that must be considered when using geocoded patient or other sensitive data. To preserve patient privacy, disassociating patient information such as names, social security numbers, health care system or other individual identifiers from the geocoded address locations is necessary.

4.0 Workflow for Conducting Public Health Research with Geographic Data

Since its inception, the CGA has provided services in support of hundreds of Harvard research projects that have involved geographic information. To handle efficiently this large volume of services, a well-defined work flow has been developed and is used by CGA personnel. This workflow involves several stages from initial consultation to final product creation, all of which are outlined below. When providing services for health researchers, there are specific considerations common to most clients that must be factored into the workflow cycle. This cycle involves the following general stages: initial consultation, project execution, and final delivery.

4.1 Initial Consultation

The initial consultation between a CGA staff member and a researcher who is interested in using geographic information usually originates from an email or at the CGA Help Desk. The CGA general contact email address (contact@help.cga.harvard.edu) is prominently displayed on the CGA website[44]. Emails sent to this address are automatically forwarded to the inboxes of several CGA staff members to ensure a response to all queries within 1-2 business days. In

addition, emails sent to this address are automatically logged into a request tracking (RT) ticketing system[45]. Each request ticket is assigned to the appropriate CGA member for resolution.

The CGA conducts a Help Desk every Tuesday afternoon both at the Harvard main campus in Cambridge and at the Harvard Medical campus in the Longwood Medical Area in Boston. During this time at least one CGA Specialist is available at a computer lab at each location. These labs have high powered desktop computers each with a variety of GIS software installed. This is an opportunity for any Harvard researcher to receive in person consultation, hands on troubleshooting assistance, or GIS data and software demonstrations from a CGA Specialist. Appointments can be scheduled, or people can "drop in" to receive help. Each Help Desk consultation or help session administered is logged into the CGA RT ticketing system.

Whether initiated through the contact email or Help Desk, the nature of requests may vary, from a quick two-minute answer to a multi-year project involving multiple stages, deliverables, and personnel requirements. The requesting client's knowledge regarding the use of geographic information and GIS may also range from someone having just heard about using GIS to an experienced GIS practitioner. This wide range of project complexity and client understanding requires a flexible, customized approach to handling each request.

During the initial consultation, the scope of work must be determined as soon as possible. This may be defined in the initial request, such as "How to convert a table of longitude, latitude coordinates into a shape file for use on a map?" or "Is there a GIS dataset available that contains hospital locations for Florida?" These clear procedural or information requests can usually be handled in a single reply; for example, sending the client a tutorial to resolve the former request, and extracting the requested data in the latter request. Typically, however, the scope of work

determination is less straightforward.

To define the scope of work properly, the project objective must first be determined. The researcher usually has a very clear objective, for example "I want a map that displays one mile buffers around hospitals in Chicago, fast food restaurants within these buffers, and a table listing the number of fast food restaurants within one mile of each hospital". Once the objective is established, then a scope of work required to achieve the objective can be developed. In defining the scope of work the following must be taken into consideration: 1) Datasets required to make the map or perform the analysis, 2) Procedures and methodologies necessary to perform the work, 3) Tools required to execute these procedures, and 4) Deliverables needed and type of each deliverable (i.e. a web map, statistical table, geodatabase, etc.). The scope of work is documented in written form such that each component of the scope is clearly communicated.

4.2 Project Execution

Once the scope of work is determined, there are several ways to execute the project. The CGA's first approach is to equip the researcher to perform the work themselves. Harvard University has several proprietary GIS software programs licensed on a site-wide level, and supports the use of many free and open source software programs, as well. Every semester the CGA conducts 6 different free, instructor-led training sessions both at the Cambridge and Boston campuses. These are intended for the student or researcher who knows little about GIS but wants to learn the basics to apply to their coursework or research. More intensive two-week "GIS Institutes" are held each summer and winter, geared at educating the graduate student level researcher. The Institute consists of lectures, lab exercises, discussions, facility tours, and culminates with each participant presenting a GIS project featuring individual work. There are

also many self-help tutorials and "how-to" documents available on the CGA website to help enable any Harvard personnel to use GIS.

Often it is not feasible for the researcher to perform the work. In these cases the CGA will schedule the work into their queue of service projects. For such projects first time clients receive 4 hours of CGA time for free, and then are charged $75 for each additional hour of work required. Using this consulting model has enables the CGA to prioritize projects in order to meet project specific deadline schedules. A project specification document is filled out for these projects, where the scope of work, project methodology, budget, timeline, and deliverables are documented and agreed upon by both parties (Figure 2).

Figure 2 – CGA Project Specification Document.

The first stage of project execution involves gathering or creating the required datasets. As outlined above in section three, there are many datasets that are readily available for public health research, and various methods to geocode tabular data that may be necessary for use. The appropriate datasets are thusly acquired as needed for the specific area of interest.

Once all datasets are acquired or created, the necessary methodology is applied to perform the GIS analysis. The methodology required to complete a project could use any number of GIS technologies and procedures. An overview of the general project methodology used in completing the major groupings of requests CGA receives will be discussed below. Detailed methodology for several specific projects will be presented further on in the article. The types of service projects CGA provides can usually be grouped into one of the following major categories: 1) Geocoding and census variable extraction, 2) Map creation for print publications, 3) Dynamic web map creation, and 4) GIS analysis and visualization.

4.2.1 Geocoding and Census Variable Extraction

Many researchers desire specific census demographic characteristics regarding their data for use in statistical regression modeling or analysis. If a researcher knows the level of geography and geographic area they need census information for (for example, all census tracts for the state of California) the CGA may provide an assisting role in pointing the researcher toward one of several methods to obtain demographic data for their area of interest, or perform the data extraction ourselves. Care is taken in this process to ensure the proper unique identifiers exist in attribute fields that are common to both datasets so that a table join can be performed. If a researcher needs to determine which census geographical unit their data are in, then CGA will first geocode the dataset using one of the methods listed in section three above. Then a spatial join using ArcGIS software will be performed between the geocoded data and the desired level of census geography. An ArcGIS spatial join evaluates the geographic location of every feature in an input dataset (the geocoded data) against the join dataset (census GIS data) and appends the attributes of both datasets together into a new dataset As a final step in this process, the data are often exported to comma separated value (.csv) format, for import into a statistical analysis software program.

4.2.2 Map Creation for Print Publications

Map creation for book, journal, dissertation, thesis, and other publications is performed nearly on a daily basis at CGA. Both general reference and thematic maps are made, depending on a researcher's request. The application of best practice cartographic principles regarding the map layout, scale, projection, color, symbology, and type are used on these maps. Often

explaining and/or demonstrating different variations of these principles with a researcher's data in necessary in order to produce the desired map. The audience for the map is taken into consideration, as is the final output media type. Printed maps are published in sizes ranging from 2" x 2" to 42" x 60". The CGA has a 42" wide plotter on which large format maps or posters can be plotted. In addition, maps are made in a wide variety of image formats and resolutions for use in presentations and publication on websites. Technology used to create the maps usually includes use of one or more GIS software programs (predominantly ArcGIS), and potentially image processing or graphic design software.

4.2.3 Dynamic Web Map Creation

The ability to customize a variety of web map application program interfaces (API) has created an increasing trend in dynamic web map requests with various levels of functionality. The CGA publishes interactive web maps using a variety of API including OpenLayers and Google Maps. Researchers readily recognize the utility of an interactive web map embedded with their own data. This allows for entire research teams to view and interact with the same custom maps from anywhere in the world, fueling thought and collaboration. Technology used to create interactive web maps is driven by user requests. The nature of the request (whether the researcher wants a handful of points with information window popup functionality, or a web map with complex symbolization capabilities and full database interaction) plays a large role in determining which software stacks are used complete the request.

Enabling researchers to create their own web maps is a major focus at the CGA. This has resulted in the CGA building a series of Google Map mashup tutorials[46], and creation of the WorldMap platform[47]. The WorldMap platform is an open source framework currently under

development by the CGA designed for viewing and interpreting maps collaboratively. One can load data and maps from various sources in multiple formats, and can have flexibility in choosing how to create, share, and mix different datasets and maps together.

4.2.4 GIS Analysis and Visualization

The CGA aids public health researchers with many forms of GIS analysis. A critical first step in the workflow methodology of every GIS analysis project is recognizing what projected coordinate system is best to use for the type of analysis to be performed. Data layers are then projected into the appropriate coordinate system. Some of the more common GIS analyses applied for public health researchers are the creation of straight line and network buffers. Buffer creation enables proximity analysis, which, as described above, is a heavily applied spatial analysis method in public health research. Performing overlay and intersect between two map layers (for example, intersecting land cover and village regions to produce land cover type percentage values per village) is another common type of analysis desired by public health researchers at Harvard. Interpolation techniques such as inverse distance weighting, kriging, point density and others are applied to datasets if needed. The CGA provides geostatistical analysis consulting as needed, and calls upon a network of professors and researchers at Harvard for referral of specific questions. For many GIS analysis projects, the methodology is such that batch processing, macros, or software programming is required to best complete the job. Different programming languages such as Python, JAVA, .NET, and PHP are used when necessary. If specific programming skills required to complete the work do not exist within CGA, sub-contactors are hired to write code. Animated PowerPoint™ slides, .gif images, and video files in all sorts of formats are produced when the temporal nature of a geographic dataset

needs to be visualized.

4.3 Final Project Delivery

After the project execution stage concludes, final product delivery occurs. The end products to be delivered are always documented in the project specification document, which is updated and communicated to the client if the scope of work changes during the project execution stage. Project results may be tabular or GIS datasets, printed hard copy maps of various sizes, static map images for inclusion into print or web publications, or interactive websites. For tabular or GIS dataset delivery, these data are sent to the researcher through email or secure file transfer. A field key explaining what the data attributes or column names are is always included. Often a document listing the input data used and procedures employed to produce the data is also included. This usually completes the project workflow for tabular or GIS data set delivery. For map, image, or website delivery there is nearly always a revision process involved, and CGA factors this into the project speci+fication. A draft map will be sent to the client for review and comment, CGA will revise the map based on the comments, and republish. This process is repeated until the researcher is satisfied with the map.

Once a dataset or map is deemed acceptable, the CGA Specialist provides necessary metadata for the product, and documents any processing procedures. This procedures document is saved into the project folder on CGA's file system, and is passed on to the researcher if requested. All project work is saved into individual project folders using a standard naming convention that includes the year the project was performed, project number, and client name. Subfolders within each project folder also have standardized names, so any CGA Specialist is able to access and understand any project data if the need arises in the future. Project data are

saved on a Netapp[48] file storage system. After two years of inactivity, project data are archived on flash memory storage media to free up space on the Netapp. In delivering webmaps hosted on the CGA web map servers, websites are hosted for clients for one year. Subsequently, the researcher may opt to ensure permanence of a custom website for a yearly fee. Once the product, metadata, and documentation are delivered to the client's satisfaction, an invoice for the work is sent. Datasets produced that may be of use to others can be published into HGL with the client's permission.

5.0 Current GIS Research in Public Health at Harvard University

Many research projects are currently underway at Harvard that involve geographic analysis in the field of public health. Five of these projects will be highlighted in this section.

5.1 The Nurses' Health Study

The Nurses' Health Studies are among the most significant and longest-running epidemiological studies of women's health, established at the Department of Nutrition of the Harvard School of Public Health. Since 1976, investigators have followed over 238,000 registered nurses to examine risk factors for major non-communicable diseases. This study brings together medical and public health researchers from the Harvard School of Public Health, Harvard Medical School, Brigham and Women's Hospital, Dana Farber Cancer Institute, Boston Children's Hospital, and Beth Israel Deaconess Medical Center. The Nurses' Health Study, established in 1976 by Dr. Frank Speizer, and the Nurses' Health Study II, established in 1989 by Dr. Walter Willett, are landmark epidemiological studies on women's health and non-communicable diseases.

GIS regression approaches and geocoded residential addresses are being used to estimate

different exposures, and investigate associations between health outcomes and the environment. The following sections will highlight some of the major analyses being conducted in this study.

5.1.1 Spatio-temporal Estimation of Particulate Matter Exposure

Public health researchers have been investigating the health effects of particulate matter (PM) air pollution effects as one of the components of the Nurses' Health Study. Yanosky, Paciorek et al. (2008) used GIS techniques to analyze data from the Nurses' Health Study; investigators used semi-empirical models to predict spatially and temporally resolved long-term average outdoor concentrations of particulate matter to successfully predict chronic fine and coarse particulate exposures for the Northeastern and Midwestern United States (Yanosky et al. 2009).

Researchers used GIS-based spatial smoothing model to predict monthly outdoor PM10 concentrations (Yanosky et al. 2008), which included monthly smooth spatial terms and smooth regression terms of GIS-derived and meteorological predictors. Final model performance was strong (cross-validation R^2=0.62), with little bias (−0.4 μg m−3) and high precision (6.4 μg m−3). The final model performed better than a model with seasonal spatial terms (cross-validation R^2=0.54), and performed well in both urban and rural areas and across seasons and years. The addition of GIS-derived and meteorological predictors improved predictive performance over spatial smoothing (cross-validation R^2=0.51) or inverse distance weighted interpolation (cross-validation R^2=0.29) methods alone, and increased the spatial resolution of predictions. The strong model performance demonstrated the suitability of these GIS-based spatial smoothing methods to estimate individual-specific chronic PM10 exposures for large populations.

5.1.2 Spatio-temporal Estimation of Particulate Matter Exposure and its Effects on Cardiovascular Disease

Researchers have applied GIS methods to investigate the associations between environmental risk factors and health outcomes, using data from the Nurses' Health Study. In an analysis conducted by Puett *et al.* (2009), investigators examined the association of chronic particulate exposures with incident nonfatal myocardial infarction, fatal coronary heart disease (CHD), and all-cause mortality, in a prospective cohort of 66,250 women from the Nurses' Health Study in northeastern United States (Puett 2009). Researchers developed a spatio-temporal model to estimate monthly PM10 and PM2.5 exposure between 1988 and 2002, using government monitoring data, geocoded residential addresses, and covariates calculated using GIS overlay analysis between addresses locations and census tract boundaries . Multivariate models included hypertension, family history of MI, hypercholesterolemia, body mass index (BMI, continuous), physical activity (< 3, 3 to < 9, 9 to < 18, 18 to < 27, or ≥ 27 metabolic equivalent (MET) hr per week), smoking status (never, former, or current), diabetes, median house value, and household income for census tract of residence, season, and state of residence and were stratified by age in months. GIS techniques were used to account for uncertainty and truncation in available data sources. In addition to the primary exposure, defined as the average exposure to $PM_{2.5}$ and $PM_{10–2.5}$ in the 12 months before the outcome of interest, other windows of exposure were considered, namely 1, 3, 24, 36, and 48 months prior to the exposure; $PM_{2.5}$ and $PM_{10–2.5}$ were also assessed in separate single- and two-pollutant models. All multivariate models were also stratified by age in months and adjusted for state of residence (indicator variables), year (linear term), and season (indicator variables), in order to adjust for large-scale

spatial mortality patterns which might be related to factors apart from pollution. Authors found increased risk of all-cause mortality [hazard ratio (HR), 1.26; 95% confidence interval (CI), 1.02-1.54] and fatal CHD (HR = 2.02; 95% CI, 1.07-3.78) associated with each 10-microg/m(3) increase in annual PM(2.5) exposure (Puett 2008; Puett 2009). Findings demonstrated that chronic PM(2.5) exposure was associated with risk of all-cause and cardiovascular mortality.

5.1.3 Association between Residences in U.S. Northern Latitudes and Rheumatoid Arthritis

In order to examine the geographic variation in the occurrence of rheumatoid arthritis (RA), investigators analyzed geocoded addresses and incident case diagnosis of rheumatoid arthritis (or censoring of controls) in the Nurses' Health Study between 1988 and 2002. Generalized additive models were used to predict a continuous surface adjusted for known risk factors. Permutation tests were conducted to test for the importance of location and identify areas with statistically significant increased risk, compared to the entire study area. Spatial analyses demonstrated that women residing in high northern latitudes may be at greater risk for rheumatoid arthritis ($P<0.05$). Findings also demonstrated the utility of applying GIS methods to traditional large-scale epidemiological studies to pose new research questions and generate hypotheses for future investigation.

5.2 The AfricaMap Project

AfricaMap is an open-source software project developed to support academic research and teaching, and bring together resources from a variety of disciplines in a single geographic environment. This open source system allows for the investigation, analysis, visualization, and communication of multi-disciplinary, multi-source and multi-format data, organized spatially

and temporally.

This project was the first application of WorldMap, developed at the Center for Geographic Analysis (Lewis and Guan 2010). Since its Beta release in November of 2008, the AfricaMap framework has been implemented in several different locations including metro Boston, Chicago, East Asia, Vermont geological sites, Harvard Forest, and Paris, France. These web mapping applications are used by individual researchers, and have been incorporated into courses at Harvard University.

AfricaMap consists of a set of public digital base maps of the continent, viewable dynamically at a range of scales, and composed of the best cartographic mapping publicly available. A gazetteer provides rapid navigation to specific locations across a vast landscape. As more detailed mapping becomes available it is added to the system. Because of its decentralized architecture, there is (in theory) no hardware or software limitation on the amount of data that can be incorporated. Although currently focused on Africa, Web-based mapping framework behind the project could be used to organize information for any region of the world. This model aggregates data using maps, rather than disciplines, authors, subjects, or indices.

AfricaMap serves the needs of researchers in multiple disciplines interested in Africa, including public health and medicine, including: a common web accessible set of current and historic maps for Africa, a comprehensive gazetteer for African place names, and a repository for spatial and non-spatial data sets for research projects on Africa.

The AfricaMap project represents a framework for organizing African data from a variety of disciplines in a single environment. This allows researchers to explore public health research questions with the breadth and depth of data from other disciplines in a single environment. For example, researchers examining malarial transmission patterns in sub-Saharan Africa can

explore environmental factors, vector sources, habitat reservoirs, soil type, land usage, urban planning, agriculture, socio-demographic data, and population health in a single environment, and develop hypotheses for further analysis and investigation. Various spatial information is also available to view and download for further, including: historic maps, topographic maps, historic trade routes, anthropological and ethnographic data, and epidemiological data.

Figure 3 - AfricaMap displaying malaria distribution.

AfricaMap supports collaborative public health research and teaching on Africa. This project brings together public health researchers from across Harvard University - including the Harvard School of Public Health, the Harvard Medical School, and affiliated hospitals, the Harvard Humanitarian Initiative, the Harvard Initiative for Global Health, and the Faculty of Arts and Sciences – from a variety of disciplines, including public health, medicine, infectious diseases, nutrition, epidemiology, biostatistics, demography, geography, sociology, and anthropology. Public health researchers can utilize the Projects layer in AfricaMap to search for ongoing or historic research projects, examine relevant publications and funding sources, and identify investigators and potential collaborators in a single environment. AfricaMap facilitates discovery and collaboration among researchers from across disciplines to examine health research questions and geographic analysis in Africa.

5.3 Child Physical Activity in the Built Environment

Childhood obesity is a public health issue that has risen to epidemic proportions in the past few decades (Koplan et al. 2005). Between 1970 and 2004, the prevalence of obesity almost tripled among U.S. preschoolers and adolescents and quadrupled among children aged 6 to 11 years (Ogden et al. 2006). Poor daily diet and lack of physical activity will frequently lead to

childhood obesity, and numerous factors figure into shaping these two behaviors including personal and cultural beliefs, environmental conditions, societal influences, health care access, and individual physiology (National Institute of Health 2006).

Recommendations by the American Academy of Pediatrics to prevent childhood obesity include getting at least 1 hour of physical activity per day, limiting high sugar beverage and high fat fast food consumption, and switching dietary habits to include low-fat dairy products, and high fiber and calcium rich foods[49]. It is common knowledge that physical activity can be increased through activities such as organized sports, but no intervention has yet sought to increase physical activity by increasing unstructured physical activities that include the use of one's built environment (Ogden et al. 2006). This includes sidewalks, open space, playgrounds parks, and other areas freely accessible to children in which physical activity can occur. Assessing an individual's comprehensive use of the built environment throughout a given day in an attempt to find possible ways to enhance physical activity has had little previous study, and is the focus of a current study[50] conducted by Harvard-affiliated medical professionals from Massachusetts General Hospital.

The study aims to collect objective information on adolescent's use of the built environment using Global Positioning System (GPS) receivers and accelerometers (devices that measure one's movement / physical activity). By using GIS overlay analysis with the GPS and accelerometer data the built environment elements most associated with physical activity in adolescents can be identified. Variations in use patterns of the built environment by age, gender, and socio-demographics can then be assessed. This objective analysis of where children are active or inactive throughout the course of an entire day may lead to discovery of methods for children to use more effectively the built environment for exercise.

The CGA is involved in this study by providing consultation regarding appropriate GPS devices to with which to equip the children, writing a script that will automatically join GPS readings and accelerometer readings by date and time (using 30 second epochs), and performing overlay analysis using GIS. During December of 2009, a preliminary study was conducted that involved middle school children wearing a GPS and accelerometer for 7 consecutive days. The GPS and accelerometer data were joined by date and time. Locations were classified into categories in GIS using basemap data provided by MassGIS[51].

Figure 4 GPS location classifications on a MassGIS orthophoto.

Once classified, activity levels and locations were graphed for visual analysis, and loaded into SAS for statistical analysis.

[FIGURE 4 GOES HERE]

Once classified, activity levels and locations were graphed for visual analysis, and loaded into SAS for statistical analysis.

[FIGURE 5 GOES HERE]

This study is in the early stages, and will collect activity and location information for the same group of children over several years.

5.4 The Surgical Safety Web Map

The Safe Surgery Saves Lives (SSSL) program has a mission of improving surgical care worldwide by ensuring adherence to proven standards of care in all countries[52]. Sponsored by the World Health Organization (WHO), the SSSL team is headquartered at the Harvard School of Public Health, and led by Dr. Atul Gawande, an endocrine surgeon from Boston's Brigham and Women's hospital. In order to improve surgical care worldwide, the team created a

checklist, (19 steps to be completed during any operation) tested the checklist in clinical settings in 8 hospitals worldwide, and found that using the checklist reduced major complications from surgery by 36% and the rate of death resulting from surgery by 43% (Gawande 2009). These study results were sufficient justification to begin a worldwide dissemination effort to get as many hospitals as possible to use the checklist. In the summer of 2007, members from the SSSL team came to the CGA Help Desk to inquire about adding a mapping component to their project.

The team wanted a web map displaying 1) distinct markers at the hospital locations that had registered to use the checklist and hospitals that were actively using the checklist, and 2) a thematic map showing surgical rates per country, normalized by population. They wanted to be able update the map as new hospitals signed on, and have popup window functionality displaying the hospital name, city, and country upon a mouse click. The team wanted to use the map both internally to track progress of the checklist dissemination and for hospital information access, and externally as a marketing tool to show to hospital administrators, chief surgeons, or other decision makers, illustrating the wide acceptance and proximity of hospitals using the checklist. The SSSL team liked the look, feel, and functionality of Google Maps, and wanted to use this as their basemap platform. In December of 2007, the CGA published the Surgical Safety Web Map with the then 25 participating hospitals displayed as a KML file, and the surgical rates map as a screen overlay.

Figure 6 - The Surgical Safety web map from November, 2008, with 384 participating hospitals.

Figure 7 - The Surgical Rates thematic map overlay from November, 2008.

The team quickly put the map to use, featuring it in presentations and referring potential checklist adaptors to it as their dissemination effort continued. New participating hospitals were

sent to the CGA for inclusion on the map each month, and a new KML file was generated and published

Team members also started using the map as a way to view the current status of the dissemination effort. Members could use the map to quickly get an idea of the participating hospitals in their area of interest, which are available worldwide to anyone with an internet connection. This use of the map prompted a second phase in the web map development. The team wanted to be able to update the map themselves, as soon as hospitals signed up for the checklist. Also desired was an ability to access more information about the hospitals through the map, such as the number of beds, number of physicians, contact person information, and whether devices such as pulse oximiters were available. The SSSL team wanted to be able to view and update this hospital information, yet restrict information access to the general public to the hospital name, city, and country. At this point the total hospital count was nearing 400, which exceeded the number of points that can be rendered on a Google Map mashup at once. Also, it was difficult to discern individual hospitals on the map in cities with many participating hospitals in close proximity. Because of this, the team wanted a list of hospitals in the viewable map area to be displayed interactively.

To accommodate the new web map functionality requests, CGA migrated the hospital data layer into a PostgresSQL database, geospatially enabled with the PostGIS module. Mapserver was used to render the hospitals, and hospitals were symbolized according to whether they were registered or actively using the checklist. To enable the SSSL team to upload new hospitals, a data import program was written in JAVA, and the SSSL team was trained on how to geocode new hospitals using Batchgeocode.com to obtain longitude, latitude coordinates. The longitude, latitude coordinates found for each new hospital were saved in columns in an MS

Excel template that matched the PostGIS database schema. Once the template was populated with the hospital information and coordinates, it was saved into .csv format, and uploaded to the database through the data import program. This upload program requires a login to access, and checks if the .csv file is in the right format. Once uploaded, the new hospitals immediately appear on the map, as each time the map is rendered it draws all of the points in the database according to their coordinate locations. In addition, when one clicked on a hospital to identify it, a "For more hospital information" message was added to the information window, and when clicked prompts the user to enter a username and password. Upon authentication a form appears listing all of the hospital information in editable text forms. The longitude, latitude forms are editable as well, enabling position refinement of hospital locations. A new sidebar was added to the map, and programmed to display hospitals in the current view in a tabular fashion, enabling one to click on a hospital name in the list and have that hospital identified on the map displaying its information.

Figure 8. Updated Surgical Safety Web Map

The final task was writing another JAVA program that allows for an export of the full database back to .csv format after logging in. This enabled any member of the team to save out the latest version of the database, and interact with it in Excel, a database, or statistical program.

Once the new functionality was tested and rolled into the production web map the CGA's role of providing monthly update services was no longer needed. The SSSL team was now fully enabled to update the database and the map on their own. The database and map continue to be updated, and as of this writing contain 3,924 hospitals that have signed up to use the checklist.

5.5 Enabling Health Researchers to use GIS: The Lancet Publication

CGA's role of enabling public health researchers to use GIS technology is showcased in The Lancet November 29, 2010 issue: "Health professionals for a new century: transforming education to strengthen health systems in an interdependent world". The cover of this issue features global cartogram maps, as do several figures within the issue. The maps were produced by Dr. Ananda S. Bandyopadhyay, a graduate of the HSPH Masters in Public Health program. In performing his research at HSPH, Dr. Bandyopadhyay came to the CGA help desk to learn how to make maps and perform spatial analysis. Through CGA workshop training courses, the CGA GIS Institute, and several help desk sessions he became proficient in using many GIS analytical and cartographic tools. Dr. Bandyopadhyay was asked to join the Lancet publication research team, and his GIS skills enabled the production of these descriptive, informative maps, creating an unanticipated benefit to the publication that became one of the focal points of the publication.

Figure 9 (*pending approval for usage*): Map illustrations as published in The Lancet Vol. 376 (December 4, 2010).

6.0 Conclusion

The GIS infrastructure at Harvard, in addition to the services that the HGL, HMC, and CGA offer, provides a robust environment in which GIS can be applied, in the realm of public health research. This article highlighted the application of geographic concepts, pertinent geographic data, the project workflow of the CGA, and case studies in public health research. These topics are prevalent in the Harvard research community today, and will continue to grow and develop into the future.

List of Figures

Figure 1: Spread of P-1 India in recent years.

Figure 2 – CGA Project Specification Document.

Figure 3 - AfricaMap displaying malaria distribution.

Figure 4: GPS location classifications on a MassGIS orthophoto

Figure 5 – Graph of physical activity and locations

Figure 6 - The Surgical Safety web map from November, 2008, with 384 participating hospitals

Figure 7 - The Surgical Rates Thematic map overlay.

Figure 8 - Updated Surgical Safety Web Map

Figure 9: Map illustrations as published in The Lancet Vol. 376 (*pending approval*) (December 4, 2010).

Notes

1. Harvard University Graduate School of Design Web site: http://www.gsd.harvard.edu/gis/manual/normalize
2. ESRI 2009 International Users Conference Proceedings Web site: http://proceedings.esri.com/library/userconf/health09/docs/monday/constructing_a_geospatial_database_of_united_states_local_health_departments.pdf
3. The Commonwealth of Massachusetts GIS Web site: http://www.mass.gov/mgis/
4. New York GIS Clearinghouse Web site: http://www.nysgis.state.ny.us/gisdata/
5. GeoNames Search Web site: http://geonames.nga.mil/ggmagaz
6. Oak Ridge National Laboratory Web site: http://www.ornl.gov/sci/landscan/
7. Demographic and Health Surveys Web site: http://www.measuredhs.com/
8. U. S. Census Bureau Web site: http://www.census.gov

9. United Kingdom Office for National Statistics Web site: http://www.ons.gov.uk/census/index.html
10. Israel Census of Population and Housing Web site: http://www.cbs.gov.il/mifkad/e-mifk.htm
11. Gridded Population of the World Web site: http://sedac.ciesin.columbia.edu/gpw/
12. United Nations Population Information Network Web site: http://www.un.org/popin/
13. PBL Netherlands Environmental Assessment Agency History Database of the Global Environment Web site: http://themasites.pbl.nl/en/themasites/hyde/index.html
14. U. S. Census Bureau International Data Base (IDB) Web site: http://www.census.gov/ipc/www/idb/index.php
15. World Health Organization Data and Statistics Web site: http://www.who.int/research/en
16. Demographic and Health Surveys Web site: http://www.measuredhs.com/
17. UNICEF Childinfo Monitoring the Situation of Children and Women Web site: http://www.childinfo.org
18. UNICEF Childinfo Multiple Indicator Cluster Surveys (MICS) Web site: http://www.childinfo.org/mics.html
19. Centers for Disease Control and Prevention Web site: http://www.cdc.gov/nchs/
20. Centers for Disease Control Behavioral Risk Factor Surveillance System Web site: http://www.cdc.gov/brfss/maps/gis_data.htm
21. Centers for Disease Control Behavioral Risk Factor Surveillance System City and County Data Web site: http://apps.nccd.cdc.gov/BRFSS-SMART/index.asp
22. Centers for Disease Control WONDER Online Database Web site:

http://wonder.cdc.gov

23. Health and Medical Care Archive of the Robert Wood Johnson Foundation Web site: http://www.icpsr.umich.edu/icpsrweb/HMCA/index.jsp
24. New York City Department of Health and Mental Hygiene Web site: http://www.nyc.gov/html/doh/html/survey/survey.shtml
25. Los Angeles County Public Health Assessment Web site: http://www.publichealth.lacounty.gov/ha/HA_DATA.htm
26. The University of North Carolina Population Center China Health and Nutrition Web site: http://www.cpc.unc.edu/projects/china
27. National Family Health Survey of India Web site: http://www.nfhsindia.org/odata.html
28. U. S. Census Bureau American FactFinder Web site: http://factfinder.census.gov/home/saff/main.html?_lang=en
29. GeoLytics Web site: http://www.geolytics.com
30. Social Explorer Web Site: http://www.socialexplorer.com/pub/home/home.aspx
31. Inter-University Consortium for Political and Social Research Web site: http://www.icpsr.umich.edu/icpsrweb/ICPSR/
32. ESRI Community Analyst Web site: http://communityanalyst.esri.com
33. All China Data Center: China Data Online Web site: http://chinadataonline.org/
34. Government of India Ministry of Home Affairs Census of India Web site: http://censusindia.gov.in
35. U. S. G. S. EarthExplorer Web site: http://edcsns17.cr.usgs.gov/NewEarthExplorer/
36. GeoEye GeoFUSE Online Maps Web site: http://geofuse.geoeye.com/maps/Map.aspx

37. Food and Agriculture Organization of the United Nations Web site: http://www.fao.org/corp/statistics/en/
38. United States Environmental Protection Agency Envirofacts Web site: http://www.epa.gov/enviro/index.html
39. U. S. Environmental Protection Agency Geospatial Data Download Service Web site: http://www.epa.gov/enviro/geo_data.html
40. U. S. EPA EnviroMapper Web site: http://www.epa.gov/emefdata/em4ef.home
41. U. S. Geological Survey Board on Geographic Names, Foreign Names Web site: http://geonames.usgs.gov/foreign/index.html
42. U. S. Geological Survey Board on Geographic Names, Domestic Names Web site: http://geonames.usgs.gov/domestic/index.html
43. GeoNames Forum Web site: http://forum.geonames.org/gforum/posts/list/5.page
44. Center for Geographic Analysis at Harvard University Web site: http://gis.harvard.edu
45. Best Practical RT: Request Tracker Web site: http://bestpractical.com/rt
46. Center for Geographic Analysis at Harvard University Google Maps Tutorial Web site: http://maps.cga.harvard.edu/gmaps_instruction/
47. WorldMap Alpha Web site: http://worldmap.harvard.edu
48. NetApp Web site: http://www.netapp.com/us
49. American Academy of Pediatrics Prevention and Treatment of Childhood Overweight and Obesity Web site: http://www.aap.org/obesity/families.html?technology=1
50. National Institute of Health Research Portfolio Online Reporting Tools Project Information Web site:

http://projectreporter.nih.gov/project_info_description.cfm?aid=7962059&icde=7202284

51. The commonwealth of Massachusetts GIS Web site: http://www.mass.gov/mgis/
52. World Health Organization Safe Surgery Saves Lives Web site: http://www.who.int/patientsafety/safesurgery/en

References

Bandyopadhyay, Ananda Sankar. 2010. Spatial Diffusion of Polio in India: Reviewing the Role of Railways. *MPH (Global Health) thesis; Department of Global Health and Population; Harvard School of Public Health*.

Bazemore, A., Phillips, R.L., Miyoshi, T. 2010. Harnessing Geographic Information Systems (GIS) to enable community-oriented primary care. *Am J Board Fam Med,* January/February 23 (1):22-31.

Chung, K., Yang, D.H., Bell, R. 2004. Health and GIS: toward spatial statistical analyses. *J Med Syst,* August 28 (4): 349-360.

Cromley, Ellen K., McLafferty, Sara L. 2002. *GIS and Public Health.* New York: The Guilford Press.

Fortney, J.C., Booth, B.M., Blow, F. C., Bunn J.Y. 1995. The effects of travel barriers and age on the utilization of alcoholism treatment aftercare. *Am J Drug Alcohol Abuse* 21:391-406.

Gawande, Atul. 2009. *The Checklist Manifesto: How to Get Things Right*. New York: Henry Holt and Co.

Gobalet, J.G., Thomas, R.K. 1996. Demographic Data and Geographic Information Systems for Decision Making: The Case of Public Health. *Population Research and Policy Review, Applied Demography: Demography and Decision-Making* 15 (5/6): 537- 548.

Guagliardo, Mark F. 2004. Spatial accessibility of primary care: concepts, methods and challenges. *International Journal of Health Geographics*. http://www.ij-healthgeographics.com/content/3/1/3.

Guan, Weihe Wendy, Burns, Bonnie, Finkelstein, Julia L. and Blossom, Jeffrey C. 2011. Enabling Geographic Research Across Disciplines: Building an Institutional Infrastructure for Geographic Analysis at Harvard University. *Journal of Map & Geography Libraries* 7 (1): 36–60.

Higgs, G. 2004. A literature review of the use of GIS-based measures of access to health care services. *Health Services & Outcomes Research Methodology* 5 (1): 119-139.

Prevention and Treatment". 2007. *National Heart Lung and Blood Institute, National Institutes of Health*. http://www.nhlbi.nih.gov/meetings/workshops/child-obesity/index.htm.

Jenks, R. H., Malecki, J.M. 2004. GIS – a proven tool for public health analysis. *J Environ Health* October 67 (3): 32-34.

Jerrett, M, Gale, S., Kontgis, C. 2010. Spatial modeling in environmental and public health research. *Int J Environ Res Public Health* April 7 (4): 1302-1329.

Koplan, Jeffrey P., Liverman, Catharyn T., Kraak, Vivica I., (Eds.). 2005. Institute of Medicine of the National Acadamies. *Preventing Childhood Obesity.,* Washington, D.C: National Academy Press.

Lewis, B., and Guan, W. 2010. Jump-starting the Next Level of Online Geospatial Collaboration: Lessons from AfricaMap. *Advances in Web-based GIS, Mapping Services and Applications* Li, Dragicevic, and Veenendaal (Eds). Boca Raton, FL: CRC Press.

Longley, P.A., Goodchild, M.F., Maguire, D.J. and Rhind, D.W. 2010. *Geographic Information Systems and Science, 3rd Edition*. Chichester: John Wiley & Sons., Ltd.

Martens, P. and Hall, L. 2000. Malaria on the move: human population movement and malaria transmission *Emerg. Infect Dis*. March/April 6 (2): 103–109.

McLafferty, S.L. 2003. GIS and health care. *Annual Review of Public Health* 24 (1): 12-25.

Meden, T., St. John-Larkin, C., Hermes, D., Sommerschield, S. 2002. Relationship between travel distance and utilization of breast cancer treatment in rural northern Michigan. *Journal of the American Medical Association* 287 (1): 111

Ogden, C.L., Carroll, M.D., Curtin, L.R., McDowell, M.A., Tabak, C.J., Flegal, K.M., 2006. Prevalence of overweight and obesity in the United States, 1999-2004. *Journal of the American Medical Association* 295 (13): 1549-1555.

Paolino, L., Sebillo, M., Cringoli, G. 2005. Geographical Information Systems and on-line GIServices for health data sharing and management. *Parassitologia* March 47 (1): 171-175.

Prothero, RM. 1977. Disease and mobility: a neglected factor in epidemiology. *Int J Epidemiol* 6 (3): 259-267.

Puett, Robin C., Hart, Jaime E., Yanosky, Jeff D., Paciorek, C., Schwartz, J., Suh, H., Speizer, F. E., Laden, F. 2009. Chronic fine and coarse particulate exposure, mortality, and coronary heart disease in the Nurses' Health Study. *Environmental Health Perspectives* November 117 (11): 1697-1701

Puett, Robin C., Schwartz, Joel, Hart, Jaime E., Yanosky, Jeff D., Speizer, Frank E., Suh, Helen, Paciorek Christopher J., Neas, Lucas M., Laden, Francine. 2008. Chronic particulate exposure, mortality, and coronary heart disease in the nurses' health study. *American journal of epidemiology* 168 (10): 1161-1168.

Ricketts, T.C. 2003. Geographic information systems and public health. *Annual Review of Public Health* 24: 1-6.

Rushton, G. 2003. Public health, GIS, and spatial analytic tools. *Annual Review of Public Health*. 24 (1): 43-56.

Slocum, Terry A., McMaster, Robert B., Kessler, Fritz C., and Howard, Hugh H. 2009. *Thematic Cartography and Geovisualization (3rd edition)*. Upper Saddle River: Pearson Prentice Hall.

Sonesson, C. and Bock, D. 2003. A review and discussion of prospective statistical surveillance in public health. *Journal of the Royal Statistical Society Series A (Statistics in Society)* 166 (1): 5–21.

The Lancet 2010. *Health professionals for a new century: transforming education to strengthen health systems in an interdependent world* 376 (1): 1-3.

Yanosky, J, Paciorek, C., Suh, H. 2009. Predicting chronic fine and coarse particulate exposures using spatio-temporal models for the northeastern and Midwestern US. *Environ Health Perspect* 117 (1): 522–529.

Yanosky, J.D., Paciorek, C.J., Schwartz, J., Laden, F., Puett, R.C., Suh, H. 2008. Spatio-temporal modeling of chronic PM10 exposure for the Nurses' Health Study. *Atmos. Environ.* 42 (18): 4047–4062.