[DA-037] GIS&T and Epidemiology

Abstract

Location plays an important role in human health. Where we live, work, and spend our time is associated with different exposures, which may influence the risk of developing disease. GIS has been used to answer key research questions in epidemiology, which is the study of the distribution and determinants of disease. These research questions include describing and visualizing spatial patterns of disease and risk factors, exposure modeling of geographically varying environmental variables, and linking georeferenced information to conduct studies testing hypotheses regarding exposure-disease associations. GIS has been particularly instrumental in environmental epidemiology, which focuses on the physical, chemical, biological, social, and economic factors affecting health. Advances in personal exposure monitoring, exposome research, and artificial intelligence are revolutionizing the way GIS can be integrated with epidemiology to study how the environment may impact human health.

Keywords: disease, epidemiology, health, modeling

Author & citation

VoPham, T. (2018). GIS&T and Epidemiology. The Geographic Information Science & Technology Body of Knowledge (1st Quarter 2018 Edition), John P. Wilson (ed). DOI:10.22224/gistbok/2018.1.1.

This entry was first published on January 8, 2018. No earlier editions exist.

Explanation

1. Definitions

Epidemiology: the study of the distribution and determinants of disease.

Descriptive epidemiology: describing disease according to characteristics related to person, place, and time.

Analytic epidemiology: studies which test hypotheses regarding the association between an exposure and disease (or other outcome of interest) using a comparison group. Refer to the Additional Resources for more in-depth discussion of epidemiologic methods.

Risk: the probability that an event will occur in a specified time period.

Risk factor: a characteristic of an individual that increases the likelihood of developing a disease. For example, smoking is a risk factor for lung cancer.

Environmental epidemiology: the study of physical, chemical, biological, social, and economic factors affecting health.

Exposure assessment: in environmental epidemiology, determining the distribution of exposure within a population.

Exposure model: a representation of an exposure variable developed using environmental measurements, knowledge of physical, chemical, and/or biological properties of features (i.e., deterministic), and/or statistical relationships between variables (i.e., stochastic).

Exposome: the totality of exposures experienced by a person during life and the health impact of those exposures.

geoAI: geospatial artificial intelligence, or the application of machine learning and deep learning methods to learn and extract information from spatial big data.

2. The Role of GIS in Epidemiology

Epidemiology is a scientific discipline within public health that is focused on studying the distribution and determinants of disease. Epidemiology is categorized as **descriptive epidemiology**, or studies describing disease according to person, place, and time; or **analytic epidemiology**, or the conduct of studies designed to test hypotheses regarding the association between an exposure and disease (or other outcome of interest). The unit of analysis in these studies can be the individual or an aggregate variable (e.g., areal geographic variable). Geographic information systems (GIS) have and continue to play an important role in many aspects of epidemiologic studies as location is a key consideration in human health. For example, the characteristics of where an individual spends their time (e.g., environmental exposures) – at home or at work, indoors or outdoors – may have a direct impact on their health. GIS enables the incorporation of location-based information to describe and visualize patterns of **risk factors** (i.e., characteristics of an individual that increase the likelihood of developing a disease) and disease; to model and estimate exposure to specific factors in the population; and to link georeferenced data to conduct epidemiologic studies examining exposure-disease associations.

3. GIS applications in descriptive epidemiology

GIS has allowed epidemiologists to use locational information available for cases (i.e., individuals with disease) to extract meaningful information regarding disease according to:

- Person: attributes of cases such as age, race, and gender
- Place: geographic variation in disease
- Time: how disease may vary over time

Incorporating GIS into descriptive epidemiologic studies is valuable for visualizing and

understanding important aspects of disease, including the attributes of the cases, the locations (and characteristics of the locations) at which disease is occurring, and temporal trends that may help shed light on the disease process. GIS-based descriptive epidemiologic studies have been instrumental as exploratory, hypothesis-generating research guiding subsequent analytic epidemiologic studies.

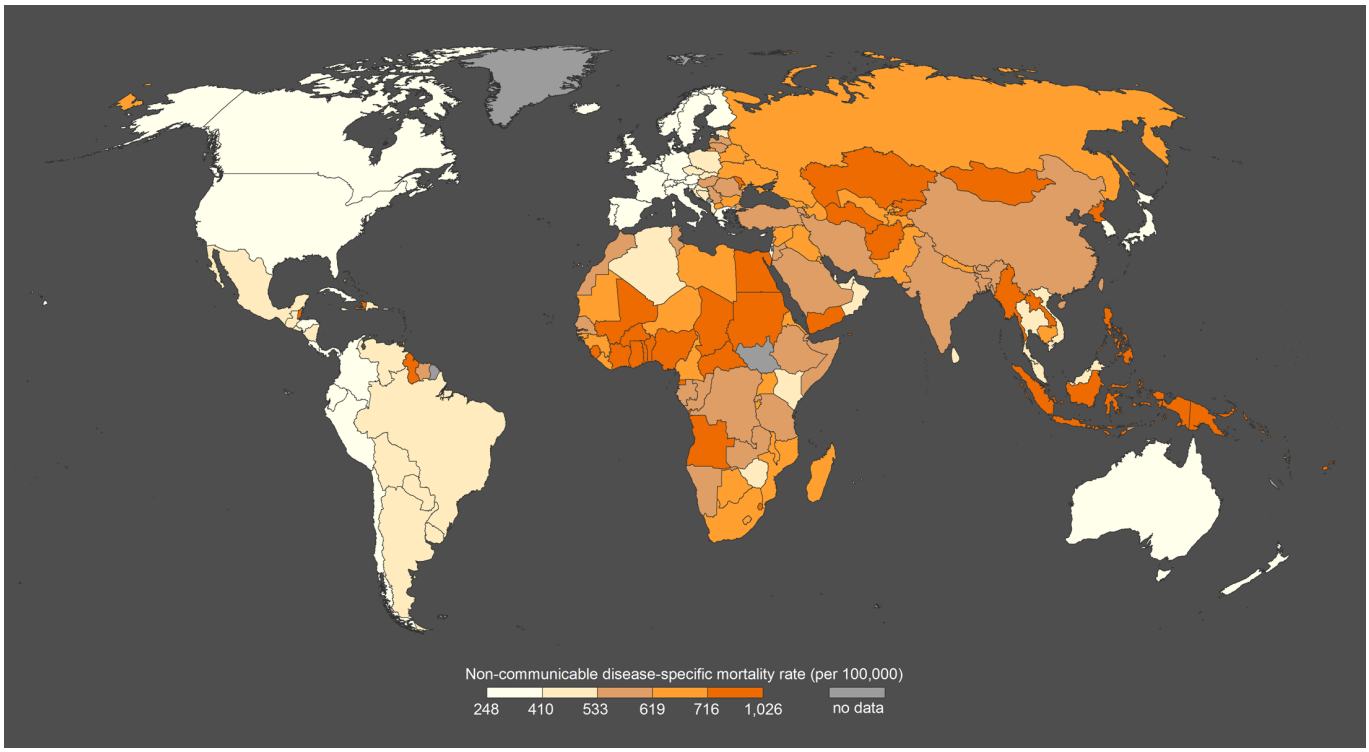

Figure 1. Age-adjusted mortality rates from non-communicable diseases in 2015 (deaths per 100,000 population) categorized according to quintiles. Mortality data were provided by the World Health Organization (WHO) Global Health Observatory (GHO) data repository.

Commonly used GIS tools for descriptive epidemiology include disease mapping and spatial cluster analyses. Epidemiologists have mapped measures of disease occurrence, including incidence (i.e., the occurrence of new cases of disease in a population over a specific period of time), mortality (i.e., the occurrence of deaths in a population over a specific period of time), and prevalence (i.e., the proportion of the population with a disease at a specific time point or period of time), to illustrate the spatial distribution of where new cases are occurring, of access and proximity to health care resources, and of the burden of disease, respectively. A hallmark example of disease mapping was conducted by John Snow in 1854 as part of a London cholera epidemic, where Snow showed a higher frequency of cholera deaths occurring near the contaminated Broad Street public water pump. In 1975, the National Cancer Institute released its first atlas of county-level choropleth maps of cancer mortality in the U.S. from 1950-1969. The overarching goals of this atlas were to identify areas with high rates of cancer mortality, generate hypotheses related to geographically varying factors that may drive these disease patterns, and to delineate high-risk areas that may represent target populations for early cancer detection and prevention. Another example is illustrated in Figure 1, which shows age-adjusted mortality rates from non-communicable diseases in 2015 provided by the World Health Organization (WHO), which continuously compiles mortality data from country civil registration systems each year.

Spatial cluster analyses are conducted to determine the presence and/or locations of clusters, or spatial aggregations of events (Pfeiffer et al., 2008). Clusters may exist as a result of the spatial distribution of the population at risk for a disease or its risk factors. For example, spatial clustering methods were used to study geographic variation of breast cancer mortality in the northeastern U.S. (Kulldorff et al., 1997). In addition, disease surveillance, or the active or passive collection, analysis, and interpretation of health events, also utilizes GIS for disease outbreak detection and to inform resource allocation for disease control and prevention.

4. GIS applications in analytic epidemiology

Analytic epidemiology tests hypotheses regarding the relationship between an exposure and disease (or other outcome of interest) using a comparison group. A common observational (i.e., exposure and disease are observed and not manipulated) study design is the cohort study, where participants are categorized according to whether or not they are exposed to a factor of interest and followed over time to determine if they develop the disease of interest. Another common observational study design is the case-control study, where participants with and without disease are sampled, and their exposure status is subsequently determined. A distinguishing feature of studies is whether they are conducted prospectively vs. retrospectively. A measure of association is calculated, such as a relative risk, which quantifies the association between an exposure and disease, conveying the direction of effect (i.e., positive, negative, or null) and the magnitude or strength of the association. Refer to the Additional Resources for a more detailed discussion on epidemiologic methods. GIS has played a prominent role in **environmental epidemiology**, or the study of physical, chemical, biological, social, and economic factors affecting health. Research in environmental epidemiology will be the focus of the remainder of this section.

GIS is applied to many fundamental aspects of analytic epidemiologic studies including:

- **Exposure assessment**: determining the distribution of exposure within a population
- Linkage of georeferenced information regarding exposure, outcome, and other relevant data: A standard example is linking spatial data for an exposure of interest with georeferenced information available for each study participant (e.g., geocoded residential address), which allows for subsequent linkage to the study participant's outcome and other relevant data to determine if there is an association between an exposure and outcome.

An important consideration in analytic epidemiologic studies using GIS is the spatial scale at which geographic variables are available. Ideally, studies are conducted at the individual level. Many existing cohorts, such as the Nurses' Health Studies, have developed GIS data infrastructures where participant addresses (usually residential) have been geocoded (i.e., georeferenced to assign a latitude and longitude), which allows investigators to propose studies using previously linked spatial data or to link new spatial data to these geocoded addresses. However, when individual-level data are unavailable, ecological studies can be conducted, where the association between an exposure and frequency of disease, each aggregated to an areal unit such as an administrative boundary (e.g., county), are investigated (Santos, 1999). Publicly available databases, such as Surveillance, Epidemiology, and End Results (SEER) and the National Health and Nutrition Examination Survey (NHANES), provide areal geographic variables with which to link georeferenced data

to facilitate the conduct of ecological studies. In addition, epidemiologic studies may be interested in investigating the impact of neighborhood-level exposures, such as socioeconomic factors, on health for which areal units would be appropriate. Time is another important consideration, as the temporal relationship of exposure preceding disease is essential to demonstrating causality in epidemiologic studies. Further, studies may need to reconstruct historical exposure, as epidemiologic studies of chronic diseases, such as cancers, are often associated with a latency period between exposure and onset of disease symptoms.

4.1 Exposure assessment and linkage with georeferenced data to conduct analytic epidemiologic studies

Direct methods of exposure assessment include biomonitoring (measuring a substance using urine, hair, nails, or blood), while indirect methods of exposure assessment include using GIS. For example, GIS has been used to measure the built environment, or our man-made surroundings that affect physical activity, an important research area for aging epidemiology as well as other fields. A walkability index was created using GIS, incorporating spatial data on land use, street connectivity, and population density, to examine its association with obesity and physical activity among older adults (Portegijs et al., 2017).

GIS has been widely applied to **exposure modeling**, or the development of a model to represent a particular environmental variable using environmental measurements, knowledge of physical, chemical, and/or biological properties of features, and/or statistical relationships between variables (Nieuwenhuijsen, 2015). Epidemiologic studies are often comprised of large numbers of individuals, rendering direct exposure assessment unfeasible. Thus, exposure modeling incorporating GIS has been used to estimate environmental exposures, which is scalable to accommodate large study populations. Further, GIS is well-suited to modeling environmental exposures as they by definition geographically vary, for example, due to spatial variability in meteorological conditions that are predictive of air pollution levels. Ultimately, exposure models can be developed and used to determine if an individual, according to their georeferenced location, is exposed to a particular environmental factor; this information is subsequently used to examine if there is an association between the exposure and outcome of interest.

GIS-based exposure modeling can consist of basic spatial analyses, such as overlays, distance/proximity measures, and buffers. For example, Lin et al. (2012) examined the association between ambient ultraviolet (UV) radiation exposure, estimated by overlaying geocoded residential addresses with satellite remote sensing (i.e., data captured from a distance) images of UV, and the **risk** (i.e., probability that an event will occur in a specified time period) of developing cancers. Remote sensing data are valuable to epidemiologic studies as they are typically associated with expansive spatial and temporal coverage that can address large study populations and time periods of interest. Hart et al. (2013) estimated residential distance to roads as a proxy for changes in traffic-related exposures (e.g., air pollution and noise) to study in relation to myocardial infarction and all-cause mortality. Rull and Ritz (2003) developed a GIS method to estimate exposure to agricultural pesticides using pesticide-treated crop fields intersecting a radial buffer centered on a geocoded residential address. This method has been used to examine the association between pesticide exposure and risk of Parkinson’s disease and other health endpoints.

GIS-based exposure modeling has also included the application of advanced spatial methods for interpolation, or estimating values at unmeasured locations using data measured at sample points. Examples include Hoek et al. (2002) estimating traffic-related air pollution exposure, calculated using an inverse distance weighting (IDW)-based measure that assigned greater weight to air pollution monitors closer to geocoded residential addresses, to study in relation to mortality. IDW has also been used to create predicted raster surfaces of environmental exposures that are subsequently linked to georeferenced data. Kriging is a popular geostatistical method that incorporates the spatial structure or correlation of the measured data for interpolation. For example, area-to-point residual kriging was used to create a UV exposure model of the contiguous U.S. using spatial data on known predictors of UV including ozone and aerosol optical depth (VoPham et al., 2016). Land use regression (LUR) is an advanced spatial method typically applied to air pollution exposure modeling, utilizing air pollution measurements and spatial data on roads, traffic, land use, population density, and elevation. For example, LUR models for particulate matter (PM) air pollution have been developed and linked with geocoded addresses from the European Study of Cohorts for Air Pollution Effects (ESCAPE) project to conduct studies examining the association between air pollution and health outcomes such as mortality (Eeftens et al., 2012).

5. Issues and Research Challenges

Locational uncertainty is an important consideration in epidemiologic studies using GIS. Residential addresses are often used to assign exposure, which do not account for mobility and time spent at other locations. Spatial mismatches in data should be addressed, for example, where individual-level geocoded addresses are linked to spatial data at a spatial resolution that is too coarse to reflect personal exposure; or use of an areal geographic variable for study participants, which may not reflect where the participants live, work, and spend their time. The modifiable areal unit problem (MAUP), or observing different results when using different scales or zones/aggregations, and its epidemiologic manifestation as the ecological fallacy, where results from ecological studies may not reflect individual-level associations, should be considered. Another key consideration in epidemiology is confounding, or distortion in study results due to a confounding factor that is associated with the exposure and outcome, but is not in the causal pathway between the exposure and outcome. Potential confounders, such as socioeconomic factors that have been shown to exhibit spatial variation, are important to account for in environmental epidemiologic studies in which exposures also spatially vary – to determine if the association between an exposure and outcome is independent of the confounding factor. Finally, validation is an important consideration in exposure modeling, which determines the extent to which a model measures what it intends to measure. For example, the Rull and Ritz (2003) buffer-based GIS pesticide exposure model described earlier was validated using serum pesticide levels as the gold standard.

6. Future directions: personal exposure monitoring, the exposome, and artificial intelligence

Advances in personal exposure monitoring, particularly using smart phone applications, Global Positioning System (GPS), wearable devices (e.g., accelerometers), and sensors, allow for the opportunity to obtain real-time, individual-level measurements regarding a person’s location, ambient exposures, and behaviors. Smart phone applications have been developed to objectively collect data on physical activity, taking advantage of smart phone

GPS capabilities to record locations, while accompanying sensors capture ambient environmental exposures such as to black carbon air pollution (Nieuwenhuijsen et al., 2015). Further, collecting GPS coordinates allows for linkage to spatial data that can be used to examine a multitude of potential exposure-disease associations. These innovations enable precise delineation of a person's activity space, or the places where their time is spent throughout the course of daily activities. This further allows epidemiologists to more accurately assess the environmental factors to which a person may be exposed based on highly granular information collected regarding their locations. At the community level, the enhanced availability of spatial data, as well as spatial technologies to collect, visualize, and analyze such data, facilitates community initiatives such as mapping for health promotion and risk communication.

Research into the **exposome**, or the totality of exposures experienced by a person during life and the health impact of those exposures, has garnered increased attention (DeBord et al., 2016). Environmental exposures as part of the general external exposome domain can be measured using GIS through linking spatial data on a variety of environmental exposures to geocoded locations. Epidemiology provides methods to conduct exposomic research to better understand the relationships between exposures and health outcomes to lead to improved disease prevention and control.

A rapidly emerging field is geospatial artificial intelligence, or **geoAI,** which has the capacity to revolutionize exposure modeling for epidemiology. In the current era of big data, and in particular spatial big data, characterized by high volume, variety, and velocity, specialized methods and technologies are required to process and analyze these data. Combined with modern-day advances in high-performance computing and storage, geoAI involves the application of methods in spatial science, machine learning (e.g., deep learning), and data mining to learn and extract information from spatial big data. For example, geoAI applications range from feature recognition to remote sensing image enhancement. geoAI offers the ability to overcome previous limitations in exposure modeling regarding the large burden of time and resources, and can utilize vast amounts of high spatial and temporal resolution big data to provide highly resolved exposure information to inform epidemiologic studies. Issues regarding confidentiality and data quality must be considered when acquiring, processing, and analyzing spatial big data.

References

DeBord, D. G., Carreon, T., Lentz, T. J., Middendorf, P. J., Hoover, M. D., & Schulte, P. A. (2016). Use of the "Exposome" in the Practice of Epidemiology: A Primer on -Omic Technologies. American Journal of Epidemiology, 184(4), 302-314.

Eeftens, M., Beelen, R., de Hoogh, K., Bellander, T., Cesaroni, G., Cirach, M., . . . de Nazelle, A. (2012). Development of land use regression models for PM2.5, PM2.5 absorbance, PM10 and PMcoarse in 20 European study areas; results of the ESCAPE project. Environmental Science & Technology, 46(20), 11195-11205.

Hart, J. E., Rimm, E. B., Rexrode, K. M., & Laden, F. (2013). Changes in Traffic Exposure and the Risk of Incident Myocardial Infarction and All-Cause Mortality. Epidemiology, 24(5), 734.

Hoek, G., Brunekreef, B., Goldbohm, S., Fischer, P., & van den Brandt, P. A. (2002). Association between mortality and indicators of traffic-related air pollution in the Netherlands: a cohort study. The Lancet, 360(9341), 1203-1209.

Kulldorff, M., Feuer, E. J., Miller, B. A., & Freedma, L. S. (1997). Breast Cancer Clusters in the Northeast United States: A Geographic Analysis. American Journal of Epidemiology, 146(2), 161-170.

Lin, S. W., Wheeler, D. C., Park, Y., Cahoon, E. K., Hollenbeck, A. R., Freedman, D. M., & Abnet, C. C. (2012). Prospective study of ultraviolet radiation exposure and risk of cancer in the United States. International Journal of Cancer, 131(6).

Nieuwenhuijsen, M. J. (2015). Exposure Assessment in Environmental Epidemiology (2nd ed.). New York, NY: Oxford University Press.

Nieuwenhuijsen, M. J., Donaire-Gonzalez, D., Rivas, I., De Castro, M., Cirach, M., Hoek, G., . . . Sunyer, J. (2015). Variability in and Agreement between Modeled and Personal Continuously Measured Black Carbon Levels Using Novel Smartphone and Sensor Technologies. Environmental Science & Technology, 49(5), 2977-2982.

Pfeiffer, D., Robinson, T. P., Stevenson, M., Stevens, K. B., Rogers, D. J., & Clements, A. C. (2008). Spatial analysis in epidemiology. New York, NY: Oxford University Press.

Portegijs, E., Keskinen, K. E., Tsai, L.-T., Rantanen, T., & Rantakokko, M. (2017). Physical Limitations, Walkability, Perceived Environmental Facilitators and Physical Activity of Older Adults in Finland. International Journal of Environmental Research and Public Health, 14(3), 333.

Rull, R. P., & Ritz, B. (2003). Historical pesticide exposure in California using pesticide use reports and land-use surveys: an assessment of misclassification error and bias. Environmental Health Perspectives, 111(13), 1582.

Santos, S. I. (1999). Cancer Epidemiology: Principles and Methods. Lyon, France: International Agency for Research on Cancer.

VoPham, T., Hart, J. E., Bertrand, K. A., Sun, Z., Tamimi, R. M., & Laden, F. (2016). Spatiotemporal exposure modeling of ambient erythemal ultraviolet radiation. Environmental Health, 15(1), 111.