J Geograph Syst (2000) 2:91–97

Journal of Geographical Systems

Spatial analysis in epidemiology: Nascent science or a failure of GIS?

Geoffrey M. Jacquez

BioMedware, Inc., 516 North State Street, Ann Arbor, MI 48104-1236, USA
(e-mail: Jacquez@BioMedware.com)

Abstract. This paper summarizes contributions of GIS in epidemiology, and identifies needs required to support spatial epidemiology as science. The objective of spatial epidemiology is to identify disease causes and correlates by relating spatial disease patterns to geographic variation in health risks. GIS supports disease mapping, location analysis, the characterization of populations, and spatial statistics and modeling. Although laudable, these accomplishments are not sufficient to fully identify disease causes and correlates. One reason is the failure of present-day GIS to provide tools appropriate for epidemiology. Two needs are most pressing. First, we must reject the static view: meaningful inference about the causes of disease is impossible without both spatial *and* temporal information. Second, we need models that translate space-time data on health outcomes and putative exposures into epidemiologically meaningful measures. The first need will be met by the design and implementation of space-time information systems for epidemiology; the second by process-based disease models.

Key words: GIS, epidemiology, spatial statistics, spatial models

1 Introduction

Public health concerns itself with the identification, monitoring, and maintenance of the health of human populations. This is accomplished by monitoring disease outcomes, by identifying health risks, and by designing and implementing interventions to ameliorate those risks (Thacker, Stroup et al. 1996). GIS is making substantive contributions in public health at a highly practical level. Applied questions such as 'Where is the best place for the new clinic?', 'Where is disease mortality high?', 'Where is there a disease cluster?' are inherently spatial and readily addressable using current GIS and statistical tools for analyzing GIS data (Gatrell and Loytonen 1998).

Epidemiology is the study of health and disease in human populations, and, because populations are inextricably bound to 'place', it seems reasonable to expect GIS to advance Epidemiology as science. Despite the many

practical applications in public health, some authors believe this expectation hasn't been fully met (Jacquez 1998). Sources for this failure are several and include the lack of epidemiologists trained to 'think spatially'; a lack of spatial studies that soundly demonstrate unique and substantial contributions of GIS in epidemiology; and the failure of commercial off-the-shelf (COTS) GIS to provide appropriate tools for spatial epidemiology. As a tool for advancing epidemiology as science, GIS thus appears to have severe limitations. To understand this perspective, and how this failure can be corrected, this paper considers GIS both as technology and as scientific tool. To begin I present several acknowledged contributions of GIS and spatial analysis in epidemiology. These contributions are deemed insufficient to advance spatial epidemiology as science, and needs that must be met to accomplish this objective are identified. The paper concludes by identifying the important research topics in spatial epidemiology.

2 Strengths of GIS and spatial analysis in epidemiology

Mapping and cartography. Visualization is one of the first steps in exploratory spatial data analysis (ESDA). GIS quickly creates maps of morbidity and mortality patterns in relation to population density, putative exposures and geographic features. The future promises a close link between statistical plots and geographic maps ('cartographic brushing', Monmonier 1996, Tweedie 1997) that will more directly support pattern identification and, it is expected, the formulation of spatio-epidemiologic hypotheses.

Location allocation. The location allocation problem addresses the need to place health facilities and services in 'the best' geographic locations. Current GIS and add-ons support the determination of where health facilities may best be located, the estimation of ambulance travel times to hospitals, and the identification of hospital catchment areas. In public health GIS has proven contributions in vector control (e.g. where to place the intervention?) to reduce malaria, Lyme disease, Shistosomiasis and other vector-borne diseases.

Characterization of populations. To outsiders, epidemiologists seem fixated on 'numerators' and 'denominators', and with good reason: In a disease rate the numerator is disease cases while the denominator is the population at risk of contracting the disease. In spatial epidemiology, the characterization of geographic populations is an important task directly supported by GIS. This entails exposure assessment (where are risk factors high?) and the identification of high-risk populations for further study (*e.g.* leukemia near power lines – Wartenberg, Greenberg et al. 1993; breast cancer in the northeastern United States – Kulldorff, Feuer et al. 1997).

Spatial statistics. Spatial statistics quantify geographic variation in geographic variables. They can identify violations of assumptions of independence required by many epidemiological statistics; and measure how populations, their characteristics, covariates and risk factors vary in geographic space (Rushton and Lolonis 1996, Haining 1998). Methodological research on disease cluster techniques has burgeoned (for recent reviews see Jacquez, Waller et al. 1996, Jacquez, Grimson et al. 1996, Kulldorff 1998) and spatial statisti-

cal software is increasingly integrated with GIS, making possible interactive Exploratory Spatial Data Analysis (Bailey and Gatrell 1995).

Spatial models. Recent developments of spatial models in public health include Bayesian smoothing of disease rates (Devine, Louis et al. 1994), geostatistical models, and advanced Monte Carlo estimation techniques (the papers in Lawson 1999 are representative of the state of the field). These models address important issues of the stabilization of disease rates, interpolation, map presentation, and the estimation of variables at 'not gauged' locations.

While a formidable list, in general these accomplishments support descriptive epidemiology and do not directly increase our ability to link human health outcomes to putative exposures and environmental risk factors. Under the best of circumstances they can identify a local excess of disease that may suggest, because of geographic proximity, a possible cause. This seems poor payoff for what is often a costly and time consuming geographic investigation. Shouldn't we expect more?

3 Needs of spatial epidemiology

This paucity of scientific payoff is rooted in technological determinism – GIS tools determining the hypotheses that can be addressed. The questions one can ask of the data depend implicitly on the data models, spatial data structures, and queries employed by ones GIS. Like a child trying to drive a nail with a wrench, we're using the wrong tool for the job. But unlike the child, we can't go to the basement and get a hammer – we're going to have to design the appropriate tool. This involves specification of the tools requirements.

Process-pattern link. We need an increased understanding of the relationships between disease processes and resulting disease patterns. Spatial disease patterns are the outcomes of space-time processes, and several alternative processes may give rise to similar spatial patterns (Schaerstrom 1996) making it very difficult to interpret spatial disease patterns in other than broadly descriptive terms ('there is a hot-spot'). The elucidation of this link requires models of process while the estimation of such a model's parameters requires space-time information systems.

Models of process. Models of process are expressed in terms of the physical and biological mechanisms underlying the system under study. This contrasts with models of data that are expressed in terms of the data's statistical properties. At present most if not all spatial models in public health are models of data, rather than models of process. While useful for prediction, the parameters of models of data have limited epidemiological utility. In contrast, the parameters of process models are by definition directly interpretable in terms of underlying diseases. Compartmental analysis (Jacquez 1996) is a powerful approach for constructing models of space-time processes that holds great promise in epidemiology.

Space-time information systems (STIS). Estimation of the parameters of such process models will require information systems capable of dealing with

spatial and temporal referencing. While several authors recognize the need for 'time GIS' in order to represent space-time data in general (e.g. Langran 1992, Peuquet 1994) and human disease data in particular (Loytonen 1998), its essential role in the estimation of parameters of space-time process models has received little recognition. Because STIS will make possible estimation of the parameters of space-time models, they are expected to be indispensible for advancing spatial epidemiologic science.

Designer statistics for spatial epidemiology. Consider the data points:

$$(x_i, y_i, z_{i1}, \ldots, z_{im}) = (x_i, y_i, \mathbf{z}_i), \quad i = 1, \ldots, N. \tag{1}$$

Here (x_i, y_i) is the spatial coordinate of the ith data point and $(z_{i1}, \ldots, z_{im})$ are measurements of attributes (e.g. disease outcomes) at that location. Several authors have noted that spatial statistics are comprised of two parts: a proximity measure (p_{ij}) quantifying geographic relationships among the sample locations, and a data measure (d_{ij}) reflecting relationships among the variable(s) observed at those locations.

$$\begin{aligned} p_{ij} &= f(x_i, y_i; x_j, y_j) \\ d_{ij} &= g(\mathbf{z}_i, \mathbf{z}_j) \end{aligned} \tag{2}$$

Examples of the function f include geographic distance, nearest neighbor relationships, spatial adjacency and spatial weight. Alternatives for the function g include genetic distance, demographic distance, and Mahalanobis distance, to name only a few. The vast majority of spatial statistics, including Moran's I, Geary's c, and many disease cluster statistics (see Jacquez and Jacquez 1999 for examples) can be expressed in the general form:

$$\Gamma = S \sum_{i=1}^{N} \sum_{j=1}^{N} p_{ij} d_{ij} \tag{3}$$

Here S is a scalar. The function f quantifies the researchers alternative hypothesis, and should be so specified. Thus a test for excess disease near a given location (*e.g.* a focused test Waller and Poquette 1998) typically is constructed by specifying f so that it increases as one gets closer to the focus. In spatial epidemiology tests of the type in Eq. 3 are appropriate only when the function f is a reasonable exposure surrogate. In practice this arises when exposure wasn't measured, and/or when the causes of the disease under study aren't known. In epidemiology this is a profound level of ignorance that seldom occurs. In most instances researchers have data on putative exposures (other than geographic location), and on covariates, as well as on the disease outcomes themselves. To reformulate the problem, the data points typically are:

$$\begin{aligned} &(x_i, y_i, z_{i1}, \ldots, z_{im}, e_{i1}, \ldots, e_{in}, c_{i1}, \ldots, c_{io}) \\ &\quad = (x_i, y_i, \mathbf{z}_i, \mathbf{e}_i, \mathbf{c}_i), \quad i = 1, \ldots, N \end{aligned} \tag{4}$$

Here the $\mathbf{e}_i$ are exposure measures at location i, and $\mathbf{c}_i$ are observations on covariates at that location. One then needs to construct measures of proximity

that correspond to spatio-epidemiological hypotheses incorporating spatial location, exposure, and covariates:

$$p_{ij} = h(x_i, y_i, \mathbf{e}_i, \mathbf{c}_i; x_j, y_j, \mathbf{e}_j, \mathbf{c}_j) \tag{5}$$

This revised framework makes possible 'designer' statistics for spatial epidemiology that incorporate relevant descriptors of 'place', namely spatial location, exposure and covariates. Further, once space-time information systems are available this framework is readily extensible by incorporating time as one of the relevant descriptors of 'place'. Currently used spatial statistics are special cases that arise when knowledge of exposure and covariates is lacking and when the function h corresponds to one of the currently used specifications of function f. To date, the need for a framework such as that represented in Eq. 5 has been recognized by only a few authors (e.g. Zhou 1998), but is expected to become more pressing as deficiencies of currently used spatial statistics for epidemiology become more widely recognized.

4 Discussion

I have painted with a broad and heavy brush to stimulate thought on first, important problems in spatial epidemiology and second, the technology required to solve them. There is a dynamic and a tension between problems and technology. In some instances a problems solution stimulates development of critical technology (call this the 'designed' mode); in other instances a technological advance results in novel solutions to problems that were not the target of the technological innovation (call this the 'vicarious' mode). To date the use of GIS in spatial epidemiology has been vicarious. Epidemiologists and public health practitioners have taken GIS, determined what it could accomplish, and applied it to problems that 'fit' the existing technology. And, as noted earlier in this paper, they have addressed a diverse array of important problems. Nonetheless, the design of present GIS technology limits its ability to represent critical aspects of epidemiologic data, such as measures of proximity that correspond to spatio-epidemiological hypotheses incorporating spatial location, exposure, and covariates (Eq. 5). Despite the attention grabbing 'failure of GIS' in this papers title, my objective is not to blame the technology, rather, it is to identify needs that have yet to be adequately addressed. These needs are best met by carefully designing GIS technology for epidemiology.

There are technological issues that, while interesting, have relatively minor implications in terms of advancing spatial epidemiology. For example, there is a clear need for distributed spatial decision support systems in public health. These systems will appear to 'run' on researchers desktops, and will have true concurrency by providing access to centralized disease data (registries, etc.) and a central repository of GIS and epidemiological tools. While their implementation is non-trivial from a computer science perspective, they will only ease the tasks we already 'do' in spatial epidemiology; and will not address the 'needs of spatial epidemiology' described earlier:

- An increased understanding of the relationships between disease processes and patterns;

- Process-based disease models;
- Space-time information systems for epidemiology; and
- Designer statistics for spatial and spatio-temporal epidemiology.

Several conclusions may be drawn. First, we must reject the static view presented by COTS GIS: meaningful inference about disease causes is impossible without both spatial *and* temporal information. It thus comes as no surprise that the 'scientific yield' of spatial epidemiological studies is typically small. Second, we must appreciate that our current spatial models are primarily models of data that are inadequate for representing epidemiological relationships. We need models of process that translate space-time data on health outcomes and putative exposures into epidemiologically meaningful measures. Addressing these four needs (above) is essential for spatial epidemiology to move forward as a scientific field.

Acknowledgements. This research was funded by grants R42CA64979 from the National Cancer Institute and R43ES10220 from the National Institute of Environmental Health Sciences. The views in this paper are those of the author and do not necessarily reflect the official views of the funding agencies.

References

Bailey TC, Gatrell AC (1995) *Interactive spatial data analysis*. Longman Scientific & Technical, Essex, England

Devine OJ, Louis TA et al. (1994) Empirical Bayes estimators for spatially correlated incidence rates. *Environmetrics* 5:381–398

Gatrell AC, Loytonen M (1998) GIS and health research: An introduction. In: Gatrell AC, Loytonen M, *GIS and health*. Taylor and Francis, London, pp 3–16

Haining R (1998) Spatial statistics and the analysis of health data. In: Gatrell AC, Loytonen M, *GIS and health*. Taylor and Francis, London, pp 29–48

Jacquez GM (1998) GIS as an Enabling Technology. In: Gatrell AC, Loytonen M, *GIS and health*. Taylor and Francis, London, pp 17–28

Jacquez GM, Grimson R et al. (1996) The analysis of disease clusters Part II: Introduction to techniques. *Infection Control and Hospital Epidemiology* 17:385–397

Jacquez GM, Jacquez JA (1999) Disease clustering for uncertain locations. In: Lawson A, Bertollini R, *Disease mapping and risk assessment for public health decision making*. Wiley, London

Jacquez GM, Waller LA et al. (1996) The analysis of disease clusters Part I: State of the art. *Infection Control and Hospital Epidemiology* 17:319–327

Jacquez JA (1996) *Compartmental analysis in biology and medicine*. BioMedware, Ann Arbor, MI

Kulldorff M (ed) (1998) *Selection of statistical methods for the analysis of spatial health data*. GIS and Health. Taylor and Francis, London

Kulldorff M, Feuer EJ et al. (1997) Breast cancer clusters in northeastern United States: A geographical analysis. *American Journal of Epidemiology* 146:161–170

Langran G (1992) *Time in Geographic Information Systems*. Taylor and Francis, London

Lawson A, (ed) (1999) *Disease mapping and risk assessment for public health*. Wiley, London

Loytonen M (1998) GIS, time geography and health. In: Gatrell AC, Loytonen M, *GIS and health*. Taylor and Francis, London

Monmonier M (1996) *How to lie with maps*. The University of Chicago Press, Chicago, IL

Peuquet DJ (1994) It's about time: A conceptual framework for the representation of temporal dynamics in GIS. *Annals of the Association of American Geographers* 84(3):441–461

Rushton G, Lolonis P (1996) Exploratory spatial analysis of birth defects in an urban population. *Statistics in Medicine* 15:717–726

Schaerstrom A (1996) *Pathogenic paths? A time geogaphical approach in medical geography*. Lund University Press, Lund, Sweden

Thacker SB, Stroup DF et al. (1996) Surveillance in environmental public health: Issues, systems and sources. *American Journal of Public Health* 86(5):633–638

Tweedie L (1997) *Characterizing interactive externalizations*. CHI'97, ACM Press, Atlanta, GA

Waller L, Poquette CA (1998) The power of focused score tests under mis-specified cluster models. In: Lawson A, *Disease surveillance and public health*. Wiley, London

Wartenberg D, Greenberg M et al. (1993) Identification and characterization of populations living near high-voltage transmission lines: A pilot study. *Environmental Health Perspectives* 101:626–632

Zhou Y (1998) GIS-based temporal and spatial modeling of shistosomiasis infection for local transmission control. *Environmental Health Sciences*. University of California Berkeley, p 222