SPATIAL DISEASE MAPPING

Lance A. Waller
Department of Biostatistics, Rollins School of Public Health, Emory University, USA

Keywords: Statistics, epidemiology, Bayesian methods, hierarchical models, random effects, small area estimation, spatial point process, geostatistics

Contents

Summary

We review statistical methods for analyzing spatially referenced disease data. Such data come in a variety of formats, depending on the mode of data collection and other issues such as patient confidentiality. Statistical methods for analyzing mapped health data were developed specific to the particular forms of data, with intent to address the spatially structured correlation often observed in such data. Developments in methodology have largely been within data type, along with associated advances in probabilistic theory. Methods for point data typically involve spatial point process stochastic models and estimates of characteristics of such models. Methods for regional data often involve random effects which effectively "borrow strength" across all or a local subset of the data in order to improve estimation for small geographic areas. Methods for geostatistical data involve statistical predictions of unobserved quantities based on similar quantities observed at nearby locations. Recent research begins to illuminate similarities in the underlying mathematical models from each data category, and the research focus in disease mapping is moving from goals specific to the type of data available toward more comprehensive modeling regarding known and suspected risk factors and using data in a variety of forms and levels of aggregation.

1. Introduction

There is much interest in analyzing the observed spatial pattern of disease, primarily in hopes that the pattern may reveal insight into the underlying etiology of the disease. In

a much cited example, Dr. John Snow mapped the residence locations of cholera deaths during the 1854 London epidemic and noted an aggregation of cases near a certain public water pump. Dr. Snow, an early advocate of a water-born mode of transmission, and others used the maps and other evidence to argue for the closing of the water pump in question.

With the advent of modern computing and geographic information systems (GISs) allowing fast and efficient linking of spatially-referenced data, there is increased interest in mapping disease data in conjunction with data relating to putative causes (e.g., environmental exposures) in hopes of investigating associations between the two. This is a geographic version of the usual statistical scatterplot wherein one seeks to visually identify correlative structure between two (or more) variables. While GISs provide computational means to link and display disease and other data over the same geographic region, they currently contain relatively rudimentary statistical capabilities limiting inference to visual assessments and simple summary measures.

Paralleling the computational developments has been growth in the field of spatial statistics which extends traditional methods of statistical inference to allow for spatial autocorrelation where observations are no longer independent (as typically assumed in many statistical methods), but rather observations are positively correlated with other observations taken nearby. Typically, one assumes such spatial dependence declines as observations are taken at locations further apart.

2. Reasons for Spatial Patterns in Disease Data.

In a health setting, spatial patterns arise for a wide variety of reasons. As mentioned above, interest in disease mapping centers around patterns induced by specific etiologic factors (e.g., some environmental exposure), but other factors also influence observed spatial patterns in disease incidence as detailed below.

Foremost, the population at risk is not distributed at random in space, rather people tend to cluster in towns and cities, and one expects some clustering of cases, even if all individuals were subject to exactly the same risk of a disease. Typically, researchers seek spatial patterns in disease *risk*, or the probability of contracting the disease in a particular time period. Patterns of incident cases identical to patterns of the population at risk offer little additional insight into the disease process or mechanism of spread. Hence, observed differences between the pattern in the population at risk and the pattern in the cases are of primary interest in disease mapping.

Secondly, the nature of disease transmission influences spatial pattern. Infectious diseases often spread by person-to-person contact, or by vectors transmitting an infectious agent between hosts. These infection processes induce direct association between nearby cases, i.e., the occurrence of a case at one location raises the risk of disease for nearby individuals.

For non-infectious diseases, spatial variation in the demographic structure of the population at risk results in spatial variation in disease outcomes. In this case, various demographic factors influence disease risk, leading to higher rates in areas with

concentrations of individuals in high-risk groups. For instance, the risk of most cancers increases with age, so one may observe higher rates of disease in neighborhoods whose populations consist of a higher proportion of older residents than other neighborhoods. In this situation, the presence of cases does not influence local increases in risk, rather spatial aggregation of individuals in similar risk groups results in local similarities in risk.

The nature of data collection may also impact observed spatial patterns of disease. In order to map diseases, cases must be reported and stored in a data set. If there are local variations in the accuracy and reliability of case reporting, these variations serve as a "filter" of the disease process.

That is, the true underlying pattern of disease is filtered through the reporting process so that most cases in areas with strong reporting programs appear in the observed pattern but a greater proportion of cases in areas with weak reporting may not.

Lastly, spatial dependence in health data may arise from underlying environmental exposures. This situation forms the basis of many geographic studies of disease wherein researchers attempt to quantify associations between disease and environment. Such associations are of interest in both infectious diseases (e.g., remote sensing of vector habitat for targeted spraying programs) and non-infectious diseases (e.g., health effects in residents near hazardous waste sites). For non-infectious diseases, cases may occur independently of other cases, but individuals residing in locations near to each other may share similar environmental exposures impacting their respective risk of contracting the disease in similar ways, resulting in spatial similarities in local disease risk. Many of the published statistical approaches limit attention to such non-infectious disease applications, as do the methods outlined below.

In short, spatial dependence in health data may be based on an underlying infectious process (i.e., cases actually cause nearby cases), a spatially heterogeneous reporting process, an underlying environmental component (i.e., nearby cases are independent of one another but share local increases or decreases in risk), or any combination of these scenarios.

3. Types of Spatial Disease Data

Statistical methods for the analysis of spatially referenced health data depend on the type of data available, motivating a brief summary of three broad categories of spatial disease data.

3.1. Point Data

Point data include a unique point location for each health event. Many applications of disease mapping assign cases to their residence location. Other possible locations include schools and worksites. Different patterns will appear using different locations, and researchers should consider whether the available locations reflect those best associated with the scientific questions of interest. For example, occupational locations may be more appropriate than residential locations in a study of the impact of environmental exposures for individuals receiving the majority of their personal

exposure on the job. Analytic methods typically involve comparisons between the spatial "point patterns" defined by the cases and a set of suitably chosen non-cases or "controls", as outlined in Section 4.1.

-
-
-

Bibliography

Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems (with discussion). *Journal of the Royal Statistical Society, Series B* **36,** 192-236. [Contains the mathematics underlying conditionally autoregressive models, where the conditional distribution of a random variable at one location depends only on its spatial neighbors.]

Besag, J., York, J.C., and Mollié, A. (1991). Bayesian image restoration, with two applications in spatial statistics (with discussion). *Annals of the Institute of Statistical Mathematics* **43,** 1-59. [Proposes a fully hierarchical Bayesian version of the Clayton and Kaldor (1987) model with important corrections. Includes a disease mapping example. This paper presents one of the more popular Bayesian disease mapping models currently in use.]

Brody, H., Rip, M.R., Vinten-Johansen, P., Paneth, N., and Rachman, S. (2000). Map-making and myth-making in Broad Street: the London cholera epidemic, 1854. *Lancet* **356,** 64-68. [A historical review of the role of maps in the analysis and administrative response to the 1854 London cholera epidemic.]

Clayton, D.G., and Kaldor, J. (1987). Empirical Bayes estimates of age-standardized relative risks for use in disease mapping. *Biometrics* **43,** 671-681. [The authors propose the empirical Bayes estimation procedure and the Poisson regression model with spatially correlated random effects detailed in the text above.]

Cressie, N.A.C. (1993). *Statistics for Spatial Data, Revised Edition* New York: Wiley. [A detailed compendium of statistical methods for the analysis of spatially referenced data. Few health examples, but many examples from other fields.]

Diggle, P.J. (2000). Overview of statistical methods for disease mapping and its relationship to cluster detection. In *Spatial Epidemiology: Methods and Applications* P. Elliott, J.C. Wakefield, N.G. Best, and D.J. Briggs, eds. Oxford: Oxford University Press. 87-103. [An overview of case-control point process methods in spatial epidemiology and their relationship to disease mapping approaches. Includes a detailed bibliography. In addition, other chapters in the same text offer further historical and technical information regarding disease mapping, as well as several case studies.]

Diggle, P.J., Moyeed, R.A., and Tawn, J.A. (1998). Model-based geostatistics (with discussion). *Applied Statistics* **47,** 299-350. [Introduces a hierarchical Bayesian approach to geostatistics. Includes application to spatial variation in the risk of campylobacter infections.]

Kelsall, J.E. and Diggle, P.J. (1995). Nonparametric estimation of spatial variation in relative risk. *Statistics in Medicine* **14,** 2335-2342. [Details kernel density estimation of relative risk surfaces for point data.]

Oliver, M.A., Muir, K.R., Webster, R., Parkes, S.E., Cameron, A.H., Stevens, M.C.G., Mann, J.R. (1992). A geostatistical approach to the analysis of pattern in rare disease. *Journal of Public Health Medicine*, **14,** 280-289.

Statistics in Medicine (2000). (Special issue devoted to "Disease Mapping with a Focus on Evaluation"). **19,** 2201-2594. [Includes examples of the methods outlined in the text above, as well as many alternative approaches.]

Biographical Sketch

Lance A. Waller is an Associate Professor in the Department of Biostatistics, Rollins School of Public Health at Emory University. His interests involve statistical analysis of spatially referenced data. Examples include tests of spatial clustering of disease cases, for example around a hazardous waste site; small area estimation; hierarchical models with spatially structured random effects; and spatial point process models. Recent applications include spatiotemporal mapping of disease rates, statistical methods for assessing environmental justice, the analysis of spatial trends in Lyme disease incidence and reporting, spatial modelling of the spread of raccoon rabies, and point process analysis of sea turtle nesting locations in Florida. He is interested in both the statistical methodology, and the environmental and epidemiologic models involved in the analysis of this type of data.