Applied Spatial Statistics for Public Health Data

LANCE A. WALLER

Emory University
Department of Biostatistics
Atlanta, Georgia

CAROL A. GOTWAY

National Center for Environmental Health
Centers for Disease Control and Prevention
Atlanta, Georgia

A JOHN WILEY & SONS, INC., PUBLICATION

Applied Spatial Statistics for Public Health Data

WILEY SERIES IN PROBABILITY AND STATISTICS

Established by WALTER A. SHEWHART and SAMUEL S. WILKS

A complete list of the titles in this series appears at the end of this volume.

Applied Spatial Statistics for Public Health Data

LANCE A. WALLER

Emory University
Department of Biostatistics
Atlanta, Georgia

CAROL A. GOTWAY

National Center for Environmental Health
Centers for Disease Control and Prevention
Atlanta, Georgia

A JOHN WILEY & SONS, INC., PUBLICATION

Published by John Wiley & Sons, Inc., Hoboken, New Jersey.
Published simultaneously in Canada.

Library of Congress Cataloging-in-Publication Data:

Waller, Lance A., 1965–
Applied spatial statistics for public health data / Lance A. Waller, Carol A. Gotway.
p. cm.—(Wiley series in probability and statistics)
Includes bibliographical references and index.
ISBN 0-471-38771-1 (coth)
1. Public health—Statistical methods. 2. Spatial analysis (Statistics) I. Gotway, Carol A., 1961– II. Title. III. Series.

RA440.85.W34 2004
614′.07′27—dc22

2003066065

Printed in the United States of America.

10 9 8 7 6 5 4 3 2 1

Statistics, too, have supplied us with a new and powerful means of testing medical truth. . . .

Dr. Benjamin Babbinton
President of the London Epidemiological Society, 1850
Lancet, Volume 2, p. 641

Dedicated with love to

Dr. Alisha A. Waller
Allyn, Matthew, and Adrian Waller
Dr. Clement A. and Mrs. Patricia L. Gotway

Contents

Preface

> Spatial statistical analysis has never been in the mainstream of statistical theory. However, there is a growing interest both for epidemiologic studies, and in analyzing disease processes.

The above is a quote one of us (L.A.W.) received on a grant review in 1997, and it outlines succinctly our motivation for this book. Topics in spatial statistics are usually offered only as special-topic elective courses, if they are offered at all. However, there is growing interest in statistical methods for the analysis of spatially referenced data in a wide variety of fields, including the analysis of public health data. Yet, there are few introductory, application-oriented texts on spatial statistics. For practicing public health researchers with a general background in applied statistics seeking to learn about spatial data analysis and how it might play a role in their work, there are few places to turn.

Our goal is to provide a text that moves from a basic understanding of multiple linear regression (including matrix notation) to an application-oriented introduction to statistical methods used to analyze spatially referenced health data. This book is less an effort to push the methodological frontier than an effort to gather and consolidate spatial statistical ideas developed in a broad variety of areas and discuss them in the context of routinely occurring spatial questions in public health. A complication in this effort is the wide variety of backgrounds among this interest group: epidemiologists, biostatisticians, medical geographers, human geographers, social scientists, environmental scientists, ecologists, political scientists, and public health practitioners (among others). In an effort to provide some common background, in Chapters 1 to 3 we provide an overview of spatial issues in public health, an introduction to typical (nonspatial) analytic methods in epidemiology (for geographers who may not have encountered them previously), and an introduction to basic issues in geography, geodesy, and cartography (for statisticians and epidemiologists who may not have encountered them previously). In Chapter 4 we merge ideas of geography and statistics through exploration of the methods, challenges, and approaches associated with mapping disease data. In Chapter 5 we provide an introduction to statistical methods for the analysis of spatial point patterns, and in Chapters 6 and 7 we extend these to the particular issue of identifying disease clusters, which is often of interest in public health. In Chapter 8 we explore statistical methods for mapping environmental exposures and provide an

introduction to the field of geostatistics. Finally, in Chapter 9 we outline modeling methods used to link spatially referenced exposure and disease data.

Throughout, we provide "data breaks" or brief applications designed to illustrate the use (and in some cases, misuse) of the methods described in the text. Some sequences of data breaks follow the same data set, providing bits and pieces of a broader analysis to illustrate the steps along the way, or simply to contrast the different sorts of insights provided by different methods. In general, we collect methods and ideas around central questions of inquiry, then explore the particular manner in which each method addresses the question at hand. We also include several case studies, wherein we provide a start-to-finish look at a particular data set and address the components of analysis illustrated through the data breaks in a new (and often more involved) setting.

Finally, since spatial statistics is often out of the mainstream of statistical theory, it is often also out of the mainstream of statistical software. Most of the analyses in this book utilized routines in SAS (Littell et al. 1996), the S+SpatialStats module for S-plus (Kaluzny et al. 1998), and various libraries in the freely available R (Ihaka and Gentleman 1996) language. For particular applications, we made use of the freely available software packages WinBUGS (Spiegelhalter et al. 1999), SaTScan (Kulldorff and International Management Services, Inc. 2002), and DMAP (Rushton and Lolonis 1996), and used the geographic information system (GIS) packages ArcGIS and ArcView (Environmental Systems Research Institute 1999, including the spatial autocorrelation scripts for ArcView by Lee and Wong 2001). Regarding Internet addresses, we decided to provide references and detailed descriptions of particular data sets and software packages since links often shift and go out of date. However, we do post related links, the tabulated data sets, and most of our R and SAS codes relating to the data breaks on the book's Web site, linked from *www.wiley.com*. This code should allow readers to duplicate (and hopefully expand!) many of the analyses appearing throughout the book, and perhaps provide a launching point for the analyses of their own data.

L. A. WALLER
C. A. GOTWAY CRAWFORD

Atlanta, Georgia

Acknowledgments

To begin at the beginning, we have to go back to 1999. Owen Devine had substantial input into the original outline and we wish he had been able to continue with us to the end. Brad Carlin, Linda Young, and Ray Waller provided good advice on choosing a publisher and defining the contract terms. Graduate students at the University of Minnesota and Emory University heard the early rumblings of some of the material here and their reactions, frustrations, and enthusiasm drove much of the layout and structure of the text.

Particular individuals provided very helpful review comments on certain chapters. Linda Pickle provided valuable suggestions updating our original ideas on visualization in Chapter 4, and Dan Carr constructed the linked micromap example and suggested several references on current research in cartographic visualization. Felix Rogers was nice enough to let us play with his data on very low birthweights that we used extensively in Chapter 4 and then again in Chapter 9. Owen Devine gave us many editorial suggestions on Chapter 4 which improved our writing. Aila Särkkä provided very thoughtful comments improving the original presentation of point process material in Chapter 5. In 1997, Richard Wright provided Lance with one of his all-time favorite data sets, the medieval grave site data that are featured predominately in Chapters 5 and 6. Dick Hoskins, Bob McMaster, Gerry Rushton, Dan Griffith, and Luc Anselin gamely entertained off-the-cuff questions in a patient and helpful manner. Betsy Hill, Traci Leong, DeMarc Hickson, and Monica Jackson all provided valuable comments regarding Chapters 6 and 7. David Olson, Ricardo Olea, and Konstantin Krivoruchko provided detailed reviews of Chapter 8, and many of their suggestions greatly improved this chapter. Brad Carlin and Alan Gelfand provided valuable insight and a great sounding board for the section on Bayesian models for disease mapping in Chapter 9. Eric Tassone expertly implemented multiple versions of the hierarchical Bayesian Poisson regression model, providing the conclusion to the series of New York leukemia data breaks. Andy Barclay deserves mention for his encouragement and example. We have yet to find a statistical computation that he can't decipher or improve. Throughout, Steve Quigley at Wiley offered patient encouragement by often noting that completed texts tended to sell better than works in progress.

Noel Cressie merits special mention. In particular, this book would not be a Wiley book if not for his unconditional support of the project. He has been there for us in many ways throughout our careers, and with respect to this endeavor, he convinced Wiley to take us on as authors, provided suggestions on the original table of contents, and reviewed a draft of the entire book. His review of the entire book was especially important to us. We needed someone objective, yet knowledgeable, to check on the overall flow and vision. His specific comments on Chapter 9 helped us focus this chapter and better link it to the earlier material. We are very grateful for his help and very lucky to be able to count on him.

On the production side, we thank Connie Woodall for her wonderful help with some of our complex graphics. We are very thankful to David Olson for his ability to buffer Carol from the noise at work and allow her time to write. Charlie Crawford provided excellent technical writing advice, and we thank Charlie, Kevin Jones, and Blackstone and Cullen, Inc. for our ftp site, which was very helpful in allowing us to exchange files easily. Alisha Waller did us a tremendous favor by technically editing the entire book just before it was submitted for publication. Alisha and Charlie were both more than wonderful during the entire process, but especially toward the end, when finishing this book became the focus of our lives. We appreciate their support more than we can say. We also appreciate the support and understanding of our friends and colleagues (most notably Linda Pickle and the other coauthors on the "best practices in spatial statistics" guide, Oli Schabenberger, and Linda Young) as we greatly delayed joint work with them while we were working on this book. We thank the staff at the Caribou Coffee, LaVista at North Druid Hills, Atlanta, Georgia, for their hospitality and the gallons of coffee and wonderful lattes we've enjoyed over the last five years while we met to discuss the book.

All of the people listed above provided comments that improved the presentation of the material found here, but we ultimately retain the final responsibility for any errors, typos, or muddy descriptions that remain.

Finally, we owe a debt of gratitude to each other. At one point in the final months of writing, Lance remarked,"It's a miracle we don't hate each other yet!" While we knew we shared a similar philosophy on applied statistics, we have different backgrounds and different research areas. Thus, we needed to blend our ideas into something of which we could both be proud. Through our collaboration, the book contains far more details than if Lance wrote it alone, and far fewer than if Carol wrote it alone, yielding a balance we hope is much better for the readers than anything either of us could have produced without the other.

L. A. W.
C. A. G. C.

CHAPTER 1

Introduction

Time, space, and causality are only metaphors of knowledge, with which we explain things to ourselves.

Friedrich Nietzsche (1844–1900)

It is part of human nature to try to discover patterns from a seemingly arbitrary set of events. We are taught from an early age to "connect the dots," learning that if we connect the right dots in the right way, a meaningful picture will emerge. People around the world look to the night sky and create patterns among the stars. These patterns allow navigation and provide a setting for a rich variety of mythologies and world views. In scientific studies, formalized methods for "connecting the dots" provide powerful tools for identifying associations and patterns between outcomes and their putative causes. In public health, identification and quantification of patterns in disease occurrence provide the first steps toward increased understanding and possibly, control of that particular disease.

As a component of the pattern observed, the location *where* an event happens may provide some indication as to *why* that particular event occurs. Spatial statistical methods offer a means for us to use such locational information to detect and quantify patterns in public health data and to investigate the degree of association between potential risk factors and disease. In the nine chapters of this book, we review, define, discuss, and apply a wide variety of statistical tools to investigate spatial patterns among data relating to public health.

1.1 WHY SPATIAL DATA IN PUBLIC HEALTH?

The literature uses the phrases *geographical epidemiology*, *spatial epidemiology*, and *medical geography* to describe a dynamic body of theory and analytic methods concerned with the study of spatial patterns of disease incidence and mortality. Interest in spatial epidemiology began with the recognition of maps as useful tools for illuminating potential "causes" of disease.

Applied Spatial Statistics for Public Health Data, by Lance A. Waller and Carol A. Gotway
ISBN 0-471-38771-1

Dr. John Snow's study of London's cholera epidemic in 1854 provides one of the most famous examples of spatial epidemiology. Snow believed that cholera was transmitted through drinking water, but at the time, this theory was met with extreme skepticism (Snow 1855; Frerichs 2000). Although the cholera deaths appeared to be clustered around the Broad Street public water pump, Snow could not find any evidence of contamination at that particular pump. His contemporary critics noted that people tended to live close to public drinking water supplies, so the clustering observed could simply have been due to the population distribution: outbreaks occur where people are. However, by considering a few carefully selected controls (i.e., people nearby that did not have cholera) and by interviewing surviving members of almost every household experiencing a cholera death, Snow eventually gathered support for his theory. Brody et al. (2000) provide a detailed history of the role of maps (by Snow and others) in the investigation of the 1854 outbreak.

Other early examples of spatial epidemiology include the study of rickets made by Palm (1890), who used maps to delineate the geographical distribution of rickets. Palm observed the greatest incidence in industrial urban areas that had a cold and wet climate. Today we know that rickets is caused by a vitamin D deficiency, which in turn can be caused by a lack of ultraviolet radiation. In a related but more recent study, Blum (1948) surmised sunlight as a causal factor for skin cancer, again based primarily on the geographical distribution of disease cases observed.

Clearly, where people live can be of great importance in identifying patterns of disease. However, spatial analyses in public health need not pertain solely to *geographical distributions* of *disease*. The spatial distributions of the sociodemographic structure, occupational patterns, and environmental exposures of a population are also of particular interest.

1.2 WHY STATISTICAL METHODS FOR SPATIAL DATA?

Although best known among spatial analysts for the Broad Street maps, it was Dr. Snow's careful case definition and analysis of cholera deaths in a wider area of London that placed him among the founders of epidemiology rather than from his maps per se (Lilienfeld and Stolley 1984, pp. 28–29; Hertz-Picciotto 1998, pp. 563–564; Rothman and Greenland 1998, p. 73). Central to this analysis was Snow's "natural experiment," wherein he categorized cholera deaths by two water companies, one drawing water upstream from London (and its sewage), the other downstream. The water company service was so intermingled that "in many cases a single house has a supply different from that on either side" (Snow 1936, p. 75). Thus, in addition to maps, study design and simple statistics were important tools in Snow's analysis.

The analysis of spatial public health data involves more than just maps and visual inference. Medical science provides insight into some specific causes of disease (e.g., biological mechanisms of transmission and identification of infectious agents); however, much remains unknown. Furthermore, not all persons experiencing a

suspected causal exposure contract the disease. As a result, the analysis of public health data often builds from the statistical notion of each person having a *risk* or probability of contracting a disease. The analytic goal involves identification and quantification of any exposures, behaviors, and characteristics that may modify a person's risk. The central role of probabilities motivates the use of *statistical methods* to analyze public health data and the use of *spatial statistical methods* to (1) evaluate differences in rates observed from different geographic areas, (2) separate pattern from noise, (3) identify disease "clusters," and (4) assess the significance of potential exposures. These methods also allow us to quantify *uncertainty* in our estimates, predictions, and maps and provide the foundations for statistical inference with spatial data. Some spatial statistical methods are adaptations of familiar nonspatial methods (e.g., regression). However, other methods will most likely be new as we learn how to visualize spatial data, make meaningful maps, and detect spatial patterns.

Applying statistical methods in a spatial setting raises several challenges. Geographer and statistician Waldo Tobler summarized a key component affecting any analysis of spatially referenced data through his widely quoted and paraphrased *first law of geography*: "Everything is related to everything else, but near things are more related than far things" (Tobler 1970). This law succinctly defines the statistical notion of (positive) *spatial autocorrelation*, in which pairs of observations taken nearby are more alike than those taken farther apart. Weakening the usual assumption of independent observations in statistical analysis has far-reaching consequences. First, with independent observations, any spatial patterns are the result of a spatial *trend* in the probabilistic expected values of each observation. By allowing spatial correlation between observations, observed spatial similarity in observations may be due to a spatial trend, spatial autocorrelation, or both. Second, a set of correlated observations contains less statistical information than the same number of independent observations. Cressie (1993, pp. 14–15) provides an example of the reduction in *effective sample size* induced by increasing spatial autocorrelation. The result is a reduction in statistical precision in estimation and prediction from a given sample size of correlated data compared to what we would see in the same sample size of independent observations (e.g., confidence intervals based on independent observations are too narrow to reflect the appropriate uncertainty associated with positively correlated data). Ultimately, all statistical methods for spatial data have to take the spatial arrangement, and the resulting correlations, of the observations into consideration in order to provide accurate, meaningful conclusions.

1.3 INTERSECTION OF THREE FIELDS OF STUDY

We focus this book on statistical methods and assume that our readers have a familiarity with basic probabilistic concepts (e.g., expectation, variance, covariance, and distributions) and with statistical methods such as linear and logistic regression (including multivariate regression). Most of the methods presented in the book

Table 1.1 Representative List of Journals That Regularly Contain Articles on Spatial Statistical Methods Useful in the Analysis of Public Health Data

Field of Study	Journals
Statistics/Biostatistics	*Applied Statistics* *Biometrics* *Biometrika* *Environmetrics* *Journal of the American Statistical Association* *Journal of the Royal Statistical Society, Series A* *Journal of the Royal Statistical Society, Series B* *Statistics in Medicine* *Statistical Methods in Medical Research*
Epidemiology	*American Journal of Epidemiology* *Epidemiology* *International Journal of Epidemiology* *Journal of Epidemiology and Community Health*
Geography/Geology	*Annals of the Association of American Geographers* *Environment and Planning A* *Health and Place* *International Journal of Geographic Information Science* *Journal of Geographical Systems* *Mathematical Geology* *Social Science and Medicine*

build from these concepts and extend them as needed to address non-Gaussian distributions, transformations, and correlation assumptions.

Even though our focus is on statistical methods, we recognize that the analysis of spatially referenced public health data involves the intersection of at least three traditionally separate academic disciplines: statistics, epidemiology, and geography. Each field offers key insights into the spatial analysis of public health data, and as a result, the literature spans a wide variety of journals within each subject area. Table 1.1 lists several journals that regularly contain articles relating to the spatial analysis of health data.

Although by no means exhaustive, the journals listed in Table 1.1 provide a convenient entry point to the relevant literature. In our experience, journal articles tend to reference within a subject area more often than between subject areas, so searches across disciplines will probably reveal a wider variety of related articles than searches conducted on journals within a single discipline.

At times, the relationship between statistics and the fields of both epidemiology and geography is less than cordial. Often, a backlash occurs when statisticians attempt to transfer a family of methods wholesale into a new area of application without input from the subject-matter experts regarding the appropriateness of assumptions, the availability of requisite data, and even the basic questions of interest. We refer readers interested in such debates to Bennett and Haining (1985),

Openshaw (1990), and Rothman and Greenland (1998, Chapters 2 and 12). As always, there are two sides to the story. An equal amount of criticism also occurs when epidemiologists and geographers use and extend statistical methods without fully appreciating the assumptions behind them or the theoretical foundations on which their validity is based. Often, this just results in inefficiency (and underutilized and annoyed statisticians!) but there are times when it also produces strange inconsistencies in analytical results and erroneous or unsubstantiated conclusions. As applied spatial statisticians, we appreciate both sides and attempt to walk a fine line between emphasizing important assumptions and theoretical results and focusing on practical applications and meaningful research questions of interest.

1.4 ORGANIZATION OF THE BOOK

Many spatial statistics books (e.g., Upton and Fingleton 1985; Cressie 1993; Bailey and Gatrell 1995) organize methods based on the type of spatial data available. Thus, they tend to have chapters devoted to the analysis of spatially continuous data (e.g., elevation and temperature, where we can potentially observe a point anywhere on Earth), chapters devoted to statistical methods for analyzing random locations of events (e.g., disease cases), and chapters devoted to the analysis of *lattice data*, a term used for data that are spatially discrete (e.g., county-specific mortality rates, population data).

Although the data type does determine the applicable methods, our focus on health data suggests an alternative organization. Due to the variety of disciplines interested in the spatial analysis of public health data, we organize our chapters based on particular questions of interest. In order to provide some common ground for readers from different fields of study, we begin with brief introductions to epidemiologic phrases and concepts, components and sources of spatial data, and mapping and cartography. As statisticians, we focus on reviews of statistical methods, taking care to provide ongoing illustrations of underlying concepts through *data breaks* (brief applications of methods to common data sets within the chapters outlining methodologies). We organize the methods in Chapters 2–9 based on the underlying questions of interest:

- *Chapter 2*: introduction to public health concepts and basic analytic tools (*What are the key elements of epidemiologic analysis?*)
- *Chapter 3*: background on spatial data, basic cartographic issues, and geographic information systems (*What are the sources and components of spatial data, and how are these managed?*)
- *Chapter 4*: visualization of spatial data and introductory mapping concepts (*How do we map data effectively to explore patterns and communicate results?*)
- *Chapter 5*: introduction to the underlying mathematics for spatial point patterns (*How do we describe patterns mathematically in spatially random events?*)

- *Chapter 6*: methods for assessing unusual spatial clustering of disease in point data (*How do we tests for clusters in collections of point locations for disease cases?*)
- *Chapter 7*: methods for assessing spatial clustering in regional count data (*How do we test for clusters in counts of disease cases from geographically defined areas?*)
- *Chapter 8*: methods for exposure assessment and the analysis of environmental data (*How do we spatially interpolate measurements taken at given locations to predict measurements at nonmeasured locations?*)
- *Chapter 9*: methods for regression modeling using spatially referenced data (*How do we quantify associations between spatially referenced health outcomes and exposures?*)

Collectively, we hope these questions and the methodology described and illustrated in each chapter will provide the reader with a good introduction to applied spatial analysis of public health data.

CHAPTER 2

Analyzing Public Health Data

Disease generally begins that equality which death completes.
Samuel Johnson (London, September 1, 1750),
quoted in the *Columbia Encyclopedia*

Any important disease whose causality is murky, and for which treatment is ineffectual, tends to be awash in significance.
Susan Sontag, *Illness as Metaphor*, 1979, Vintage Books, Ch. 8

The results of studies of health and related risk factors permeate the public health literature and the popular press. We often read of associations between particular diseases (e.g., cancers, asthma) and various "exposures" ranging from levels of various environmental pollutants, to lifestyle factors such as diet, to the socioeconomic status of persons at risk. Although some studies involve carefully controlled experiments with random assignment of exposures to individuals, many involve *observational* data, where we observe disease outcomes and exposures among a subset of the population and want to draw inferences based on the patterns observed.

The analysis of public health data typically involves the concepts and tools of *epidemiology*, defined by MacMahon and Pugh (1970) as the study of the distribution and determinants of disease frequency. In this chapter we provide a brief review of assumptions and features of public health data, provide an outline of the basic toolbox for epidemiological analysis, and indicate several inferential challenges involved in the statistical analysis of such data.

2.1 OBSERVATIONAL VS. EXPERIMENTAL DATA

In most cases, epidemiological analyses are based on observations of disease occurrence in a population of people "at risk." Typically, we want to relate occurrence patterns between collections of people experiencing different levels of exposure to some factor having a putative impact on a person's risk of disease. Such *observational studies* differ in several important ways from *experimental studies* common in other fields of scientific inquiry. First, experimental studies attempt to control all factors that may modify the association under study, while observational studies

Applied Spatial Statistics for Public Health Data, by Lance A. Waller and Carol A. Gotway
ISBN 0-471-38771-1

cannot. Second, most experimental studies randomize assignment of the factors of interest to experimental units to minimize the impact of any noncontrolled concomitant variables that may affect the relationship under study. Observational studies step in where experimental studies are infeasible due to expense or ethical concerns. For example, studying a very rare disease experimentally often involves huge recruitment costs; withholding a treatment with measurable impact often violates ethical research standards. Whereas controlled randomization of assignment of a potential treatment within the confines of a clinical trial may be a justifiable use of human experimentation, random assignment of exposure to a suspected carcinogen for the purposes of determining toxicity is not.

The presence of controlled environments and randomization in experimental studies aims to focus interpretation on a particular association while limiting the impact of alternative causes and explanations. Observational studies require more care in analysis and interpretation, since controlled environments and randomization often are not possible. Consequently, observational studies involve potential for a wide variety of misinterpretation. The nature of observational studies, particularly of epidemiological studies in the investigation of determinants of disease, provides a framework for interpretation for most spatial analyses of public health data. Central to this framework is the quantification of patterns in the frequency of disease occurrence among members of the population under observation.

2.2 RISK AND RATES

The study of disease in a population begins by addressing the occurrence of a particular outcome in a particular population over a particular time. A common goal of an epidemiological study is to determine associations between patterns of disease occurrence and patterns of exposure to hypothesized risk factors. Due to the central nature of disease occurrence summaries in epidemiology, the related literature contains very specific nomenclature for such summaries. We outline the basic ideas here, referring interested readers to epidemiology texts such as Selvin (1991, Chapter 1) or Rothman and Greenland (1998, Chapter 3) and the references therein for more detailed discussion.

2.2.1 Incidence and Prevalence

The first distinction contrasts disease incidence and disease prevalence. *Incidence* refers to the occurrence of *new* cases within a specified time frame and provides a view of onset within a relatively narrow window of time. *Prevalence* refers to the total number of *existing* cases over a specific time frame and provides a summary of the current burden of the disease under study within the population. For a given disease, incidence and prevalence differ when diseased individuals survive for long periods of time, so that prevalent cases include people who recently contracted the disease (incident cases) and people who contracted the disease some time ago. For diseases with a high likelihood of subsequent mortality in a relatively short time span, incidence and prevalence will be similar. Most epidemiological applications

favor incidence over prevalence as an outcome in order to assess factors influencing disease onset, since both onset and duration influence prevalence. However, in cases where onset is difficult to ascertain (e.g., congenital malformations, infection with HIV), researchers may use prevalence coupled with assumptions regarding disease duration as a surrogate for incidence (Rothman and Greenland 1998, pp. 44–45).

2.2.2 Risk

The *risk* of contracting a disease represents the probability of a person contracting the disease within a specified period. We stress that in this context, risk is an attribute of a person, determined and modified by characteristics such as age, gender, occupation, and diet, among other *risk factors*. Risk is an unobserved and dynamic quantity that we, the researchers, wish to estimate. A primary goal of an epidemiological study is to summarize the level of risk of a particular disease in a particular population at a particular time. Associated with this goal is that of identifying factors influencing risk, and quantifying their impact, through observation of disease occurrence within a study population. The statistical question becomes one of estimating risk and related interpretable quantities from observations taken across this study population.

2.2.3 Estimating Risk: Rates and Proportions

In general use the term *rate* defines the number of occurrences of some defined event per unit time. However, application to disease incidence raises some complications, and the epidemiologic literature is quite specific in definitions of disease rates (Elandt-Johnson 1975; Rothman and Greenland 1998, pp. 31–37). Unfortunately, the literature on spatial data analysis applied to health data is not similarly specific, resulting in some potential for misunderstanding and misinterpretation. Although we review relevant issues here, our use of the term *disease rate* in this book falls somewhere between the strict epidemiologic definition(s), and the general use in the spatial epidemiological literature, for reasons outlined below.

In an observational setting, subjects under study may not be at risk for identical times. People move from the study area, are lost to follow-up, or die of causes unrelated to the disease under study. As a result, the epidemiological definition of *incidence rate* is the number of incident (new) cases observed in the study population during the study period divided by the sum of each person's observation time. We often refer to the denominator as a measure of *person-time*, reflecting the summation of times over the persons under observation. Rothman and Greenland (1998, p. 31) note that person-time differs from calendar time in that person-time reflects time summed over several people during the same calendar time rather than a sequential observation of people. In epidemiological studies of chronic, nonrecurring diseases, a person's contribution to person-time ends at onset, since at that point, the person is no longer among the population of people at risk for contracting the disease.

Under the person-time definition, a disease rate is not an estimate of disease risk. In fact, the person-time rate is expressed in inverse time units (often written

as "cases/person-year") and, technically, has no upper limit. Although a population of 100 persons can only experience 100 cases of a nonrecurring disease, these cases could happen within any person-time period (e.g., 10, 100, or 10,000 person-years), affecting the magnitude of the incidence rate.

In contrast to the precise epidemiological use of *rate*, the spatial epidemiology literature (including journals in statistics, biostatistics, and geography) tends to use *disease rate* to refer to the number of incident cases expected per *person* rather than per unit of *person-time*. That is, this use of *disease rate* refers to the total number of cases observed divided by the total number of people at risk, both within a fixed time interval. Technically, this usage corresponds to a incidence *proportion* rather than a *rate*, but is very common because this incidence proportion is a population-based estimate of (average) individual risk within the study population. We note that the time interval provides critical context to interpretation of an incidence proportion, as we expect very different values from data collected over a single year and that collected over a decade (since the numerator of the ratio increases each year but the number at risk is fairly stable and often assumed constant).

The primary differences between the incidence proportion and the incidence rate lie in assumptions regarding each person's contribution to the denominator of the ratio under consideration. In a closed population (no people added to or removed from the at-risk population during the study period) where all subjects contribute the same observation time, the incidence proportion would be equal to the incidence rate multiplied by the length of the (common) observation time for each person. Some difference between the two quantities always remains since a person stops contributing person-time to the denominator of the incidence rate the moment that person contracts the disease. However, this difference between the incidence rate and incidence proportion diminishes with rare diseases in the population at risk and/or short observation time per person (i.e., with less loss of observed person-time per diseased person). This feature represents one of several instances outlined in this chapter where the assumption of a *rare disease* (disease with low individual risk) provides convenient numerical approximations. (See the exercises at the end of this chapter to assess the impact of the precise rarity of a disease on the performance of some approximations.)

For the remainder of the book we take care to clarify our use of the term *disease rate* in any given instance. In most cases we follow the spatial literature in using the term to refer to incidence proportion and appeal to an assumption of a rare disease to justify this use for most of our examples. However, applications of the spatial statistical techniques outlined in subsequent chapters to more common diseases require a more careful wording and interpretation of results.

2.2.4 Relative and Attributable Risks

Incidence proportions provide an estimate of the average disease risk experienced by members of a study population. Often, analytic interest centers around comparing risks between individuals with and without a certain exposure. We define