Chapter 7

SPATIAL ANALYSIS OF DISEASE

Linda Williams Pickle
National Cancer Institute

1. INTRODUCTION

Cancer rate comparisons around the world suggest clear geographic differences that have only recently been appreciated and evaluated by statistical methods. The goal of this chapter is to briefly review the progression of the spatial analysis of disease from simple dot maps and crude rate comparisons to the complex hierarchical spatial models used today. After providing a historical background and necessary epidemiologic fundamentals, we summarize available methods for the exploration, hypothesis testing, and modeling of spatial data. Although the focus here is on methods appropriate for cancer research, other related methods will be mentioned.

2. HISTORY OF THE SPATIAL ANALYSIS OF DISEASE, WITH AN EMPHASIS ON CANCER

2.1 An Early History of Mapping

The earliest maps of disease were produced over two hundred years ago. Although John Snow's dot maps of cholera cases in London, first published in

1855 (Snow 1855), are the most well known, research has uncovered a "spot map" of yellow fever in New York published by Seaman in 1798 and an unpublished disease map of the world produced by Finke in 1792 (Barrett 2000).

These first disease maps identified the locations of the residences of cases by either a dot or a small bar, e.g., a version of Snow's map has a small stacked bar graphic at each street address representing the number of cholera victims at that house (McLeod 2000). Patterns were discerned by visual inspection and the case locations were compared to those of suspected risk factors, such as the locations of London water pumps during the cholera epidemic. Differential shading was used to indicate levels of cholera mortality in 1852 (Peterman 1852). The first maps of cancer appeared in 1875 (Haviland 1875), notable because of the use of color, although the use of red for low rates and blue for high rates is the reverse of today's convention. For more details of this fascinating history, the reader is referred to a review by Howe (Howe 1989).

2.2 Beginnings of Disease Pattern Analysis

Spatial analysis through the early twentieth century was hampered not only by a lack of appropriate statistical methods, but by a lack of data. Identification of cases for the early maps described above was made by individual physicians, and rates could not be calculated in the absence of area-wide population enumeration. Stocks first adjusted cancer rates for 1921-23 by age, sex, and urban distribution in English counties and later showed standardized mortality ratios on choropleth maps (Stocks 1928; Stocks 1936). When health outcomes data became available on a national level, e.g., required certification of each death in the U.S. beginning in 1933 and the start of the National Health Service in the U.K. in 1948, statistical methods for their analysis soon followed.

The following decade saw the development of statistical methods to evaluate clustering (Moran 1948; Geary 1954), to measure disease-risk associations (e.g., Bross 1954) and their asymptotic variances (e.g., (Woolf 1955)), the development of the logistic model (reviewed in Cox 1970), the application of the relative risk concept to case-control data (Cornfield 1951), and the definition of the Poisson cluster process (Neyman and Scott 1958). This early work was extended to the detection of disease clustering (Mantel 1967) and space-time interactions (Knox 1964) and the logistic model was extended to more complex studies (Walker and Duncan 1967).

2.3 Cancer Atlases

The modern-day atlas began with Howe's *National Atlas of Disease Mortality in the United Kingdom* (Howe 1963). The first cancer atlases in the U.S. mapped 34 types of cancer at the small area level (Mason et al. 1975). Burbank tested U.S. state rates for space-time clustering (Burbank 1972) and several later atlases also included a measure of spatial clustering (Ohno and Aoki 1981; Kemp et al. 1985; Le et al. 1996). The second generation of cancer atlases included results of model-based procedures, such as a time trends map based on a straightforward Poisson regression model (Pickle et al. 1987; Pickle et al. 1990) and an atlas which mapped empirical Bayes estimates (Riggan et al. 1987). An injury atlas presented rates predicted by a constrained empirical Bayes procedure developed to improve the fit of the model to the observed rates (Devine et al. 1991). All of the U.S. atlases have presented age-adjusted rates but a recent atlas of the leading causes of death in the U.S. also included predicted age-specific maps and regional rates resulting from a mixed effects model (Pickle et al. 1996a). The most recent cancer atlas focused on the presentation of observed rates (Devesa et al. 1999).

With a few exceptions noted above, the early atlases had relied upon the readers' perception of visual patterns to identify salient features on the maps. Only recently has attention turned to the evaluation of map design features (Pickle and Herrmann 1995), but clearly the characteristics of a map, such as color and cutpoint choices, can have an important impact on its apparent patterns. Thus reliance on maps alone could lead to different interpretations of the same data, depending on the presentation method. A review of data visualization methods is beyond the scope of this chapter, but the reader should be aware of the potential impact of a map's design on perceived spatial patterns.

2.4 Epidemiologic Studies of Geographic Patterns

2.4.1 Ecologic Studies

Because the first U.S. atlas showed surprising concentrations of high cancer rates in certain regions of the country, it was followed by a series of ecologic regression studies that identified associations between cancer death rates and various sociodemographic and occupational factors. Although these studies

found plausible associations between cancer and purported risk factors, such as certain manufacturing industries, etiologic field studies often found other explanations for the high cancer rates. For example, lung cancer death rates among white men were high during 1950-69 in scattered cities along the south Atlantic and Gulf coasts (Mason et al. 1975). A correlation study showed an association with the paper, chemical, petroleum and transportation industries (Blot and Fraumeni Jr. 1976), but several case-control interview studies also found an association with the shipbuilding/ship repair industry. In addition to cigarette smoking patterns, employment in this industry during World War II was a major risk factor for lung cancer in these port cities, most likely because of the workers' asbestos exposure (Blot et al. 1978). The war had occurred between censuses of U.S. manufacturers, and so data available for the correlation studies did not record the high shipyard employment in these cities. Thus the failure of many of the correlational studies to pinpoint the causes of the high cancer rates may have been due in part not only to the ecological fallacy, i.e., where associations between disease and risk factor exposure differ among individuals compared to those among aggregated groups, but also to unmeasured risk factors for small geographic areas.

2.4.2 Etiologic Studies

The etiologic studies that followed avoided the potential ecological bias by gathering data from individual cancer cases and controls. The analysis of these data was generally by logistic models, where $\log\left(\frac{\pi}{1-\pi}\right) = \mathrm{X}'\beta + \log(OR)$, where π = P(individual was exposed | individual was a case), X is a matrix of confounding variables, β is a vector of corresponding coefficients, and *OR* is the odds ratio, an approximation to the relative risk of disease due to this exposure. Parameters were estimated by maximum likelihood or the least squares approach. In a few of these studies, distance to a suspected carcinogenic polluting source was calculated, but generally these regression models did not account for spatial adjacency or nearness.

2.4.3 Cancer Clusters

The clustering of cancer cases has long been suspected by the public, but the confirmation and search for causes of such clusters have typically been disappointing. For example, a cluster of childhood leukemia cases in Seascale, U.K., has been widely studied, but no clear cause has been identified (Draper et al. 1993). Statistical power is low to detect these small clusters, unless the underlying risk is quite high, and relevant historic environmental and personal exposure measures often are not available (Najem and Cappadona 1991). Cancer surveillance around suspected carcinogenic point sources has proven more fruitful, despite the time and expense involved. For example, followup studies of the atomic bomb survivors in Japan have yielded extensive information on radiation carcinogenesis (Beebe G.W. et al. 1971). Likewise, health studies of residents near the Chernobyl nuclear power plant have found an excess of childhood thyroid cancer (Astakhova et al. 1998), but no cancer clusters were confirmed around the Love Canal, NY, toxic waste site (Janerich et al. 1981). In a study currently underway on Long Island, NY, the first Congressionally-mandated Geographic Information System will be used in an attempt to discover the reason for an excess of breast cancer there (Kulldorff 1997; Varon 1998). Statistical methods for cluster detection will be discussed in Section 6.

2.5 Related Statistical Methods

Other statistical methods for spatial analysis developed in parallel to those in epidemiology. Geostatistical methods arose from the need to interpolate and predict in the geologic sciences, for example, to produce a surface rendition of soil content or to predict where oil drilling would be successful. Trend surface analysis by polynomial regression surfaces and kriging will be discussed in Section 5. These methods initially were for lattice point data, such as the soil content from regularly-spaced samples. Extensions allowed application to irregularly-spaced data. Prediction models were also developed for small area (e.g., state) estimation from national survey data (for a review of these methods, see Ghosh and Rao 1994). The goal of small area estimation is to predict responses in non-sampled areas, similar to geostatistics, but the method

includes explanatory covariates in the regression model and ignores any spatial correlation in the data.

2.6 A Convergence of Methods

The recent dramatic improvements in computational speed have made possible a merging of features of these methods from epidemiology, geostatistics, and survey sampling to provide powerful new methods for the spatial analysis of disease patterns (for example, see Ghosh et al. 1998). Fully Bayesian estimation employing Monte Carlo techniques can be used to predict multi-dimensional disease patterns and to provide more realistic significance levels of statistical tests. After considering the features and limitations of health data, the remainder of this chapter will examine statistical methods for the estimation, exploratory analysis, hypothesis testing and modeling of disease data.

3. CHARACTERISTICS OF DISEASE DATA

3.1 Data Description

Spatial data may consist of point data, such as the locations of breast cancer patients, area data, such as the breast cancer rates by county, or line data, such as locations of roadways. Point data are usually irregularly spaced, but are sometimes aggregated to a regularly-spaced grid ("binning") for convenience or to maintain confidentiality of the data (see section 3.2). Environmental exposures may be represented as spatially continuous data or as data points at monitoring locations. Line data are rarely relevant for cancer research. An obvious analytic problem is how to handle combinations of different types of data, e.g., point locations of cancer cases, area-level demographic data and spatially continuous environmental data. A related problem is spatial mis-alignment, when variables are available at different geographic scales. Some interesting work is currently underway regarding how to correct for this, such as when population data must be measured on the same scale as the number of cases to permit the calculation of rates (Zhu and Carlin 2000).

The ecologic fallacy noted in section 2.4.1 is also a scaling problem, but where the data are available at geographic units larger than those for which we wish to make inferences. Typically, we wish to draw inferences about individuals but only have data about aggregated sets of individuals, such as the cancer incidence rate for each county. A slightly different scaling problem occurs when different variables in an analysis are available for different levels of geographic aggregation. Multilevel models can be constructed that take this geographic hierarchy into account. This will be discussed in section 7.

The health data of interest here are observational, i.e., no experiments have been performed to control for the many potential confounding variables related to human health. Experimental clinical trials are an exception, but spatial pattern is rarely of interest in these studies. Spatial sampling is a method often used in the earth sciences to ensure geographic representativeness, but this is often impractical for human studies. Thus the analyst needs to determine the representativeness of the health data for inferential purposes. For example, a sample of hospital patients may not be representative of all residents of an area but may be acceptable as a sample of all hospital patients. Epidemiologic case-control studies attempt to match the distribution of important individual characteristics of controls to cases. Even in these situations, the individuals may not be spatially representative, so that estimates made for aggregated geographic units may be biased; this is why we are unable to make direct inferences from survey data about geographic units that are smaller than those included in the sample design. For example, the National Health Interview Survey is a nationwide survey which samples from over two thirds of U.S. states. Resulting estimates from this survey are only provided for four broad regions and, in the most recent design, for large metropolitan areas but not for states.

3.2 Data Limitations

A serious limitation of health data for spatial analyses is the restriction placed on these data because of confidentiality and privacy concerns. Actual addresses of patients are often not available for geocoding (assigning a specific spatial location), so the geographic locations of cases are often known only to a small administrative unit, such as zip code or census tract. Furthermore, even aggregated health data may not be released if there is a concern that, because

of small numbers, the identity of the patient could be determined. For example, the National Center for Health Statistics will not publish mortality data at the county level for single years for this reason. Some methods to mask the identity of the patient while still providing some measure of geographic accuracy are discussed in the accompanying chapter.

An additional problem is the lack of appropriate covariate data for the spatial analysis of cancer. Information on lifestyle factors is not available in hospital records, cancer registries or other common sources of cancer data. The Behavioral Risk Factor Surveillance System provides information on smoking, obesity, and other health risk factors at the state level (Nelson et al. 1998); recently county-level maps of these data have been developed (Pickle and Su 2001). Some relevant environmental data are available from monitoring sites but models of the dispersion of potentially hazardous pollutants through the air, water, or soil are needed to determine the potential for exposure by individuals living in a certain area. Even when such models exist, unmeasured personal habits can negate the exposure, such as the regular use of sunscreen and protective clothing when exposed to strong sunlight outdoors. A further problem for cancer studies is that for most types of cancer the lag between exposure and cancer diagnosis can be decades. Thus we need historic exposure data for analysis, something not available at a small area level even for the long-recognized risk factor cigarette smoking.

Data availability and quality are usually the limiting factors in a health data analysis. The quality of data is important for any study, but historical exposure data, if available at all, may be derived from administrative databases (e.g., hospital patient records) that were not designed for accurate collection of these data. Even the most direct sources of some data, the patients themselves or their next of kin, may not remember or report the necessary information accurately, or may refuse to provide it at all. New techniques utilizing satellite imagery, dispersion models for environmental pollutants, and Geographic Information Systems may soon offer a method to estimate individuals' environmental exposures, information heretofore unknown to them (Xiang et al. 2000).

3.3 Potential Analytic Problems

3.3.1 Stationarity and Isotropy

In order to draw valid statistical inferences, we must make certain assumptions about the spatial structure of the data. If the data-generating spatial process can vary at each data point, we have no repeated measurements from which to make inferences. In addition, the random variables in a spatial process are spatially correlated, at least locally, and so cannot be assumed to be independent. These difficulties may be avoided if stationarity can be assumed, i.e., that the process has a constant mean across the space and the covariance between random variables at two locations depends on their relative, not absolute, locations. Specifically, spatial structure may be described in terms of first and second order effects, or large-scale and small-scale effects, respectively: $\mu(s) = Z(s) + \varepsilon(s)$ where $\mu(s)$ represents the mean value at location *s*. If the large-scale effect $Z(s)$ is not a constant across the entire space, then the data may first be detrended by subtracting the large scale effect so that the resulting differences have a constant mean $(\mu(s) - Z(s) = \alpha)$. Under the stationarity assumption, the observed data are replicates of the same spatial process, and so can be used for statistical inference. The local effects (residuals from the mean process), $\varepsilon(s)$, are usually correlated because of the influence of common factors in a small neighborhood around *s*. This correlation structure is assumed, under stationarity, to be a function only of the distance and direction between two points, *s* and *s'*, not of their actual locations, that is $\text{cov}(Y(s), Y(s')) = C(s - s')$ where *C* is some covariance function. Note that a Gaussian spatial process is completely specified by these assumptions. A weaker form of stationarity, intrinsic stationarity, is defined as a spatial process with constant mean where $\text{Var}(Y(s), Y(s')) = 2\gamma(s - s')$ is again a function only of the relative locations of the two points. The semi-variogram, $\gamma(s - s')$, may be modeled to provide an estimate of the covariance structure (Cressie

1993). If in addition to these stationarity conditions, the covariance is not dependent on the direction between the two points, the process is said to be isotropic. In some situations, anisotropy can be corrected by a data transformation, in others the dependence on direction can be built into the covariance model.

3.3.2 Effects of Topography

In defining spatial neighbors, trends, distances and correlations, we usually ignore the real topography of the region. For example, is it correct to call Mississippi and Arkansas adjacent neighbors when the Mississippi River separates them? Should distances in mountainous areas be computed "as the crow flies" or along the more circuitous roadways? For airborne exposures, the former is probably best, but for measuring access to medical care, the latter seems more reasonable. There is no universal answer, but the analyst must consider these questions for each study.

Another problem in defining neighbors arises at the edge of the study area or at a coastline - no neighbors exist in one or more directions. Some non-parametric smoothing algorithms extrapolate data so as to create neighbors where there are none, but some of these algorithms are more adversely affected by edge effects than others (Kafadar 1994). These algorithms will be discussed further in section 5.

4. BASIC EPIDEMIOLOGIC ANALYSIS

4.1 Notation

In this section, notation and basic statistical measures for cancer rates and risks are described. Exploring patterns in point data will be discussed in section 6. For illustration of areal data, we consider cancer incidence, although the same methods pertain for mortality. Let d_{ij} represent the number of new cancer cases in place i, i=1,2,...,I, age group j, j=1,2,...,J, and let n_{ij} represent the corresponding population at risk. Then the observed age-place-specific cancer

incidence rate is $r_{ij} = d_{ij} / n_{ij}$. The crude place-specific rate, ignoring age, is $r_{i.} = d_{i.} / n_{i.}$ where the period (.) denotes summation over that subscript.

4.2 Adjusted rates and risks

The crude cancer rate, $r_{i.}$, is highly dependent on the age distribution of the population because most types of cancer occur predominantly in the elderly. A comparison of crude cancer rates for Utah and Florida would be meaningless because more cancer cases would be expected in Florida's older population. An age-adjusted rate is preferred to put the rates for different places on the same scale for proper comparison. The actual value of any standardized rate is only meaningful in comparison to other rates that have been standardized in the same way. The two methods most often used to adjust epidemiologic rates are the direct and indirect methods (Fleiss 1981).

The directly adjusted rate is $R_{(dir)i} = \sum_j u_j r_{ij}$ where u_j is the proportion of the standard population that is in the jth age group. Various standards are used, e.g., the total U.S. population in 1970 or a constructed world population. A relative measure, termed the Comparative Mortality (or Morbidity) Ratio, may be derived by dividing the directly adjusted rate by the standard's rate.

An alternative method for age adjustment is indirect standardization, where $R_{(ind)i} = \sum_j c_j n_{ij}$, where the age-specific rates in the standard population (c_j) are applied to the area-specific populations. From this, the Standardized Mortality (or Morbidity) Ratio (SMR) is calculated as $SMR_i = R_{(ind)i} / R_s$, where R_s is the rate in the standard population. The SMR may be interpreted as the ratio of observed to expected numbers of cases, where the expected number is determined from the standard population. SMRs for two places are directly comparable only if the proportionality condition holds , i.e., $r_{ij} = \alpha_j \theta_i$, otherwise, the SMRs are standardized to different populations (n_{ij} and $n_{i'j}$) (Pickle and White 1995a).

The advantage of the indirect method is that it may be used for sparsely populated areas which would have age-specific rates too unreliable for the direct method of standardization. However, as illustrated in the table below, the direct standardization method retains the rank order and the proportional differences of the age-specific rates between places. In this example, the age-specific rates for place B are three times those of place A for every age group. The ratio of the directly adjusted rates (B:A) is 3, as one would expect. The SMR calculations show that place B has a 50% excess of cases over what would be expected, whereas place A has no excess. This comparative inference is correct but it would not be correct to say that place B has a 50% greater risk than place A, as implied by the SMRs, when there is a threefold ratio of age-specific rates.

Table 1. Illustration of directly and indirectly standardized rates

	Populations			Rates			Observed # of cases	
age	A	B	standard	A	B	standard	A	B
1	10,000	40,000	250,000	0.001	0.003	0.005	10	120
2	20,000	30,000	200,000	0.002	0.006	0.004	40	180
3	30,000	20,000	150,000	0.003	0.009	0.003	90	180
4	40,000	10,000	50,000	0.004	0.012	0.002	160	120
Total:	100,000	100,000	650,000				300	600
	directly adjusted rates: (Per 100,000 population)			200	600			
	indirectly adjusted rates (SMR):			1.0	1.5			

The relative risk is the measure of disease risk due to a particular exposure, i.e., $\theta_i = \dfrac{\text{P(disease in place } i \mid \text{exposure in place } i)}{\text{P(disease in place } i \mid \text{no exposure in place } i)}$. The relative risk may be estimated directly from prospective studies, where persons who were and others who were not exposed to some risk factor are followed for some period of time to determine the probability that they will become diseased. For

case-control studies, where study subjects are chosen on the basis of their disease status, the odds ratio is used as an approximation to the relative risk when the disease is rare (see section 2.4.2). The relative risk or odds ratio is typically estimated from logistic regression models which adjust for potential confounders, i.e., variables that alter the association between disease and the risk factor of interest.

4.3 Overdispersion

Assume that the number of cases d_{ij} is a Poisson random variable with mean $\lambda_{ij} n_{ij}$. Poisson regression may be used to model the effect of explanatory variables on the rates, i.e., $\log\left(\lambda_{ij}\right) = \mathrm{X}_{ij}'\beta_{ij}$ and homogeneity of rates over age and/or place may be modeled by a suitably reduced parameter vector. Overdispersion of the d_{ij} is common and can result, for example, from rate heterogeneity, spatial correlation, or other clustering in the underlying population. Under an assumption that the data result from clusters and that the cluster size *k* is fixed, McCullagh and Nelder have shown that the d_{ij} are approximately binomially distributed. The degree of overdispersion can then be estimated by , $\tilde{\sigma}^2 = \frac{1}{IJ - p}\sum_{ij}\frac{(d_{ij} - E(d_{ij}))^2}{E(d_{ij})(1-\hat{\lambda})}$ where p is the number of parameters estimated in the model (McCullagh and Nelder 1983) p.127).

Because cancer rates are relatively low in the general population $(1-\hat{\lambda}_{ij}) = 1$, so that $\tilde{\sigma}^2$ reduces to the familiar form of a goodness-of-fit statistic for Poisson counts. This estimator may be used to scale the likelihood function and to adjust for overdispersion in hypothesis tests.

5 EXPLORATORY ANALYSIS

5.1 Basic Tools

As a first step in the analysis, plots and maps of disease data can show differences across the geographic units. For example, boxplots of small area rates by region can point to a broad spatial trend or differences in variation by region. Maps can be tied directly to plots, e.g., to show the rank of state cancer rates and their spatial distribution in a single graphic (Carr and Pierson 1996). Maps conditioning on potential risk factors and confounders can suggest the need for interactions in subsequent models. Now that mapping software has become commonplace, such plots and maps of the data can be quickly generated. In the remainder of this section, a number of smoothing methods are described that can highlight the large scale patterns in the data.

5.2 General Smoothing Methods

The primary purpose of two-dimensional smoothing algorithms is to remove background noise from the data so that the underlying spatial pattern can be seen. These methods can also be used to identify outliers, sometimes called “hot spots”, by subtraction of the smoothed surface from the original map, although this differencing method highlights random extreme points as well as truly unusual clusters. Nonparametric methods for estimating spatial trend were developed for point data, but these methods have also been applied to areal data by assigning the area’s value to its centroid and proceeding as if the random variable occurred at that point. In general, smoothing methods borrow information from neighboring places to improve the estimated value for each point.

A problem common to most smoothing methods is how to define spatial neighbors. Defining a neighbor is straight forward in a unidimensional problem, such as a time series, because there is a clear ordering of points to one side or the other of the point to be smoothed. In two dimensions, neighbors can be defined in a number of ways, such as those areas having centroids within a specified distance of the center or those that share a border with the area to be smoothed. These subjective neighborhood definitions can impact the analysis, particularly when areas vary greatly in size and shape.

The methods described in this section smooth values from a single random variable; none account for possible explanatory variables. Most do not permit inverse variance weights, making them inappropriate for rate and count data except perhaps as a crude first look at the patterns or for areas such as census tracts where population sizes are roughly equal. We review both linear and non-linear two-dimensional smoothers that are commonly recommended for spatial data.

5.3 Linear Smoothers

The simplest two-dimensional smoother is an average of all values within a distance *h* of each data point, repeated in turn for every point in the entire space. A similar disk averaging method that includes inverse distance weights, i.e., weights equal to the inverse of the distance h_{ij} (or squared distance) between two points *i* and *j*, provides a more gradual decline in weights with distance than the simple unweighted disk average method. Squared distance weights have been recommended for data with little structure because of the more rapidly declining weights, but this method trades off increased bias for decreased prediction variance (Kafadar 1994; (Cressie 1993), p.189). Another commonly-used method is LOESS, locally weighted linear regression, with weights constructed from a cubic function of distance (Cleveland and Devlin 1988). The proportion of points to be included as neighbors is a tuning constant set by the analyst for any of these methods.

5.4 Non-linear Smoothers

5.4.1 Response Surface Analysis

In contrast to the local smoothers described above, the purpose of response surface analysis is to model the entire spatial area as a continuous surface. This can be used for interpolation and prediction of non-sampled locations, for example to provide values of explanatory environmental variables for a regression model of disease rates. Given measurements at sampled points $\{Y_{ij}$, where *i* and *j* identify the spatial location$\}$, assume that the three-dimensional surface can be represented as $Y = A'\beta + e$ where A is a vector of location coordinates and e is the prediction error. Polynomial and spline functions (two-

dimensional piece-wise polynomials with a smoothing penalty) of the coordinates have been proposed to fit the trend surface, usually by ordinary least squares. The error covariance matrix V may be specified so that stationarity is not required for inference, but usually Y is assumed to be a multivariate normal random variable generated by a stationary, but not necessarily isotropic, spatial process (see section 3.3.1). The number of parameters required to fit the data well using these methods can be high, resulting in an ill-conditioned problem, particularly if the observed data points are not well spaced throughout the area to be modeled. For more details on these methods, see Haining (Haining 1990).

5.4.2 Kriging

Historically, the most commonly used global smoother, or trend surface analysis, has been kriging, which arose from a need to smooth geologic data for mining applications. Hence, kriging falls into a class of methods referred to as "geostatistics". Following the notation from the previous section, "simple" kriging assumes that E(e)=0 with known covariance matrix V and computes the weighted least squares estimate $\hat{\beta}$; i.e., this is also a weighted average type of smoother, with weights chosen to minimize the mean square error. This prediction is unbiased under the normal assumption, but is biased and sensitive to outliers if this assumption is violated (Haining 1990).

These are the theoretical underpinnings of kriging, but the covariance matrix is rarely known, so simple kriging is of little use. Simplifying assumptions must be made in order to estimate the covariance matrix and then fit the surface. "Ordinary" kriging assumes that Y has a constant mean ($A'\beta = \mu$) and a stationary spatial process (see section 3.3.1). "Universal" kriging allows a spatial trend in the data or ordinary kriging may be used after "detrending" the data by subtracting the mean values and then kriging the residuals.

5.4.3 An Illustration of Variogram Modeling

Even though the assumptions required for kriging are usually violated by health data, empirical variogram modeling may prove useful in exploring (a) data transformations that would reduce non-stationarity and anisotropy, (b) appropriate covariance structures for more complex modeling, and (c) the

spatial correlation that remains in model residuals (illustrated in the next chapter). We provide a simple example examining the stationarity assumption here; a more difficult problem is illustrated in the next chapter. For a more detailed discussion of kriging and variogram modeling, see Cressie (Cressie 1993).

As noted in section 3.3.1, the semi-variogram may be modeled to provide an estimate of the covariance matrix V. Models that are frequently used include the power, exponential, Gaussian, linear, sine wave, spherical and log-linear functions of distance. For illustration, we calculated the empirical semi-variogram for the rates of mortality due to all cancer, age-adjusted to the 1940 U.S. population, for white males, 1988-92 (Pickle et al. 1996a; Pickle et al. 1996b). All of the above models were fit to these data, up to a maximum of 1250 miles, weighting by $\sqrt{n/\hat{\gamma}}$ as Cressie suggests (Cressie 1993), where n is the number of observations in that distance bin. The variogram of the original data (Figure 1, top) shows a linearly increasing γ with increasing distance, strongly indicative of non-stationary data. After crudely removing the spatial trend in the data by using a generalized additive model of smoothed (LOESS) functions of latitude and longitude (Kaluzny et al. 1998), the log-linear distance function fit the data well (r^2=92%), i.e., $\gamma(h;\varphi) = 91.8 + 43.5\{\log(h)\}$ for distance h>0 (Figure 1, bottom). Because of the continuing increase of variance with increasing distance, this plot indicates slight non-stationarity, but is much improved over the original data plot. The spherical (stationary) function fit the data reasonably well overall (r^2=75%), but provided a poor fit to points for distances over 750 miles.

5.5 Other Non-linear Smoothers

Whereas the goal of response surface analysis is to fit an entire surface by a single parametric function, median polish and headbanging are two nonparametric methods proposed by Tukey for locally smoothing spatial data. For median polish, a grid is placed over the map and the values on the map are assigned to a grid cell, averaging if there are multiple values per cell (Cressie and Read 1989; Tukey 1977). Assuming independence of row and column effects, the smoothed value is the sum of the overall mean, row and column

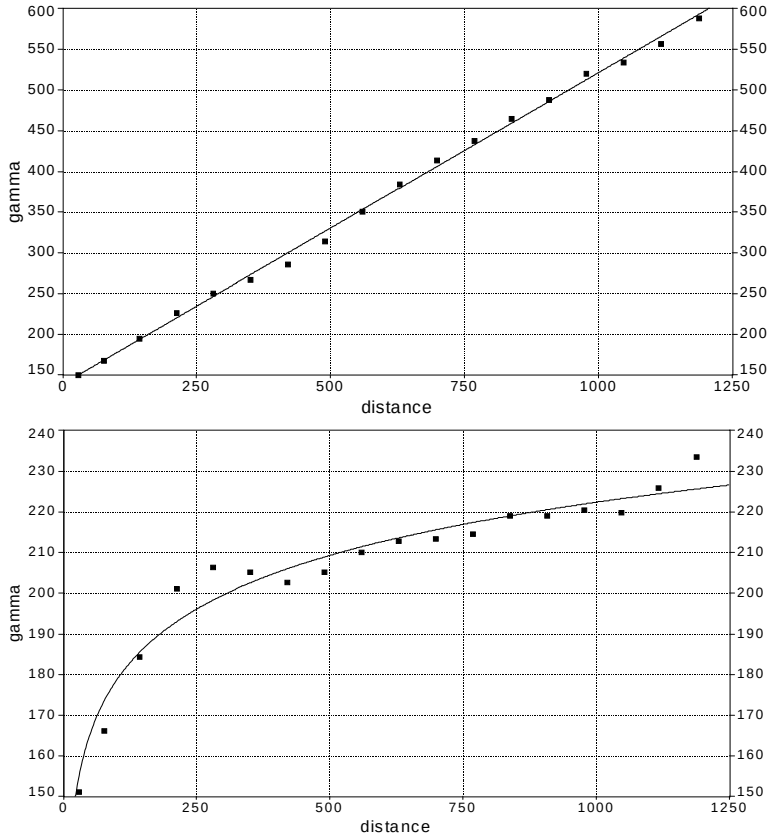

Figure 1. Empirical variogram of death rates due to all cancer among white men, 1988-92, before (top) and after (bottom) detrending the data. Note gamma scaling differences.

effects. This method, like the linear LOESS method described above, can be sensitive to the orientation of the map (Kafadar 1994).

Headbanging is a median-based smoothing algorithm where for every point a set of "triples" is identified consisting of that point plus two neighbors that are constrained to be nearly linear with the center point. The value of the center point is compared to the median of the lower-valued neighbors and the median of the higher-valued neighbors, and is changed ("smoothed") if it falls outside this range (Tukey and Tukey 1981; Hansen 1991).

5.6 Comparison of Smoothing Methods

Kafadar (Kafadar 1994) compared the performance of several of these linear and non-linear local smoothing algorithms. The simple weighted average

methods generally recovered the patterns in simulated data well, but the non-linear smoothers performed at least as well for data with ridge, peak and depression patterns and were less likely to falsely identify data structure (Kafadar 1994). Headbanging has also been shown to retain edge effects better than a linear method (Hansen 1991).

None of these smoothers, as originally published, account for hetero-skedasticity of the random variable to be smoothed. That is, the variance of both the number of cases and the disease rate in each place varies as a function of the population $n_{i.}$. For this reason these methods are inappropriate for smoothing disease counts or rates unless, as previously noted, populations are roughly equal for all areas. Mungiole et al. (Mungiole et al. 1999) added weights to the headbanging algorithm to overcome this deficiency.

Figure 2 illustrates both the effects of smoothing and the importance of including inverse variance weights when applicable. Figure 2a is a map of the observed age-adjusted rates of death due to HIV infection among white men during 1988-92 for 805 Health Service Areas (Pickle et al. 1996a; Pickle et al. 1996b). Counties are shaded according to quintiles of their rates on each map. High rates are scattered throughout coastal states and the midwest. After applying an unweighted headbanging algorithm, smoothing each area (centroid) using up to 30 nearest neighbors, rates are high in the urban northeast, along the Gulf coast, and in the southwest (Figure 2b). Figure 2c shows the weighting effect on smoothing, where the weights are approximately inversely proportional to the variance of the rates. Although the general pattern is similar to that of the unweighted map, the relative level of rates in several areas, e.g., Texas, differs according to whether weights are included.

High rates in urban places are particularly affected; after weighted smoothing, these rates remain high because they are considered reliable, while high rates in sparsely populated areas are smoothed toward those of neighboring areas. Considering rates in Minnesota as an example, the observed rate in Minneapolis-St. Paul was in the highest rate class but the unweighted algorithm smoothed it into the lowest class, similar to the rest of the state (Figures 2a, 2b).

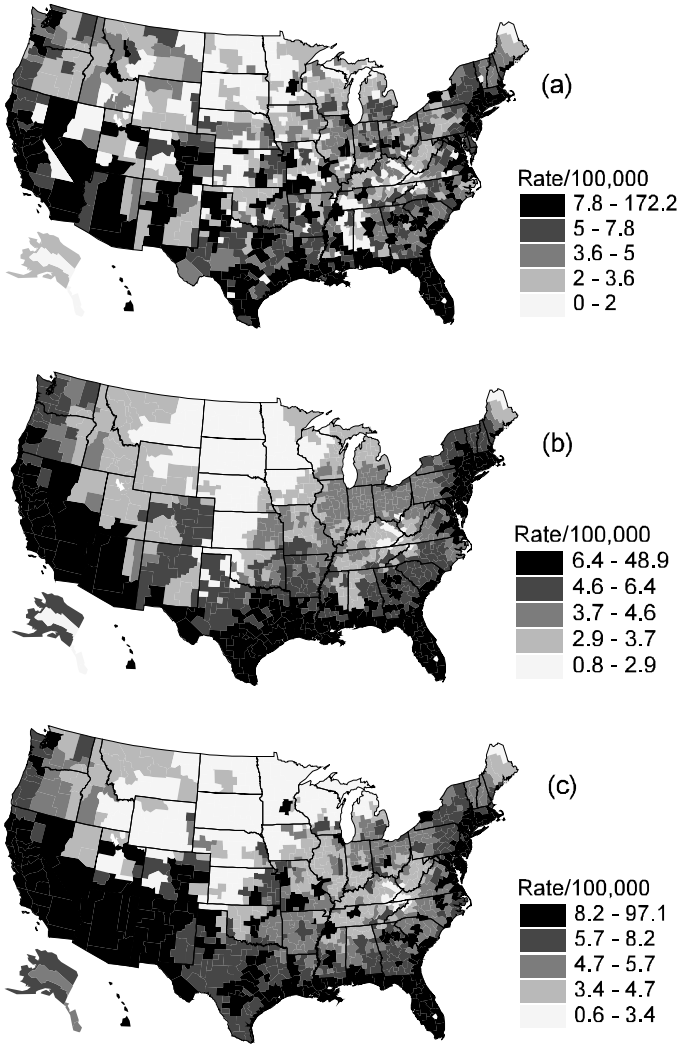

Figure 2. Age-adjusted mortality rates due to HIV infection among white males, 1988-92. (a) original data, (b) smoothed without weights, (c) smoothed with weights.

The inclusion of inverse variance weights, however, leaves the reliable Minneapolis-St. Paul rate in the highest rate class but smooths the less populated remainder of the state to the lowest class (Figure 2c). Similar improvements using other smoothing algorithms would be expected if their weights can be modified to include inverse variance as well as distance weights.

For example, a WEIGHT statement can be added to SAS PROC MIXED to implement weighted kriging (Littell et al. 1996a).

The methods described in this section can provide useful initial looks at both point and area data by removing background noise in order to reveal underlying large scale patterns. For this purpose, the weighted average methods and headbanging have been shown to perform well and are easy to implement. Any smoothing of rates or counts should include appropriate inverse variance weights to avoid smoothing away highly reliable values.
Response surface analysis seems less useful for health data, except perhaps for estimating environmental exposures over a space from discrete samples, although a counterexample is provided in the next chapter. In general, health outcomes as well as sociodemographic and risk factor data vary by characteristics of the population and so are better fit by regression models than by overly simplistic smoothing methods. In addition, the heteroskedasticity of disease rates and counts is not accommodated by the traditional implementation of geostatistical methods. However, the variogram plot and its modeling, which arose as an adjunct to these methods, can be a useful exploratory tool as illustrated in this chapter and the next.

6. HYPOTHESIS TESTING OF SPATIAL PATTERN

6.1 Introduction

Methods in the previous section can help to clarify the underlying patterns in the data but do not provide measures of significance of these patterns. Even simulated random spatial data can appear to be clustered, so visual inspection of maps must be supplemented by a statistical measure of the strength of clustering. The identification of clusters is an important tool for cancer surveillance but the term “cluster” itself is an imprecise term. Clustering has been variously defined as

- -- the presence of “a geographically bounded group of occurrences of sufficient size and concentration to be unlikely to have occurred by chance” (Knox 1989),
- -- a non-independence of case locations (Diggle 2000),
- -- the observation of a significantly greater number of cases (or relative risk) in an area than expected,

-- areas with at least 5 cases and relative risks of at least 20 (Neutra 1990), and
-- "residual spatial variation in risk" (Wakefield et al. 2000) which the authors note does not imply that cases are close in space.

The identification of clusters does not provide any causal explanation for the pattern. Clustering of disease is often, but not exclusively, a result of clustering of host (genetic) susceptibility or environmental (risk) factors. Spatial clustering of cause-specific mortality, for example, may also occur due to geographic differences in diagnostic methods, treatment patterns or accuracy of death certification. More in-depth studies are required to distinguish among these causes once disease clusters are identified.

There have been hundreds of tests proposed to assess clustering (Kulldorff 2001). Some attempts have been made to compare methods and to recommend optimal tests on a theoretical basis (e.g., Tango 1999) and in particular situations (Zoellner and Schmidtmann 1999) but more work is needed in this area. In this section we list several of the most popular methods to test for spatial randomness (or conversely, for clustering) and to identify significant clusters. For an extensive review of this topic, see chapters 7-12 in Lawson et al. (Lawson et al. 1999)or Wakefield et al. (Wakefield et al. 2000).

6.2 Tests of Randomness

6.2.1 For Counts

Tests of complete spatial randomness are often conducted as the first step in spatial data exploration when there is no point source suspected as a risk factor *a priori*. Following the notation of section 4, the simplest test for count data is the index of dispersion test: $T = \sum_{i=1}^{I} \frac{(d_i - \overline{d})^2}{\overline{d}}$ where the observed counts, d_i, are the number of cases in each sub-area i. If the number of sub-areas is sufficiently large, $T \sim \chi^2_{1-1}$. This quadrat (grid) method is not appropriate for disease data when the population varies over the sub-areas. An alternative method to use is Pearson's chi-square statistic: $T = \sum_i \frac{(d_i - E_i)^2}{E_i}$, where

$E_i = \sum_j c_j n_{ij}$, the expected number of cases computed by indirect standardization, i.e., by applying the stratum j-specific rates in the standard population (c_j) to the stratum j-specific population in area *i*. If the data are randomly distributed, $T \sim \chi^2_{1-1}$ as before.

Potthoff and Whittinghill (Potthoff and Whittinghill 1996) have shown that the locally most powerful test of random pattern against heterogeneity (specifically, a gamma-Poisson alternative hypothesis) is: $T = E_{\cdot} \sum_i \frac{d_i(d_i - 1)}{E_i}$, where $E_{\cdot} = \sum_i E_i$. When the expected counts are all constant and $E_{\cdot}$ is assumed fixed, then this is equivalent to Pearson's chi-square statistic (Alexander and Cuzick 1992, p. 242).

Tango's method (Tango 1995) utilizes a measure of "closeness" between areas: $T = (r-p)'A(r-p) = \sum_i \left\{ \sum_j a_{ij}(r_i - p_i)(r_j - p_j) \right\} = \sum_i U_i$, where A is a matrix of distance or adjacency measures and *r* and *p* are vectors of relative frequencies of observed and expected cases, respectively, in each area *i*. Recently, Tango (Tango 2000) extended this statistic to adjust for cluster size and multiple comparisons and now suggests the use of $a_{ij} = \exp\{-4(h_{ij}/\lambda)^2\}$ as the distance measure, where h_{ij} is the distance between case *i* and case *j* and λ is the maximum distance between cases that are considered to be in the same cluster. The values of λ are varied from near 0 to about half the distance across the entire study area, and the significance level is determined by Monte Carlo methods.

Bonetti (Bonetti and Pagano 2001) recently extended an earlier interpoint distance method by Whittemore (Whittemore et al. 1987) to compare the cumulative distribution function of distances between cases with that of the population at risk. This test allows for differing population density across the region. In a simulation study, its power was at least as good as that of Tango's statistic and much better than that of Whittemore.

The Moran autocorrelation statistic may also be used to compare observed and expected cases (Moran 1948), although it ignores heteroskedasticity due to varying populations. Let a_{ij} be a measure of the closeness of areas *i* and *j* as

above. Then $I_{Moran} = \frac{I\sum_i\sum_j a_{ij}(Z_i - \bar{Z})(Z_j - \bar{Z})}{(\sum_i\sum_j a_{ij})\sum_k (Z_k - \bar{Z})^2}$, where $Z_i=d_i/E_i$, the SMR for area i (i=1,...,I), measures the similarity of relative risk in nearby areas; I_{Moran} ranges from 0 (no clustering) to 1.

6.2.2 Comparing Cases and Controls

Cuzick and Edwards' test (Cuzick and Edwards 1990) examines whether cases are clustered nearer to each other rather than evenly interspersed with controls: $T_k = \sum_i \sum_{j \neq i} \delta_{ij}$ where $\delta_{ij} = 1$ if case j is one of the k nearest neighbors of case i. Cuzick and Edwards (Cuzick and Edwards 1996) suggest an adjustment for multiple comparisons using several values of k. Diggle (Diggle 1983) (Diggle 1983) proposed a similar test which is the scaled difference of the mean number of cases within distance k of an arbitrary case and the similar mean for controls.

6.3 Tests to Identify Specific Clusters

Individual terms of the global clustering tests, e.g., Pearson's chi-square, Potthoff-Whittinghill's or Tango's method, have been used to identify specific areas with an unusually high number of cases. Moving window methods are most often used to identify individual clusters.

Besag and Newell (Besag and Newell 1991) proposed an improvement to Openshaw's (Openshaw et al. 1987) moving average method in which a circle is centered at each individual case with a radius that includes the k nearest neighboring cases. The circle defines a cluster if the expected number of cases in the population at risk in that area is significantly less than k. The circles are comparable because each contains k cases but the method tends to detect small rural clusters and the choice of k is arbitrary. An adjustment for multiple comparisons can be made if various values of k are tried. The individual circle statistics can be summed for an overall test of randomness.

Scan statistics improve upon this approach by computing the number of cases that occur within windows of constant size. Turnbull (Turnbull et al. 1990) (Turnbull et al. 1990) defined windows to contain a constant population,

centered on each area centroid, then calculated the maximum number of cases across windows. Kulldorff (Kulldorff 1997) defines the circles to contain up to a pre-specified fraction of the total population in the entire region. The maximum likelihood ratio statistic is calculated as $T = \max_{\text{j where } d_j > E_j} \left(\frac{d_j}{E_j} \right)^{d_j} \left(\frac{d_\cdot - d_j}{d_\cdot - E_j} \right)^{d_\cdot - d_j}$ where d_j and E_j are the observed and expected numbers of cases in circle j, respectively, and $d_\cdot$ and $E_\cdot$ are the corresponding sums over all the circles and the significance level of T is determined by Monte Carlo methods. As was the case for the moving window methods, there are no guidelines for the choice of the maximum fraction of the population to include in each circle. Although Kulldorff's test was designed to identify the single most significant cluster in a region, its exact properties are not known when multiple clusters are identified (Wakefield et al. 2000). However, it can be shown to be a conservative significance test of each secondary cluster (Kulldorff, personal communication).

6.4 Recommendations

More complete studies comparing these various methods are needed to provide specific recommendations for general and specific tests of clustering. However, it is clear that clustering tests should not be used for health data unless they account for varying population sizes across areas. In addition, some adjustment for multiple comparisons should be made whenever necessary, such as when different sized moving windows are tried. Whenever possible, it seems that use of a Monte Carlo method to compute the significance level of the test is to be preferred over asymptotic results based on questionable assumptions. These tests can provide a useful initial evaluation of clusters in an area, but should be followed by careful field investigation to verify the existence and importance of the identified patterns.

7. SPATIAL MODELING

7.1 Introduction

As noted in section 2.6, a number of methods from epidemiology, geostatistics and small area modeling have converged to provide powerful new, but complex, models with which to analyze health data. This section will describe the background necessary to apply these new methods to cancer mortality data for illustration. However, only slight modifications are necessary to apply the same principles to other types of data, such as survival data using proportional hazards models, late stage versus early stage of cancer detection using logistic models, and prevalence data (Verdecchia et al. 1989). The next chapter will illustrate the application of the methods described here.

Readers interested in details of the historical development of these models are referred to Lawson et al., chapters 13-16 (Lawson et al. 1999). Briefly, theoretical developments over the past 25 years have included the consideration of spatial correlation for lattices, first regularly spaced (Besag 1974) then irregularly spaced (Besag 1975), the application of Bayesian techniques to health data (Clayton and Kaldor 1987; Manton et al. 1989), the extension of regression methods to non-normal data through generalized linear models (McCullagh and Nelder 1983), and the recognition of overdispersion in disease rates (summarized in (Brillinger 1986). Of course, this work is built upon earlier developments in regression models and maximum likelihood and Bayesian estimation methods.

7.2 The Fixed Effects Model

Following the notation in section 4, let d_{ij} represent the number of cancer deaths in place *i*, *i*=1,2,...,I, age group *j*, *j*=1,2,...,J, and let n_{ij} represent the corresponding population at risk. Assume further that d_{ij} is a Poisson random variable with mean $\lambda_{ij} n_{ij}$ and that, in the simplest case, the effect of fixed explanatory covariates on the rates may be modeled as a log-linear function: $\ln\left(\lambda_{ij}\right) = \mathrm{X}_{ij}'\beta_{ij}$. The basic Poisson variance may be generalized to include potential overdispersion: $\mathrm{Var}(d_{ij} \mid \beta; n_{ij}) = \varphi\lambda_{ij} n_{ij}$.

As written, this is a saturated model, i.e., *IJ* parameters to be estimated from *IJ* stratified counts, but of greater interest is a reduced model that will allow inferences across places and/or age groups. U.S. age-specific cancer mortality rate curves have similar shapes for blacks and whites, males and females; a cubic spline function of age has been found to fit these data well (Pickle et al. 1996a; Pickle et al. 1996b). One choice might be to assume the same age-specific rates everywhere: $\ln\left(\lambda_{ij}\right) = \mathrm{X}_{ij}'\beta_j$. This does not fit the data well because there are geographic differences in rates across the U.S.

The reader should note that this is not the conventional setup for disease rate models, such as those introduced by Clayton and Kaldor (Clayton and Kaldor 1987). A currently popular approach is to first compute the expected number of cases (E_i) in place *i* from age-specific rates $\{r_{0j}\}$ in a standard population. Then the analysis focuses on the age-adjusted SMRs of the areas, rather than their age-specific counts. If the proportionality model for age and place effects holds everywhere, i.e., if $\lambda_{ij} = r_{0\,j}\exp(\mu_i)$ for all *i=1,...I*, then $\mathrm{E}(d_i \mid \mu_i) = E_i \exp(\mu_i)$ and this age-adjusted model is equivalent to the age-specific one given above. If the age-specific data are extremely sparse, it may be necessary to use this approach. However, the proportional rate assumption does not always hold (Pickle and White 1995b), and so we prefer to begin with the age-specific model. If, in the final reduced model, no terms remain that depend on both place and age, then these two approaches will yield the same results. Otherwise there are different age effects by place and epidemiologists would not advise using any type of age-standardized rate that masks these effects (Fleiss 1981).

The models described in this section may be extended to include temporal trends, spatio-temporal interactions, age-period-cohort effects and others.

7.3 Adding Random Effects

7.3.1 Rationale

Recognizing that there are many types of models that will yield similar results, we illustrate the extension from fixed effects to hierarchical models using the age-specific model defined in the previous section. As noted, a fixed

effects model that assumes a constant disease rate across all areas is unrealistic and uninformative. We could try to estimate parameters for smaller areas, such as regions or states, but at some point the data are too sparse to support estimation at a smaller area level using the fixed effects approach. By considering the small area effects to vary randomly within larger areas, information is "borrowed" from other relevant areas, thus stabilizing the small area estimates. In addition, this extra source of variation can help to explain the overdispersion typically seen in disease rate and count data.

7.3.2 Hierarchical Models

To illustrate the principles, consider a simple version of the fixed effects model of section 7.2: $\ln\left(\lambda_{ij}\right) = \mathrm{X}_{ij}'\beta_{ij} = \beta_{0i} + \mathrm{age}_j\beta_1$. If we are willing to assume that the age specific curves for the small areas are parallel to but not necessarily identical to the regional rates, as illustrated in Figure 3a, we could add a random intercept to the fixed effects model. That is, we would assume that $\mathrm{E}(\beta_{0i} \mid \beta_0) = \beta_0$ with some variance σ_0^2. If the slopes as well as the intercepts may vary within the region (as in Figure 3b), then we may substitute β_{1i} for β_1 above and assume that $\mathrm{E}(\beta_{1i} \mid \beta_1) = \beta_1$ and $\mathrm{Var}(\beta_{0i}, \beta_{1i} \mid \beta_0, \beta_1) = G$. It is usually acceptable to assume that these log-linear parameters are normally distributed (Littell et al. 1996a). Until recent improvements in estimation algorithms, Bayesians would choose a gamma distribution as the "conjugate prior" for *G* so that the posterior distribution would be of a convenient form. Now a distribution that is consistent with the data may be chosen without as much regard to the method of estimation. For a more extensive description of the evolution of hierarchical models, see Waller (Waller forthcoming)

Other explanatory covariates may be added to the model. The general form of this "mixed effects" or "hierarchical" model is usually written as $\mu_i = x_i'\beta_i + Z_i'\gamma_i$ where β_i and γ_i are fixed and random effects parameters, respectively, the conditional variance of γ_i is *G* and the residual variance is *R*. Because our underlying model is for counts $\{ d_{ij} \}$, *R* has the appropriate Poisson form although excess heterogeneity of rates can be

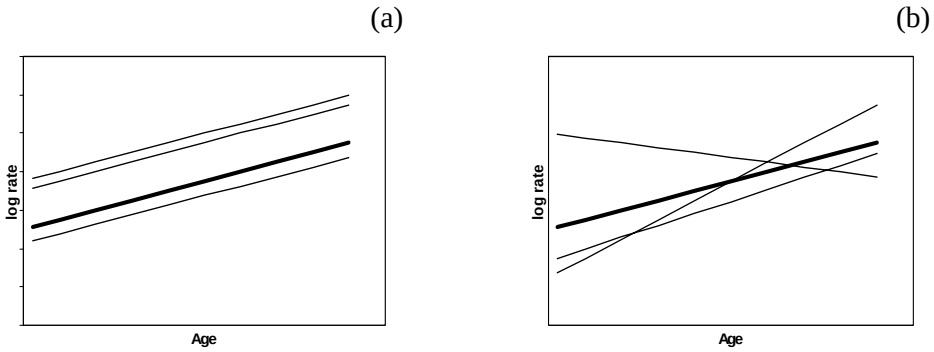

Figure 3. Examples of age-specific log rates with random intercepts (a) and random intercepts and slopes (b). Heavy line denotes regional rate, standard lines denote small area rates within that region.

accommodated through a scaling parameter. The goal of the models described here is to explain the similarities of rates and counts in neighboring places through the covariate effects, leaving uncorrelated residuals. If residuals from the final model are spatially correlated, there are important covariates or interactions missing from the model.

The geographic hierarchy described above can be extended to more levels, such as models of small area effects within state and state effects within region, each conditional on the larger unit effects. For example, studies of mortality data have shown that even after accounting for characteristics of individual decedents, there are community effects on the small area death rates (LeClere et al. 1998; Cubbin et al. 2000). Often covariate information is available at different levels of geography, such as demographic data for the small areas but health risk data only for states. "Multilevel" models refer to those hierarchical models that include covariates at different geographic scales (Rasbash et al. 2000).

In addition to random effects that describe intra-regional variation, covariates that are imprecisely measured may also be considered to vary randomly about a true but unknown mean value. The analysis of these "errors-in-covariates" models requires some additional information, such as a separate validation study that provides information about the distribution of the measurements in relation to the correct values (Carroll et al. 1995; Bernardinelli et al. 1997).

7.4 Modeling Spatial Dependence

The spatial similarity of the observations is modeled through the covariance structure of either *R* or *G*, the variance matrices of the residuals and random effects, respectively. Note that for a typical analysis of repeated measures, it is the residuals that are assumed to be spatially correlated (see section 3.3.1). However, for disease rates or counts, the residual errors (R) are of the binomial or Poisson form and the structure of the random effects variance matrix (G) reflects the spatial similarity of rates in nearby small areas. Spatial autocorrelation may be modeled in a number of ways, such as an exponential function of distance between each pair of points, $\rho = \exp(-h^2 / \varphi)$. Other possible functions are spherical, Gaussian, log linear and power functions (Littell et al. 1996a; Littell et al. 1996b). "Closeness" may be defined in terms of distance between point events or centroids of areas or adjacency. Clues as to the most appropriate covariance function may be gained from examination of the empirical variogram, as illustrated in section 5.4.3.

The spatial autoregressive structures may be estimated for each random effect, conditional on the observed values of all the others (the conditionally autoregressive (CAR) model) or for all of the random effects simultaneously (the simultaneous autoregressive (SAR) model). The SAR model produces spatially correlated residuals, resulting in inconsistent least-squares estimators (for further discussion, see Besag and Kooperberg 1995 and Cressie 1993).

7.5 Parameter Estimation

Models of the complex structure described here may be fit using maximum likelihood or restricted maximum likelihood, e.g., using SAS or S+ software (Littell et al. 1996a; MathSoft 1999)or empirical or full Bayesian methods (see (Carlin and Louis 2000), e.g., using WinBUGS (Spiegelhalter et al. 1995). Except for empirical Bayes methods where the choice of conjugate distribution yields a simple posterior distribution, all of these methods require iterative estimation processes such as Markov Chain Monte Carlo (MCMC) methods.

7.6 Model Checking

It is beyond the scope of this chapter to detail methods used to check these models but a few guidelines will be offered and the next chapter will illustrate their application. First, prior to fitting the model, stationarity and potential functions for modeling spatially autoregressive covariance structures can be checked by examining the empirical variogram (see section 5.4.3). The

association of covariates with the log rates or counts should also be verified, as for any regression analysis, to determine whether a linearizing transformation or covariate categorization is needed. The proportional rate assumption may be checked if SMR models are preferred.

After estimation of the model parameters, plots of standardized residuals are helpful in pointing to covariate strata or geographic areas that are not fit well. No simple statistics are available to judge the adequacy of these complex models; more work on diagnostics for hierarchical models is needed. Because of the geographic nature of the observations, it is usually helpful to supplement typical regression diagnostic plots by maps of predicted observations, their standard errors and standardized residuals.

8. Summary

In this chapter, we have reviewed the history of the spatial analysis of disease and the statistical methods used for the exploratory analysis, testing and modeling of spatial patterns. In the next chapter, the principles described here will be illustrated.

REFERENCES

Alexander F. E., Cuzick J. (1992). Methods for the assessment of disease clusters. In: Elliott P., Cuzick J., English D., Stern R., editors. *Geographical & Environmental Epidemiology: Methods for Small-Area Studies*. New York: Oxford University Press; p 238-50.

Astakhova L. N., Anspaugh L. R., Beebe G. W., Bouville A., Drozdovitch V. V., Garber V., Gavrilin Y. I., Khrouch V. T., Kuvshinnikov A. V., Kuzmenkov Y. N., Minenko V. P., Moschik K. V., Nalivko A. S., Robbins J., Shemiakina E. V., Shinkarev S. (1998). Chernobyl-related thyroid cancer in children of Belarus: A case-control study. *Radiation Research* 150(3):349-56.

Barrett F. A. (2000). Finke's 1792 map of human diseases: the first world disease map? *Social Science & Medicine* 50(7-8):915-21.

Beebe G.W., Kato H., Land C. E. (1971). Studies of the mortality of A-bomb survivors. 4. Mortality and radiation dose, 1950-1966. *Radiat Res* 48(3):613-49.

Bernardinelli L., Pascutto C., Best N. G., Gilks W. R. (1997). Disease mapping with errors in covariates. *Statistics in Medicine* 16(7):741-52.

Besag J., Newell J. (1991). The detection of clusters in rare diseases. *Journal of the Royal Statistical Society, Ser A* 154:143-55.

Besag J. (1974). Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society, Series B* 36:192-236.

Besag J. (1975). Statistical analysis of non-lattice data. *The Statistician* 24(No.3):179-95.

Besag J., Kooperberg C. (1995). On conditional and intrinsic autoregression. *Biometrika* 82(4):733-46.

Blot W. J., Fraumeni Jr. J. F. (1976). Geographic patterns of lung cancer: Industrial correlations. *American Journal of Epidemiology* 103(6):539-50.

Blot W. J., Harrington J. M., Toledo A., Hoover R., Heath C. W., Jr., Fraumeni J. F., Jr. (1978). Lung cancer after employment in shipyards during World War II. *N Engl J Med* 299(12):620-4.

Bonetti M., Pagano M. (2001). Detecting clustering in regional counts. *Proceedings of the Section on Survey Research Methods of the ASA, 2000*.

Brillinger D. R. (1986). The natural variability of vital rates and associated statistics. *Biometrics* 42:693-734.

Bross I. D. J. (1954). A confidence interval for a percentage increase. *Biometrics* 10:245-50.

Burbank F. (1972). A Sequential space-time cluster analysis of Cancer mortality in the United States: Etiologic Implications. *Journal of Epidemiology* 95(5):393-417.

Carlin B. P., Louis T. A. (2000). *Bayes and Empirical Bayes Methods for Data Analysis*. Second ed. New York: Chapman & Hall/CRC.

Carr D. B., Pierson S. M. (1996). Emphasizing statistical summaries and showing spatial context with micromaps. *Statistical Computing & Statistical Graphics Newsletter* 7(3):16-23.

Carroll R. J., Ruppert D., Stefanski L. A. (1995). *Measurement Error in Nonlinear Models*. 1st ed. ed. London: Chapman & Hall.

Clayton D., Kaldor J. (1987). Empirical bayes estimates for age-standardized relative risks for use in disease mapping. *Biometrics* 43:671-81.

Cleveland W. S., Devlin S. J. (1988). Locally weighted regression - An approach to regression analysis by local fitting. *Journal of the American Statistical Association* 83(403):596-610.

Cornfield J. (1951). A method of estimating comparative rates from clinical data. *Journal of the National Cancer Institute* 11:1269-75.

Cox D. R. (1970). *The Analysis of Binary Data*. 1st ed. London: Methuen & Co. Ltd.

Cressie N. A. C. (1993). *Statistics for Spatial Data*. Revised ed. New York: J. Wiley.

Cressie N. A. C., Read T. R. C. (1989). Spatial data-analysis of regional counts. *Biometrical Journal* 31(6):699-719.

Cubbin C., Pickle L. W., Fingerhut L. (2000). Social Context and Geographic Patterns of Homicide Among U.S. Black and White Males. *American Journal of Public Health* 90(4):579-87.

Cuzick J., Edwards R. (1990). Spatial clustering for inhomogeneous populations. *Journal of the Royal Statistical Society, Series B* 52:73-104.

Cuzick J., Edwards R. (1996). Cuzick-Edwards one-sample and inverse two-sampling statistics. In: Alexander F. E., Boyle P., editors. *Methods for investigating localized clustering of disease*. Lyon: International Agency for Research on Cancer; p 200-2.

Devesa S., Grauman D. J., Blot W. J., Pennello G. A., Hoover R. N., Fraumeni Jr J. F. (1999). *Atlas of Cancer Mortality in the United States: 1950-94*. Bethesda, MD: National Cancer Institute.

Devine O. J., Annest J. L., Kirk M. L., Holmgreen P., Emrich S. S. (1991). *Injury Mortality Atlas*. Atlanta, GA: U.S. Department of Health and Human Services.

Diggle P. (1983). *Statistical Analysis of Spatial Point Patterns*. London: Academic Press.

Diggle P. J. (2000). Overview of statistical methods for disease mapping and its relationship to cluster detection. In: Elliott P., Wakefield J. C., Best N. G., Briggs D. J., editors. *Spatial Epidemiology: Methods and Applications*. Oxford: Oxford University Press; p 87-103.

Draper G. J., Stiller C. A., Cartwright R. A., Craft A. W., Vincent T. J. (1993). Cancer in Cumbria and in the vicinity of the Sellafield nuclear installation, 1963-90. *British Medical Journal* 306(6870):89-94.

Fleiss J. L. (1981). *Statistical Methods for Rates and Proportions*. Second ed. New York: John WIley & Sons.

Geary R. C. (1954). The contiguity ratio and statistical mapping. *The Incorporated Statistician* 5:115-45.

Ghosh M., Natarajan K., Stroud T. W. F., Carlin B. P. (1998). Generalized linear models for small-area estimation. *Journal of the American Statistical Association* 93(441):273-82.

Ghosh M., Rao J. N. K. (1994). Small-area estimation - an appraisal. *Statistical Science* 9(1):55-76.

Haining R. P. (1990). *Spatial Data Analysis in the Social and Environmental Sciences*. Cambridge England: Cambridge University Press.

Hansen K. M. (1991). Head-banging: robust smoothing in the plane. *IEEE Transactions on Geoscience and Remote Sensing* 29(3):369-78.

Haviland A. (1875). *The geographical distribution of heart disease and dropsy, cancer in females and phthisis in females in England and Wales*. London: Swan Schonnenschein.

Howe G. M. (1963). *National Atlas of Disease Mortality in the United Kingdom*. London: Nelson.

Howe G. M. (1989). Historical evolution of disease mapping in general and specifically of cancer mapping. In: Boyle P., Muir C. S., Grundmann E., editors. *Cancer Mapping*. New York: Springer-Verlag; p 1-21.

Janerich D. T., Burnett W. S., Feck G., Hoff M., Nasca P., Polednak A. P., Greenwald P., Vianna N. (1981). Cancer incidence in the Love Canal area. *Science* 212(4501):1404-7.

Kafadar K. (1994). Choosing among two-dimensional smoothers in practice. *Computational Statistics & Data Analysis* 18:419-39.

Kaluzny S. P., Vega S. C., Cardoso T. P., Shelly A. A. (1998). *S+ Spatial Stats: User's manual for Windows® and UNIX®*. New York: Springer.

Kemp I., Boyle P., Smans M., Muir C. S. (1985). *Atlas of cancer in Scotland, 1975-1980, incidence and epidemiological perspective*. Lyon: International Agency for Research on Cancer.

Knox E. (1989). Detection of clusters. In: Elliott P., editor. *Methodology of enquiries into disease clustering*. London: Small Area Health Statistics Unit.

Knox G. (1964). Detection of space-time interactions. *Appl Stat* 13:25-9.

Kulldorff M. (2001). Tests for spatial randomness adjusted for an inhomogeneity: A general framework. *Statistics in Medicine* .

Kulldorff M. (1997). A spatial scan statistic. *Communications in Statistics-Theory and Methods* 26:1481-96.

Lawson A., Biggeri A., Bohning D., Lesaffre E., Viel J.-F., Bertollini R. (1999). *Disease Mapping and Risk Assessment For Public Health*. Chichester: Wiley.

Le N. D., Marrett L. D., Robson D. L., Semenciw R. M., Turner D., Walter S. D. (1996). *Canadian cancer incidence atlas*. Ottawa: Government of Canada.

LeClere F. B., Rogers R. G., Peters K. (1998). Neighborhood social context and racial differences in women's heart disease mortality. *J Health Soc Behav* 39(2):91-107.

Littell R. C., Milliken G. A., Stroup W. W., Wolfinger R. D. (1996b). *SAS® System for Mixed Models*. Cary, NC: SAS Institute Inc.

Littell R. C., Milliken G. A., Stroup W. W., Wolfinger R. D. (1996a). *SAS® System for Mixed Models*. Cary, NC: SAS Institute Inc.

Mantel N. (1967). The detection of disease clustering and a generalized regression approach. *Cancer Research* 27:209-20.

Manton K. G., Woodbury M. A., Stallard E., et al. (1989). Empirical Bayes procedures for stabilizing maps of U.S. cancer mortality rates. *Journal of the American Statistical Association* 84:637-50.

Mason T. J., McKay F. W., Hoover R., Blot W. J., Fraumeni Jr J. F. (1975). *Atlas of cancer mortality for U.S. counties: 1950-1969*. Washington, D.C.: U.S. Government Printing Office.

MathSoft. (1999). *S-Plus 2000 User's Guide*. Seattle, WA: Data Analysis Products Division, MathSoft.

McCullagh P., Nelder J. A. (1983). *Generalized Linear Models*. Second ed. London: Chapman and Hall.

McLeod K. S. (2000). Our sense of Snow: the myth of John Snow in medical geography. *Social Science & Medicine* 50(7-8):923-35.

Moran P. A. P. (1948). The interpretation of statistical maps. *Journal of the Royal Statistical Society, Series B* 10:243-51.

Mungiole M., Pickle L. W., Simonson K. H. (1999). Application of a weighted head-banging algorithm to mortality data maps. *Statistics in Medicine* 18(23):3201-9.

Najem R. G., Cappadona J. L. (1991). Health effects of hazardous chemical waste disposal sites in New Jersey and in the United States: A review. *Statistics in Medicine* 7(6):352-62.

Nelson D. E., Holtzman D., Waller M., Leutzinger C. L., Condon K. (1998). Objectives and design of the Behavioral Risk Factor Surveillance System. *Proceedings of the Section on Survey Research Methods, ASA*:214-8.

Neutra R. R. (1990). Counterpoint from a cluster buster. *American Journal of Epidemiology* 132:1-8.

Neyman J., Scott E. L. (1958). Statistical approach to problems of cosmology. *Journal of the Royal Statistical Society, Series B* 20:1-29.

Ohno Y., Aoki K. (1981). Cancer deaths by city and county in Japan (1959-1971): A test of significance for geographic clusters of disease. *Social Science & Medicine* 15:251-8.

Openshaw S., Charlton M., Wymer C., Craft A. W. (1987). A mark I geographical analysis machine for the automated analysis of point data sets. *International Journal of Geographical Information Systems* 1:335-58.

Peterman AH. 1852. Cholera map of the British Isles showing the districts attacked in 1831, 1832, and 1833. Constructed from official documents. London: Betts; Available from.

Pickle LW, Herrmann D. 1995. Cognitive Aspects of Statistical Mapping. National Center for Health Statistics; Report nr 18. Available from.

Pickle L. W., Mason T. J., Howard N., Hoover R., Fraumeni Jr. J. F. (1987). *Atlas of U.S. cancer mortality among whites:1950:1980*. Washington: Dept. of Health and Human Services.

Pickle L. W., Mason T. J., Howard N., Hoover R., Fraumeni J. F. (1990). *Atlas of U.S. Cancer Mortality Among Nonwhites*. Washington D.C.: U.S. Government Printing Office.

Pickle L. W., Mungiole M., Jones G. K., White A. A. (1996b). *Atlas of United States Mortality*. Hyattsville, MD: National Center for Health Statistics.

Pickle L. W., Mungiole M., Jones G. K., White A. A. (1996a). *Atlas of United States Mortality*. Hyattsville, MD: National Center for Health Statistics.

Pickle L. W., Su Y. (2001). Geographic patterns of health insurance coverage and health risk factors for U.S. counties. *American Journal of Public Health*.

Pickle L. W., White A. A. (1995b). Effects of the choice of age-adjustment method on maps of death rates. *Statistics in Medicine* 14(5-7):615-27.

Pickle L. W., White A. A. (1995a). Effects of the choice of age-adjustment method on maps of death rates. *Statistics in Medicine* 14(5-7):615-27.

Potthoff R. F., Whittinghill M. (1996). Testing for homogeneity in the Poisson distribution. *Biometrika* 53:183-90.

Rasbash J., Browne W., Goldstein H., Yang M., Plewis I., Healy M., Woodhouse G., Draper D., Langford I., Lewis T. (2000). *A user's guide to MLwiN*. London: Multilevel Models Project, Institute of Education, University of London.

Riggan W. B., Creason J. P., Nelson W. C., Manton K. G., Woodbury M. A., Stallard E., Pellom A. C., Beaubier J. (1987). *U.S. Cancer Mortality Rates and Trends, 1950-1979*. Research Triangle Park, NC: U.S. Environmental Protection Agency.

Snow J. (1855). *On the Mode of Communication of Cholera*. 2nd ed. London: John Churchill.

Spiegelhalter D., Thomas A., Best N., Gilks W. (1995). *BUGS: Bayesian inference using GIBBS Sampling, Version 0.50*. Cambridge: MRC Biostatistics Unit.

Stocks P. 1928. On the evidence for a regional distribution of cancer prevalence in England and Wales. London: British Empire Cancer Campaign; Available from.

Stocks P. 1936. Distribution in England and Wales of cancer of various organs. London: British Cancer Campaign; Report nr 13. Available from.

Tango T. (1995). A class of tests for detecting 'general' and 'focused' clustering of rare diseases. *Statistics in Medicine* 14:2323-34.

Tango T. (1999). Comparison of general tests for spatial clustering. In: Lawson A., Biggeri A., Bohning D., Lesaffre E., Viel J.-F., Bertollini R., editors. *Disease Mapping and Risk Assessment for Public Health*. Chichester: John Wiley and Sons, Ltd; p 111-7.

Tango T. (2000). A test for spatial disease clustering adjusted for multiple testing. *Statistics in Medicine* 19(2):191-204.

Tukey J. W. (1977). *Exploratory Data Analysis.* Reading, Mass.: Addison-Wesley Pub. Co.

Tukey P. A., Tukey J. W. (1981). Graphical display of data sets in 3 or more dimensions. In: Barnett V., editor. *Interpreting Multivariate Data*. New York: John Wiley and Sons.

Turnbull B., Iwano E. J., Burnett W. S. (1990). Monitoring for clusters of disease: application to leukemia incidence in upstate New York. *Am J Epidemiol* 132:S136-S143.

Varon E. 1998 Sep 7. Connecting the dots: NCI hopes advanced GIS tools will help trace causes of breast cancer. Federal Computer Week;1,76.

Verdecchia A., Capocaccia R., Egidi V., Golini A. (1989). A method for the estimation of chronic disease morbidity and trends from mortality data. *Statistics in Medicine* 8(2):201-16.

Wakefield J. C., Kelsall J. E., Morris S. E. (2000). Clustering, cluster detection, and spatial variation in risk. In: Elliott P., Wakefield J. C., Best N. G., Briggs D. J., editors. *Spatial Epidemiology: Methods and Applications*. Oxford: Oxford University Press; p 128-52.

Walker S. H., Duncan D. B. (1967). Estimation of the probability of an event as a function of several independent variables. *Biometrika* 54:167-79.

Waller, L. A. Hierarchical models for disease mapping. Encyclopedia of Environmetrics. Forthcoming.

Whittemore A. S., Fiend N., Brown Jr B. W., Holly E. (1987). A test to detect clusters of disease. *Biometrika* 74(3):631-5.

Woolf B. (1955). On estimating the relation between blood group and disease. *Annals of Human Genetics* 19:251-3.

Xiang H. Y., Nuckols J. R., Stallones L. (2000). A geographic information assessment of birth weight and crop production patterns around mother's residence. *Environmental Research* 82(2):160-7.

Zhu L., Carlin B. P. (2000). Comparing hierarchical models for spatio-temporally misaligned data using the deviance information criterion. *Statistics in Medicine* 19(17-18):2265-78.

Zoellner I. K., Schmidtmann I. M. (1999). Empirical studies of cluster detection - Different cluster tests in application to German cancer maps. In: Lawson A., Biggeri A., Bohning D., Lesaffre E., Viel J.-F., Bertollini R., editors. *Disease Mapping and Risk Assessment for Public Health*. Chichester: John Wiley and Sons, Ltd.; p 169-78.