Rate Transformations and Smoothing

Article · January 2006

3 authors, including:

Luc Anselin
University of Chicago
300 PUBLICATIONS   **61,730** CITATIONS

Julia Koschinsky
Arizona State University
43 PUBLICATIONS   **1,325** CITATIONS

Rate Transformations and Smoothing

Luc Anselin

Nancy Lozano

Julia Koschinsky

Spatial Analysis Laboratory

Department of Geography

University of Illinois, Urbana-Champaign

Urbana, IL 61801

http://sal.uiuc.edu/

Revised Version, January 31, 2006

Acknowledgments

1   Introduction and Background

This report is part of a series of activities to review and assess methods for exploratory spatial data analysis and spatial statistics as they pertain to the study of the geography of disease in general, and prostate cancer in particular. It focuses on the intrinsic instability of rates or proportions and how this affects the identification of spatial trends and outliers in an exploratory exercise in the context of disease mapping. It provides an overview of some technical (statistical) aspects of a range of approaches that have been suggested in the literature to deal with this issue. The review does not attempt to be comprehensive, but focuses on commonly used techniques (in the research literature) as well as on methods that can be readily implemented as software tools. The objective is to provide a description that can form the basis for the development of such tools. References to the literature are provided for a detailed coverage of the technical details.

Note that this is an evolving document, and it is updated as new techniques are reviewed and implemented in the software.

The report is supplemented by an extensive and searchable annotated bibliography of over fifty articles, available from the SAL web site.[1] It is also supplemented by software tools available from the SAL site (or through links on the SAL site). Some of the software is developed in the Java language and is included in the GeoVISTA Studio package, distributed by the Pennsylvania State University.[2] The bulk of the software development, however, is carried out in the Python language and included in the PySAL library of spatial

---

[1]http://sal.uiuc.edu/stuff/stuff-sum/abstracts-on-spatial-analysis-health

[2]This is available from Sourceforge at http://sourceforge.net/projects/geovistastudio.

analytical tools (available from the SAL site), and disseminated through the STARS software package for space-time exploration.[3]

The remainder of this report consists of nine sections. We begin with a definition of terms, outline the problem of intrinsic variance instability and provide a brief overview of suggested solutions in Section 2. Next we move on to a review of specific methods, organized in six sections. First, in Section 3, rate transformations are covered. This is followed by an overview of four categories of smoothing methods: mean and median based smoothers in Section 4, non-parametric smoothers in Section 5, Empirical Bayes smoothers in Section 6 and fully Bayes (model based) smoothers in Section 7.

To keep the scope of our discussion manageable, we do not cover methods that can be considered "model fitting," i.e., predictive models for rates of incidence or mortality, where the predictive value is taken as the smoothed rate. This predicted value takes into account that part of the spatial variation of a disease distribution that can be explained by known factors. Such methods include local regression, such as local trend surface regressions, loess regression, general additive models (GAM), median polish smoothers, and various interpolation methods (including thin plate splines, and centroid free smoothing). Overviews of these techniques and additional references can be found in Kafadar (1994) and Waller and Gotway (2004), among others.

Finally, in Section 8, an alternative set of approaches is covered, where the problem of rate instability is approached by changing the spatial scale of observation. This includes techniques for regionalization and spatial aggregation. We close with a summary and assessment in Section 9.

---

[3]This is available from http://stars-py.sourceforge.net.

2 Rates, Risk and Relative Risk

2.1 Rates and Risk

Proportions or rates are often not of interest in and of themselves, but serve as estimates for an underlying *risk*, i.e., the probability for a particular event to occur. For example, one would be interested in a measure of the risk of dying from a type of cancer and the extent to which this risk may vary across space, over time, or across other characteristics of individuals. Typically, the risk is estimated by the ratio of the number of people who actually experienced the event (like a death or a disease occurrence) during a given period over the total number of the *population at risk* to whom the event might have occurred.

The point of departure is that the risk (probability) is constant and one wants to assess the extent to which this may not be the case. Since the events happen with uncertainty, the same risk may yield different counts of events in actual samples. Simply concluding that the risk varies because the counts vary is not correct, since a range of different outcomes are compatible with the same underlying risk. More precisely, given a risk π of being exposed to an event for a given population P during a given time period, the number of people O *observed* as experiencing the event can be taken to follow a binomial distribution. The probability of observing x events is then:

$$\text{Prob}[O = x] = \binom{P}{x} \pi^x (1 - \pi)^{P-x}, \text{for } x = 0, 1, \ldots, P. \qquad (1)$$

The mean and variance of this distribution are, respectively, πP and $\pi(1 - \pi)P$. In other words, with a risk of π and a population of P one may

3

expect πP events to occur on average. This is the basis for the estimation of the risk by means of a rate:

$$\hat{\pi} = O/P, \tag{2}$$

where O (the numerator) and P (the denominator) are as above. This simple estimator is sometimes referred to as the *crude rate* or *raw rate*.

Given the mean and variance for the random variable O, it can be seen that the crude rate is an unbiased estimator for the unknown risk π. More precisely:

$$E[O/P] = \frac{E[O]}{P} = \frac{\pi P}{P} = \pi. \tag{3}$$

The variance of the rate estimator then follows as:

$$\text{Var}[O/P] = \frac{\text{Var}[O]}{P2} = \frac{\pi(1-\pi)P}{P2} = \pi(1-\pi)/P. \tag{4}$$

In practice, the rate or proportion is often rescaled to express the notion of risk more intuitively, for example, as 10 per 100,000, rather than 0.00010. This scaling factor (S) is simply a way to make the rate more understandable. The scaled rate would then be $(O/P) \times S$. Different disciplines have their own conventions about what is a base value, resulting in rates expressed as per 1,000 or per 10,000, etc. In reporting cancer rates, the convention used in the U.S. is to express the rate as per 100,000 (e.g., Pickle et al. 1996).

The rate is computed for a given time period, so that the population at risk must be the population observed during that time period. This is expressed in terms of person-years. When the events are reported for a period that exceeds one year (as is typically the case for cancer incidence and mortality), the population at risk must be computed for the matching

4

period. For example, with events reported over a five year period, say 1995-2000, the population at risk must be estimated for the five year period:

$$\hat{\pi}_{95-00} = O_{95-00}/P_{95-00}. \tag{5}$$

In practice, in the absence of precise information on the population in each year, this is often done by taking the average population and multiplying it by the time interval:[4]

$$\hat{\pi}_{95-00} = O_{95-00}/\{[(P_{95} + P_{00})/2] \times 5\}. \tag{6}$$

This yields the rate as an estimate of the risk of experiencing the event during a one year period.

2.2   Relative Risk

Rather than considering a rate in isolation, it is often more useful to compare it to a benchmark. This yields a so-called *relative risk* or excess risk as the ratio of the observed number of events to the expected number of events. The latter are computed by applying a reference estimate for the risk, say $\widetilde{\pi}$, to the population, as:

$$E = \widetilde{\pi} \times P, \tag{7}$$

yielding the relative risk as

$$r = O/E. \tag{8}$$

[4]In the absence of population totals for different points in time, the base population is multiplied by the proper number of years to yield person-years compatible with the event counts.

A relative risk greater than one suggests a higher incidence or mortality in the observed case than warranted by the benchmark risk. A map of the relative risks is sometimes referred to as an *excess risk* map.

When considering several geographical units, the estimate $\widetilde{\pi}$ is typically based on the aggregate of those units (i.e., the total number of observed events and the total population). This is sometimes referred to as *internal* standardization, since the reference risk is computed from the same source as the original data (for a more detailed discussion of standardization, see Section 2.3). For example, consider the observed events O_i and populations P_i in regions $i = 1, \ldots, N$. The reference risk estimate can then be obtained as:

$$\widetilde{\pi} = \frac{\sum_{i=1}^{N} O_i}{\sum_{i=1}^{N} P_i}. \tag{9}$$

Note that $\widetilde{\pi}$ is not the same as the average of the region-specific rates, $\bar{\hat{\pi}}$:

$$\widetilde{\pi} \neq \bar{\hat{\pi}} = \sum_{i=1}^{N} \hat{\pi}_i / N, \tag{10}$$

where $\hat{\pi}_i = O_i / P_i$. Instead, it is the weighted average of these region-specific rates, each weighted by their share in the overall population:

$$\widetilde{\pi} = \sum_{i=1}^{N} \hat{\pi}_i \times \frac{P_i}{\sum_{i=1}^{N} P_i}. \tag{11}$$

Only in the case where each region has the same population will the average of the region-specific rates equal the region-wide rate.

For small event counts, a common perspective is to consider the number of events in each region as a realization of a count process, modeled by a Poisson distribution. In this sense, each observed count O_i is assumed

to be an independent Poisson random variable with a distribution function expressing the probability that x events are observed:

$$\text{Prob}[O_i = x] = \frac{e^{-\lambda}\lambda^x}{x!}, \tag{12}$$

where λ is the mean of the distribution. The mean is taken as the expected count of events, i.e., $\lambda = E_i = \widetilde{\pi} \times P_i$. The underlying assumption is that the risk is constant across regions. The Poisson distribution provides a way to compute how *extreme* the observed count is relative to the mean (the expected count) in either direction.

For observed values less than the mean, i.e., $O_i \leq E_i$, this probability is the cumulative Poisson probability:

$$\rho_i = \sum_{x=0}^{x=O_i} \frac{e^{-E_i}E_i^x}{x!}, \quad O_i \leq E_i. \tag{13}$$

For observed values greater than the mean, i.e., $O_i > E_i$, it is the complement of the cumulative Poisson probability:

$$\rho_i = 1 - \sum_{x=0}^{x=O_i} \frac{e^{-E_i}E_i^x}{x!}, \quad O_i > E_i. \tag{14}$$

Choynowski (1959) refers to a choropleth map of the ρ_i as a *probability map* (see also Cressie 1993, p. 392). In practice, only regions with extreme probability values should be mapped, such as $\rho_i < 0.05$ or $\rho_i < 0.01$.

The probability map suffers from two important drawbacks. It assumes independence of the counts, which precludes spatial autocorrelation. In practice, spatial autocorrelation tends to be prevalent. Also, it assumes a Poisson distribution, which may not be appropriate in the presence of over- or under-dispersion.[5]

---

[5]A Poisson distribution has a mean equal to the variance. Over-dispersion occurs when

7

2.3 Age Standardization

The crude rate in equation (2) carries an implicit assumption that the risk is constant over all age/sex categories in the population. For most diseases, including cancer, the incidence and mortality are not uniform, but instead highly correlated with age. In order to take this into account explicitly, the risk is estimated for each age category separately (or, age/sex category).[6] The overall rate is then obtained as a weighted sum of the age-specific rates, with the proportion of each age category in the total population serving as weight.

Consider the population by age category, $P_h, h = 1, \ldots, H$, and matching observed events, O_h. The age-specific rate then follows as:

$$\hat{\pi}_h = O_h/P_h. \tag{15}$$

The overall summary rate across all age groups is obtained as:

$$\hat{\pi} = \sum_{h=1}^{H} \hat{\pi}_h \times \frac{P_h}{P}. \tag{16}$$

Note that in this summary, the detail provided by the age-specific rates is lost. However, it is often argued that it is easier to compare these summaries and that the information in the age-specific rates may be overwhelming.

In the case where rates are considered for different regions, the summary

---

the variance is greater than the mean, under-dispersion in the reverse case. In either case, the probabilities given by the Poisson distribution will be biased.

[6]The number of categories varies by study and depends on the extent to which the risks vary by age. Typically, 18 categories are used, with five year intervals and all 85+ grouped together. For example, this is the case for data contained in the SEER registries.

rate can be expressed as:

$$\hat{\pi}_i = \sum_{h=1}^{H} \hat{\pi}_{hi} \times \frac{P_{hi}}{P_i}, \tag{17}$$

where the subscript i pertains to the region. The region-specific rate thus combines age-specific risks with the age composition. Different regional rates can result either from heterogeneity in the risks or heterogeneity in the age distribution. To facilitate comparisons, it is often argued that rate differentials due to the latter should be controlled for through the practice of *age-standardizing* the rates. Two approaches are in common use, *direct* standardization and *indirect* standardization. They are considered in turn. For a more in-depth discussion, see, among others, Kleinman (1977), Pickle and White (1995), Goldman and Brender (2000),Rushton (2003, p.47), as well as Waller and Gotway (2004, pp. 11–18).

2.3.1 Direct Standardization

When the main objective is to control for rate differentials that are due to differences in the age distribution, *direct* standardization is appropriate.[7] The principle consists of constructing an overall rate by weighting age-specific risk estimates by the proportion of that age group in a reference population, rather than in the population at hand. More precisely, the population proportions from equation (16) are replaced by the age proportion in a *standard* population, often referred to as the *standard million*. Commonly used standard populations for cancer studies in the U.S. are those for 1940, 1970, and,

---

[7]For example, the National Center for Health Statistics (NCHS) uses direct standardization in its reports of cancer mortality rates. For an illustration, see Pickle et al. (1996, 1999).

more recently, 2000. Formally:

$$\hat{\pi}_{ds} = \sum_{h=1}^{H} \hat{\pi}_h \times \frac{P_{hs}}{P_s}, \tag{18}$$

where P_{hs}/P_s is the share of age category h in the standard population.

To compute this rate in practice, one needs data on age-specific events (O_h), the age-specific population at risk (P_h), as well as the age distribution in the standard population (P_{hs}/P_s). Note that for small regions, the estimate of age-specific risk may be highly imprecise, due to the small number problem (see further Section 2.4).

When the interest is in mapping the rates for geographic areas, a problem arises when maps are compared that use different standard populations. When the age distribution of an area differs greatly from the standard million (e.g., in an area undergoing rapid immigration or population growth), the choice of a standard will affect the relative ranking of the rate. Since most choropleth maps are based on the ranks of the observations, this may yield misleading impressions. The effect of the choice of standard population on directly age-standardized rates is discussed in detail in Krieger and Williams (2001) and Pamuk (2001).

2.3.2   Indirect Standardization

Whereas the motivation for direct standardization is to correct for variability in the age distribution across areas, indirect standardization addresses the potential imprecision of age-specific rates due to the small number problem. An indirectly standardized rate is obtained by using estimates for age-specific risk from a reference population, rather than for the population observed.

This is also the suggested approach when age-specific rates are not available for an area (for example, due to disclosure problems). When the reference rates are obtained by aggregating the areas under consideration (e.g., age-specific rates for a state in a study of the counties in the state), this is referred to as *internal* standardization. The alternative is when the reference age-specific rates are from a different source, such as national or international estimates. This approach is referred to as *external* standardization.

Formally, the indirectly standardized rate is obtained as:

$$\hat{\pi}_{is} = \sum_{h=1}^{H} \hat{\pi}_{hs} \times \frac{P_h}{P}, \tag{19}$$

where $\hat{\pi}_{hs}$ is the reference rate for age group h.

To compute an indirectly age standardized rate in practice, one needs age-specific risks $\hat{\pi}_{hs}$, either computed internally or given externally, as well as the age distribution for each observation (P_h).

The numerator in equation (19) yields the expected number of events if the reference risk applied to the population at risk, $E_h = \hat{\pi}_{hs} P_h$, as in equation (7). The ratio of the observed number of events (O) to the expected events obtained from indirect age standardization is referred to as the standardized mortality ratio (SMR):

$$SMR = \frac{O}{E}, \tag{20}$$

with $E = \sum_h E_h$. This is essentially the same concept as the relative risk given in equation (8), but based on age-specific reference rates. As for the relative risk, an $SMR > 1$ suggests an elevated risk, and vice versa. The SMR is an unbiased estimator of relative risk, but since it only uses a sample

11

of one, it is referred to as a *saturated model* estimate (e.g., Lawson et al. 2000, p. 2219).

While rates derived from indirect age standardization lend themselves readily to mapping and GIS analysis (see, e.g., Waller and McMaster 1997), they are based on the assumption that the influence of age (so-called age effects) and location (so-called area effects) on the estimate of risk are independent, which is often not the case in practice. As a result, indirectly adjusted rates tend not to be comparable across areas. A formal analysis of the conditions under which maps based on direct and indirect age-standardized rates yield similar results (i.e., similar rankings) is outlined in Pickle and White (1995).

2.3.3 Confidence Intervals

The rate estimates considered so far provide point estimates of the underlying risk, but do not contain an indication of the precision of that point estimate. It is standard practice in the public health community to publish direct age standardized rates with a *confidence interval*.

The point of departure is that the age standardized rate in equation (18) can also be written as:

$$\hat{\pi}_{ds} = \sum_{h=1}^{H} \frac{O_h}{P_h} \times \frac{P_{hs}}{P_s}, \tag{21}$$

or,

$$\hat{\pi}_{ds} = \sum_{h=1}^{H} O_h w_h, \tag{22}$$

with the weights $w_h = P_{hs}/(P_h P_s)$. The standard assumption is that the event counts O_h are distributed as independent Poisson random variables

with mean θ_h. Consequently, the unknown rate π can be considered as a weighted sum of independent Poisson random variables, with mean $\mu = \sum_h w_h \theta_h$. A confidence interval for the unknown parameter μ can be constructed by using $\hat{\pi}_{ds}$ as an estimate for its mean and $\nu = \sum_h O_h w_h^2$ as an estimate for its variance.

Early approaches were based on a normal approximation (see, e.g., Clayton and Hills 1993). However, this approximation requires large observed counts in each age interval, which is typically not the case for a disease like cancer. Current practice is based on an approximation by means of a Gamma distribution, due to Fay and Feuer (1997).[8] In this approach, the confidence interval is constructed for a random variable distributed as Gamma(a, b), with a and b as, respectively, the shape and scale parameters, where:[9]

$$a = \frac{\hat{\pi}_{ds}^2}{\nu}, \ b = \frac{\nu}{\hat{\pi}_{ds}}. \tag{23}$$

The lower confidence limit for a significance level of α is obtained as the $\alpha/2$ quantile of the Gamma distribution with parameters a and b from equation (23):

$$L = G(a, \ b)^{-1}(\alpha/2). \tag{24}$$

For the upper limit an adjustment factor k is required, which Fay and Feuer (1997, p. 795) suggest to set to the maximum of the w_h. Then, the adjusted parameters a and b become

$$a_k = \frac{(\hat{\pi}_{ds} + k)^2}{\nu + k^2}, \ b_k = \frac{\nu + k^2}{\hat{\pi}_{ds} + k}, \tag{25}$$

---

[8]See also Dobson et al. (1991) and Swift (1995) for extensive technical discussions.

[9]The Gamma density with shape parameter a and scale parameter b, for a random variable x is $\Gamma(a, b) = [x^{(a-1)}e^{(-x/b)}]/[b^a \Gamma(a)]$, where Γ is the Gamma function. The corresponding mean $E[x] = a/b$ and the variance $Var[x] = a/b^2$.

with the upper confidence limit as the $(1 - \alpha/2)$ percentile of the corresponding Gamma distribution:

$$U = G(a_k, \ b_k)^{-1}(1 - \alpha/2). \tag{26}$$

In practice, a slightly simpler approximation is suggested by NCHS, where the upper and lower confidence limits are first found for a standardized Gamma distribution, using $G(a, \ 1)$ and $G(a + 1, \ 1)$, with a as in equation (23).[10] The confidence limits for the rate are then obtained by rescaling (multiplying) these results by b from equation (23).

One final complication occurs when there are no events and the rate is zero. The lower confidence limit is then set to zero, and only the upper confidence limit is considered. For the count (a Poisson random variable), this simplifies to:

$$U_{O=0} = (1/2)(\chi^2)_2^{-1}(1 - \alpha/2), \tag{27}$$

the $(1 - \alpha/2)$ quantile of a χ^2 distribution with two degrees of freedom. The upper limit for the corresponding rate is found as $U_{O=0}/P$.

2.4 Variance Instability

As shown in equation (4), the precision of the *crude rate* O/P estimate depends on the size of the population at risk in each spatial unit. Unless this population is constant (or very large everywhere) the crude rates become difficult to compare and may spuriously suggest differences in the underlying risk. This is referred to as *variance instability*.

---

[10]http://www.hsph.harvard.edu/thegeocodingproject/webpage/monograph/step%205.htm.

The variance $\pi(1 - \pi)/P$ of $\hat{\pi}$ has two non-standard features. First, the unknown mean π appears, which is referred to as *mean-variance dependence.* For small values of π (which is always less than 1), the second term in the numerator $\pi - \pi^2$ becomes negligible, so that the variance is approximately proportional to the mean. This suggests the use of a simple square root transformation to eliminate the mean-variance dependence (see Cressie 1993, p. 395). Other transformations to deal with this have been suggested as well (see Section 3).

In addition, equation (4) shows how the variance is an inverse function of the size of the population at risk, P. In other words, the smaller P, the less precise the crude rate will be as an estimator for π. This is referred to as the *small number problem.* Also, when the observations have highly varying P, the resulting crude rates will have highly variable precision and be difficult to compare. As a consequence, any map suggesting "outliers" may be spurious, since the extreme values may simply be the result of a higher degree of variability of the estimate.

The same problem also pertains to the SMR as an estimate of relative risk. One can assume the observed counts of events in a region (O_i) to follow a Poisson distribution, as in equation (12). The mean is the expected count, assuming constant risk. However, allowing for heterogeneity in the risk, one can take the mean to be $\theta_i E_i$, where E_i is the expected count assuming constant risk, and θ_i is the relative risk, or:

$$O_i \sim \text{Poisson}(\theta_i E_i), \tag{28}$$

$$\text{Prob}[O_i = x] = \frac{e^{-\theta_i E_i}(\theta_i E_i)^{O_i}}{O_i!}. \tag{29}$$

15

Using standard principles, the maximum likelihood estimator for the relative risk follows from this as:[11]

$$\hat{\theta}_i^{ML} = \frac{O_i}{E_i},$$
(30)

with as variance,

$$\mathrm{Var}[\hat{\theta}_i^{ML}] = \frac{\theta_i}{E_i}.$$
(31)

Since the expected count $(E_i = \tilde{\pi} \times P_i)$ depends on the population at risk, the precision of the SMR estimate will vary inversely with the size of the population and suffer from the same problems as the crude rate.

2.5 Correcting for Variance Instability

The prevalence of variance instability in rates due to the variability of populations across spatial units has received extensive attention in the field of disease mapping. Overviews of the main issues and techniques can be found in, among others, Marshall (1991), Cressie (1992), Gelman and Price (1999), Lawson et al. (1999), Bithell (2000), Lawson (2001b), Lawson and Williams (2001), Lawson et al. (2003), Waller and Gotway (2004, pp.86–104), and Ugarte et al. (2006).

In this report, we group the methods for dealing with this problem into three broad categories: transformations, smoothing and regionalization.

The first category consists of techniques that change the original rate into a different variable through a transformation that removes some of the

---

[11]The log likelihood is, apart from the constant $O_i!$, $lnL = -\theta_i E_i + O_i ln(\theta_i E_i)$. The maximum likelihood estimate follows from the first order condition, setting $\partial lnL/\partial \theta_i = 0$. The variance follows by taking the negative inverse of the second partial derivative and replacing O_i by its expected value $\theta_i E_i$.

mean-variance dependence and/or variance instability.

A second category of techniques, smoothing methods, improves the properties of the estimate of risk (or relative risk) by borrowing information from sources other than the observation at hand. The goal of this exercise is to achieve a better mean squared error of the estimate, trading off bias for greater precision.[12] We distinguish between four types of such techniques. First, we consider mean or median based techniques which obtain a new estimate by applying a moving window to the data. Second, nonparametric approaches smooth the rate estimate by using weighted averages for both the numerator and denominator constructed from observations at neighboring locations. By far the most commonly used smoothing methods are derived from Bayesian principles, in particular the Empirical Bayes approach and the full Bayesian approach, also referred to as model based smoothing. We consider Empirical Bayes as our third set of smoothing techniques, and model based smoothing as the fourth. As mentioned in the introduction, we do not consider smoothing methods that result in a predicted value from a model fit.

The third category of techniques takes a totally different approach. Instead of focusing on the rate, it tackles the problem of instability by increasing the population at risk of a spatial unit, through aggregation with neighboring units. Such regionalization techniques accomplish greater precision at the cost of changing the spatial layout of the observations.

---

[12]The crude rate estimator is unbiased, so this aspect cannot be improved upon. However, other, biased estimators may have a smaller variance, which yields a smaller overall mean squared error (the sum of the variance and the squared bias).

The literature on rate transformations and smoothing in the context of disease mapping is voluminous and still an area of active research. Our coverage can therefore not claim to be complete, but aims at providing a representative sample of the breadth of techniques available. Some recent comparisons among competing techniques can be found in, among others, Kafadar (1994), Gelman et al. (2000), Lawson et al. (2000), and Richardson et al. (2004).[13]

3   Rate Transformations

To deal with the problem of mean-variance dependence and the intrinsic variance instability of rates, a number of transformations have been suggested in the literature. The principle behind these transformations is to obtain a new random variable that no longer suffers from these problems. Note, however, that these transformed variates need not be in the same (or similar) scale as the original variable, so that their actual values may be hard(er) to interpret.

Formally, the problem can be described as one of finding a functional transformation $g(T)$ for a random variable T that removes the dependence of its variance on the mean (Rao 1973, p.426). Following Rao (1973), if

$$\sqrt{n}(T_n - \theta) \to X \sim N[0, \sigma^2(\theta)]^2, \tag{32}$$

then

$$\sqrt{n}[g(T_n) - g(\theta)] \to X \sim N[0, [g'(\theta)\sigma(\theta)]^2], \tag{33}$$

---

[13]See also Section 9.

18

where g is a differentiable function and $g'(\theta) \neq 0$. This can be exploited to find a function g such that the variance in equation (33) becomes independent of θ, or

$$g'(\theta)\sigma(\theta) = c, \tag{34}$$

with c as a constant, independent from θ. Using differential calculus, the function g can be found as a solution to

$$g = c \int \frac{d\theta}{\sigma(\theta)}. \tag{35}$$

In practice, the transformed random variable $g(T_n)$ can be further *standardized* by subtracting its expected value and dividing by its (approximate) standard error,

$$Z = \frac{g(T) - g(\theta)}{\sqrt{c}}, \tag{36}$$

yielding an (approximate, asymptotic) standard normal variate.

A simple transformations is the square root transformation, which is easy to implement. More complex transformations include the *Freeman-Tukey* transformation (Freeman and Tukey 1950), the *arcsin* (Rao 1973, p. 427), *Anscombe* (Anscombe 1948), and the *empirical Bayes (EB)* standardizations (Assunção and Reis 1999). They are briefly considered in turn.

3.1   Freeman-Tukey Transformation

The Freeman-Tukey transformation (FT) (Freeman and Tukey 1950) is a variance controlling transformation that takes out the dependence of the variance on the mean (π).[14] However, it does not control for the variance instability due to unequal populations at risk in the observational units (P_i).

---

[14]For an application in exploratory spatial data analysis, see Cressie (1993, p. 395).

Formally, with O_i as the observed count of events in unit i, and P_i as the population at risk, the FT transformation is:

$$Z_i = \sqrt{O_i/P_i} + \sqrt{(O_i + 1)/P_i}, \tag{37}$$

and, similarly, when a multiplier S is used,

$$Z_i^* = Z_i\sqrt{S}. \tag{38}$$

The variance of the transformed variate is approximately τ^2/P_i, where τ^2 is a constant that no longer depends on π. As pointed out, the dependence on P_i in the denominator has not been eliminated. Often, multiplying Z_i with $\sqrt{P_i}$ will yield the desired result. The transformed values are no longer in the same scale as the original rates and should not be interpreted as such. The main objective of the transformation is to alter the moments so that standard statistical techniques (like OLS regression) can be applied without modification. Alternatively, one could keep the crude rates and deal with the non-standard assumptions with appropriate estimation methods (for example, heteroskedasticity robust regression).

3.2   ArcSin Standardization

The arcsin transformation of the square root of a binomial random variable was suggested in Anscombe (1948) as a procedure to eliminate the dependence of the variance on the mean. In the notation of equation (35), $g = arcsin\sqrt{T}$. The transformed rates are thus:

$$X_i = arcsin\sqrt{\frac{O_i}{P_i}}. \tag{39}$$

20

The asymptotic variance of this random variate is $1/4P_i$. Therefore, a *standardized* variate can be obtained as

$$Z_i = \frac{(X_i - arcsin\sqrt{\hat{\pi}})}{\sqrt{1/4P_i}}, \tag{40}$$

or

$$Z_i = 2(X_i - arcsin\sqrt{\hat{\pi}})\sqrt{P_i}, \tag{41}$$

where $\hat{\pi}$ is the estimated mean rate (risk), $\sum_i O_i / \sum_i P_i$.

3.3 Anscombe Standardization

In addition to the straight arcsin-square root transformation in (39), Anscombe (1948) also suggested a slight variation:

$$X_i = arcsin\sqrt{\frac{O_i + 3/8}{P_i + 3/4}}, \tag{42}$$

which is identical to (39) except for the adjustments in the numerator and denominator. The asymptotic variance of the transformed variate is $1/(4P_i + 2)$. A *standardized* variate is then obtained as

$$Z_i = \frac{(X_i - arcsin\sqrt{\hat{\pi}})}{\sqrt{1/(4P_i + 2)}}. \tag{43}$$

3.4 Empirical Bayes Standardization

An Empirical Bayes (EB) standardization was recently suggested by Assunção and Reis (1999) as a means to correct Moran's I spatial autocorrelation test statistic for varying population densities across observational units, when the variable of interest is a proportion. This standardization borrows ideas from the Bayesian *shrinkage* estimator outlined in Section 6

21

on Empirical Bayes smoothing. However, it should be distinguished from the smoothing in that the original rate is not smoothed, but transformed into a standardized random variable. In other words, the crude rate is turned into a new variable that has a mean of zero and unit variance, thus avoiding problems with variance instability. The mean and variance used in the transformation are computed for each individual observation, thereby properly accounting for the instability in variance. The basis for the estimate of the mean and variance is in a Bayesian model that assumes a prior distribution for the unknown risk π. This is considered in further technical detail in Section 6.

Formally (following Assunção and Reis 1999, pp. 2156–2157), the true rate in each location i is π_i, which is estimated by $\hat{\pi}_i = O_i/P_i$. Using Bayesian terminology, the estimator has a conditional mean

$$\mathrm{E}[\hat{\pi}_i|\pi_i] = \pi_i, \tag{44}$$

and conditional variance

$$\mathrm{Var}[\hat{\pi}_i|\pi_i] = \pi_i/P_i. \tag{45}$$

The underlying rates π_i are assumed to have a prior distribution with mean β and variance α. Taking expectations over the conditioning parameter, and using the law of iterated expectations, this yields the unconditional marginal expectation of $\hat{\pi}_i$ as β and the unconditional marginal variance as $\alpha + \beta/P_i$. The incorporation of the prior structure removes the instability in the mean, but retains the instability in the variance.

The rationale behind the Assunção-Reis approach is to standardize each

22

$\hat{\pi}_i$ as

$$Z_i = \frac{\hat{\pi}_i - \hat{\beta}}{\sqrt{\hat{\alpha} + (\hat{\beta}/P_i)}}, \tag{46}$$

using an estimated mean $\hat{\beta}$ and standard error $\sqrt{\hat{\alpha} + \hat{\beta}/P_i}$.

Estimation of the α and β parameters is typically based on a method of moments approach due to Marshall (1991) (see also Bailey and Gatrell 1995, pp. 304-306, for a detailed description):

$$\hat{\beta} = O/P, \tag{47}$$

and

$$\hat{\alpha} = [\sum_i P_i(p_i - \hat{\beta})^2]/P - \hat{\beta}/(P/N), \tag{48}$$

with $E = \sum_i E_i$ and $P = \sum_i P_i$ as the totals for the events and population at risk, and where N is the total number of observations.[15]

One problem with the method of moments estimator is that equation (48) could yield a negative value for $\hat{\alpha}$. In that case, typically $\hat{\alpha} = 0$, as in Bailey and Gatrell (1995, p. 306). In Assunção and Reis (1999), the $\hat{\alpha}$ is only set to zero when the resulting estimate for the variance is negative, that is, when $\hat{\alpha} + \hat{\beta}/P_i < 0$ (Assunção and Reis 1999, p. 2157). Slight differences in the standardized variates may result from this.

An application of this technique in the analysis of lung cancer rates is illustrated in Goovaerts and Jacquez (2004).

---

[15]Hence, P/N is the population each location would have if they all had an equal share of the total, or, the average population.

4 Mean and Median Based Smoothing

Straightforward smoothing of rates can be obtained by constructing a local average of rates or a local median of rates, either weighted or unweighted (see, e.g., Waller and Gotway 2004, pp. 87–88). A special case of a weighted local median smoother is the weighted headbanging technique, used in several recent U.S. cancer mortality atlases (e.g., Pickle et al. 1999). These techniques are briefly considered in turn.

4.1 Locally Weighted Averages

A locally weighted rate average is an example of disk smoothing or a moving average window, where a value at a location is replaced by an average based on observations that include surrounding locations. For each location, a neighborhood set needs to be established, e.g., by means of a spatial weight w_{ij} that equals 1 for each neighbor j of i, and zero otherwise. Note, that unlike standard practice in the analysis of spatial autocorrelation, the location itself is included in the weights, such that $w_{ii} = 1$.[16] The neighbor relation can be based on a number of different criteria, such as a distance threshold ($w_{ij} = 1$ for $d_{ij} < \delta$), k nearest neighbors, or contiguity ($w_{ij} = 1$ for i and j sharing a common boundary).

With $\hat{\pi}_i$ as the usual unbiased estimate for location i ($\hat{\pi}_i = O_i/P_i$), the window weighted average rate becomes:

$$\bar{\pi}_i = \frac{\sum_j w_{ij}\hat{\pi}_j}{\sum_j w_{ij}}. \tag{49}$$

---

[16]In spatial autocorrelation analysis, the standard approach is to set $w_{ii} = 0$.

Similarly, the window weighted average SMR or relative risk estimate (with the estimated SMR at i as $\hat{\theta}_i = O_i/E_i$) can be obtained as:

$$\bar{\theta}_i = \frac{\sum_j w_{ij} \hat{\theta}_j}{\sum_j w_{ij}}. \tag{50}$$

In what follows, the examples will be given for the crude rate, but it should be kept in mind that all smoothers can equally well be applied to the SMR, with the expected population E_i playing the same role as the population at risk P_i for the crude rate.

When the weights take binary values of 0, 1, the smoothed rate is a simple average of the original rates in the window. This ignores the difference in precision of the rates included in the average.

This problem can be addressed by using a general spatial weight, which combines both the spatial characteristic (the definition of the neighbors in the window) with a value, such as the population. For example, with $w_{ij} = P_j$ for $j \in J_i$ with J_i as the neighbor set centered on location i (and including i), and $w_{ij} = 0$ elsewhere, the window average becomes the unbiased rate estimate for the region covered by the window.[17] This is equivalent to a spatial rate smoother, further considered in Section 5.2.

Other weights have been suggested as well, such as inverse distance ($w_{ij} = 1/d_{ij}$, with d_{ij} as the distance between i and j), a power of the inverse distance (e.g., Kafadar 1994, p. 426), or a kernel smoother (e.g., Lawson et al. 2000, p.2220).[18]

---

[17]With population weights, the average becomes $\sum_j (P_j/P_{J_i}) \times \hat{\pi}_j$, for $j \in J_i$ and P_{J_i} as the total population at risk within the window centered on i. See also equation (11).

[18]See also Section 5.2.1 on the use of kernel smoothers.

4.2 Locally Weighted Medians

Disk smoothing can also be implemented by taking the median of the rates within the window. The smoothed rate then becomes:

$$\bar{\pi}_i = \text{median}(\hat{\pi}_j), \text{ for } j \in J_i, \tag{51}$$

where, as before, J_i is the neighbor set for the window centered on location i (and including i). This neighbor set can be based on a distance threshold, k nearest neighbors, or contiguity.

This procedure can be iterated, by including the value of $\bar{\pi}_i$ from the previous iteration instead of $\hat{\pi}_i$ in the computation of the median, until no further change occurs. This is referred to as iteratively *resmoothed* medians.

To compensate for variance instability, the pure median can be replaced by a *weighted median*. As weights for crude rates, one could use the corresponding population, and weights for SMR would be the inverse standard error (e.g., as in Mungiole et al. 1999).

To compute a weighted median, the original values are first sorted and matched with a cumulative sum of the weights. Consider the index $h = 1, \ldots, N$, giving the rank of the original data, and matching weights w_h. For each rank, the cumulative sum of the weights is computed, say $s_h = \sum_{i \leq h} w_i$, with $s_{tot} = \sum_i w_i$ as the total cumulative sum of weights. The weighted median is the value matching position $m = \min\{h | s_h \geq s_{tot}/2\}$.[19]

To illustrate this smoothing procedure, consider the example of an imaginary city with a crude rate of 10 (per 100,000) and population of 100 (thou-

---

[19]By convention, if $s_m > s_{tot}/2$, there is an unambiguous assignment of the weighted median. If $s_m = s_{tot}/2$, then the weighted median is the average of the value at rank m and rank $m + 1$ (see, e.g., Mungiole et al. 1999, p. 3203).

sand), surrounded by smaller rural areas with respective rates and populations of: 9-50, 25-10, 30-20, and 32-10. The sorted rates are: 9, 10, 25, 30, 32. The median smoother would thus be 25. The matching sorted weights (populations) are: 50, 100, 10, 20, 10. The associated cumulative sums are: 50, 150, 160, 180, 190. Since $150 \geq 190/2$, the weighted median smoother is 10, retaining the original rate for the city under consideration. In other words, the possibly spurious effect of the surrounding small areas is eliminated, whereas in the simple median smoother a much higher smoothed rate would be assigned to the city. As for the simple median smoother, a weighted median smoother can be iterated.[20]

4.2.1 Headbanging

An interesting variant of median disk smoothing is the so-called *headbanging* method, initially introduced for irregular spatial data (as opposed to gridded data) by Hansen (1991). Rather than considering all "neighbors" contained in a moving window, the headbanging method focuses on so-called *triples* of near collinear points, that contain the location under consideration in the center. The selection of collinear points enables directional trends to be recognized and enhanced, whereas a traditional disk smoother would tend to eliminate these effects. The objective of headbanging is to preserve insight into broad regional trends and edges while removing small scale local variability. While originally developed for points, this smoother can easily be applied to regional data as well, where the region centroids form the relevant

---

[20]In an iterated weighted median smoother, the smoothed rate replaces the original rate, but the weights are left unchanged.
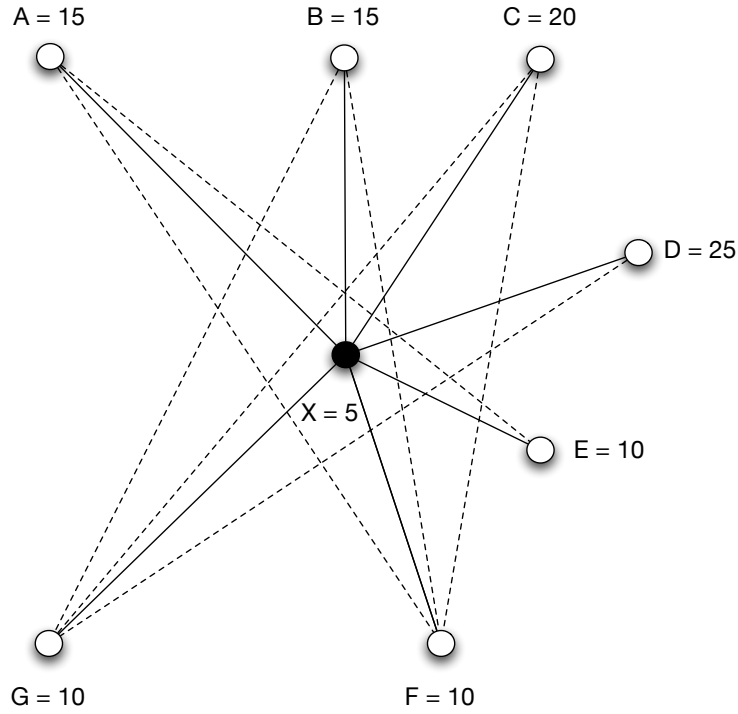
Figure 1: Headbanging smoothing example

points.

The headbanging algorithm outlined in Hansen (1991) consists of two main stages: the selection of triples of points to be used in the smoothing and the smoothing itself. To illustrate the rationale behind the algorithm, consider the simple example in Figure 1. The central point is surrounded by seven nearest neighbors, with their values listed. The value for the central point is 5.

The nearest neighbors constitute a candidate list from which the appropriate triples are selected. This is one of the parameters to be chosen when

implementing the algorithm in practice. In order to retain directional effects, only those triples are selected (containing two nearest neighbors of X as endpoint and X in the middle) that form an angle of more than 135° (i.e., $90^o + 45^o$).[21] In our example, these would be the triples A-X-E, A-X-F, B-X-F, B-X-G, C-X-F, C-X-G, and D-X-G.

A second criterion (mostly to enhance computational speed) is the number of triples to be used in the smoothing. Only those triples for which the distance between X and the line segment connecting the endpoints (the dashed line in Figure 1) is the smallest are kept. In our example, using 3 as the number of triples to be used for smoothing, the selected ones would be A-X-E, B-X-F and C-X-G, with values 15-5-10, 15-5-10 and 20-5-10.

The second step in the algorithm is the actual smoothing process. Two lists are constructed from the values for the triples, one containing the lowest values (low_i), the other containing the highest values ($high_i$). In our example, the lists would be: low $= \{10, 10, 10\}$, and high $= \{15, 15, 20\}$. A low screen and high screen are defined as the median in each list, e.g., 10 for low and 15 for high. The smoothed value for X is the median of the original value at X, the low screen and the high screen, i.e., the median of 5, 10, 15, or 10. The value at X is replaced by 10.[22]

This process is carried out for each point on the map, after which all the points are updated and the process is repeated in the next iteration, up to the specified number of iterations.

[21]The choice of this angle is arbitrary, but 135°, the value suggested by Hansen (1991) seems to be the value most often used in practice.

[22]Compare this to the median smoothing using the complete window, which would yield (10+15)/2=12.5, or the window average, which would yield 13.75.
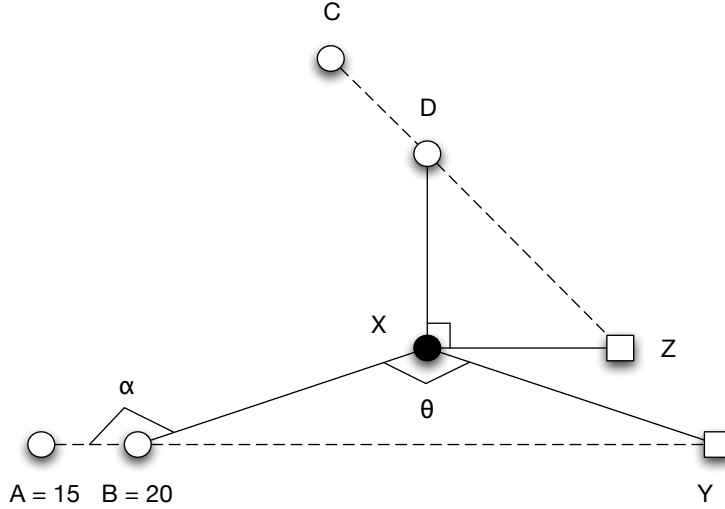
Figure 2: Headbanging extrapolation for edge effects

For edge points, the standard way to construct triples breaks down, since there will be no nearest neighbors outside the convex hull of the data set. Hansen (1991) suggests an extrapolation approach to create artificial triples in order to carry out the smoothing for edge points. Consider the example in Figure 2 with the edge point X and nearest neighbors A through D. The basic idea behind the extrapolation approach is to create artificial endpoints by linear extrapolation from two "inner" nearest neighbors to a location on the other side of X. However, in order to be valid, this location must satisfy the angular constraint for triples. Consider specifically the points A and B relative to X. A linear extrapolation yields the point Y at the same distance from X as B and on the line connecting A and B. The angle θ must satisfy the 135° constraint. Alternatively, this constraint boils down to $\alpha \geq 90^o + \theta/2$. This is satisfied for the triple B-X-Y in Figure 2, but clearly not for another

candidate triple, D-X-Z, which uses extrapolation through C-D. The value assigned to the artificial endpoint is a simple linear extrapolation between the values of A and B: $Y = B + (B - A)(d_{BY}/d_{AB})$ (Hansen 1991, p. 373). In our example, this would be $Y = 20 + (20 - 15)(6/1) = 50$.

The headbanging smoother is compared to a number of other approaches in Kafadar (1994) and performs well. However, it does not address the variance instability encountered in rates. To incorporate this aspect into the smoothing algorithm, Mungiole et al. (1999) introduce a weighted headbanging method, suggesting the use of the inverse standard error as the weights (e.g., for crude rates this would be the population size).

This proceeds along the same lines as the standard headbanging technique, except that a weighted median is used in the selection of the high screen and the low screen. If the center point value is between these two screens, its value is unchanged. If, on the other hand, the center point value is below the low screen or above the high screen, the weights are used to decide on the adjustment. When the sum of the weights of the triples' endpoints is greater than the weight of the center point times the number of triples, the adjustment is made (i.e., the low screen or high screen is assigned to the center point). However, the weight of the center point is not adjusted.

The weighted headbanging smoother is applied in several mortality atlases (e.g., Pickle et al. 1996, 1999) and is compared to other smoothers in Kafadar (1999) (for an alternative illustration, see also Wall and Devine 2000). In a detailed simulation study, Gelman et al. (2000) assessed the performance of unweighted and weighted headbanging for a range of simulated data sets. They caution against the uncritical use of these smoothers, since artifacts

can be introduced that do not reflect the true underlying spatial structure.

5 Nonparametric Smoothing

Nonparametric smoothers provide estimates for the unknown risk or relative risk without making distributional assumptions. We consider two broad classes of this technique: spatial filters, and their generalization in the form of spatial rate smoothers.

5.1 Spatial Filters

The original use of spatial filters or disk smoothers occurred in the context of dealing with point data, such as cases and controls for a particular disease. Rather than estimating risk based on aggregates of these events to arbitrary administrative units, the estimate in a spatial filter is the ratio of the count of events to the count of population at risk within a given disk. Early suggestions of this approach can be found in Bithell (1990) and Openshaw et al. (1987, 1990). It was popularized as *spatial filtering* in Rushton and Lolonis (1996) (see also, among others, Rushton et al. 1996, Rushton and West 1999, Rushton 2003). It is also referred to as *punctual kriging* in the geostatistical literature (Carrat and Valleron 1992, Webster et al. 1994).

The idea behind this technique is to create a series of circular moving windows or disks centered on point locations on a regular grid that covers the study area. For each of the disks, the rate is computed as the crude rate of events to population, or, alternatively, as the ratio of cases to controls. The resulting values at the grid point are then represented in smooth form

by means of contour plots, a three-dimensional surface, or similar techniques.

In Rushton and Lolonis (1996), both the events and the population at risk are completely enumerated, so that the rate computed from the data included in a circular area is an actual estimate of the risk in that area. In contrast, in some applications, only the locations of controls are available, yielding a *pseudo rate*, since the controls are not the actual population at risk, but only a sample representative of the spatial distribution of the population (e.g., Paulu et al. 2002). In the latter instance, a more appropriate approach is to compute an odds ratio (similar to a standardized incidence rate) by dividing the disk pseudo rate by the pseudo rate of the entire study area (for details, see Paulu et al. 2002).

While the typical application centers the disks on regular grid points, sometimes the actual locations of the events are used as the center of the circular areas (e.g., Paulu et al. 2002). Also, individual event or population locations are not necessarily available, and instead they may be represented by the centroids (or population weighted centroids) of administrative areas, such as zip code zones (e.g., Talbot et al. 2000).

As such, the spatial filter does not address the variance instability inherent in rates, since the population at risk included in the circular area is not necessarily constant. The spatial filter with fixed population size and the spatial filter with nearly constant population size attempt to correct for this by adaptively resizing the circular area such that the population at risk is near constant. The resulting variable spatial filter, also referred to as *adaptive k smoothing* yields risk estimates that are equally reliable (Talbot et al. 2000, 2002, Paulu et al. 2002).

In practice, the near constant population size is achieved by incrementally adding population locations or population counts in centroids in increasing order of nearest neighbor to the grid reference point, until an acceptable base population size is reached (e.g., 500, 1000, 2500). When working with aggregate data, it is not always possible to achieve the target exactly, and an apportioning method is sometimes used. In this approach, the same percentage of cases in an aggregate unit is counted as the percentage of the population needed to reach the target size (see Talbot et al. 2000).

5.2   Spatial Rate Smoother

A spatial rate smoother is a special case of a nonparametric rate estimator based on the principle of locally weighted estimation (e.g., Waller and Gotway 2004, pp. 89–90). Rather than applying a local average to the rate itself, as in Section 4, the weighted average is applied separately to the numerator and denominator. As a result, the smoothed rate for a given location i becomes:

$$\bar{\pi}_i = \frac{\sum_{j=1}^{N} w_{ij} O_j}{\sum_{j=1}^{N} w_{ij} P_j}, \tag{52}$$

where the w_{ij} are weights. Similarly, for the SMR, the smoothed relative risk estimate at i is obtained as:[23]

$$\bar{\theta}_i = \frac{\sum_{j=1}^{N} w_{ij} O_j}{\sum_{j=1}^{N} w_{ij} E_j}. \tag{53}$$

Different smoothers are obtained for different spatial definitions of neighbors and when different weights are applied to those neighbors.

---

[23]As before, we will primarily use the crude rate in the discussion that follows, although it should be clear that the rate smoothing principles equally apply to relative risk estimates.

A straightforward example is the spatial rate smoother outlined in Kafadar (1996), based on the notion of a spatial moving average or *window average*. This is similar to the locally weighted average in Section 4.1. Both numerator and denominator are (weighted) sums of the values for a spatial unit together with a set of "reference" neighbors, S_i. As before, the neighbors can be defined in a number of different ways, similar to the way in which spatial weights matrices are specified. In the simplest case, the smoothed rate becomes:

$$\bar{\pi}_i = \frac{O_i + \sum_{j=1}^{J_i} O_j}{P_i + \sum_{j=1}^{J_i} P_j}, \tag{54}$$

where $j \in S_i$ are the neighbors for i. As in the case of a spatial filter, the total number of neighbors for each unit, J_i is not necessarily constant and depends on the definition used (the type of "spatial weights"). Consequently, the smoothed rates do not necessarily have constant variance, although by increasing the denominator, the small number problem has been attenuated.

A map of spatially smoothed rates tends to emphasize broad trends and is useful for identifying general features of the data (much less "spiked" than the crude rates). However, it is not useful for the analysis of spatial autocorrelation, since the smoothed rates are autocorrelated by construction. It is also not as useful for identifying individual outliers, since the values portrayed are really "regional" averages and not specific to an individual location. By construction, the values shown for individual locations are determined by both the events and the population size of adjoining spatial units, which can give misleading impressions.

The simple spatial rate smoother can be readily extended by implementing distance decay or other weights, which lessens the effect of neighboring

units on the computed rate. Examples are inverse distance and inverse distance squared (Kafadar 1996), or weights associated with increasing order of contiguity (e.g., Ministry of Health 2005, use 1 as the weight for first order neighbors, 0.66 for second order and 0.33 for third order neighbors). As for the spatial filter, the spatial range for the smoother can be adjusted to achieve a constant or nearly constant population size for the denominator (e.g., in Ministry of Health 2005).

The spatial rate smoother is easy to implement in software systems for exploratory spatial data analysis, as illustrated in Anselin et al. (2004, 2006), and Rey and Janikas (2006), among others.

5.2.1   Kernel Smoothers

A general approach to combine the choice of neighbors with differential weights in a rate smoother is the use of a two-dimensional *kernel* function. In essence, a two-dimensional kernel is a symmetric bivariate probability density function with zero mean. As such, it integrates to 1 over the area that it covers. All kernel functions are distance decay functions, where the rate and range of the decay is determined by the functional form of the kernel and the bandwidth. The bandwidth is the threshold distance beyond which the kernel is set to zero.

Formally, a kernel function can be expressed as $K_{ij}(d_{ij}/h)$, where K stands for a particular functional form, d_{ij} is the distance between i and j, and h is the bandwidth. A commonly used example is the Gaussian ker-

36

nel:[24]

$$K_{ij} = \frac{1}{\sqrt{2\pi h}} \exp[-d_{ij}^2/2].$$ (55)

A kernel smoother for a rate is then obtained as:[25]

$$\bar{\pi}_i = \frac{\sum_{j=1}^{N} K(d_{ij}/h)O_i}{\sum_{j=1}^{N} K(d_{ij}/h)P_i},$$ (56)

with $K(d_{ij}/h)$ as a kernel function. In Kelsall and Diggle (1995) it is shown how a risk surface can be estimated non-parametrically as the ratio of two kernel functions, which provides the theoretical basis for the use of kernel smoothers (see also Diggle 2003, pp. 133–234).[26]

5.2.2 Age-Adjusted Rate Smoother

A smoother that is particularly useful when direct age-standardization is used is suggested by Kafadar (1996)(see also Kafadar 1999). This approach combines a smoothing across space of age-specific counts with a direct age-standardization across the smoothed age-specific events and populations.

Consider the direct age-standardized rate in location i as:

$$\hat{\pi}_{si} = \sum_h p_{hs} \frac{O_{hi}}{P_{hi}},$$ (57)

---

[24]An extensive discussion of kernel smoothers can be found in Simonoff (1996, Chapters 3–4).

[25]This is sometimes referred to as a Nadaraya-Watson kernel smoother, e.g., in (Lawson et al. 2000, p. 2220).

[26]In Kelsall and Diggle (1998), this is further extended to allow covariates in a regression approach. The log odds of a disease can be estimated by using a non-parametric logistic regression for a binary random variable that takes on the value of 1 for an event and 0 for a control. Since this falls under a "modeling" approach, we do not consider this further in this report.

where h stands for the age group, p_{hs} is the share of age group h in the standard population, and O_{hi} and P_{hi} are age-specific events and population at risk.

The age-specific weighted ratio smoother first smoothes the age-specific events and age-specific populations separately:

$$O_{hi}^* = \sum_j w_{ij} O_{hj}, \tag{58}$$

and

$$P_{hi}^* = \sum_j w_{ij} P_{hj}, \tag{59}$$

where w_{ij} are the weights.

In a second step, these smoothed counts are combined into a rate, using the population shares in the standard population:

$$\bar{\pi}_i = \sum_h p_{hs} \frac{O_{hi}^*}{P_{hi}^*}. \tag{60}$$

Kafadar (1999, p. 3171) suggests the use of weights that include both a distance decay effect and a correction for variance instability:

$$w_{ij} = (\sum_h P_{hj})^{1/2} [1 - (d_{ij}/d_{max})^2]^2, \tag{61}$$

for locations j within the maximum distance threshold d_{max}, and zero otherwise. The specific distance decay function is known as the bisquare kernel function, and is only one of many options that could be considered.

A specific application of this approach to prostate cancer rates is given in Kafadar (1997).

6 Empirical Bayes Smoothers

The Empirical Bayes smoother uses Bayesian principles to guide the adjustment of the crude rate estimate by taking into account information in the rest of the sample. The principle is referred to as *shrinkage*, in the sense that the crude rate is moved (shrunk) towards an overall mean, as an inverse function of the inherent variance. In other words, if a crude rate estimate has a small variance (i.e., is based on a large population at risk), then it will remain essentially unchanged. In contrast, if a crude rate has a large variance (i.e., is based on a small population at risk, as in small area estimation), then it will be "shrunk" towards the overall mean. In Bayesian terminology, the overall mean is a *prior*, which is conceptualized as a random variable with its own "prior" distribution.

We begin the review of EB methods with a brief discussion of the general principle behind the Bayesian approach. This is followed by a description of two examples of a fully parametric approach. We consider two different classes of distributions as the prior for the risk or relative risk, the Gamma distribution and the Normal and related distributions. Next, we review estimation of the EB smoothers, starting with the method of moments estimation (referred to as the global linear Bayes and local linear Bayes smoothers). We also briefly review a number of alternative approaches to obtain estimates for the moments of the prior distribution. We next outline ways in which the precision of the estimates has been assessed in the literature. We close the section with a cursory discussion of nonparametric mixture models.

6.1 Principle

In Bayesian statistical analysis, both the data and parameters are considered to be random variables. In contrast, in classical statistics, the parameters are taken to be unknown *constants*.[27] The crucial concept is Bayes' rule, which gives the connection between the prior distribution, the likelihood of the data, and the posterior distribution. Formally, interest centers on a parameter η, and data are available in the form of a vector of observations y. All information about the distribution of the parameter *before* the data are observed is contained in the so-called *prior* distribution, $f(\eta)$. The likelihood of the data is then seen as a conditional distribution, conditional upon the parameter η, $f(y|\eta)$. Bayes' rule shows how the posterior distribution of η, conceptualized as the new information about the uncertainty in η, *after* the data is observed, is proportional to the product of the likelihood and the prior:

$$f(\eta|y) \propto f(y|\eta) \times f(\eta). \tag{62}$$

In the context of disease mapping, the observed number of events in a spatial unit i, O_i (disease counts, counts of deaths) is assumed to follow a Poisson distribution with mean (and variance) either $\pi_i P_i$ or $\theta_i E_i$. More precisely, the distribution of O_i is *conditional* on the unknown risk or relative risk parameters. The former case is relevant when attention focuses on estimates of risk (π_i) through the crude rate $\hat{\pi}_i$, the latter when interest centers on the relative risk (θ_i), through the SMR $\hat{\theta}_i$. In terms of the Empirical

---

[27]An extensive review of Bayesian principles is outside the current scope. Recent overviews can be found in, among others, Carlin and Louis (2000), Congdon (2001, 2003), Gill (2002), and Gelman et al. (2004).

Bayes (and Bayes) smoothing, the formal treatment of both cases follows an identical logic, using either P_i or E_i in the relevant expressions.

It follows then that:

$$O_i|\pi_i \sim \text{Poisson}(\pi_i P_i), \tag{63}$$

or,

$$O_i|\theta_i \sim \text{Poisson}(\theta_i E_i). \tag{64}$$

These conditional distributions are assumed to be independent. In other words, the conditional independence implies that any spatial patterning in the counts of events is introduced through patterning in either the parameters (the π_i or θ_i) or in the population at risk (the P_i), but not directly in the counts once these parameters are known (the conditional distribution).

The Empirical Bayes approach towards rate smoothing, originally spelled out in Clayton and Kaldor (1987), consists of obtaining estimates for the parameters of the prior distribution of π_i or θ_i from the marginal distribution of the observed events O_i. With these estimates in hand, the mean of the posterior distribution of π_i or θ_i can then be estimated. This mean is the Empirical Bayes estimate or smoother for the underlying risk or relative risk.

Formally, the prior distribution $f(\pi_i)$ or $f(\theta_i)$ can be taken to have mean μ_i and variance σ_i^2. A theoretical result due to Ericson (1969) shows how under a squared error loss function, the best linear Bayes estimator is found to be a linear combination of the prior mean μ_i and the crude rate or SMR. For example, for the SMR, this yields:

$$\hat{\theta}_i^{EB} = \mu_i + \frac{\sigma_i^2}{(\mu_i/E_i) + \sigma_i^2}(\hat{\theta}_i - \mu_i), \tag{65}$$

41

or, alternatively:

$$\hat{\theta}_i^{EB} = w_i\hat{\theta}_i + (1 - w_i)\mu_i, \tag{66}$$

with

$$w_i = \frac{\sigma_i^2}{\sigma_i^2 + (\mu_i/E_i)}. \tag{67}$$

The EB estimate is thus a weighted average of the SMR and the "prior," with weights inversely related to their variance. By assumption, the variance of the prior distribution is σ_i^2. The variance of the SMR is found by applying standard properties relating marginal variance to conditional mean and conditional variance, and follows as $\sigma_i^2 + (\mu_i/E_i)$.[28]

The EB estimate for the underlying risk is similarly:

$$\hat{\pi}_i^{EB} = w_i\hat{\pi}_i + (1 - w_i)\mu_i, \tag{68}$$

but now with the weights as:

$$w_i = \frac{\sigma_i^2}{\sigma_i^2 + (\mu_i/P_i)}, \tag{69}$$

where μ_i and σ_i^2 are the corresponding moments of the prior distribution for π.

When the population at risk is large (and similarly, E_i), the second term in the denominator of (67) or (69) becomes near zero, and $w_i \to 1$, giving all

[28]From equation (64) it follows that $E[O_i|\theta_i] = \text{Var}[O_i|\theta_i] = \theta_iE_i$. Consequently, the conditional mean of the SMR is $E[\hat{\theta}_i|\theta_i] = E[O_i|\theta_i]/E_i = \theta_i$, and the conditional variance is $\text{Var}[\hat{\theta}_i|\theta_i] = \text{Var}[O_i|\theta_i]/E_i^2 = \theta_i/E_i$. The marginal (unconditional) mean follows from the law of iterated expectations as $E[\hat{\theta}_i] = E_{\theta_i}[E[\hat{\theta}_i|\theta_i]] = E_{\theta_i}[\theta_i] = \mu_i$, the prior mean. The marginal variance follows similarly from the variance decomposition into the variance of the conditional mean and the mean of the conditional variance: $\text{Var}[\hat{\theta}_i] = \text{Var}_{\theta_i}[E[\hat{\theta}_i|\theta_i]] + E_{\theta_i}[\text{Var}[\hat{\theta}_i|\theta_i]]$. The first term in this expression is $\text{Var}_{\theta_i}[\theta_i] = \sigma_i^2$, the prior variance. The second term is $E_{\theta_i}[\theta_i/E_i] = \mu_i/E_i$.

the weight in (66) to the SMR and in (68) to the crude rate estimate. As P_i (and thus E_i) gets smaller, and thus the variance of the original estimate grows, more and more weight is given to the second term in these equations, i.e., to the prior.

Empirical Bayes estimators differ with respect to the extent to which distributional assumptions are made about the prior distribution (i.e., parametric, as in Sections 6.2 and 6.3, vs. non-parametric, as in Section 6.7), and how the moments of the prior distribution are estimated (e.g., maximum likelihood, method of moments, estimating equations, or simulation estimators). We now turn to these aspects more specifically, beginning with the consideration of two commonly used parametric models. Recent reviews of the technical aspects of Empirical Bayes smoothers can be found in Meza (2003) and Leyland and Davies (2005), among others.

6.2   The Gamma Model

In the Gamma model, or Gamma-Poisson model, the prior distribution for the risk or relative risk is taken to be a Gamma distribution with shape parameter α and scale parameter β:

$$\pi_i \sim \text{Gamma}(\alpha, \beta), \tag{70}$$

and,

$$\theta_i \sim \text{Gamma}(\alpha, \beta). \tag{71}$$

Therefore, the mean and variance of the prior distribution for the risk and for the SMR can be expressed in terms of the shape and scale of the

Gamma density as $\mathrm{E}[\pi_i] = \mu = \alpha/\beta$, and $\mathrm{Var}[\pi_i] = \sigma^2 = \alpha/\beta^2$, and similarly for the SMR.

Using standard Bayesian principles, the combination of a Gamma prior with a Poisson density for the likelihood (hence Gamma-Poisson model) yields a posterior density for the risk or relative risk that is also distributed as Gamma.[29] Formally,

$$\pi_i | O_i, \alpha, \beta \sim \mathrm{Gamma}(O_i + \alpha, P_i + \beta), \tag{72}$$

and,

$$\theta_i | O_i, \alpha, \beta \sim \mathrm{Gamma}(O_i + \alpha, E_i + \beta), \tag{73}$$

where the parameters in parentheses for the Gamma distribution are respectively the shape and scale parameters. Consequently, the mean and variance of the posterior distribution follow as:

$$\mathrm{E}[\pi_i | O_i, \alpha, \beta] = \frac{O_i + \alpha}{P_i + \beta}, \tag{74}$$

$$\mathrm{Var}[\pi_i | O_i, \alpha, \beta] = \frac{O_i + \alpha}{(P_i + \beta)^2}, \tag{75}$$

and,

$$\mathrm{E}[\theta_i | O_i, \alpha, \beta] = \frac{O_i + \alpha}{E_i + \beta}. \tag{76}$$

$$\mathrm{Var}[\theta_i | O_i, \alpha, \beta] = \frac{O_i + \alpha}{(E_i + \beta)^2}. \tag{77}$$

With estimates for α and β in hand, these expressions can be used as smoothed estimates for the risk and SMR, and associated standard errors.

[29]See, e.g., Gelman et al. (2004, pp. 51–55). Note that the marginal distribution of the counts O_i is no longer Poisson, but becomes negative binomial in this case.

Alternatively, the full posterior density can be used to make inference about the risk and SMR.

The mean of this posterior distribution corresponds to the Empirical Bayes estimator (in the sense of minimizing expected square loss). For example, using equation (69) directly, with $\mu_i = \alpha/\beta$ and $\sigma_i^2 = \alpha/\beta^2$ yields:

$$w_i = \frac{\alpha/\beta^2}{(\alpha/\beta^2) + \alpha/(\beta P_i)}, \tag{78}$$

or,

$$w_i = \frac{P_i}{P_i + \beta}. \tag{79}$$

The corresponding weight for the SMR case is $w_i = E_i/(E_i+\beta)$. Substituting (79) into the expression for the Empirical Bayes estimator for π_i, yields:

$$\hat{\pi}_i^{EB} = \frac{P_i}{P_i + \beta}(O_i/P_i) + \frac{\beta}{P_i + \beta}(\alpha/\beta), \tag{80}$$

which simplifies to:

$$\hat{\pi}_i^{EB} = \frac{O_i + \alpha}{P_i + \beta}, \tag{81}$$

and similarly for the SMR, with P_i replaced by E_i.

Different estimators for α and β result in different Empirical Bayes estimates, which is revisited in Sections 6.4 and 6.5. An early application of this method can be found in Manton et al. (1989).

6.3   Priors Related to the Normal Distribution

6.3.1   The Log-Normal Model

Clayton and Kaldor (1987) also suggested the use of a log-normal distribution as an alternative prior for the risk or relative risk. Formally:

$$\pi_i = e^{\gamma_i}, \text{or}, \gamma_i = log(\pi_i), \tag{82}$$

or,

$$\theta_i = e^{\gamma_i}, \text{or}, \ \gamma_i = log(\theta_i), \tag{83}$$

with

$$\gamma \sim MVN(\mu, \Sigma). \tag{84}$$

Here γ is an N by 1 vector of the γ_i, μ is a matching vector of prior means, and Σ is a variance covariance matrix.

In this specification, the mean of the posterior distribution cannot be expressed in closed form. Instead, Clayton and Kaldor (1987) use a quadratic approximation.

One further complication is that neither $log(\hat{\theta}_i)$ nor $log(\hat{\pi}_i)$ are defined for $O_i = 0$. Clayton and Kaldor (1987) suggest the use of a bias-corrected version, with $\hat{\pi}_i = (O_i + 0.5)/P_i$, and $\hat{\theta}_i = (O_i + 0.5)/E_i$. The analysis then focuses on $\hat{\gamma}_i = log[(O_i+0.5)/P_i]$, or $\hat{\gamma}_i = log[(O_i+0.5)/E_i]$ and its Empirical Bayes estimator.

For the simple case with an i.i.d. prior $N(\mu, \sigma^2)$, the Empirical Bayes estimator for γ_i can be found as:[30]

$$\hat{\gamma}_i^{EB} = \frac{\mu + \sigma^2(O_i + 0.5)\hat{\gamma}_i - 0.5\sigma^2}{1 + \sigma^2(O_i + 0.5)} \tag{85}$$

With estimates for μ and σ in hand, equation (85) can be evaluated, and the EB smoothed rate or relative risk follows as $\exp(\hat{\gamma}_i^{EB})$.

The prior (84) also provides a natural means to introduce explanatory variables in the model, by setting

$$\mu_i = x_i\beta, \tag{86}$$

_____

[30]For technical details, see Clayton and Kaldor (1987, pp. 674–675) and Meza (2003, p. 49), among others.

where x_i is a row vector of observations on explanatory variables, and β is a matching coefficient vector. This is equivalent to specifying:

$$\gamma_i = x_i\beta + \xi_i, \tag{87}$$

where the error terms ξ_i follow a MVN distribution (see, e.g., Meza 2003, p. 49).

6.3.2 The Normal Model

Rather than using the log-normal transformation (82) or (83), Cressie (1992, 1995) uses the normal distribution directly. As a result, the prior for the unknown risk vector π becomes:

$$\pi \sim MVN[X\beta, \Sigma(\gamma)], \tag{88}$$

with X as a matrix of observations on explanatory variables, β as a matching vector of coefficients, and Σ as the variance-covariance matrix, itself parameterized as a function of γ. The latter allows for the introduction of spatial effects, such as spatial autocorrelation in the form of a conditional autoregressive model. A detailed discussion of this model and its estimation is beyond the current scope (for details, see, in particular, Cressie 1992).

6.4 Method of Moments Estimation

6.4.1 Global Linear Empirical Bayes Smoother

The EB approach consists of estimating the moments of the prior distribution from the data, rather than taking them as a "prior" in a pure sense. A number of different estimators have been suggested, but by far the easiest one to apply

in practice is the so-called method of moments applied to the Gamma-Poisson model (Marshall 1991). In the *global* model, the prior mean and variance are assumed to be constant. They are estimated by their corresponding sample moments.

Using a risk estimate as the example, this yields, for the mean:

$$\hat{\mu} = \frac{\sum_{i=1}^{N} O_i}{\sum_{i=1}^{N} P_i}, \tag{89}$$

and, similarly for the SMR, with E_i replacing P_i in equation (89). This estimate is simply the average rate based on the total sum of all events over the total sum of all the populations (where N is the total number of spatial units).[31]

For the variance, the estimator is (using the risk case as an example):

$$\hat{\sigma}^2 = \frac{[\sum_{i=1}^{N} P_i(\hat{\pi}_i - \hat{\mu})^2]}{\sum_{i=1}^{N} P_i} - \frac{\hat{\mu}}{\bar{P}}, \tag{90}$$

where $\hat{\mu}$ is the estimate for the mean, and $\bar{P} = \sum_{i=1}^{N} P_i/N$ is the average population at risk. The estimate for the prior variance for the SMR is found by using $\hat{\theta}_i$ instead of $\hat{\pi}_i$ and E_i instead of P_i in equation (90).

In contrast to the EB standardization (section 3.4), the moment estimates (89) and (90) are not used to transform the original crude rate by subtracting a mean and dividing by a standard deviation. Instead, they are used in equations (69) and (67) to compute the weights for the shrinkage estimator in equations (68) and (66). Alternatively, the estimates $\hat{\mu}$ and $\hat{\sigma}^2$ can be used directly in the expressions for the posterior mean, equations (74) and (76), with:

$$\hat{\alpha} = \hat{\mu}^2/\hat{\sigma}^2, \tag{91}$$

[31]This is not the same as the average of the rates for the individual spatial units.

and,

$$\hat{\beta} = \hat{\mu}/\hat{\sigma}^2. \tag{92}$$

Note that it is possible for equation (90) to yield a negative estimate for the variance. In that case, standard practice is to set $\hat{\sigma}^2$ to zero, effectively equating the local estimate to the prior mean $\hat{\mu}$.

This approach is arguably the most commonly used EB smoother in practice. For further details, see also Bailey and Gatrell (1995, pp. 303–308), and, for a recent illustrative example, see, e.g., Berke (2004). This smoother is also implemented in several software packages, such as Anselin et al. (2004, 2006), Bivand (2006), and Rey and Janikas (2006).

6.4.2 Local Linear Empirical Bayes Smoother

The local EB smoother, also referred to as the *spatial* EB smoother, is based on the same principle as the "global" EB smoother, except for the computation of the priors. Instead of estimating a constant mean and variance from all the data points, only a limited subsample is used. This subsample is different for each location, allowing for greater flexibility in modeling heterogeneity.

Specifically, the prior mean for the crude rate at i is estimated as:

$$\hat{\mu}_i = \frac{\sum_{i \in J_i} O_i}{\sum_{i \in J_i} P_i}, \tag{93}$$

where J_i is the "local" neighborhood set for i (including i), and, similarly for the SMR, with E_i replacing P_i

The local estimator for the prior of the variance is (using the risk case as

an example):

$$\hat{\sigma}_i^2 = \frac{[\sum_{i \in J_i} P_i (\hat{\pi}_i - \hat{\mu}_i)^2]}{\sum_{i \in J_i} P_i} - \frac{\hat{\mu}_i}{\bar{P}_i}, \tag{94}$$

where $\hat{\mu}_i$ is the local prior for the mean, and $\bar{P}_i = \sum_{i \in J_i} P_i / N$ is the local average population at risk. The estimate for the prior variance for the SMR is found by using $\hat{\theta}_i$ instead of $\hat{\pi}_i$ and E_i instead of P_i in equation (94).

As in the estimation of the global priors, it is possible (common) to encounter negative estimates for the variance, in which case $\hat{\sigma}_i^2$ is set to zero. The local priors are then used to replace μ and σ^2 in equations (69) and (67).

The local linear EB smoother is also implemented in software packages, such as Anselin et al. (2004, 2006), Bivand (2006), and Rey and Janikas (2006).

6.5 Other Estimators

Estimators other than the method of moments have been suggested for the EB smoother, although they are less commonly applied in practice. We will review them briefly and refer to the original sources for technical details.

For the Gamma model, Clayton and Kaldor (1987) also outlined a maximum likelihood estimator.[32] This approach exploits the property that the marginal distribution of event counts is negative binomial. The associated log likelihood is an expression in the parameters α and β, which yields ML estimates as the solution of the first order conditions. As given in Clayton and Kaldor (1987, p. 673), those conditions are:

$$\sum_{i=1}^{n} \sum_{h}^{O_i - 1} (\frac{1}{\alpha + h}) + (n \log \beta) - \sum_{i=1}^{n} \log(P_i + \beta) = 0, \tag{95}$$

---

[32]See also Marshall (1991), Bailey and Gatrell (1995, pp. 303–308) and Meza (2003).

and,

$$\alpha/\beta = (1/m) \sum_{i=1}^{n} \frac{O_i + \alpha}{P_i + \beta}, \tag{96}$$

with E_i replacing P_i when the interest is in relative risk estimates.[33] This set of equations can be solved by means of iterative techniques.

A computationally efficient way to proceed is a combination of a method of moments rationale with the second ML condition. The latter is equivalent to the ratio α/β equalling the average of the EB estimates. A simple iterative procedure then proceeds by calculating the EB estimates using equation (80) or (81) for given starting values of α and β. Updated estimates for α and β are obtained from two auxiliary equations:

$$c = (1/n) \sum_{i} \hat{\pi}_i^{EB}, \tag{97}$$

the average of the EB estimates, and:

$$d = (1/n-1) \sum_{i} (1 + \beta/P_i)[\hat{\pi}_i^{EB} - (\alpha/\beta)]^2. \tag{98}$$

The updates are then obtained as $\alpha = c^2/d$ and $\beta = c/d$ and the method continues until a criterion of convergence is met.

An alternative estimation approach can be based on the principle of estimating equations, originally introduced by Lahiri and Maiti (as cited in Meza 2003).

For the log-normal model, Clayton and Kaldor (1987) outlined an EM estimation procedure.

While EB estimates may provide an effective estimate for the disease rate in each geographic area, they do not necessarily satisfy other important

---

[33]For $O_i = 0$, the sum to $O_i - 1$ is set to zero.

goals of disease mapping, such as representing the "histogram" of the true underlying rates, and a correct ranking of rates (see also Shen and Louis 1998, and Section 9.1). Several improvements to the EB estimator have been suggested to take some of these other criteria into account. Constrained Bayes estimates were outlined in Louis (1984) and Ghosh (1992), and in a slightly different context, in Cressie (1992). Applications to disease mapping can be found in Devine and Louis (1994), and Devine et al. (1994, 1996), among others. In the context of developing estimators that meet all three goals, Shen and Louis (1998) further generalized the constrained approach.

A separate concern pertains to the assumption of a particular prior density, especially when the assumed parametric family for the prior is misspecified (see, for example Yasui et al. 2000). Laird and Louis (1991) and Shen and Louis (1999) suggest a computational approach, referred to as "smoothing by roughening," which uses an EM algorithm to produce a non parametric maximum likelihood estimate of the prior (see also Shen and Louis 2000, for a comparison of several EB estimators).

6.6   Precision of EB Estimates

The EB estimates are point estimates and it is often desirable to have some measure of precision associated with them. A naive approach is to use the variance of the posterior distribution. For example, for the Gamma model, as in equation (75), with estimates $\hat{\alpha}$ and $\hat{\beta}$ substituted, this would yield:

$$\mathrm{Var}[\hat{\pi}_i^{EB}] = \frac{O_i + \hat{\alpha}}{(P_i + \hat{\beta})^2}, \tag{99}$$

52

and similarly for $\text{Var}[\hat{\theta}_i^{EB}]$, using E_i instead of P_i. This measure will tend to underestimate the true uncertainty, since it takes the parameter estimates $\hat{\alpha}$ and $\hat{\beta}$ as known, and ignores the uncertainty associated with their estimation.[34]

An alternative to the analytical approximation of the variance of the estimate is to use a simulation estimator, such as the parametric bootstrap (e.g., Laird and Louis 1987, Meza 2003). The parametric bootstrap creates a series of artificial data sets, using the estimated parameters as the basis for a random number generation. Specifically, in the case of counts of events, a bootstrap "sample" of the O_i can be obtained by generating n independent Poisson random variates, each with mean $\hat{\pi}_i^{EB} P_i$ or $\hat{\theta}_i^{EB} E_i$. These "data" are then used to obtain EB estimates for the risk or relative risk parameters. The process is repeated multiple times, such that the empirical distribution of the EB estimates from the bootstrap samples can form the basis for an estimate of variance. While computational intensive, this procedure is relatively easy to implement.[35]

---

[34]In addition, it ignores any uncertainty associated with imprecise values for the denominator, the P_i and E_i. Especially for inter-censal years and for small areas, it may be overly optimistic to consider these as error free.

[35]The parametric bootstrap method is also increasingly used to obtain estimates of precision in more complex models for risk or relative risk, such as non-parametric mixture models (see, for example Ugarte et al. 2003, and Section 6.7), and fully Bayesian CAR models (e.g., MacNab and Dean 2000, MacNab et al. 2004).

6.7 Non-Parametric Mixture Models

As an alternative to the parametric specification of the prior distribution for the risk or relative risk, as in Sections 6.2 and 6.3, Clayton and Kaldor (1987) also suggested a non-parametric approach, in which the distribution is left in unspecified form. This leads to non-parametric mixture models, in which the underlying risk is assumed to be represented by a discrete number of subpopulations, each corresponding to a different level of risk or relative risk. In addition to yielding an EB estimate for the risk or relative risk, mixture models provide a natural way to classify each observation into a given risk category (for overviews, see, e.g., Schlattmann and Böhning 1993, Böhning and Schlattmann 1999).[36]

As before, the conditional distribution of $O_i|\pi_i$ or $O_i|\theta_i$ is assumed to be Poisson. Instead of focusing on the posterior distribution of the risk parameters, the marginal likelihood is constructed by integrating over all the possible values of the prior distribution. In the mixture case, this boils down to a sum over the values taken by a discrete non-parametric distribution over π or θ.

For example, assume that π takes on k different values in the underlying unobserved *latent* distribution, each with a probability $p_h, h = 1, \ldots, k$. This is represented as:

$$P = \begin{bmatrix} \pi_1 & \pi_2 & \ldots & \pi_k \\ p_i & p_2 & \ldots & p_k \end{bmatrix} \tag{100}$$

---

[36]Extensions to space-time disease mapping are given in Böhning (2003).

The likelihood for the marginal distribution of the O_i then becomes:

$$L = \prod_i \sum_h \text{Poisson}(O_i|\pi_h P_i)p_h. \qquad (101)$$

The non-parametric maximum likelihood estimator (NPLMLE) yields estimates of both π_h and p_h.[37]

With those estimates in hand, an EB estimate for the risk or relative risk at i follows as:

$$\hat{\pi}_i^{EB-NP} = \frac{\sum_h \hat{\pi}_h \text{Poisson}(O_i|\hat{\pi}_h P_i)\hat{p}_h}{\sum_h \text{Poisson}(O_i|\hat{\pi}_h P_i)\hat{p}_h}, \qquad (102)$$

or,

$$\hat{\theta}_i^{EB-NP} = \frac{\sum_h \hat{\theta}_h \text{Poisson}(O_i|\hat{\theta}_h E_i)\hat{p}_h}{\sum_h \text{Poisson}(O_i|\hat{\theta}_h E_i)\hat{p}_h}, \qquad (103)$$

i.e., a weighted average of EB estimates, weighted by the probabilities associated with each discrete category.

In addition, it is possible to associate a given risk or relative risk category to each location, by using a decision rule that assigns location i to category h if h maximizes the posterior density, or:

$$\max_h \frac{\text{Poisson}(O_i|\hat{\pi}_h P_i)\hat{p}_h}{\sum_h \text{Poisson}(O_i|\hat{\pi}_h P_i)\hat{p}_h}, \qquad (104)$$

or,

$$\max_h \frac{\text{Poisson}(O_i|\hat{\theta}_h E_i)\hat{p}_h}{\sum_h \text{Poisson}(O_i|\hat{\theta}_h E_i)\hat{p}_h}, \qquad (105)$$

The computations to obtain these estimates are not trivial, but special computer software has been developed to carry this out, such as C.A. MAN (see Böhning et al. 1992), and DismapWin (Schlattmann 1996).

---

[37]Technical discussions of estimation by means of non-parametric maximum likelihood, originally based on the work of Laird (1978), can be found in Böhning (1995, 2000), among others.

Using a similar nonparametric approach, Böhning et al. (2002) and Böhning et al. (2004) develop measures for the variability of estimates of SMR, such as EB estimates (see also Section 6.6). Other extensions of mixture models consist of incorporating them into Bayesian hierarchical frameworks, as in Lawson (2001a) and Lawson and Clark (2002). A detailed discussion of these extensions is beyond the current scope, as they are clearly aspects of sophisticated models of risk and relative risk, rather than focused on simple smoothing.

7   Fully Bayes Smoothers

Even though, strictly speaking, fully Bayes or model-based approaches to smoothing are beyond the scope of this review, we provide a brief overview of the main aspects of the most commonly used approach, the so-called BYM (Besag-York-Mollie) model (e.g., Besag et al. 1991, 1995). We include this for the sake of completeness, since the application of Bayesian hierarchical modeling to disease mapping has seen explosive growth in recent years, both in terms of methodology as well as in terms of applications (for a recent review of methodological aspects, see Banerjee et al. 2004). This has further been fueled by the ready availability of easy to use software, such as WinBUGS, which has popularized its use in applied work (e.g., Lawson et al. 2003).

The goal of fully Bayes smoothing is to develop a model for the variability in risk or relative risk that includes known covariates as well as random effects. The latter allow for the incorporation of complex patterns of spatial and non-spatial heterogeneity, as well as forms of spatial autocorrelation.

The end result of this exercise is a predicted value for the risk or relative risk (e.g., the mean of the posterior distribution), which then forms the basis for a map of smoothed rates. In addition, the posterior distribution provides a natural way to assess the precision of the smoothed estimate.

Here, we limit ourselves to outlining the basic principle behind the BYM model. It is important to note that the main difference between the Empirical Bayes and the "fully" Bayes approach lies in how the randomness of the parameters in the prior is treated (e.g., the α and β in the Gamma model for Section 6.2). In the Empirical Bayes approaches covered in the previous section, these parameters are replaced by a point estimate obtained from the data. In contrast, in the fully Bayes approach, a complete distribution is specified for those parameters as well, itself a function of so-called *hyper-parameters*. This naturally leads to a hierarchical structure, which, while often analytically intractable, can be estimated by means of Markov Chain Monte Carlo (MCMC) methods, such as the Gibbs sampler and the Metropolis-Hastings sampler (Gilks et al. 1996).

The point of departure is typically a Poisson model for the conditional distribution of the observed event counts O_i, conditional upon the unknown risk parameter π_i, or relative risk parameter θ_i:

$$O_i|\pi_i \sim \text{Poisson}(\pi_i P_i), \text{or}, \ O_i|\theta_i \sim \text{Poisson}(\theta_i E_i). \tag{106}$$

The next level in the hierarchy specifies a model for the risk or relative risk parameter, in terms of a distribution and a set of hyper-parameters (i.e., parameters in the prior for the prior). The BYM approach decomposes the randomness in the risk into three parts, typically assuming a lognormal

57

distribution for π_i or θ_i, such that:

$$\log \pi_i = x_i \beta + \nu_i + \phi_i, \tag{107}$$

or,

$$\log \theta_i = x_i \beta + \nu_i + \phi_i. \tag{108}$$

The first component, $x_i \beta$, pertains to systematic variation (known covariates, including age structure), the other two are random effects. One random effect (ν_i) incorporates non-spatial overdispersion, and is assumed to be distributed as mean-zero normal with variance σ_ν^2:

$$\nu_i \sim \mathrm{N}(0, \sigma_\nu^2). \tag{109}$$

The second random component, ϕ_i, incorporates overdispersion due to spatial autocorrelation and is typically modeled as a conditional spatial autoregressive model (CAR). This relates the value of ϕ at a location i to that of its immediate neighbors, specified by means of a spatial weights matrix with elements w_{ij}.[38] The conditional distribution of ϕ_i, given the neighboring values, is specified as normal:

$$\phi_i | \phi_{j \neq i} \sim \mathrm{N}(\bar{\phi}_i, \psi_i^2), \tag{110}$$

where,

$$\bar{\phi}_i = \frac{\sum_{j \neq i} w_{ij} \phi_j}{\sum_{j \neq i} w_{ij}}, \tag{111}$$

and,

$$\psi_i^2 = \frac{1}{\gamma \sum_{j \neq i} w_{ij}}, \tag{112}$$

_____

[38]For an extensive discussion of the specification of spatial models and the use of spatial weights, see, e.g., Banerjee et al. (2004), and Anselin (2006b).

58

where γ is a hyperprior parameter included to model the degree of spatial autocorrelation. Typically, the structure of the spatial dependence is further simplified as a function of a spatial autoregressive parameter ρ and a specific spatial weights structure. Estimation of the parameters of this model has to be carried out by means of Markov Chain Monte Carlo methods.

A wide range of variants of this specification have been suggested, incorporating different combinations of spatial and non-spatial heterogeneity and dependence, different distributions for the priors, and models for covariates.

Detailed discussions of the methodological issues can be found in Bernardinelli and Montomoli (1992), Maiti (1998), Ghosh et al. (1999), Best et al. (1999), Gangnon and Clayton (2003), among others. Extensions to space-time models are developed by Waller et al. (1997a,b), Xia and Carlin (1998), Knorr-Held (2000), and MacNab (2003). Some illustrative examples of applications to disease mapping include Xia et al. (1997), Short et al. (2002), Thomas and Carlin (2003), and Johnson (2004).

The merits of different assumptions, such as the sensitivity of the results to the inclusion of different priors and different hierarchical structures, the potential for oversmoothing, the stability of the estimates for relative risk, problems with model identification, and the extent to which the "true" risk surface can be discovered have received considerable attention as well, for example, in Bernardinelli et al. (1995), Pascutto et al. (2000), Colonna (2004), and Richardson et al. (2004)

A more detailed discussion of these various aspects is beyond the current scope of this document.

8 Regionalization

8.1 Principle

The final set of techniques considered to address variance instability in rates approach the problem by focusing on the denominator, rather than the numerator. Since the variance of the rate estimate in equation (4) depends inversely on the size of the population at risk, stable estimates may also be obtained by increasing the spatial scale of the units of observations. The objective is to determine the smallest spatial area that provides meaningful and stable rate estimates for the disease under study. These techniques are referred to as *regionalization* methods.

We briefly review four classes of techniques that have been suggested to accomplish this objective. We start with the overlay of regular grids or quadrats for use in analysis, consider a form of cartogram known as density equalizing map projections (DEMP), and a method to assess the effect of arbitrary spatial aggregation on the measure of risk (iterative random partitioning). We close with an overview of generic techniques to construct spatial clusters in which the components are constrained to be contiguous.

8.2 Quadrat Counts

In the spatial filtering approach considered in Section 5.1, variance instability was addressed by constructing a moving window on a set of regularly spaced locations. A similarly inspired alternative is to aggregate individual observations on cases and controls to a system of regular square grids. The analysis is then carried out using the aggregate counts and rates computed

for the grid layout, similar to the quadrat count technique in point pattern analysis. This can be carried out for different scales of spatial resolution, allowing for the explicit analysis of spatial scale (for details, see Paulu et al. 2002).

8.3 Density Equalizing Map Projections

Density equalizing map projections (DEMP) are a special case of a cartogram, in which the shape of the original spatial units is adjusted such that the population density is equalized over the study area.[39] More specifically, the boundaries of each spatial unit of observation, such as a census tract, are transformed (projected) such that the resulting area is proportional to the population at risk in the unit. As a result, a map of events or cases on the transformed spatial layout should have a uniform distribution if the underlying risk is constant.

The implementation of this technique requires specialized computer algorithms, as outlined in Merrill et al. (1996) and Merrill (2001) (see also Bithell 2000, for further discussion). While primarily exploratory, the DEMP maps can also be used to test hypotheses about disease clusters.

8.4 Iterative Random Partitioning

Iterative random partitions, suggested in Cislaghi et al. (1995) study dynamically how the spatial organization of the disease rates changes with different levels of data aggregation. The underlying idea is to remove the effect of

---

[39]For recent reviews of the cartogram technique, see, e.g., Gastner and Newman (2004) and Tobler (2004).

arbitrary administrative boundaries by constructing a summary estimate for each location based on a large number of partionings or spatial aggregations of the data.

In principle, all possible groupings of elementary units are considered, but in practice this is limited to a set constructed from randomly sampling K so-called pivotal units and assigning the remaining observations to those "cluster centers." The number of regions K is referred to as the smoothing parameter. Once the pivotal units are selected, the remaining units are assigned to the nearest pivotal unit on the basis of a distance criterion. For each of the K areas, the value of the standardized indices (crude rate, SMR) is assigned to all elementary units that belong to the area. This procedure is repeated several times, after which the values across all spatial layouts are summarized for each elemental unit (e.g., as a median or mean). The latter are then used to construct the disease map (for technical details, see Cislaghi et al. 1995, p. 2365).

In this method, no inference is possible, in the sense that no standard errors or confidence intervals can be computed. It is therefore primarily exploratory.

8.5   Contiguity Constrained Clustering

A fourth approach to regionalization consists of building up aggregate entities from elementary spatial units, such that a target size is reached for the population at risk. The motivation is similar to that of a spatial filter (see Section 5.1), but the implementation is different, in that the spatial units themselves are being aggregated. As a result, the original geography is

altered.

The regionalization problem is a special case of multivariate clustering, and, as such, it has received considerable attention in the operations research literature. In addition to the construction of clusters related to the value of attributes, the geographical location must be taken into account as well. In a recent review by Duque et al. (2006), it is shown how this can be approached by two-stage aggregation (a non-spatial clustering exercise followed by the imposition of contiguity constraints), the explicit inclusion of the contiguity information as part of the classification variables, and by the application of additional instruments to enforce contiguity (for other recent reviews, see Murtagh 1985, Gordon 1996, 1999).

An important practical issue is the choice of the target population value. In Morris and Munasinghe (1993), this is selected as the size necessary to be able to detect a rate that is twice the national rate with a power of 90 percent, assuming a binomial distribution for the national rate, and a Type I error of 5 percent. Clearly, the assumption of a different distribution and/or Type I error would lead to different target population sizes, but this approach provides an explicit link between the aggregation procedure and the statistical decision for which it is designed.

Contiguity constrained regionalization procedures have been included in operational software, such as the SAGE system for spatial data analysis (Haining et al. 1996, 2000, Wise et al. 2001), as well as the latest version of STARS.

9 Summary and Assessment

9.1 Empirical Comparisons

A wide array of techniques has been suggested in the literature to deal with
the instability of rates in disease maps, and the number of examples of em-
pirical applications employing these methods is ever growing. Two recent
studies have explicitly addressed the comparative performance of a range of
different methods in carefully controlled settings, i.e., Kafadar (1994) and
Lawson et al. (2000).[40]

Before briefly summarizing the main findings of these two studies, it is
useful to note that no method can satisfy all requirements that are expected
from a risk estimate. As outlined in Shen and Louis (1998, 2000), there are
three important but conflicting goals relevant for the estimation and mapping
of disease rates. They are: the accurate estimation of disease rates for small
areas; the proper estimation of the spatial distribution (and cumulative dis-
tribution) of the disease rates; and the appropriate ranking of observations
so that locations with extreme risks (outliers) can be correctly identified.
As argued by Shen and Louis (1998, 2000), no single set of estimates can
simultaneously satisfy these three goals. Consequently, it is important to
consider which objective is used to assess the relative performance of the
methods, as illustrated by some empirical comparisons that have appeared

---

[40]Many empirical studies also apply and compare more than one method to a particular
case study data set, but they tend to focus on similar techniques, such as Empirical Bayes
vs. Fully Bayes (Bernardinelli and Montomoli 1992), or headbanging smoothing under
different conditions (Gelman et al. 2000). A comprehensive review of these comparisons
is beyond the current scope. For a summary, see Lawson (2001a, pp. 180-182).

in the literature.

In Kafadar (1994), 16 different smoothers are compared on artificially generated data sets that use the population weighted centroids of 86 counties within a 500 miles radius of San Francisco as the spatial layout. The techniques evaluated include three disk averages with varying radius, as well as inverse distance and inverse distance squared weighted averages (see Section 4.1), a local and global Empirical Bayes smoother (see Section 6), six local trend surface regression using loess for different local ranges, median polish, headbanging (see Section 4.2.1), and resmoothed medians (see Section 4.2).

Different spatial distributions are generated by combining a trend surface function $f(x, y)$ with an error process. The function is not specific to rates and no account is taken of the special form of variance instability present in rates. Two basic structures are compared, one consisting of a mean value (no pattern), the other a combination of a ridge, a peak and a circular depression. The error term adds three levels of noise in the form of uncorrelated Gaussian error with increasing variance, for a total of six combinations. Spatial autocorrelation is not considered in this design.

The smoothers are evaluated by means of two criteria: a global measure of fit (mean squared error) and a threshold criterion related to the number of times peaks and depressions are correctly classified.[41]

Kafadar finds that the inverse distance weighted average performs well on both unstructured and structured patterns. Headbanging is more appro-

---

[41]For the particular design chosen, this primarily boils down to how well the ridge in the data can be identified and smoothed (see Kafadar 1994, p. 429).

priate in the presence of higher error variance and outliers. The Empirical Bayes method was found to only perform well in situations with minor error variance, with a slight edge for the local EB smoother. These results are overall comparisons of the smoothing methods, but may not be readily generalizable to patterns of rates, since both the variance instability and the spatial autocorrelation typically found in disease rates were not incorporated in the research design.

In contrast, the simulation experiments reported in Lawson et al. (2000) do explicitly address the variance instability of rates and relative risk estimates. In a large number of experiments that cover 154 different data generating processes, six methods are compared in terms of overall fit (using different criteria) and remaining residual spatial autocorrelation.[42]

The data are generated on the spatial layout of counties in former East Germany (a geography used in many empirical illustrations in the literature). The experiment uses the actually observed death rate from lip cancer to obtain the expected counts in each location. The simulated counts are then obtained by combining the expected counts with a measure of relative risk from one of the data generating processes. A multinomial model is used to enforce the constraint that the total number of observed cases should equal the total number of expected cases (Lawson et al. 2000, p. 2225). The simulated models range from the constant risk model and deterministic trend surfaces, to random effects models, overdispersion models, and mixture

---

[42]The specific measures of fit used are a Pearson and a Spearman correlation coefficient, a residual sum of squares, and the difference in the Bayesian Information Criterion (BIC) between the fitted model and the null model (see Lawson et al. 2000, pp. 2228–2230, for technical details).

66

models, including models with spatial structure and spatial autocorrelation.[43]

Six methods are evaluated, a kernel smoother (see Section 5.2.1), Empirical Bayes with both global and local method of moments estimates for the parameters, as well as an Empirical Bayes method using the Gamma-Poisson model (see Section 6), a non-parametric mixture model (see Section 6.7), and a fully Bayesian model, using the Besag-York-Mollie specification (BYM, see Section 7).

The models are compared in terms of how well they fit the counts as well as the relative risk. Overall, the performance of techniques varies considerably, and some techniques are more appropriate to deal with some data generating processes than others. None of the models do well in relative risk analysis when there is considerable heterogeneity. Lawson et al. (2000) recommend against the use of kernel smoothers or to limit their application to data exploration. Mixture models also do not do well for relative risks. They suggest that the BYM model is both fairly robust to misspecification, performs best for counts and near best for relative risks. The global EB method is found to be superior to its local counterpart.

Both Kafadar (1994) and Lawson et al. (2000) use a relatively small number of replications (100) of the randomly generated data sets. This may create some sampling error in the results, although it probably has little effect on the qualitative results of the comparisons.

---

[43]For the complete list of model specifications and parameters, see Lawson et al. (2000, pp. 2234–2241).

9.2 Concluding Remarks

It is obvious from this review that a wide range of methods is available to address variance instability in rates. To some extent, this may seem bewildering, and it may lead to confusion among uninformed practioners. This is particularly relevant, since in many instances different methods may yield widely different results.

The sensitivity of the outcomes to the assumptions made is still not fully understood. Several comparative analyses have been carried out, but they tend to focus on a narrow range of techniques, and, to date, there is no consensus as to which is the preferred practice.

An important aspect to keep in mind is that the various smoothing techniques are forms of modeling, in which a delicate balance is obtained between assumptions imposed by the model (e.g., distributional assumptions, structure of priors) and what the data support. Whereas, in practice, one method is often selected at the expense of all others, sensitivity analysis from the comparison of outcomes from different methods is often crucial to gain further insight into the tradeoffs involved.

Finally, any method can only be as good as the data that support the analysis. When information is scarce, as in the case of rare events in small areas, the insight that can be gained will necessarily be limited and unreliable (see also Anselin 2006a).

References

Anscombe, F. J. (1948). The transformation of Poisson, binomial and negative-binomial data. *Biometrika*, 35:246–254.

Anselin, L. (2006a). How (not) to lie with spatial statistics. *American Journal of Preventive Medicine*, 30:S3–S6.

Anselin, L. (2006b). Spatial econometrics. In Mills, T. and Patterson, K., editors, *Palgrave Handbook of Econometrics: Volume 1, Econometric Theory*, pages 901–969. Palgrave Macmillan, Basingstoke.

Anselin, L., Kim, Y.-W., and Syabri, I. (2004). Web-based analytical tools for the exploration of spatial data. *Journal of Geographical Systems*, 6:197–218.

Anselin, L., Syabri, I., and Kho, Y. (2006). GeoDa, an introduction to spatial data analysis. *Geographical Analysis*, 38:5–22.

Assunção, R. and Reis, E. A. (1999). A new proposal to adjust Moran's I for population density. *Statistics in Medicine*, 18:2147–2161.

Bailey, T. C. and Gatrell, A. C. (1995). *Interactive Spatial Data Analysis*. John Wiley and Sons, New York, NY.

Banerjee, S., Carlin, B. P., and Gelfand, A. E. (2004). *Hierarchical Modeling and Analysis for Spatial Data*. Chapman & Hall/CRC, Boca Raton.

Berke, O. (2004). Exploratory disease mapping: Kriging the spatial risk function from regional count data. *International Journal of Health Geographics*, 3.

Bernardinelli, L. and Montomoli, M. (1992). Empirical Bayes versus fully Bayesian analysis of geographical variation in disease risk. *Statistics in Medicine*, 14:983–1007.

Bernardinelli, M., Clayton, D., and Montomoli, M. (1995). Bayesian estimates of disease maps: how important are priors? *Statistics in Medicine*, 14:2411–2431.

Besag, J., Green, P., Higdon, D., and Mengersen, K. (1995). Bayesian computation and stochastic systems. *Statistical Science*, 10:3–66.

Besag, J., York, J., and A.Mollié (1991). Bayesian image restoration with two applications in spatial statistics. *Annals of the Institute of Statistical Mathematics*, 43:1–59.

Best, N. G., Arnold, R. A., Thomas, A., Waller, L. A., and Conlon, E. M. (1999). Bayesian models for spatially correlated disease and exposure data. In Bernardo, J., Berger, J., Dawid, A., and Smith, F., editors, *Bayesian Statistics 6*, pages 131–156. Oxford University Press, New York, NY.

Bithell, J. F. (1990). An application of density estimation to geographical epidemiology. *Statistics in Medicine*, 9:691–701.

Bithell, J. F. (2000). A classification of disease mapping methods. *Statistics in Medicine*, 19:2203–2215.

Bivand, R. (2006). Implementing spatial data analysis software in R. *Geographical Analysis*, 38:23–40.

Böhning, D. (1995). A review of reliable maximum likelihood algorithms for semiparametric mixture models. *Journal of Statistical Planning and Inference*, 47:5–28.

Böhning, D. (2000). *Computer Assisted Analysis of Mixtures and Applications: Disease Mapping, Meta-Analysis and Others.* Chapman & Hall/CRC, Boca Raton, FL.

Böhning, D. (2003). Empirical Bayes estimators and non-parametric mixture models for space and time-space disease mapping and surveillance. *Environmetrics*, 14:431–451.

Böhning, D., Malzahn, U., Sarol, J., Rattanasiri, S., and Biggeri, A. (2002). Efficient non-iterative and nonparametric estimation of heterogeneity variance for the standardized mortality ratio. *Annals of the Institute of Statistical Mathematics*, 54:827–839.

Böhning, D., Sarol, J., Rattanasiri, S., Viwatwongkasem, C., and Biggeri, A. (2004). A comparison of non-iterative and iterative estimators of heterogeneity variance for the standardized mortality ratio. *Biostatistics*, 5:61–74.

Böhning, D. and Schlattmann, P. (1999). Disease mapping with hidden structures using mixture models. In Lawson, A., Biggeri, A., Böhning, D., Lesaffre, E., Viel, J., and Bertollini, R., editors, *Disease Mapping and Risk Assessment for Public Health*, pages 49–60. John Wiley, New York.

Böhning, D., Schlattmann, P., and Lindsay, B. (1992). Computer-Assisted

Analysis of Mixtures (C.A.MAN): Statistical algorithms. *Biometrics*, 48(1):283–303.

Carlin, B. P. and Louis, T. A. (2000). *Bayes and Empirical Bayes Methods for Data Analysis.* Chapman & Hall/CRC, Boca Raton, FL.

Carrat, F. and Valleron, A.-J. (1992). Epidemiologic mapping using the "kriging" method: application to an influenza-like illness epidemic in France. *American Journal of Epidemiology*, 135:1293–1300.

Choynowski, M. (1959). Maps based on probabilities. *Journal of the American Statistical Association*, 54:385–388.

Cislaghi, C., Biggeri, A., Braga, M., Lagzio, C., and Marchi, M. (1995). Exploratory tools for disease mapping in geographical epidemiology. *Statistics in Medicine*, 14:2363–2381.

Clayton, D. and Hills, M. (1993). *Statistical Models in Epidemiology.* Oxford University Press, New York, NY.

Clayton, D. and Kaldor, J. (1987). Empirical Bayes estimates of age-standardized relative risks for use in disease mapping. *Biometrics*, 43:671–681.

Colonna, M. (2004). Bootstrap investigation of the stability of disease mapping of Bayesian cancer relative risk estimations. *European Journal of Epidemiology*, 19:761768.

Congdon, P. (2001). *Bayesian Statistical Modelling.* John Wiley & Sons, Chichester, UK.

Congdon, P. (2003). *Applied Bayesian Modelling.* John Wiley & Sons, Chichester, UK.

Cressie, N. (1992). Smoothing regional maps using Empirical Bayes predictors. *Geographical Analysis*, 24:75–95.

Cressie, N. (1993). *Statistics for Spatial Data.* Wiley, New York.

Cressie, N. (1995). Bayesian smoothing of rates in small geographic areas. *Journal of Regional Science*, 35:659–673.

Devine, O. J. and Louis, T. A. (1994). A constrained Empirical Bayes estimator for incidence rates in areas with small populations. *Statistics in Medicine*, 13:1119–1133.

Devine, O. J., Louis, T. A., and Halloran, M. E. (1994). Empirical Bayes estimators for spatially correlated incidence rates. *Environmetrics*, 5:381–398.

Devine, O. J., Louis, T. A., and Halloran, M. E. (1996). Identifying areas with elevated disease incidence rates using Empirical Bayes estimators. *Geographical Analysis*, 28:187–199.

Diggle, P. (2003). *Statistical Analysis of Spatial Point Patterns (Second Edition).* Arnold, London.

Dobson, A., Kuulasmaa, K., Eberle, E., and Scherer, J. (1991). Confidence intervals for weighted sums of Poisson parameters. *Statistics in Medicine*, 10:457–462.

Duque, J. C., Ramos, R., and Suriñach, J. (2006). Supervised regionalization methods: a survey. *International Regional Science Review*, 29. Forthcoming.

Ericson, W. A. (1969). A note on the posterior mean of a population mean. *Journal of the Royal Statistical Society B*, 31:332–334.

Fay, M. P. and Feuer, E. J. (1997). Confidence intervals for directly standardized rates: A method based on the Gamma distribution. *Statistics in Medicine*, 16:791–801.

Freeman, M. F. and Tukey, J. W. (1950). Transformations related to the angular and the square root. *Annals of Mathematical Statistics*, 21:607–611.

Gangnon, R. E. and Clayton, M. K. (2003). A hierarchical model for spatially clustered disease rates. *Statistics in Medicine*, 22:3213–3228.

Gastner, M. T. and Newman, M. E. J. (2004). Diffusion-based methods for producing density-equalizing maps. *Proceedings of the National Academy of Sciences*, 101(20):7499–7504.

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (2004). *Bayesian Data Analysis, Second Edition*. Chapman and Hall, Boca Raton, FL.

Gelman, A. and Price, P. N. (1999). All maps of parameter estimates are misleading. *Statistics in Medicine*, 18:3221–3234.

Gelman, A., Price, P. N., and Lin, C. (2000). A method for quantifying

artefacts in mapping methods illustrated by application to headbanging. *Statistics in Medicine*, 19:2309–2320.

Ghosh, M. (1992). Constrained Bayes estimates with applications. *Journal of the American Statistical Association*, 87:533–540.

Ghosh, M., Natarajan, K., Waller, L. A., and Kim, D. (1999). Hierarchical Bayes GLM for the analysis of spatial data: an application to disease mapping. *Journal of Statistical Planning and Inference*, 75:305–318.

Gilks, W., Richardson, S., and Spiegelhalter, D. (1996). *Markov Chain Monte Carlo in Practice*. Chapman and Hall, London, UK.

Gill, J. (2002). *Bayesian Methods. A Social and Behavioral Sciences Approach*. Chapman & Hall/CRC, Boca Raton, FL.

Goldman, D. A. and Brender, J. D. (2000). Are standardized mortality ratios valid for public health data analysis? *Statistics in Medicine*, 19:1081–1088.

Goovaerts, P. and Jacquez, G. M. (2004). Accounting for regional background and population size in the detection of spatial clusters and outliers using geostatistical filtering and spatial neutral models: the case of lung cancer in Long Island, New York. *International Journal of Health Geographics*, 3:1–23.

Gordon, A. (1996). A survey of constrained classification. *Computational Statistics and Data Analysis*, 21:17–29.

Gordon, A. (1999). *Classification, Second Edition*. Chapman & Hall, Boca Raton, FL.

Haining, R. F., Ma, J., and Wise, S. (1996). Design of a software system for interactive spatial statistical analysis linked to a GIS. *Computational Statistics*, 11:449–466.

Haining, R. F., Wise, S., and Ma, J. (2000). Designing and implementing software for spatial statistical analysis in a GIS environment. *Journal of Geographical Systems*, 2(3):257–286.

Hansen, K. M. (1991). Head-banging: Robust smoothing in the plane. *IEEE Transactions on Geoscience and Remote Sensing*, 29:369–378.

Johnson, G. D. (2004). Small area mapping of prostate cancer incidence in New York State (USA) using fully Bayesian hierarchical modelling. *International Journal of Health Geographics*, 3(1):29.

Kafadar, K. (1994). Choosing among two-dimensional smoothers in practice. *Computational Statistics and Data Analysis*, 18:419–439.

Kafadar, K. (1996). Smoothing geographical data, particularly rates of disease. *Statistics in Medicine*, 15:2539–2560.

Kafadar, K. (1997). Geographic trends in prostate cancer mortality: An application of spatial smoothers and the need for adjustment. *Annals of Epidemioliogy*, 7:35–45.

Kafadar, K. (1999). Simultaneous smoothing and adjusting mortality rates in U.S. counties: Melanoma in white females and white males. *Statistics in Medicine*, 18:3167–3188.

Kelsall, J. E. and Diggle, P. J. (1995). Nonparametric estimation of spatial variation in relative risk. *Statistics in Medicine*, 14:2335–2342.

Kelsall, J. E. and Diggle, P. J. (1998). Spatial variation in risk of disease: A nonparametric binary regression approach. *Applied Statistics*, 47(4):559–573.

Kleinman, J. (1977). Age-adjusted mortality indexes for small areas: Applications to health planning. *American Journal of Public Health*, 67:834–840.

Knorr-Held, L. (2000). Bayesian modelling of inseparable space-time variation in disease risk. *Statistics in Medicine*, 19:25552567.

Krieger, N. and Williams, D. (2001). Changing to the 2000 standard million: Are declining racial/ethnic and socioeconomic inequalities in health real progress or statistical illusion. *American Journal of Public Health*, 91:1209–1213.

Laird, N. (1978). Nonparametric maximum likelihood estimation of a mixing distribution. *Journal of the American Statistical Association*, 73:805–811.

Laird, N. and Louis, T. A. (1987). Empirical Bayes confidence intervals based on bootstrap samples. *Journal of the American Statistical Association*, 82:739–750.

Laird, N. and Louis, T. A. (1991). Smoothing the non-parametric estimate of a prior distribution by roughening: a computational study. *Computational Statistics and Data Analysis*, 12:27–37.

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

Lawson, A. B. (2001a). *Statistical Methods in Spatial Epidemiology.* John Wiley & Sons, New York, NY.

Lawson, A. B. (2001b). Tutorial in biostatistics: Disease map reconstruction. *Statistics in Medicine*, 20:2183–2204.

Lawson, A. B., Biggeri, A. B., Böhning, D., Lesaffre, E., Viel, J.-F., Clark, A., Schlattmann, P., and Divino, F. (2000). Disease mapping models: an empirical evaluation. *Statistics in Medicine*, 19:2217–2241.

Lawson, A. B., Browne, W. J., and Rodeiro, C. L. V. (2003). *Disease Mapping with WinBUGS and MLwiN.* John Wiley, Chichester.

Lawson, A. B. and Clark, A. (2002). Spatial mixture relative risk models applied to disease mapping. *Statistics in Medicine*, 21:359–370.

Lawson, A. B. and Williams, F. L. (2001). *An Introductory Guide to Disease Mapping.* John Wiley & Sons, Chichester.

Leyland, A. H. and Davies, C. A. (2005). Empirical Bayes methods for disease mapping. *Statistical Methods in Medical Research*, 14:17–34.

Louis, T. A. (1984). Estimating a population of parameter values using Bayes and Empirical Bayes methods. *Journal of the American Statistical Association*, 79:393–398.

MacNab, Y. C. (2003). Bayesian spatial modelling of small-area rates of non-rare disease. *Statistics in Medicine*, 22:1761–1773.

MacNab, Y. C. and Dean, C. B. (2000). Parametric bootstrap and penalized quasi-likelihood inference in conditional autoregressive models. *Statistics in Medicine*, 19:2421–2435.

MacNab, Y. C., Farrell, P. J., Gustafson, P., and Wen, S. (2004). Estimation in Bayesian disease mapping. *Biometrics*, 60:865–873.

Maiti, T. (1998). Hierarchical bayes estimation of mortality rates for disease mapping. *Journal of Statistical Planning and Inference*, 69:339–348.

Manton, K., Woodbury, M., Stallard, E., Riggan, W., Creason, J., and Pellom, A. (1989). Empirical Bayes procedures for stabilizing maps of U.S. cancer mortality rates. *Journal of the American Statistical Association*, 84:637–650.

Marshall, R. J. (1991). Mapping disease and mortality rates using Empirical Bayes estimators. *Applied Statistics*, 40:283–294.

Merrill, D. W. (2001). Use of a density equalizing map projection in analysing childhood cancer in four California counties. *Statistics in Medicine*, 20:14991513.

Merrill, D. W., Selvin, S., Close, E., and Holmes, H. (1996). Use of density equalizing map projections (DEMP) in the analysis of childhood cancer in four California counties. *Statistics in Medicine*, 15:1837–1848.

Meza, J. L. (2003). Empirical Bayes estimation smoothing of relative risks in disease mapping. *Journal of Statistical Planning and Inference*, 112:43–62.

Ministry of Health (2005). *Atlas of Cancer Mortality in New Zealand 1994-2000*. Ministry of Health, Wellington, New Zealand.

Morris, R. D. and Munasinghe, R. L. (1993). Aggregation of existing geographic regions to diminish spurious variability of disease rates. *Statistics in Medicine*, 12:1915–1929.

Mungiole, M., Pickle, L. W., and Simonson, K. H. (1999). Application of a weighted head-banging algorithm to mortality data maps. *Statistics in Medicine*, 18:3201–3209.

Murtagh, F. (1985). A survey of algorithms for contiguity-constrained clustering and related problems. *The Computer Journal*, 28:82–88.

Openshaw, S., Charlton, M. E., Wymer, C., and Craft, A. (1987). A mark I geographical analysis machine for the automated analysis of point data sets. *International Journal of Geographical Information Systems*, 1:359–377.

Openshaw, S., Cross, A., and Charlton, M. (1990). Building a prototype geographical correlates exploration machine. *International Journal of Geographical Information Systems*, 4:297–311.

Pamuk, E. (2001). Cautiously adjusting to the new millennium: Changing to the 2000 population standard. *American Journal of Public Health*, 91:1174–1176.

Pascutto, C., Wakefield, J. C., Best, N. G., Richardson, S., Bernardinelli, L., Staines, A., and Elliott, P. (2000). Statistical issues in the analysis of disease mapping data. *Statistics in Medicine*, 19:24932519.

Paulu, C., Aschengrau, A., and Ozonoff, D. (2002). Exploring associations between residential location and breast cancer incidence in a case control study. *Environmental Health Perspectives*, 110(5):471478.

Pickle, L. W., Mungiole, M., Jones, G. K., and White, A. A. (1996). Atlas of United States mortality. Technical report, National Center for Health Statistics, Hyattsville, MD.

Pickle, L. W., Mungiole, M., Jones, G. K., and White, A. A. (1999). Exploring spatial pattern of mortality: The new atlas of United States mortality. *Statistics in Medicine*, 18:3211–3220.

Pickle, L. W. and White, A. A. (1995). Effects of the choice of age-adjustment method on maps of death rates. *Statistics in Medicine*, 14:615–627.

Rao, C. (1973). *Linear Statistical Inference and its Applications*. Wiley, New York, 2nd edition.

Rey, S. J. and Janikas, M. V. (2006). STARS: Space-time analysis of regional systems. *Geographical Analysis*, 38:67–86.

Richardson, S., Thomson, A., Best, N., and Elliott, P. (2004). Interpreting posterior relative risk estimates in disease-mapping studies. *Environmental Health Perspectives*, 112:1016–1025.

Rushton, G. (2003). Public health, GIS and spatial analytic tools. *Annual Review of Public Health*, 24:43–56.

Rushton, G., Krishnamurthy, R., Krishnamurti, D., Lolonis, P., and Song, H. (1996). The spatial relationship between infant mortality and birth defect rates in a U.S. city. *Statistics in Medicine*, 15:1907–1919.

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

Rushton, G. and West, M. (1999). Women with localized breast cancer selecting mastectomy treatment, Iowa, 1991-1996. *Public Health Reports*, 114:370371.

Schlattmann, P. (1996). The computer package DismapWin. *Statistics in Medicine*, 15:931.

Schlattmann, P. and Böhning, D. (1993). Mixture models and disease mapping. *Statistics in Medicine*, 12:943–950.

Shen, W. and Louis, T. A. (1998). Three-goal estimates in two-stage Hierarchical models. *Journal of the Royal Statistical Society, Series B (Statistical Methodology)*, 60(2):455471.

Shen, W. and Louis, T. A. (1999). Empirical Bayes estimation via the smoothing by roughening approach. *Journal of Computational and Graphical Statistics*, 8:800–823.

Shen, W. and Louis, T. A. (2000). Triple-goal estimates for disease mapping. *Statistics in Medicine*, 19:22952308.

Short, M., Carlin, B. P., and Bushhouse, S. (2002). Using hierarchical spatial models for cancer control planning in Minnesota (United States). *Cancer Causes and Control*, 13:903–916.

Simonoff, J. S. (1996). *Smoothing Methods in Statistics*. Springer, New York, NY.

Swift, M. (1995). Simple confidence intervals for standardized rates based on the approximate bootstrap method. *Statistics in Medicine*, 14:1875–1888.

Talbot, T. O., Forand, S. P., and Haley, V. B. (2002). Geographic analysis of childhood lead exposure in New York state. In *Geographic Information Systems in Public Health, Third National Conference*, pages 249–265.

Talbot, T. O., Kulldorff, M., Forand, S. P., and Haley, V. B. (2000). Evaluation of spatial filters to create smoothed maps of health data. *Statistics in Medicine*, 19:2399–2408.

Thomas, A. J. and Carlin, B. P. (2003). Late detection of breast and colorectal cancer in Minnesota counties: an application of spatial smoothing and clustering. *Statistics in Medicine*, 22:113–127.

Tobler, W. (2004). Thirty five years of computer cartograms. *Annals of he Association of American Geographers*, 94:58–73.

Ugarte, M., Ibáñez, B., and Militino, A. (2006). Modelling risks in disease mapping. *Statistical Methods in Medical Research*, 15:21–35.

Ugarte, M., Militino, A., and Ibáñez, B. (2003). Confidence intervals for relative risks in disease mapping. *Biometrical Journal*, 45:410–425.

Wall, P. and Devine, O. (2000). Interactive analysis of the spatial distribution of disease using a geographic information system. *Journal of Geographical Systems*, 2(3):243–256.

Waller, L., Carlin, B., and Xia, H. (1997a). Structuring correlation within hierarchical spatio-temporal models for disease rates. In Grégoire, T., Brillinger, D., Russek-Cohen, P., Warren, W., and Wolfinger, R., editors, *Modeling Longitudinal and Spatially Correlated Data*, pages 309–319. Springer-Verlag, New York.

Waller, L., Carlin, B., Xia, H., and Gelfand, A. (1997b). Hierarchical spatio-temporal mapping of disease rates. *Journal of the American Statistical Association*, 92:607–617.

Waller, L. A. and Gotway, C. A. (2004). *Applied Spatial Statistics for Public Health Data*. John Wiley, Hoboken, NJ.

Waller, L. A. and McMaster, R. B. (1997). Incorporating indirect standardization in tests for disease clustering in a GIS environment. *Geographical Systems*, 4:327–342.

Webster, R., Oliver, M., Muir, K., and Mann, J. (1994). Kriging the local risk of a rare disease from a register of diagnoses. *Geographical Analysis*, 26:168–185.

Wise, S., Haining, R., and Ma, J. (2001). Providing spatial statistical data analysis functionality for the GIS user: the SAGE project. *International Journal of Geographic Information Science*, 15(3):239–254.

Xia, H. and Carlin, B. P. (1998). Spatio-temporal models with errors in covariates: Mapping Ohio lung cancer mortality. *Statistics in Medicine*, 17:2025–2043.

Xia, H., Carlin, B. P., and Waller, L. A. (1997). Hierarchical models for mapping Ohio lung cancer rates. *Environmentrics*, 8:107–120.

Yasui, Y., Liu, H., Benach, J., and Winget, M. (2000). An empirical evaluation of various priors in the Empirical Bayes estimation of small area disease risks. *Statistics in Medicine*, 19:2409–2420.