Bayesian Areal Interpolation, Estimation, and Smoothing: An Inferential Approach for Geographic Information Systems

By Andrew S. Mugglin, Bradley P. Carlin, Li Zhu, and Erin Conlon[1]
Division of Biostatistics, University of Minnesota, Minneapolis, Minnesota 55455, USA

June 25, 1998

Abstract

Geographic information systems (GIS's) offer a powerful tool to geographers, foresters, statisticians, public health officials, and other users of spatially referenced regional datasets. However, as useful as they are for data display and trend detection, they typically feature little ability for statistical *inference*, leaving the user in doubt as to the significance of the various patterns and "hot spots" identified. Unfortunately, classical statistical methods are often ill-suited for this complex inferential task, dealing as it does with data which are multivariate, multilevel, misaligned, and often nonrandomly missing. In this paper we describe a Bayesian approach to this inference problem which simultaneously allows interpolation of missing values, estimation of the effect of relevant covariates, and spatial smoothing of underlying causal patterns. Implemented via Markov chain Monte Carlo (MCMC) computational methods, the approach automatically produces both point and interval estimates which account for all sources of uncertainty in the data. After describing the approach in the context of a simple, idealized example, we illustrate it with a dataset on leukemia rates and potential geographic risk factors in Tompkins County, New York, summarizing our results with numerous maps created using the popular GIS `Arc/INFO`.

[1]Email addresses for the authors: Mugglin, `andy@biostat.umn.edu`; Carlin, `brad@biostat.umn.edu`; Zhu, `liz@biostat.umn.edu`; Conlon, `erin@biostat.umn.edu`.

1 Introduction

Geographical information systems (GIS's) have increasingly found application in varied disciplines, from marketing to forestry to public health, as more and more researchers see spatial components to their research problems. Geographically referenced data are becoming increasingly available, but for reasons of confidentiality or convenience they are frequently reported as aggregate counts over regions that partition a parcel of land. Different reporting agencies often use partitions convenient for their own purposes, with few or no regional boundaries coinciding across agencies. For instance, hospital admissions might be tracked by zip code, while disease cases may be known by county, counts of the population at risk by census tract, and environmental hazard exposure assessed in regions determined by the wind or groundwater flow patterns near a particular waste site. When it is desired to draw meaningful statistical inferences from data available only in these "misaligned" zoning systems, it is necessary to take statistical information from one zonal system of an area and convert it to estimates of that statistic for a different zonation of that same area. This process is known as *areal interpolation*.

Areal interpolation is closely related to *kriging*, which is the smoothing of a spatially-indexed response surface given its exact values at only a few locations. While the statistical literature on kriging is vast (see e.g. Cressie, 1993 for a review), it appears that only Tobler (1979) deals directly with the areal interpolation problem. By contrast, the problem has made frequent entries into the geographical literature, including for example the papers by Goodchild and Lam (1980), Lam (1983), Flowerdew and Green (1989, 1990, 1991, 1992), Flowerdew et al. (1991), and Langford et al. (1991). Goodchild et al. (1993) review various methods for areal interpolation, and propose a unified and generalized framework for their solution. Most recently, Fisher and Langford (1995) describe three major types of areal interpolation:

- Cartographic methods, including simple areal interpolation (the allocation of data to subregions proportionately to their areas) and the so-called *dasymetric* map, which uses knowledge of a locality's characteristics to identify homogeneous regions within the zones (see Flowerdew and Openshaw, 1987).

- Regression methods, which model the desired statistic as a function of covariates (often referred to as *control variables*). Linear regressions (constrained and unconstrained) as well as Poisson regressions fall into this category. Flowerdew (1988), Flowerdew and Green (1989, 1991) and Flowerdew et al. (1991) suggest that Poisson regressions are more appropriate for modeling count data than are linear regressions.

- Surface methods, which posit that the statistic of choice can be modeled as a density surface, and measurements in any region can then be obtained by integrating the surface over the regional boundaries. Tobler's (1979) method is perhaps foremost among these. Cartographic methods can be viewed as special cases of these surface methods, where the surface is not continuous but rather piecewise constant.

Fisher and Langford (1995) employ a Monte Carlo scheme based on modifiable areal units to assess the accuracy of several of these methods.

Designed specifically to handle and combine sources of data collected over spatially misaligned grids, many modern GIS's can automatically perform one or more of these areal interpolation methods. In public health, GIS's are also widely used for highlighting "hot spots" on a map, e.g., subregions having high incidence of a particular disease as well as certain sociodemographic characteristics of interest. However, as useful as they are for data display and trend detection, they typically do not include capability for statistical *inference*, thus precluding answers to questions of "significance" of the various findings. For example, when is a disease "hot spot" truly "hot," and

when is it merely the result of natural random fluctuations – say, due to a small sample size? (In a thinly populated region, it would take relatively few disease events during a "bad year" to create an observed rate that was the "hottest" of the bunch.) Which predictors in a spatial regression model are statistically significant, and which are not? What is the true underlying relative risk of disease in a given region (or subregion)? What will this risk be *next* year?

Inferential tools to answer these and other questions require both a statistical framework for modeling the underlying physical process, and techniques for estimating the variability of the various estimates and predictions obtained using this framework. In this paper, we argue on behalf of a hierarchical Bayesian approach, which captures both of these requirements in a conceptually simple way, and at the same time is quite natural for our misaligned data setting. Over the last decade or so, Bayesian methods have enjoyed increasing use in statistical applications involving complex datasets, producing exact answers where classical (or *frequentist*) methods are either infeasible or reliant on unrealistic assumptions or approximations. The emergence of Bayesian methods in actual practice was facilitated by the simultaneous emergence of *Markov chain Monte Carlo* (MCMC) computational methods for obtaining samples from the associated posterior distributions of the parameters of interest; several recent textbooks (Gelman et al., 1995; Carlin and Louis, 1996; Gilks et al., 1996) describe the Bayesian approach and its implementation via MCMC. Bayesian techniques have also made inroads into the geography literature; most notable are the papers by Hepple (1995a,b), which give detailed descriptions of the Bayesian methodological and computing tools, respectively, useful in spatial econometrics.

Recently, Mugglin and Carlin (1997) proposed a hierarchical Bayesian areal interpolation method appropriate for count data. Building on the work of Flowerdew (1988), their method incorporates covariate information available at a refined scale to allocate areal counts to subregions, and then aggregates the subregional counts according to the boundaries of a different zoning system. Since

their approach is MCMC-Bayesian, it produces not only a point estimate for each interpolated count, but in fact an estimate of the entire posterior distribution for each count. In particular, variance estimates associated with each point estimate are automatically produced, so no extra theory (or computer coding) is required.

This paper extends the setting of Mugglin and Carlin (1997) to include not only interpolation but also estimation and smoothing of underlying map characteristics (in our case, the relative risk of disease in the subregions). We also supplement the covariate information with subregion-specific random effects assigned a spatially smoothing *prior* distribution, reflecting our knowledge or intuition about the regions and their similarity before having seen the data. In particular, we use a conditionally autoregressive (CAR) prior to incorporate similarities in neighboring regions not already explained by our covariates, leading to the desired smoothing over the misaligned grid (Clayton and Kaldor, 1987; Besag et al., 1991). The CAR prior enables us to control the amount of smoothing desired, and whether this smoothing is *local* (anticipating small clusters of subregions with similar risk) or *global* (smoothing all risks toward the same grand mean).

Section 2 lays out our approach, illustrating with a simple idealized example to fix concepts, and showing how interpolation and smoothing are accomplished simultaneously, rather than sequentially. Section 3 then applies the method to a dataset of leukemia case counts (by census tract) and potential exposure covariates (by census block group) in Tompkins County, New York. GIS plots of the fitted underlying relative risks in the subregions are shown for various degrees of local and global smoothing. Finally, Section 4 summarizes our findings and offers directions for future research in this important area.

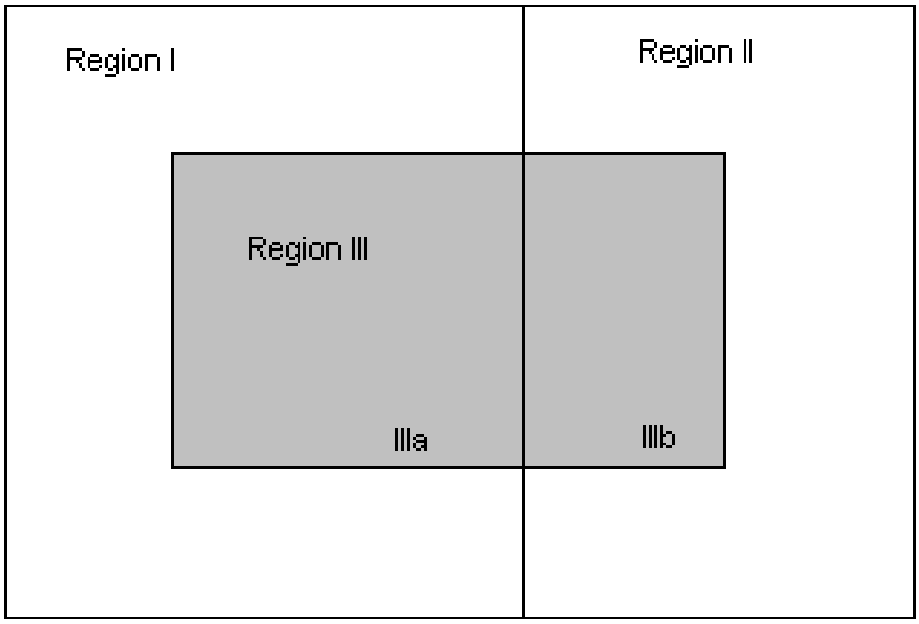

Figure 1: An idealized land parcel, showing source and target regions

2 Bayesian modeling of misaligned data

2.1 Idealized example

Our model begins by assuming that aggregated count data is available over *source* zones, but estimates of these counts are desired according to the boundaries of some (misaligned) *target* zones. Consider the idealized representation originally described by Mugglin and Carlin (1997) and shown in Figure 1. Regions I and II are taken to be source zones, while Region III is the target zone. We label the two subregions of Region III (created by the misalignment with Regions I and II) as Regions IIIa and IIIb, respectively. We take the known population sizes of Regions I and II to be y_1 and y_2, and seek to estimate Y_3, a random variable representing the population size of Region III. The simplest approach would obviously be to allocate the counts proportional to area, i.e., to allocate $y_1 \times [area(IIIa)/area(I)]$ to Region IIIa and $y_2 \times [area(IIIb)/area(II)]$ to Region IIIb, with their sum estimating Y_3. However, we instead assume that the entire region can be subdivided into the finer partition shown in Figure 2, and that covariate information is available

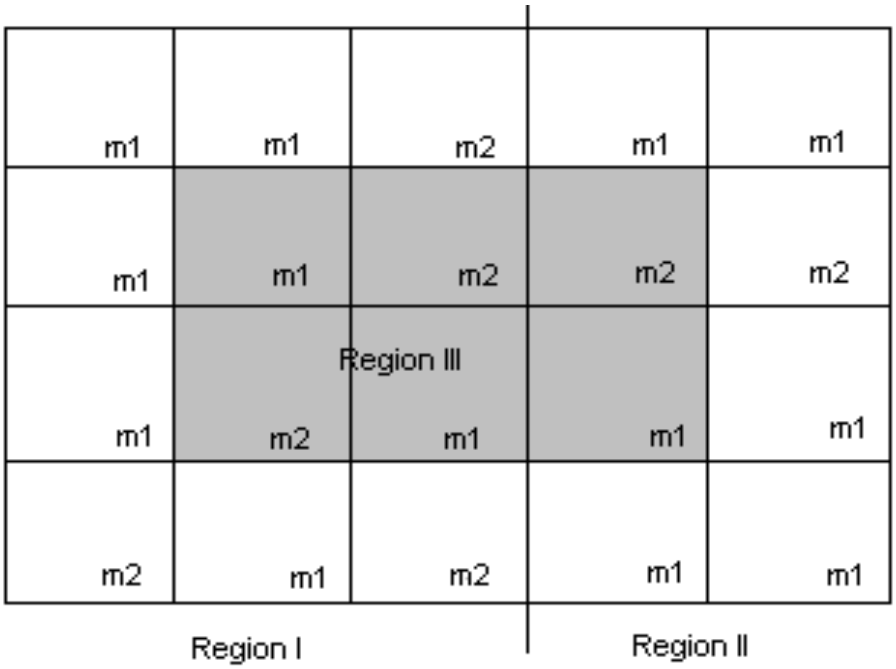

Figure 2: A refined (subregional) partition of the idealized land parcel in Figure 1

for the various subregions created. As indicated in the figure, we assume for simplicity that this covariate measurement is binary (taking the value m_1 or m_2), but this is not necessary for our method to be applicable. Note that groups of the subregions in Figure 2 can be aggregated to form any of Regions I, II, or III.

2.2 Bayesian interpolation

We now assume that on each subregion, the population is a Poisson variable, conditionally independent of the other subregional counts given the covariate values. Letting i index the region (I or II) and j index the subregion, $j = 1, \ldots, J_i$ (so that $J_1 = 12$ and $J_2 = 8$ in our example), a common approach is to assume

$$Y_{ij} \mid \delta_{ij} \overset{ind}{\sim} \text{Poisson}(E_{ij} e^{\delta_{ij}}) \,, \tag{1}$$

where E_{ij} is an *expected* count in region ij, and $e^{\delta_{ij}}$ is defined below. In disease mapping, expected counts are sometimes computed by applying an age- and sex-specific table of morbidity (or mortality) rates to the corresponding population sizes in each group and summing the results within each subregion, a process called *external standardization*. In many cases, however, a standard table is unavailable, and we must resort to *internal standardization*, e.g., setting $E_{ij} = n_{ij}\eta$, where n_{ij} is the total population count in the subregion and η is the overall probability of disease. In either case, since E_{ij} is what we "expect" for Y_{ij}, $e^{\delta_{ij}}$ is therefore the relative risk of contracting the disease in subgroup ij. Since the log-relative risk δ_{ij} can theoretically be any number between plus and minus infinity, this is the scale on which the covariates are typically modeled, i.e.

$$\delta_{ij} = \theta_0 + \theta_1 x_{1ij} + \theta_2 x_{2ij} + \cdots + \theta_K x_{Kij} \ , \tag{2}$$

where the $x_{kij}, k = 1, \ldots, K$, denote the K covariates associated with subregion ij. Here we assume these covariates are predictive of Y and available at the subregional level, thus enabling improved interpolated estimates.

The Poisson assumption is somewhat restrictive, but it also carries two key modeling advantages. First, as we aggregate counts of several subregions together, we are in effect adding conditionally independent Poisson random variables, which produces another Poisson variable. Thus in Figures 1 and 2 we have that Y_{3a} and Y_{3b} are again Poisson distributed given the θ's. Second, when conditioned on the known population values y_1 and y_2, a standard result in probability theory implies that Y_{3a} and Y_{3b} become binomial random variables which cannot exceed the values of y_1 and y_2, respectively. This ensures that population estimates of any conglomeration of subregions cannot exceed the total known population size of their parent regions, and in particular, that the sum of *all* the subregional estimates in a given region must equal the known regional total (often referred

to as the *pycnophylactic property*).

The regression form of equation (2) allows us to use covariate information in the form of the x_{kij}'s to estimate the parameters θ_k, which in turn allows the desired population interpolation, as well as assessments of which covariates are statistically significant. In the Bayesian approach, we assume the θ_k are themselves random variables from some *prior* distribution which we must specify. We may choose the prior to be minimally informative, letting the data be the principal force in producing the model estimates, or we can choose to incorporate external knowledge or perhaps shape constraints to add desirable features to the fitted values. In any case, estimates of the θ_k are obtained from their *posterior* distribution, obtained via Bayes' Rule as

$$p(\boldsymbol{\theta}|\mathbf{y}) \propto f(\mathbf{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta}) \,, \tag{3}$$

where $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_K)$, $\mathbf{y}$ is the observed data, f is the likelihood function, and π is the prior distribution. Note that given f and π, we will always have a functional form *proportional to* the posterior distribution; we require only a computational method for finding the normalized form of $f \cdot \pi$. As mentioned earlier, MCMC methods are ideally suited to this task; in particular, the Metropolis-Hastings algorithm (Metropolis et al., 1953; Hastings, 1970; Carlin and Louis, 1986, Sec. 5.4.3) provides an easily implemented way to obtain samples $\{\boldsymbol{\theta}^{(g)}, g = 1, \ldots, G\}$ from $p(\boldsymbol{\theta}|\mathbf{y})$. These samples may then be summarized in any way desired; e.g., $\bar{\theta}_1 = \frac{1}{G}\sum_{g=1}^{G} \theta_1^{(g)}$ provides an estimate of $E(\theta_1|\mathbf{y})$, the posterior mean (Bayes point estimate) of θ_1. Finally, these samples also serve as input for the final, interpolation step of the process, since for any subregion ij, the posterior distribution of the population size is

$$p(y_{ij}|\mathbf{y}) = \int p(y_{ij}|\boldsymbol{\theta}, \mathbf{y}) p(\boldsymbol{\theta}|\mathbf{y}) d\boldsymbol{\theta} \;\approx\; \frac{1}{G}\sum_{g=1}^{G} p(y_{ij}|\boldsymbol{\theta}^{(g)}, \mathbf{y}) \,,$$

a Monte Carlo integration. Point estimates of y_{ij} can thus also be obtained, and subsequently aggregated to whatever scale is desired.

In the context of the example described above and in Figures 1 and 2, since the subregions are of equal size, suppose $E_{ij} = E$ for all ij (i.e., the expected counts are equal in all subregions). Then using equations (1) and (2) and setting

$$x_{1ij} = \begin{cases} 0\ , & \text{if subregion } ij \text{ has covariate value } m_1 \\ 1\ , & \text{if subregion } ij \text{ has covariate value } m_2 \end{cases} ,$$

the likelihood for Y_1, the population size of Region I, is the sum of 7 independent Poisson(Ee^{θ_0}) random variables and 5 independent Poisson($Ee^{\theta_0+\theta_1}$) random variables, i.e. a Poisson($Ee^{\theta_0}[7+5e^{\theta_1}]$) distribution. Similarly, the likelihood for Y_2 is Poisson($Ee^{\theta_0}[6+2e^{\theta_1}]$). After choosing appropriate priors for the θ_i (say, normal distributions with mean 0 and some variance σ^2), Metropolis sampling then produces the necessary samples from the posterior distribution (3). Again looking at the subregional map in Figure 2, the two binomial distributions for Y_{3a} and Y_{3b} are

$$Y_{3a}|\boldsymbol{\theta}, \mathbf{y} \sim Bin\left(y_1\ ,\ \frac{2+2e^{\theta_1}}{7+5e^{\theta_1}}\right)\ , \tag{4}$$

$$\text{and}\quad Y_{3b}|\boldsymbol{\theta}, \mathbf{y} \sim Bin\left(y_2\ ,\ \frac{1+e^{\theta_1}}{6+2e^{\theta_1}}\right)\ . \tag{5}$$

The sum $Y_3 = Y_{3a}+Y_{3b}$ is unfortunately not itself another binomial, since the success probabilities in (4) and (5) need not be equal. However, a sampling-based estimate of Y_3 is again routine by drawing $Y_{3a}^{(g)}$ from $p(y_{3a}|\boldsymbol{\theta}^{(g)}, y_1)$ in (4), $Y_{3b}^{(g)}$ from $p(y_{3b}|\boldsymbol{\theta}^{(g)}, y_2)$ in (5), and defining $Y_3^{(g)} = Y_{3a}^{(g)} + Y_{3b}^{(g)}$. The sample mean of these values, $\bar{Y}_3 = \frac{1}{G}\sum_{g=1}^{G} Y_3^{(g)}$, provides a point estimate of the interpolated value Y_3, while the sample variance, $s^2_{Y_3} = \frac{1}{G-1}\sum_{g=1}^{G}(Y_3^{(g)} - \bar{Y}_3)^2$, provides a good corresponding variance estimate provided the autocorrelation in the induced chain of Metropolis-sampled values

$Y_3^{(g)}$ is not too high (say, less than 0.9). In any case, both of these estimates can be made arbitrarily accurate simply by taking a sufficiently large MCMC sample size G.

The choice of prior distribution $\pi(\boldsymbol{\theta})$ can have substantial impact on the outcome. If the data are highly informative, a uniform or other "noninformative" prior can be used. However, for small or otherwise weakly informative datasets, a more informative prior may be required. In the Bayesian MCMC setting, weak identification of model parameters often manifests itself as slow or erratic convergence of the MCMC algorithm to its stationary distribution (i.e., the true posterior). In our work, we typically select prior distributions that are as vague as possible while still allowing reasonably good MCMC convergence with the data and model at hand.

2.3 Bayesian smoothing

Choropleth maps of interpolated counts often appear quite "bumpy," in that regions of high count can border directly on regions of low count. This is to be expected, since the interpolated values in the subregions must sum to the fixed, observed regional totals, which may themselves be irregular. However, if the goal is to map not the interpolated counts but the underlying relative risks (or log-relative risks) of disease in the same regions, here we would likely prefer a *smoothed* map, having no stark contrasts between neighboring regions. Such a smoothed map is consistent with a belief that disease risk does not change sharply over short distances (though of course the observed counts may, especially if collected over a short time interval or over thinly populated areas). Unobserved regional covariates, such as race, income level, environmental contamination, and the like, can create "clusters" of subregions having similar disease rates. Smoothed risk maps facilitate identification of such clusters (as well as broad patterns) of disease risk. Several modern disease atlases use some form of smoothing; an excellent recent example is the U.S. mortality atlas by Pickle et al. (1996), which smoothes crude rates via a weighted "headbanging" algorithm (a

nonparametric spatially-oriented median smoother).

Of course, there are situations where smoothing would not be appropriate – for example, if all disease clusters were entirely contained within individual subregions. Alternatively, disease rates might plausibly be expected to vary sharply across subregional boundaries; Knorr-Held and Rasser (1998) show sharp distinctions in male oral cavity cancer mortality rates across the former East-West German border. In our data example, however (leukemia rates in the subregions of a county in upstate New York), we would be surprised if disease risk changed sharply after crossing a regional boundary, so some sort of smoothing seems both practical and appropriate.

In the Bayesian framework, smoothing of the log-relative risks is easily accommodated simply by adding a smoothing term to our model for δ_{ij}. Specifically, we replace equation (2) by

$$\delta_{ij} = \theta_0 + \theta_1 x_{1ij} + \theta_2 x_{2ij} + \cdots + \theta_K x_{Kij} + \phi_{ij} \ , \tag{6}$$

where the ϕ_{ij} are subregion-specific random effects assigned a prior distribution specifically tuned to capture the sort of spatial similarities we expect in the fitted risk map. For instance, we might simply assume

$$\phi_{ij} \overset{iid}{\sim} N(0, 1/\tau) \, , \tag{7}$$

where τ is a precision parameter that is either specified, or itself assigned a prior distribution (sometimes called a *hyperprior*). Note that this model smoothes the δ_{ij} toward a global (map-wide) mean value (determined by the θ's), with τ controlling the degree of smoothing (larger τ implies more smoothing).

As an alternative, we might choose a prior designed to smooth toward *local* mean values, i.e., a prior which anticipates clusters of regions having similar risks located throughout the map. Such a prior often chosen by spatial biostatisticians is the *conditionally autoregressive* (CAR) distribution

(Besag, 1974). This prior assigns a joint density proportional to $\exp(-\frac{\lambda}{2}\boldsymbol{\phi}^T B\boldsymbol{\phi})$ to the vector of random effects $\boldsymbol{\phi} \equiv \{\phi_{ij}\}$. That is, the ϕ_{ij} are multivariate normal, where the symmetric, positive definite matrix B determines the relationships among the parameters by relating them to their regions' respective positions on the map. The precision information in the B matrix is determined by the sample sizes in each region, while the correlation structure between two regions is defined by whether or not they are *neighbors*. More specifically, we follow Besag, York, and Mollié (1991), who for each region ij define a neighbor set ∂_{ij}. This set might contain all regions within a certain distance of region ij, or simply those regions adjacent to region ij. This definition may need to be adjusted (especially in the adjacency case) when regions are of greatly differing size or are separated by significant natural landmarks (lakes, mountain ranges, etc.).

Letting n_{ij} be the number of neighbors of region ij, the CAR prior produces the conditional prior distribution of ϕ_{ij} as

$$\phi_{ij} \mid \phi_{k \neq ij} \sim N\left(\bar{\phi}_{ij}\,,\ \frac{1}{\lambda n_{ij}}\right)\,, \tag{8}$$

where $\bar{\phi}_{ij} = n_{ij}^{-1} \sum_{k \in \partial_{ij}} \phi_k$. That is, the conditional prior for ϕ_{ij} is normal, centered around the mean value of its neighbors, and with variance decreasing in the number of neighbors. The parameter λ, common to all of these neighbor structures, controls the degree of smoothing imposed. Thus λ's role is similar to that of τ in the independence prior above, only now the smoothing is local instead of global. Also note λ's location in a *conditional* prior specification, while τ's prior is specified unconditionally, muddying comparisons between the two prior forms (though see Bernardinelli et al., 1995 for recent simulation work attempting to calibrate across the two scales).

Finally, we note that the CAR prior as specified in (8) is translation invariant; adding a constant to each of the ϕ_{ij} will not change the value of the prior. Thus, to preserve the estimability of θ_0 in equation (6) we impose the constraint $\sum_{ij} \phi_{ij} = 0$. Despite its awkward appearance, this

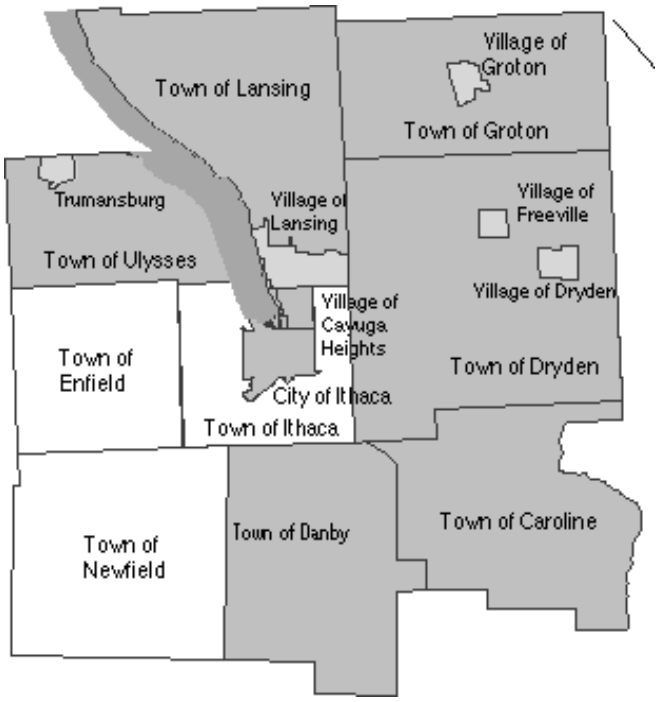

Figure 3: Map of Tompkins County, New York.

constraint is easy to implement "on the fly" following each iteration of our MCMC algorithm simply by recentering the sampled ϕ_{ij} values around their own mean.

3 Illustration: Leukemia in Tompkins County, New York

3.1 Description of dataset

Having summarized the Bayesian methodological approach to map interpolation and smoothing, we turn to an illustration using a dataset originally presented and analyzed by Waller et al. (1994), which reports the incidence (1978–1982) of leukemia in Tompkins County, New York. As seen in Figure 3, this county, located in west-central New York State, is roughly centered around the city of Ithaca. Tompkins County is divided into 23 census tracts (using the 1980 U.S. Census), with each tract further subdivided into between 1 and 5 block groups, for a total of 51 such subregions.

In this dataset, leukemia counts are available only at the census tract level; we seek to obtain

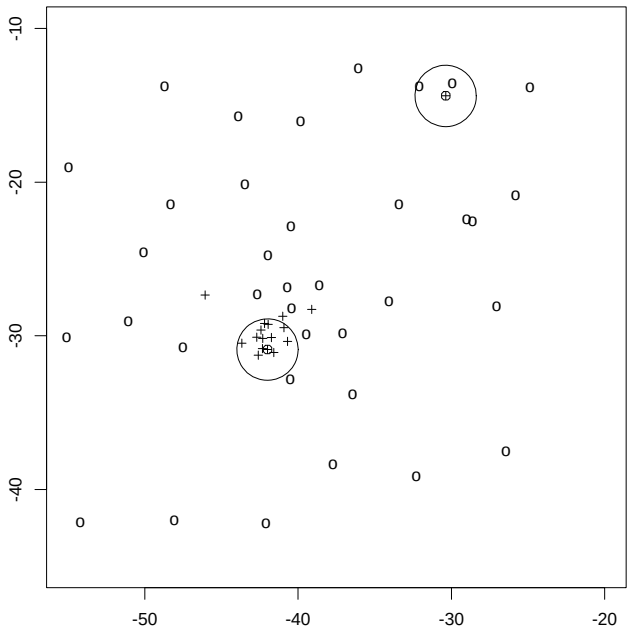

Figure 4: Schematic map of block group centroids, Tompkins County, NY. Here, "+" denotes an urban block group, "o" denotes a rural block group, and "⊕" denotes a waste site.

maps of interpolated counts and smoothed underlying relative risks at the finer, block group level, plotting both using a GIS. Fortunately, two binary covariates are available to assist us. The first is a binary indicator of whether the block group is primarily urban or rural in character (1=urban, 0=rural), with the Ithaca city block groups being those assigned the "urban" designation. The second covariate is also binary, and indicates whether or not the block group's population-weighted centroid is located within 2 km of a trichloroethylene (TCE) waste site. These are inactive hazardous waste sites listed by the New York Department of Environmental Conservation as containing TCE, a volatile organic compound putatively associated with increased leukemia risk. Waller et al. (1994) provide a review of the epidemiologic literature relating TCE exposure and leukemia, as well as an analysis of other TCE sites in upstate New York. In particular, the original analysis by Waller et al. (1994, p.10) considered exposure radii of 1, 2, and 4 km; we considered all three values but for simplicity chose to report only those results based on the middle value, 2 km.

Figure 4 provides a rough schematic of Tompkins County drawn in `S-Plus`, a popular and easy-to-use statistical package with limited mapping capabilities. While unable to show regional boundaries, this map does show the block group centroids and the two waste sites, one in the northeast corner of the county near the village of Groton, and the other in the southern part of Ithaca. Our covariate data are only moderately informative, and a quick glance at Figure 4 reveals why. Most of the urban block groups (the "+" signs on the map) are within 2 km of the Ithaca waste site; at the same time, only two rural block groups (the two in the northeast corner) are near a waste site. Statistically speaking, it is as if we have substantial information only for the diagonal entries in a two-by-two table of data, where the row designations are "urban" and "rural," and the columns are "exposed" and "unexposed." This necessitates a moderately informative prior on the θs, since noninformative priors will lead to MCMC algorithm convergence failure (the MCMC-Bayesian analogue of collinearity in traditional regression modeling). In what follows, we used a N(2,1) prior for each θ. We experimented with other priors (e.g., N(0,1) priors) and obtained broadly similar results.

Regarding neighborhood structure, we defined block groups to be neighbors if their respective population-weighted centroids were within 7 km of each other. The choice of 7 km radius is admittedly rather arbitrary, but was chosen to ensure that every block had at least one neighbor (the nearest neighbor to the most isolated block group centroid was roughly 6.5 km away). Again, alternate definitions might simply rely on regional contiguity. Best et al. (1998) investigate the effect of changing the distributional form of the smoothing prior, as well as the definition of adjacency.

Figure 5 shows a choropleth map (drawn in the GIS `Arc/INFO`) of the crude log-relative risks (method of moments estimates) computed by census tract (though block group-level boundaries are often still visible). Darker regions have higher leukemia risk. Note the preponderance of white and black regions with few gray regions. Clearly the shading scale here is too narrow to detect

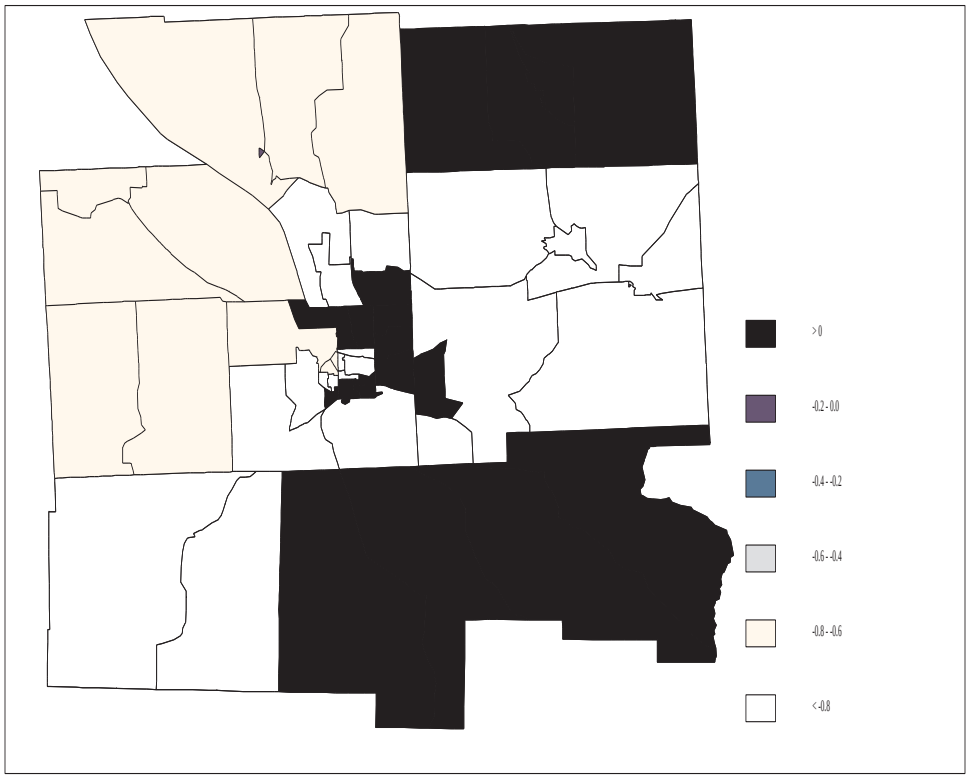

Figure 5: Log relative risk of leukemia by census tract, unsmoothed. Values actually range from –3.89 to 1.35; narrower classification key is used for consistency with subsequent smoothed maps.

subtle differences in risk, but was chosen for consistency with our later, smoothed risk maps. We remark that our log relative risks are not centered around 0 because they are standardized relative to the grand rate in the 8-county region studied by Waller et al. (1994), an area which includes Tompkins County but has slightly higher risk overall. Figure 5 summarizes the input data for our Bayesian smoothing procedure.

3.2 Parameter estimation

We begin by applying nonsmoothing interpolation model to the Tompkins County data, where equation (2) is replaced by

$$\delta_{ij} = \theta_0 + \theta_1 u_{ij} + \theta_2 w_{ij} + \theta_3 u_{ij} w_{ij} \ , \tag{9}$$

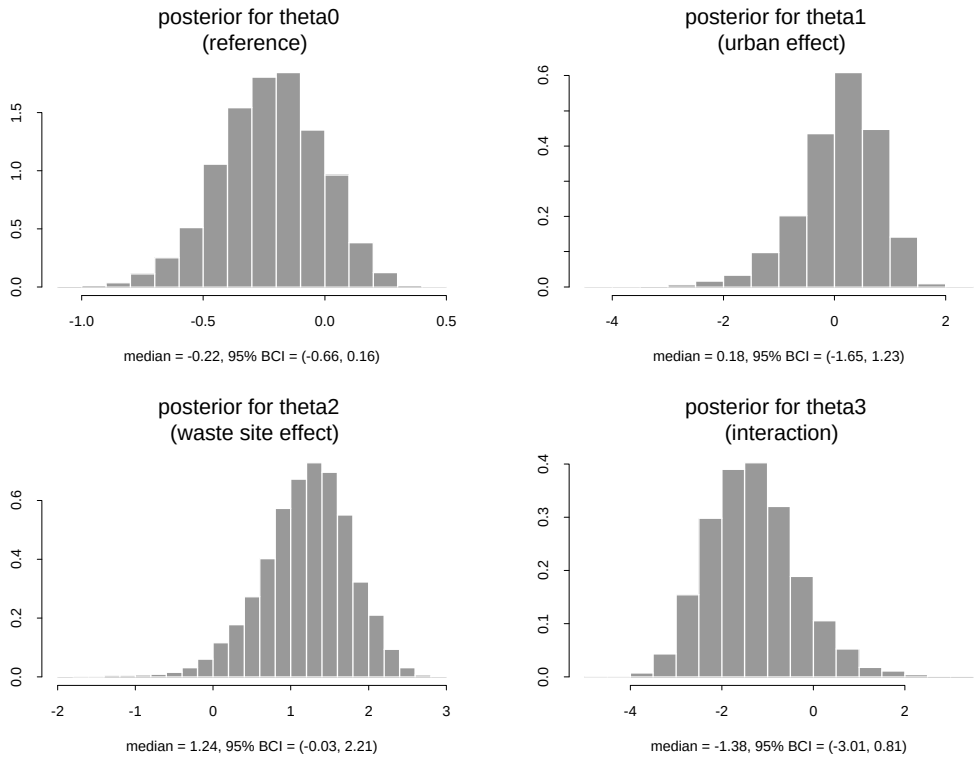

Figure 6: Posterior histograms of sampled log-relative risk parameters, Tompkins County dataset

where u_{ij} is the urban/rural covariate, and w_{ij} is the exposed/unexposed covariate. We ran 5 parallel MCMC sampling chains for 2000 iterations each, discarding the first 200 from each chain as pre-convergence "burn-in." The remaining $5 \times 1800 = 9000$ $\boldsymbol{\theta}^{(g)}$ samples are summarized in Figure 6, which shows the posterior histograms for θ_i, $i = 0, \ldots, 3$. We note that θ_0, θ_1, and θ_3 are not significantly different from 0 as judged by the 95% Bayesian confidence interval (abbreviated BCI in the figure), while θ_2 is marginally significant at this level. This suggests the data are sufficient to detect some harmful effect of residing within 2 km of a waste site, but not sufficient to detect any effect associated with living in an urban area (in this case, the city of Ithaca).

The preponderance of negative $\theta_3^{(g)}$ samples is surprising, since it suggests a *protective* effect of living both in an urban area and near a waste site. Statistically speaking, this is apparently due to the previously mentioned unbalanced nature of the data itself. Most of the block groups in

Ithaca are "exposed," including many with relatively low observed disease risk. Combined with the presence of a few rural block groups *not* near a waste site with relatively high observed rates, the interaction term θ_3 must adjust to bring the fitted rates in Ithaca back down somewhat. A possible explanation for this phenomenon involves the exposure pathway for TCE, which is normally via groundwater. If persons residing near the Groton waste site drink well water, while persons near the Ithaca site instead drink water from Lake Cayuga, this would explain the difference in observed rates between the urban and rural waste sites.

3.3 Global smoothing

Since Figure 5 captures regional mean risk but not its variability (i.e., regional sample size), its estimates are occasionally misleading. For instance, the high variance associated with the high risk estimate for the thinly populated census tract in the southeastern corner is not conveyed by the map. Furthermore, the map is drawn at the census tract level, rather than the finer block group level at which we have relevant covariate information. In any case, the map does not present a plausible picture of underlying relative leukemia risk; we would expect such a risk map to be much more smoothly varying in space. These observations motivate application of the smoothing model in equation (6).

We add the random effects ϕ_{ij} to our log relative risk model (9), and assign them the global smoothing prior (7). (Recall we expect the leukemia rates to vary smoothly over our study region, and we prefer smooth maps for identifying clusters and broad patterns in these rates.) Figure 7 maps the fitted posterior mean estimates of log relative risk, $E[\delta_{ij}|\mathbf{y}]$, obtained from our MCMC algorithm. Note that the collection of log-relative risks has been "shrunk" back towards the grand mean value, as indicated by the presence of various shades of gray in the map. Elevated risks are evident near the two waste sites, but local "clustering" is not particularly evident (see for example

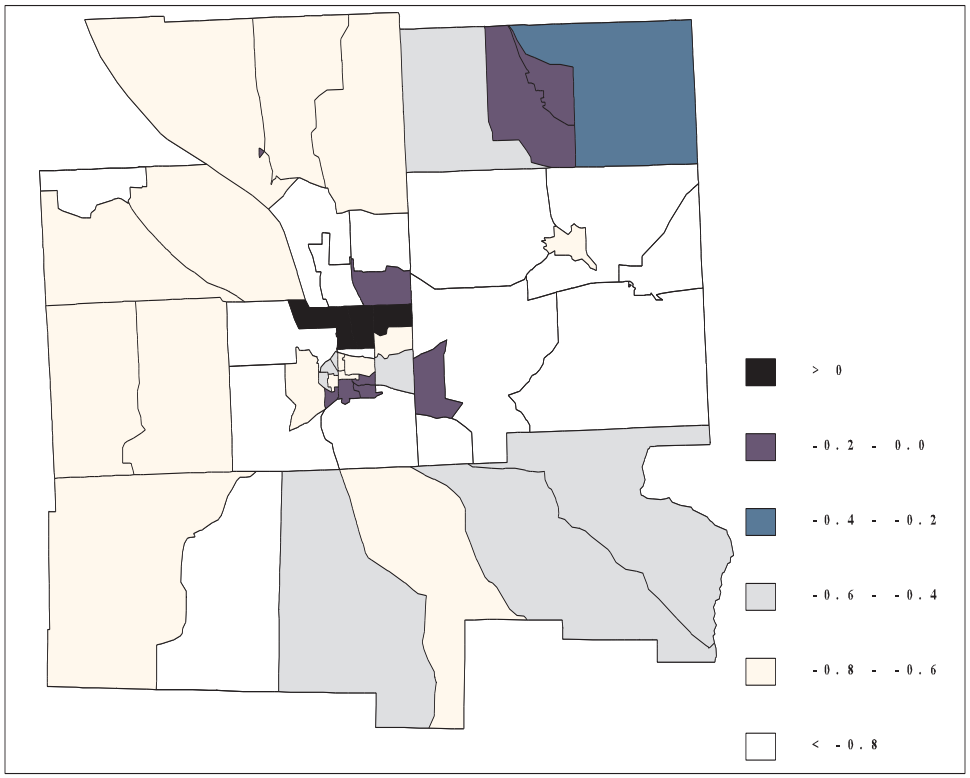

Figure 7: Log-relative risks of leukemia, global smoothing prior ($\tau = 1$)

the group of unshaded regions separating Ithaca from the western and southeastern block groups).

3.4 Local smoothing

In this section we now change the prior on ϕ_{ij} to the CAR (local smoothing) prior shown in equation (8). Following Bernardinelli et al. (1995), we begin by attempting a specification that is roughly comparable to that in Figure 7 by setting $\lambda = 0.1$. The resulting fitted risks are shown in Figure 8. Now some local clustering is apparent, with the highest fitted risks located near the two waste sites and in northern Ithaca, and with the next highest risks in regions adjacent to these. The county-wide degree of smoothness does seem comparable to that in Figure 7, with the high risk in the rural southeastern corner shrunk even more dramatically back towards the grand mean.

In Figure 9, we increase the amount of smoothing by increasing the value of the CAR parameter to $\lambda = 1$. The two block groups near the northeast waste site emerge even more prominently, and

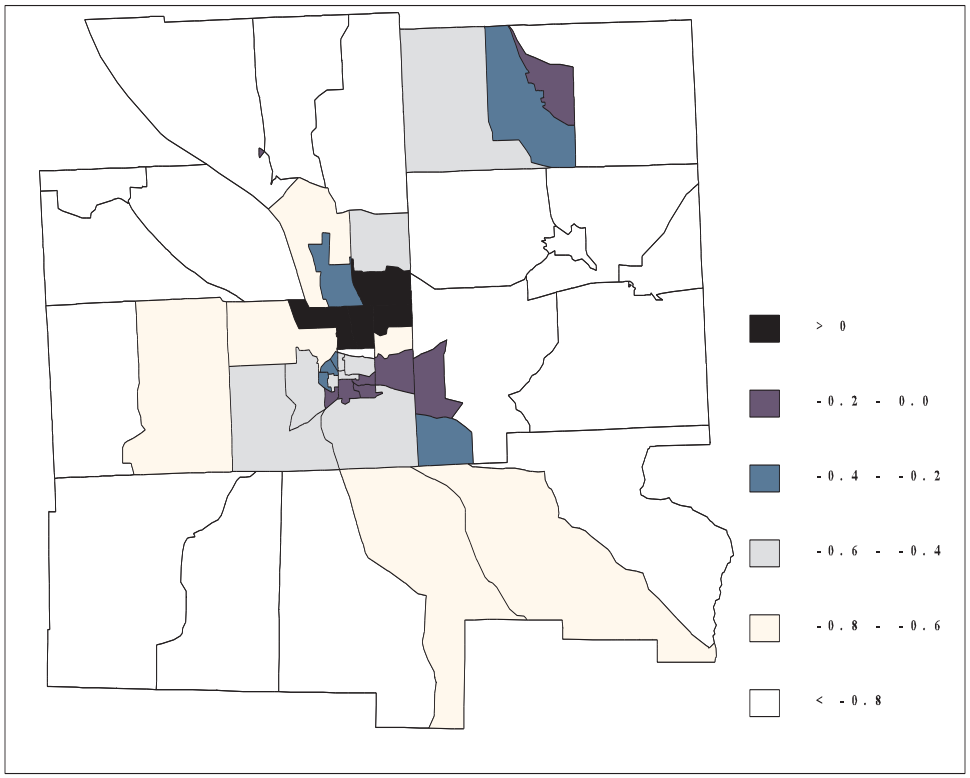

Figure 8: Log-relative risks of leukemia, local smoothing (CAR) prior ($\lambda = 0.1$)

only the two southwesternmost block groups remain unshaded. Finally, Figure 10 shows the results using $\lambda = 100$, a very large value which oversmoothes the fitted risks. The local clustering visible in Figure 9 is now obscured by the excessive level of smoothing; a plot of the fitted rates using a comparable global smoothing prior (e.g., with $\tau = 1000$) produces essentially the same picture. Despite this oversmoothing, the elevated risk near the two waste sites in Groton and Ithaca is still visible, a comforting feature of the CAR technology.

As mentioned earlier, an alternative to fixing a single value for λ (or τ) in our priors at the outset is to specify a hyperprior distribution for this parameter, and subsequently add it into the MCMC sampling order. Since λ and τ are precision parameters in normal distributions, the gamma distribution emerges as a convenient choice for such a hyperprior. In our case, however, the sparseness of our dataset required such hyperpriors to be fairly informative. As a result, we prefer

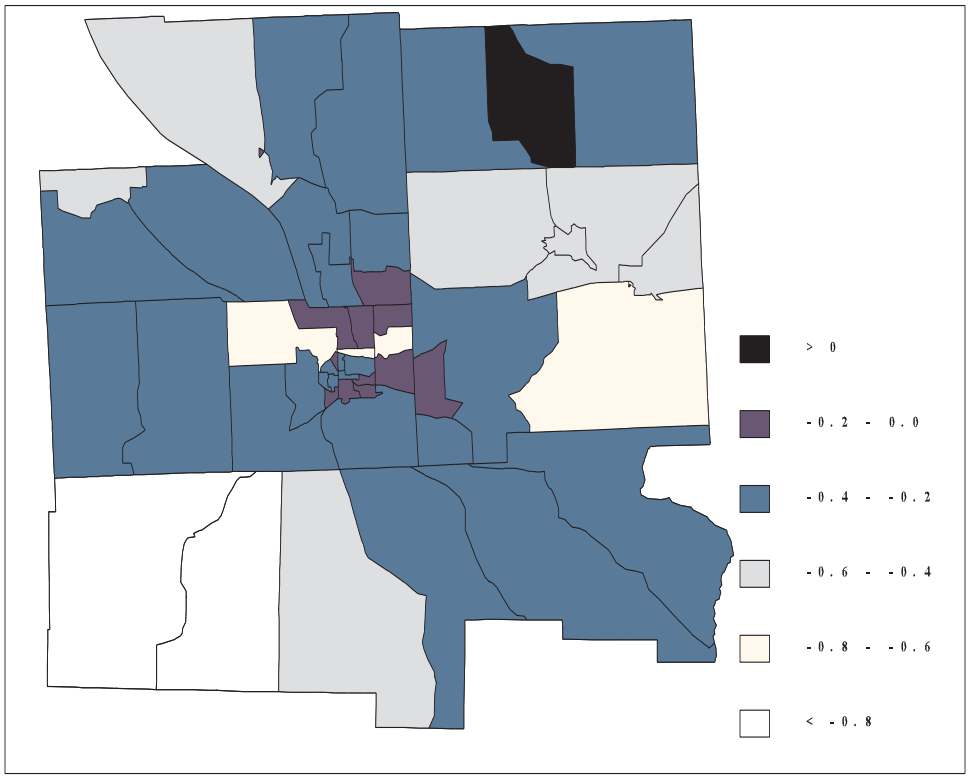

Figure 9: Log-relative risks of leukemia, local smoothing (CAR) prior ($\lambda = 1$)

to compare the fitted maps obtained under a variety of smoothing constants, as in Figures 8–10.

Finally, Figures 11 and 12 map the posterior medians of the ϕ_{ij} under the CAR smoothing priors used in Figures 8 and 9, respectively. These are often referred to as "spatial residuals," since they can be viewed as surrogates for underlying unobserved spatially varying covariates. These two figures reveal little additional spatial variability beyond that already explained by the spatially varying covariates in the model (here, urban/rural status and waste site proximity), with most values in the interval (–0.5, 0.5). The cluster of darker areas just southeast of Ithaca appear to be the result of two nonzero disease counts in the corresponding census tracts, which are not near a waste site. Adjacency seems to us to figure a bit too prominently in Figures 9 and 12, so we might conclude that the CAR($\lambda = 0.1$) model, which demands less local smoothing, is a superior fit for these data. Alternatively, more formal Bayesian model choice tools (such as deviance or other

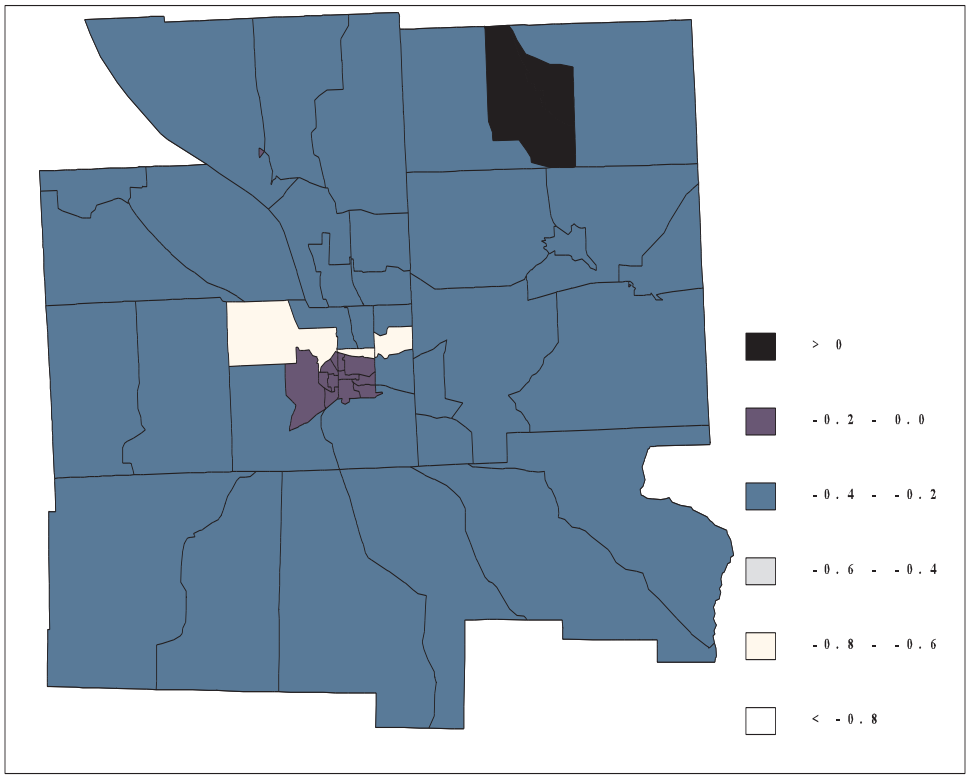

Figure 10: Log-relative risks of leukemia, local smoothing (CAR) prior ($\lambda = 100$)

predictive model discrepancy scores) could be used; see e.g. Carlin and Louis (1996, Sec. 6.4).

4 Discussion

The Bayesian approach to areal interpolation and smoothing illustrated in this paper extends an earlier model and approach by Flowerdew (1988) and Flowerdew and Green (1991). The method is flexible in that it can incorporate any number of either continuous or discrete covariates, easily implementable via Markov chain Monte Carlo computational techniques, and comprehensive in that it yields entire distributions for parameters of interest, thereby allowing assessment of the precision of its estimates. The resulting smoothed spatial disease maps are of course useful for spotting broad patterns of disease risk that might be obscured in crude maps like Figure 5, and are also more accurate for estimation purposes, since they employ the so-called Bayesian "borrowing

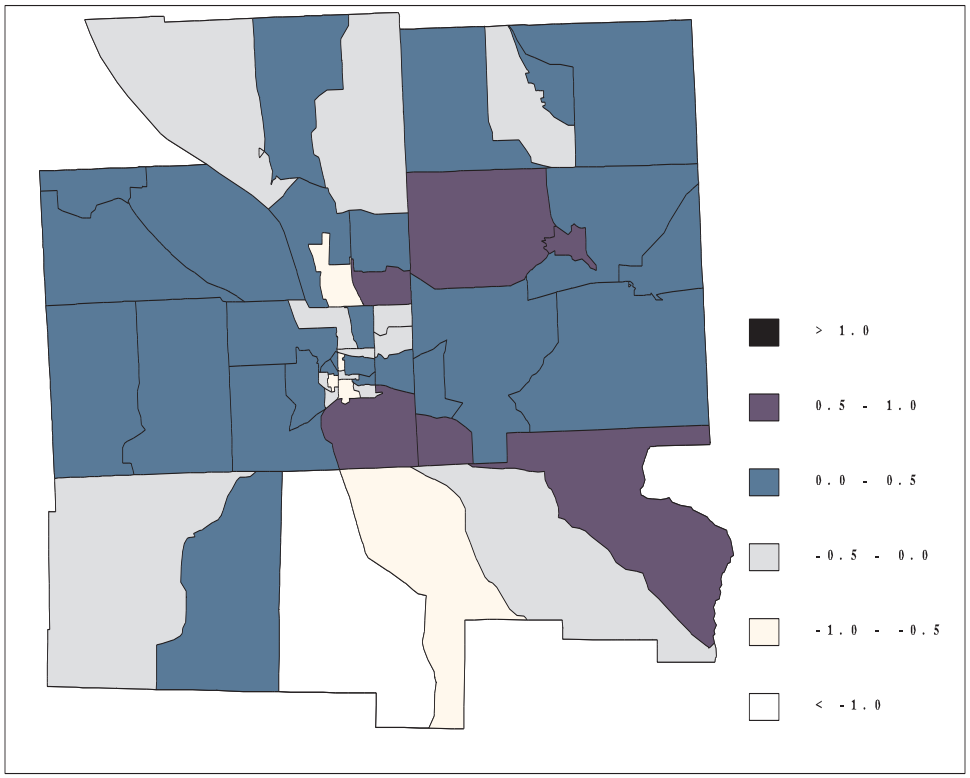

Figure 11: Posterior medians of the ϕ_{ij} (spatial residuals) from the local (CAR) smoothing model with $\lambda = 0.1$.

of strength" across regions (i.e., each region's estimate is informed by data from other, neighboring regions). Our maps can also suggest areas where a case-control study of the disease in question is justified, which in turn might motivate a careful prospective study.

Two different methods of smoothing have been presented, determined by the choice of prior distribution for a particular additive term in the model. The CAR (conditionally autoregressive) prior (8) seems particularly effective, both because of the plausibility of its assumption that neighboring regions are similar (e.g., due to the presence of unmeasured spatially varying covariates) and because the maps it produces are intuitively appealing and useful for spotting broad trends. The smoothing parameters assigned the CAR prior capture *residual* spatial correlation not already accounted for by the covariates (in our example, urban/rural designation and waste site proximity). As a result, patterns in maps of their fitted values (such as Figures 11 and 12) may suggest the

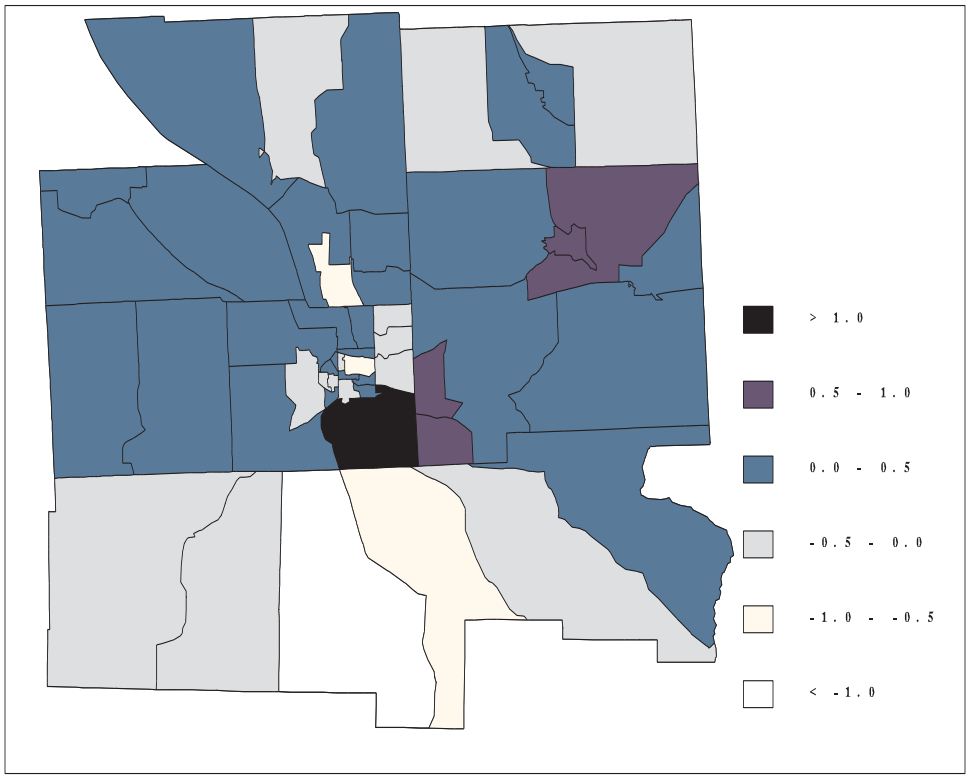

Figure 12: Posterior medians of the ϕ_{ij} (spatial residuals) from the local (CAR) smoothing model with $\lambda = 1$.

presence of spatially-correlated covariates missing from the model.

Though the census tract level information mapped in Figure 5 was taken as the input data in the previous section, block group level data are in fact available. A map of crude log-relative risks calculated at this level is shown in Figure 13. We present these data at this late point in our analysis only to illustrate the point that, while this figure is based on more refined data than our algorithm was able to see, it still does not constitute a "right answer" to the problem of mapping underlying disease risk, due to its strong lack of smoothness. For example, the extremely dark region in the southwest corner of the map is the result of just a single leukemia case in a sparsely populated block group. Clearly the smoothed maps in Figures 7, 8, and 9 offer more plausible pictures of underlying leukemia risk, while still preserving substantial geographic resolution in the areas near the waste sites.

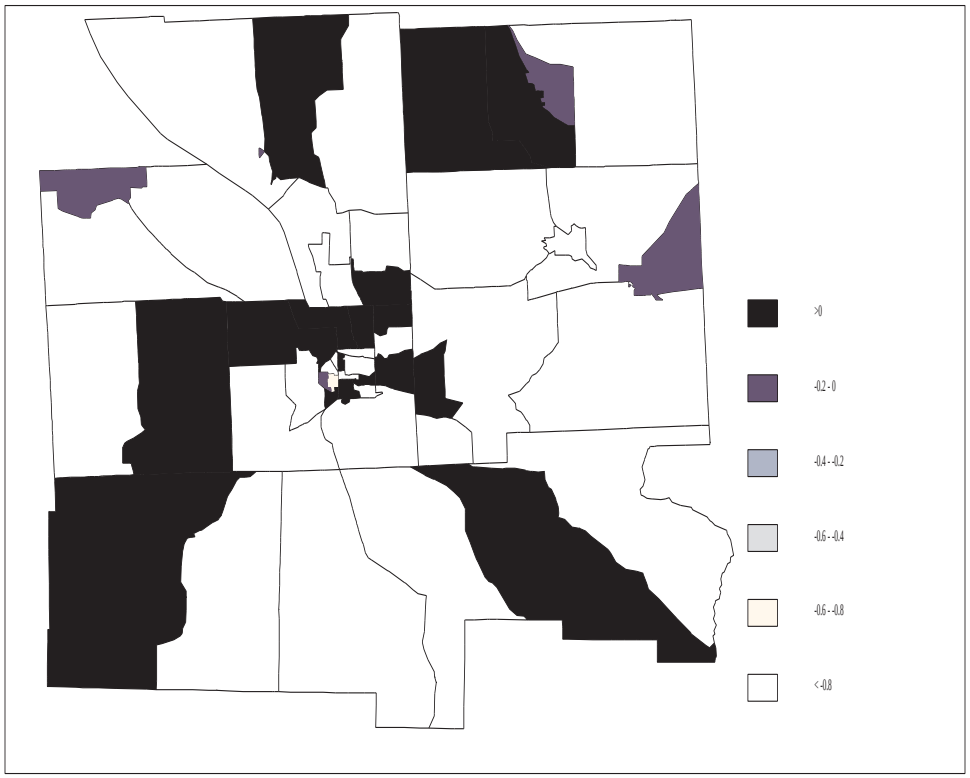

Figure 13: Log relative risk of leukemia by census block group, unsmoothed. Values actually range from –3.89 to 1.35; narrower classification key is used for consistency with previous smoothed maps.

Directions for future work in this area are many. The first concerns *nonnested* grid structure. In many problem settings, the response data and covariate information will be available over totally separate zonal systems (e.g., census tracts versus zip codes), as opposed to the nested situation (census block groups within census tracts) considered here. Bayesian methods again provide a useful blueprint for structuring the resulting model, though of course the notation and computing will be substantially more complex. Second, Bayesian methods for non-grid data (e.g., point source exposures) must be developed. The distance-based method of exposure measurement is often criticized; better models would account for geographic features (hills, lakes, etc.) as well as the transportation mechanism of the pollutant in question (hydrologic, aerosol, etc.). Third, more attention should perhaps be paid to the objective of the inference, rather than just the technical aspects of the model. For example, a strongly informative CAR prior may be entirely appropriate

when the objective is to produce a smoothed disease map for identifying broad patterns of disease in space; however, it may be less so if the goal is to identify small clusters of elevated risk, estimate the effect of a particular covariate, or obtain optimal risk predictions for a particular subregion. Finally, an important practical goal of our development must be to create a statistically "smart" GIS system which incorporates the MCMC smoothing routines described above. The developing link between `Arc/INFO` (for mapping and data management) and `S-Plus` (for computing, possibly aided by subroutines written in a compiled language such as `C++` or `Fortran 90`) offers a promising direction for investigation.

Acknowledgements

The work of all four authors was supported in part by National Institute of Environmental Health Sciences (NIEHS) Grant 1-R01-ES07750. The contents of this paper are solely the responsibility of the authors and do not necessarily represent the official views of the NIEHS or NIH. The authors are grateful to Prof. Lance Waller for providing the Tompkins County data set, as well as substantial analytic and editorial advice.

References

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

Besag, J. (1974), "Spatial interaction and the statistical analysis of lattice systems" (with discussion), *J. Roy. Statist. Soc., Ser. B*, **36**, 192–236.

Besag, J., York, J.C., and Mollié, A. (1991), "Bayesian image restoration, with two applications in spatial statistics" (with discussion), *Ann. Inst. Statist. Mathematics*, **43**, 1–59.

Best, N.G., Waller, L.A., Thomas, A., Conlon, E.M., and Arnold, R.A. (1998), "Bayesian models for spatially correlated disease and exposure data," to appear in *Bayesian Statistics 6*, J.M. Bernardo, J.O. Berger, A.P. Dawid and A.F.M. Smith, eds., Oxford: Oxford University Press.

Carlin, B.P. and Louis, T.A. (1996), *Bayes and Empirical Bayes Methods for Data Analysis*, London: Chapman and Hall.

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

Cressie, N.A.C. (1993), *Statistics for Spatial Data*, 2nd ed., New York: John Wiley.

Fisher, P.F. and Langford M. (1995), "Modeling the errors in areal interpolation between zonal systems by Monte Carlo simulation," *Environment and Planning A*, **27**, 211–224.

Flowerdew, R. (1988), "Statistical methods for areal interpolation: predicting count data from a binary variable," Research Report 16, Northern Regional Research Laboratory, Lancaster and Newcastle.

Flowerdew, R. and Green, M. (1989), "Statistical methods for inference between incompatible zonal systems," in *The Accuracy of Spatial Databases*, Eds. M.F. Goodchild and S. Gopal (Taylor and Francis, London), pp. 239–247.

Flowerdew, R. and Green, M. (1990), "Inference between incompatible zonal systems using the EM algorithm," Research Report 6, North West Regional Research Laboratory, Lancaster University.

Flowerdew, R. and Green, M. (1991), "Data integration: statistical methods for transferring data between zonal systems," in *Handling Geographical Information: Methodology and Potential Applications*, Eds. I. Masser, M. Blakemore (Longman, Harlow, Essex) pp. 38–54.

Flowerdew, R. and Green, M. (1992), "Developments in areal interpolating methods and GIS," *Annals of Regional Science*, **26**, 67–78.

Flowerdew, R., Green, M., and Kehris, E. (1991), "Using areal interpolation methods in geographic information systems," *Papers in Regional Science*, **70**, 303–315.

Flowerdew, R., and Openshaw, S. (1987), "A review of the problem of transferring data from one set of areal units to another incompatible set," Research Report 4, Northern Regional Research Laboratory, Universities of Lancaster and Newcastle upon Tyne.

Gelman, A., Carlin, J., Stern, H., and Rubin, D.B. (1995), *Bayesian Data Analysis*, London: Chapman and Hall.

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

Goodchild, M.F., Anselin, L, and Deichmann, U. (1993), "A framework for the areal interpolation of socioeconomic data," *Environment and Planning A*, **25**, 383–397.

Goodchild, M.F. and Lam, N.S. (1980), "Areal interpolation: a variant of the traditional spatial problem," *Geoprocessing*, **1**, 297–312.

Hastings, W.K. (1970), "Monte Carlo sampling methods using Markov chains and their applications," *Biometrika*, **57**, 97–109.

Hepple, L.W. (1995a), "Bayesian techniques in spatial and network econometrics: 1. Model comparison and posterior odds," *Environment and Planning A*, **27**, 447–469.

Hepple, L.W. (1995b), "Bayesian techniques in spatial and network econometrics: 2. Computational methods and algorithms," *Environment and Planning A*, **27**, 615–644.

Knorr-Held, L. and Rasser, G. (1998), "Bayesian detection of clusters and discontinuities in disease maps," Discussion Paper 107, Technische Universität München, Universität Regensburg.

Lam, N.S. (1983), "Spatial interpolation methods: a review," *The American Cartographer*, **10**, 129–149.

Langford, M., Maguire, D.J., and Unwin, D.J. (1991), "The areal interpolation problem: estimating population using remote sensing in a GIS framework," in *Handling Geographical Information: Methodology and Potential Applications*, Eds. I. Masser, M. Blakemore (Longman, Harlow, Essex) pp. 55–77.

Metropolis, N., Rosenbluth, A.W., Rosenbluth, M.N., Teller, A.H., and Teller, E. (1953), "Equations of state calculations by fast computing machines," *Journal of Chemical Physics*, **21**, 1087–1091.

Mugglin, A.S. and Carlin, B.P. (1997), "Hierarchical modeling in geographic information systems: population interpolation over incompatible zones," Research Report 97–004, Division of Biostatistics, University of Minnesota.

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

Tobler, W.R. (1979), "Smooth pycnophylactic interpolation for geographical regions" (with discussion), *Journal of the American Statistical Association*, **74**, 519–536.

Waller, L.A., Turnbull, B.W., Clark, L.C., and Nasca, P. (1994), "Spatial pattern analyses to detect rare disease clusters," in *Case Studies in Biometry*, N. Lange, L. Ryan, L. Billard, D. Brillinger, L. Conquest, and J. Greenhouse, eds., New York: Wiley, pp. 3–23.