Hierarchical Modeling in Geographic Information Systems: Popolation Interpolation over Incompatible Zones

BY ANDREW S. MUGGLIN AND BRADLEY P. CARLIN

Division of Biostatistics, School of Public Health, University of Minnesota,
Box 303 Mayo Memorial Building, Minneapolis, Minnesota 55455-0392, U.S.A.

Correspondence author: Bradley P. Carlin
telephone: (612) 624-6646
fax: (612) 626-0660
email: brad@muskie.biostat.umn.edu

February 2, 1998
Hierarchical Modeling in Geographic Information Systems:
Population Interpolation over Incompatible Zones

Summary

When inference is desired regarding some attribute of a particular geographic region, it often happens that data are not directly available for that region. However, it may be that data are available over the same general area, but reported according to a different set of regional boundaries. Recently, powerful computer programs called geographic information systems (GIS's) have enabled the simultaneous display of such “misaligned” data sets, but these systems address only the descriptive needs of the user, leaving the inferential goal unmet. In this paper we describe a hierarchical Bayes approach, implemented via Markov chain Monte Carlo methods, which provides a natural solution to this problem through its ability to sensibly combine information from several sources of data and available prior information. After presenting a simple idealized example to illustrate the method, we apply it to a data set on leukemia rates in Tompkins County, New York, wherein we use block group-level covariate information to interpolate disease counts given only aggregate (census tract-level) summaries. We display our results graphically using both statistical (S-plus) and GIS (ARC/INFO, MapInfo) software packages. The approach emerges as flexible, accurate, and suggestive of promising related methods for spatial smoothing of underlying relative risks.

Key Words: Bayesian methods; Markov chain Monte Carlo; Misaligned data; Spatial statistics.

1 Introduction

A central problem in studies of spatial processes concerns the regions for which data are available. The United States government expends considerable resources every ten years on its national census, and many other organizations, such as the Centers for Disease Control and Prevention,
track health and demographic statistics over various geographic regions. But it often happens that
the desired region of analysis is not the region for which data are available. For instance, it may
be that data are available for electoral districts, but the region of interest is an area defined by
commuting distance to a particular industrial part of town. Or perhaps disease data are known
at the county level, but hypotheses of interest pertain to sociodemographically depressed census
tracts. We refer to regions on which data are available as “source” zones and regions for which
data are needed as “target” zones.

Early studies of population interpolation among incompatible zones relied on areal interpolation,
which is the process of allocating population counts to the subregions in proportion to subregional
area. The validity of this approach obviously depends on the population being more or less evenly
distributed across each region. Tobler (1979) introduced a method that assumed population density
to be a continuous function of location; efforts then centered on estimating that function. This
method is appropriate for continuous outcome variables but is harder to justify for count data –
especially counts of human populations, since people do not generally spread out continuously over
a region of space; they tend to cluster.

Flowerdew and Green (1989) presented an approach wherein the variable of interest is count
data and which uses information about the distribution of a binary covariate in the target zone
to help estimate the counts. Their approach applies Poisson regression iteratively, using the EM
algorithm, to estimate target zone characteristics. This method takes into account any number of
other available variables and is reasonably flexible and effective. While subsequent work (Flow-
erdew and Green, 1992) extended this EM approach to continuous (typically normally distributed)
outcome variables, neither of these papers reflects a fully inferential approach to the population
interpolation problem. Various authors have shown hierarchical modeling approaches to be ideal
for combining disparate sources of information in such a way that the associated uncertainties
are correctly propagated throughout the model. Such methods are also flexible (allowing a very broad class of choices for the likelihood and prior distributions of the model parameters), easy to implement (typically via Markov chain Monte Carlo, or MCMC, computational techniques), and comprehensive (yielding entire distributions for the parameters of interest, rather than merely point or interval estimates).

Increasing interest in environmental issues and their inherent spatial considerations, combined with the concurrent emergence of geographic information systems (GIS’s) for easy data summary over incompatible zones, increases the need for inferential techniques for such data sets. This paper presents a Bayesian hierarchical modeling approach to this problem, based on the count data setting of Flowerdew and Green (1989). In Section 2 we motivate the method with a simple model, using as our basis the problem of estimating population in a specific target zone. After working through the theory and simulations we obtain an estimated posterior distribution of the target zone population. In Section 3 we generalize to a more complicated real-data situation wherein the counts to be estimated are leukemia cases in Tompkins County, NY, obtaining numerical and graphical estimates of model characteristics. Finally, in Section 4 we assess the method and suggest avenues for further research.

2 Motivating example

Consider the diagram in Figure 1. Assume that a particular rectangular tract of land is divided into two regions (I and II), and disease counts $y_1$ and $y_2$ are known for these regions (the source zones). But suppose that the quantity of interest is $Y_3$, the unknown corresponding count in Region III (the target zone), which is comprised of subsections (IIIa and IIIb) of Regions I and II.

One rather crude way to approach the problem is to assume that disease counts are distributed
evenly throughout regions I and II, and so the number of affected individuals in Region III is just

\[ y_1 \left( \frac{area(IIIa)}{area(I)} \right) + y_2 \left( \frac{area(IIIb)}{area(II)} \right) \].

This simple areal interpolation approach is available within many GIS's. However, such estimates are obviously prone to substantial error, since the assumption of homogeneous population distribution throughout the regions may well be unrealistic.

Let us now assume that the entire tract can be partitioned into smaller subsections, where on each subsection we can measure some other variable that is correlated with the disease count for that region. For instance, if we are looking at a particular tract of land, in each subsection we might record whether the land is predominantly rural or urban in character. We do this in the belief that this variable affects the likelihood of disease. Continuous covariates could also be used (say, the median household income in the subsection). Note that the subsections could arise simply as a refinement of the original scale of aggregation (e.g., if disease counts were available only by
census tract, but covariate information arose at the census block group level), or as the result of overlaying a completely new set of boundaries (say, a zip code map) onto our original map. The statistical model is easier to formulate in the former case, but the latter case is of course more general, and is the one motivated by modern GIS technology.

For simplicity, assume for the moment that subsections have equal areas and that the covariate in question is binary. Figure 2 shows one way the region in Figure 1 might be subdivided. Here we let the disease count in each subsection be Poisson-distributed with parameter $m_1$ or $m_2$, depending on which value (1 or 2) the binary measurement assumes. We further suppose that these Poisson variables are independent given the covariate, so that the disease count in any conglomeration of subsections is again Poisson-distributed with parameter equal to the sum of the parameters of the comprising subsections. (For the case of unequal areas we might instead take the weighted sum of the sub-parameters, with subregional areas as the weights.) From Figure 2 we see that $Y_1$, the population of Region I, is $Poisson(7m_1 + 5m_2)$, while $Y_2$, the population of Region II, is $Poisson(6m_1 + 2m_2)$, independently of $Y_1$ given $m_1$ and $m_2$. Following standard statistical practice
we denote the observed values of these random variables as \( y_1 \) and \( y_2 \); for the purpose of illustration we set \( y_1 = 632 \) and \( y_2 = 311 \).

The independence assumption above may seem odd in a spatial setting; counts in neighboring subregions might naturally be expected to be correlated. But we emphasize that we only assume independence conditionally given the (spatially varying) covariates \( m \equiv (m_1, m_2) \). Our hierarchical Bayesian approach now places a prior distribution on \( m \), so that marginally, the observed counts are spatially correlated as desired. Of course, it might be that not all of the spatial correlation in the data is accounted for by the covariates \( m \); we return to this issue in Section 4.

In the setting of our motivating example, let us propose rather vague and independent prior distributions for \( m_1 \) and \( m_2 \), so that the data values \( (y_1 \text{ and } y_2) \) will drive the posterior distributions. Specifically, we adopt a Gamma\((a, b)\) prior for each parameter with \( a = 0.5 \) and \( b = 100 \), so that the prior has mean 50 (roughly the average observed count per subregion) and variance 5000. The choice of the gamma distribution is convenient because of its positive support, but as we shall see, the usual motivation for pairing it with the Poisson likelihood (prior-posterior conjugacy) does not apply in our computational setting.

Denoting the likelihood by \( L \), the prior by \( p \), and writing \( y = (y_1, y_2) \), the joint posterior distribution of \( m_1 \) and \( m_2 \) is given as

\[
p(m_1, m_2 | y) \propto L(m_1, m_2; y) p(m_1, m_2)
\]

\[
\propto (7m_1 + 5m_2)^{y_1} e^{-(7m_1 + 5m_2)} (6m_1 + 2m_2)^{y_2} e^{-(6m_1 + 2m_2)} m_1^{-1} e^{-m_1/b} m_2^{-1} e^{-m_2/b} ,
\]

so that the resulting full conditional distributions for \( m_1 \) and \( m_2 \) are

\[
p(m_1 | m_2, y) \propto (7m_1 + 5m_2)^{y_1} (6m_1 + 2m_2)^{y_2} m_1^{-1} e^{-m_1(13 + b^{-1})} ,
\] (2)
and \( p(m_2|m_1, y) \propto (7m_1 + 5m_2)^{y_1}(6m_1 + 2m_2)^{y_2}m_2^{n-1}e^{-m_2(7+y_1)} \). \hspace{1cm} (3)

We see immediately that conjugacy is absent; these two expressions are not proportional to any standard distributional form. As such, we use univariate Metropolis updating (Metropolis et al., 1953; Chib and Greenberg, 1995) to obtain samples from the joint posterior distribution \( p(m_1, m_2|y) \), which can then be summarized for inferential purposes. To improve the numerical stability of this algorithm and also work with parameters having support equal to the whole real line, we transform to the log scale. That is, we reparametrize to \( \delta_1 = \log(m_1) \) and \( \delta_2 = \log(m_2) \), remembering to multiply by the Jacobian \( \exp(\delta_i), i = 1, 2 \) for each transformation. (Failing to include these Jacobians would amount to specifying different, and perhaps less noninformative, priors for the \( \delta_i \).)

Drawing our Metropolis candidates from Gaussian distributions with means equal to the current chain value and variances \((0.3)^2\) and \((0.1)^2\) for \( \delta_1 \) and \( \delta_2 \), respectively, for each parameter we ran five independent sampling chains with starting points overdispersed with respect to the suspected target distribution for 2000 iterations. The observed Metropolis acceptance rates were 45.4\% and 46.4\%, respectively, near the 50\% rate suggested by Gelman et al. (1996) as well as years of Metropolis “folklore”. The vagueness of the prior distributions coupled with the paucity of the data in this simple example (in which we are estimating two parameters from just two data points, \( y_1 \) and \( y_2 \)) leads to substantial autocorrelation in the observed chains. However, plots of the observed chains as well as the convergence diagnostic of Gelman and Rubin (1992) suggested that a suitable degree of algorithm convergence obtains after 500 iterations. The histograms of the remaining \( 5 \times 1500 = 7500 \) iterations shown in Figures 3(a) and (b) provide estimates of the marginal posterior distributions \( p(m_1|y) \) and \( p(m_2|y) \). We see that point estimates for \( m_1 \) and \( m_2 \) are 18.5 and 100.4, respectively, implying best guesses for \( 7m_1 + 5m_2 \) and \( 6m_1 + 2m_2 \) of 631.5 and 311.8, respectively — quite consistent with the observed data values \( y_1 = 632 \) and \( y_2 = 311 \). Also shown are 95\% Bayesian
credible intervals (denoted “95% BCI” in the figure legends), available simply as the 2.5 and 97.5 empirical percentiles in the ordered samples.

Our real interest lies in \( p(y_3 | y) \), the posterior distribution of \( Y_3 \). Consider the mean of this distribution, \( E(Y_3 | y) = E(Y_3 | m_1) + E(Y_3 | m_2) \). Beginning with the first term in this sum, by the Law of Iterated Expectation we may write \( E(Y_3 | m_1) = E[E(Y_3 | m_1, y)] \). Now we need the following well-known result from distribution theory:

**Lemma:** If \( X_1 \sim Po(\lambda_1), \ X_2 \sim Po(\lambda_2) \), and \( X_1 \) and \( X_2 \) are independent, then

\[
X_1 \mid (X_1 + X_2 = n) \sim Bin \left( n, \frac{\lambda_1}{\lambda_1 + \lambda_2} \right). \quad \blacksquare
\]

We apply this lemma in our setting with \( Y_3 \) playing the role of \( X_1 \), \( y_1 \) playing the role of \( n \), and the calculation conditional on \( m \). The result is

\[
E(Y_3 | y) = E[E(Y_3 | m_1, y)] = E[E(Y_3 | m_1, y_1)]
\]
\[
= E \left[ y_1 \left( \frac{2m_1 + 2m_2}{7m_1 + 5m_2} \right) \mid y_1 \right]
\]
where \{(m_1^{(g)}, m_2^{(g)}), g = 1, \ldots, G\} are the Metropolis samples drawn above. A similar calculation produces a Monte Carlo estimate of \(E(Y_{3b}|y)\), so that our final estimate of \(E(Y_3|y)\) is the sum of these two quantities. In our problem this turns out to be \(\hat{E}(Y_3|y) = 357.0\).

An expression for \(Var(Y_3|y)\) can be obtained using the Law of Iterated Variance, \(Var(Y_3|y) = E[Var(Y_3|m, y)] + Var[E(Y_3|m, y)]\), though this is substantially more complicated, as it is not in general true that \(Var(Y_3|y) = Var(Y_{3a}|y) + Var(Y_{3b}|y)\). Instead, we proceed on to obtaining an estimate of the entire posterior distribution \(p(y_3|m, y)\). Again using Monte Carlo integration, we write

\[
p(y_3|m, y) = \int p(y_3|m, y) p(m|y) dm \approx \frac{1}{G} \sum_{g=1}^{G} p(y_3|m_1^{(g)}, y_2^{(g)}, y).
\]

Using the lemma again, \(p(y_3|m, y)\) is the convolution of two independent binomials,

\[
Y_{3b}|m, y \sim Bin \left( y_1 \cdot \frac{2m_1 + 2m_2}{7m_1 + 5m_2} \right), \quad (5)
\]
and \( Y_{3a} | \mathbf{m}, y \sim Bin \left( y_2, \frac{m_1 + m_2}{6m_1 + 2m_2} \right) \). \hspace{1cm} (6)

Since these two binomials do not have equal success probabilities, this convolution is a complicated (though straightforward) calculation which unfortunately will not emerge as another binomial distribution. However, we may perform the sampling analog of this calculation simply by drawing \( Y_{3a}^{(g)} \) from \( p(y_{3a} | \mathbf{m}^{(g)}, y_1) \) in (5), \( Y_{3g}^{(g)} \) from \( p(y_{3g} | \mathbf{m}^{(g)}, y_2) \) in (6), and defining \( Y_3^{(g)} = Y_{3a}^{(g)} + Y_{3g}^{(g)} \). The resulting pairs \( \{(Y_3^{(g)}, \mathbf{m}^{(g)}), g = 1, \ldots, G\} \) are distributed according to the joint posterior distribution \( p(y_3, \mathbf{m} | y) \), so that marginally, the \( \{Y_3^{(g)}, g = 1, \ldots, G\} \) values have the desired distribution, \( p(y_3 | y) \).

In our setting, we actually drew 25 \( Y_{3a}^{(g)} \) and \( Y_{3g}^{(g)} \) samples for each \( \mathbf{m}^{(g)} \) value, resulting in 25(7500) = 187,500 \( Y_3^{(g)} \) draws from the convolution distribution. A histogram of these values (and a corresponding kernel density estimate) is shown in Figure 4. The mean of these samples is 357.2, which agrees quite well with our earlier mean estimate of 357.0 calculated just below equation (4).

Note that even when \( m_1 = m_2 \) (so that the naive allocation formula (1) produces an acceptable point estimate for \( Y_3 \)) our method still offers the advantages of accurate variance, quantile, and full posterior distributional estimates. In our case, from Figure 4 we obtain estimated 2.5% and 97.5% quantiles of 314 and 400, respectively. We remark that in general \( p(y_3 | y) \) can be bimodal (e.g., if the success probabilities in (5) and (6) are markedly different). While Figure 4 reveals that this is not the case for our data set, such bimodality, obscured by (1), would be immediately apparent in our sampling-based implementation.

We conclude this section by comparing our results to those obtained via the traditional frequentist method. Following the review of the approach for Poisson data given in Flowerdew and Green (1994), the EM algorithm alternates between computing expected subregional counts by allocating the regional totals \( y_1 \) and \( y_2 \) to the atoms in proportion to the current estimates of \( m_1 \) and \( m_2 \) (the
E step), and obtaining updated estimates \( \hat{m}_1 \) and \( \hat{m}_2 \) by fitting a Poisson regression model to the (now complete) subregional level data (the M step). As recommended by Flowerdew and Green (1994, p.127), we used the areal interpolation method to obtain starting subregional data values. The resulting EM algorithm converged very slowly, requiring 214 iterations to produce one decimal place accuracy (296 iterations for 2 good decimal places; 369 iterations for 3). The final (modal) estimates are \( \hat{m}_1 = 18.2, \hat{m}_2 = 100.9 \), which agree very closely with the Bayesian (mean) estimates (18.5, 100.4) obtained earlier. Interestingly, in this simple problem the MLE obtained via EM can be obtained virtually immediately simply by solving the system of linear equations

\[
7m_1 + 5m_2 = y_1 = 632
\]

\[
6m_1 + 2m_2 = y_2 = 311
\]

Of course, had there been more than two regional totals available to us, such a simple algebraic solution would have been unavailable.

### 3 Data example: Interpolating leukemia case counts

We turn now to a more involved application of the method using a data set originally presented and analyzed by Waller et al. (1994); namely, the incidence of leukemia in Tompkins County, New York. As seen in Figure 5, this county, located in west-central New York state, is roughly centered around the city of Ithaca, NY. Tompkins County is divided into 23 census tracts, with each tract further subdivided into between 1 and 5 block groups, for a total of 51 such subregions. We have leukemia counts available at the tract level, and we wish to predict them at the block group level with the help of population counts and covariate information available on this more refined scale. In this illustration, the two covariates we consider are whether the block group is coded as “rural” or “urban,” and whether or not the block group centroid is located within 2 kilometers of a hazardous
chemical waste site. (For this data set, we are actually unusually fortunate to have leukemia counts at the block group level, but we use only the tract totals in the model fitting process, reserving the refined information to check the accuracy of our results.) In this example, the unequal population totals in the block groups will play the weighting role that unequal areas would have played in Section 2.

Figure 6 shows a rough schematic of Tompkins County, wherein the horizontal and vertical coordinates represent kilometers from a central New York reference point used by Waller et al. (1994). Small circles and plus signs denote the population-weighted centroids of each block group, where a circle represents a rural block and a plus represents an urban block. Here “population weighted” simply means that each centroid was computed as the average of the spatial coordinates of the home addresses of every person residing in the block group. The map also shows two waste sites: one in the northeast corner of the county, and one in downtown Ithaca, near the center. Two-kilometer circles are shown around each waste site, indicating which block group centroids fall within 2 km of the site.
Figure 6: Schematic map of block group centroids, Tompkins County, NY. Horizontal and vertical coordinates are in kilometers from a central New York reference point. Here, “+” denotes an urban block group, “o” denotes a rural block group, and “□” denotes a waste site.

Figure 7 shows a census tract-level disease map produced by the GIS MapInfo. Table 1 (located on page 28) gives the block group-level population counts $n_{ij}$ and covariate values $u_{ij}$ and $w_{ij}$, where $u_{ij}$ is 1 if block group $j$ of census tract $i$ is classified as urban, 0 if rural, and $w_{ij}$ is 1 if the block group centroid is within 2 km of a waste site, 0 if not. The horizontal and vertical spatial coordinates of each block group are also given so that each can be identified in Figure 6 and subsequent schematic maps. Typical of GIS software, MapInfo permits allocation of the census tract totals to the various block groups proportional to block group area or population. We use our hierarchical Bayesian method to incorporate the covariate information, as well as to obtain variance estimates to accompany the block group-level point estimates.

Proceeding in a manner similar to that used in Section 2, we assume that

$$Y_{ij} | m_{k(i,j)} \overset{i.d.}{\sim} P \alpha (E_{ij} | m_{k(i,j)}), \ i = 1, \ldots, I, \ j = 1, \ldots, J_i,$$  

(7)
where $I = 23$, $J_i$ varies from 1 to 5, $Y_{ij}$ is the disease count in block group $j$ of census tract $i$, and $E_{ij}$ is the corresponding “expected” disease count, computed as $E_{ij} = n_{ij} \lambda$ where $n_{ij}$ is the population count in the cell and $\lambda$ is the overall probability of contracting the disease. This “background” probability could be estimated from our data; here we take $\lambda = 5.597 \times 10^{-4}$, the crude leukemia rate for the 8-county region studied by Waller et al. (1994), an area which includes Tompkins County. Hence, $m_{k(i,j)}$ is the relative risk of contracting leukemia in block group $(i, j)$, and $k = k(i, j) = 1, 2, 3, \text{ or } 4$ depending on the covariate status of the block group. Specifically, we let

$$k(i, j)= \begin{cases} 1, & \text{if } (i, j) \text{ is rural, not near a waste site} \\ 2, & \text{if } (i, j) \text{ is urban, not near a waste site} \\ 3, & \text{if } (i, j) \text{ is rural, near a waste site} \\ 4, & \text{if } (i, j) \text{ is urban, near a waste site} \end{cases}$$

Defining $\mathbf{m} = (m_1, m_2, m_3, m_4)$ and again adopting independent and minimally informative gamma priors for these four parameters, we seek estimates of $p(\mathbf{m}_k | \mathbf{y})$, where $\mathbf{y} = (y_1, \ldots, y_L)$, and $y_i =$
\[ \sum_{j=1}^{J_i} y_{ij}, \] the census tract disease count totals. We also wish to obtain block group-specific mean and variance estimates \( E[Y_{ij}|\mathbf{y}] \) and \( \text{Var}[Y_{ij}|\mathbf{y}] \), to be plotted in a disease map at the block group (rather than census tract) level. Finally, we may also wish to estimate the distribution of the total disease count in some conglomeration of block groups (say, corresponding to some village or city), analogous to our estimate of \( p(y_3|\mathbf{y}) \) in Section 2.

By the conditional independence of the block group counts we have \( Y_i|\mathbf{m} \sim i.d. \) \( Poisson(\sum_{k=1}^{4} s_k m_k) \), \( i = 1, \ldots, I \), where \( s_k = \sum_{j:k(i,j)=k} E_{ij} \), the sum of the expected cases in block groups \( j \) of region \( i \) corresponding to covariate pattern \( k \), \( k = 1, \ldots, 4 \). The likelihood \( L(\mathbf{m} : \mathbf{y}) \) is then the product of the resulting \( I = 23 \) Poisson kernels. After multiplying this by the prior distribution term \( \prod_{k=1}^{4} p(m_k) \), we can again obtain forms proportional to the four full conditional distributions \( p(m_k|m_{i\neq k}, \mathbf{y}) \) as in (2) and (3), and sample these sequentially via univariate Metropolis steps.

Once again it is helpful to reparameterize to \( \delta_k = \log(m_k), \ k = 1, \ldots, 4 \), and perform the Metropolis sampling on the log scale. We specify reasonably vague \( Gamma(a, b) \) priors for the \( m_k \) by taking \( a = 2 \) and \( b = 10 \) (similar results were obtained with even less informative Gamma priors unless \( a \) was quite close to 0, in which case convergence was unacceptably poor). For this “base prior,” convergence obtains after 200 iterations, and the remaining 1800 iterations in 5 parallel MCMC chains are retained as posterior samples from \( p(\mathbf{m}|\mathbf{y}) \).

A second reparametrization aids in interpreting our results. Suppose we write

\[ \delta_{k(i,j)} = \theta_0 + \theta_1 u_{ij} + \theta_2 w_{ij} + \theta_3 u_{ij} w_{ij}, \] (8)

so that \( \theta_0 \) is an intercept, \( \theta_1 \) is the effect of living in an urban area, \( \theta_2 \) is the effect of living near a waste site, and \( \theta_3 \) is the urban/waste site interaction. This reparametrization expresses the log-relative risk of disease as a linear model, a common approach in spatial disease mapping (Besag

15
et al., 1991; Waller et al., 1997). A simple 1-1 transformation converts our \((m_1^{(g)}, m_2^{(g)}, m_3^{(g)}, m_4^{(g)})\) samples to \((\theta_0^{(g)}, \theta_1^{(g)}, \theta_2^{(g)}, \theta_3^{(g)})\) samples on the new scale, which in turn allows direct investigation of the main effects of urban area and waste site proximity, as well as the effect of interaction between these two. Figure 8 shows the histograms of the posterior samples for \(\theta_i, i = 0, \ldots, 3\). We note that \(\theta_0, \theta_1, \) and \(\theta_3\) are not significantly different from 0 as judged by the 95\% BCI, while \(\theta_2\) is “marginally significant” (in a Bayesian sense) at this level. This suggests a moderately harmful effect of residing within 2 km of a waste site, but no effect of merely residing in an urban area (in this case, the city of Ithaca). The preponderance of negative \(\theta_3^{(g)}\) samples is somewhat surprising; we might have expected living near an urban waste site to be associated with an increased (rather than decreased) risk of leukemia. This is apparently the result of the high leukemia rate in a few rural block groups not near a waste site (block groups 1 and 2 of tract 7, and block group 2 of tract 20), forcing \(\theta_3\) to adjust for the relatively lower overall rate near the Ithaca waste site.
3.1 Individual block group estimation

To create the block group-level estimated disease map, for those census tracts having \( J_i > 1 \) we follow equation (4), obtaining a conditionally binomial distribution for \( Y_{ij} \), so that

\[
E(Y_{ij}|y) = E[E(Y_{ij}|m_i,y)] = \frac{y_i}{G} \sum_{g=1}^{G} p_{ij}^{(g)},
\]

where \( p_{ij} \) is the appropriate binomial probability found using our Section 2 lemma. For example, for \( p_{11}^{(g)} \) we have

\[
p_{11}^{(g)} = \frac{1617m_1^{(g)}}{(1617 + 702)m_1^{(g)} + (1526 + 1368)m_3^{(g)}},
\]

as determined by the covariate patterns in the first four rows of Table 1. Note that when \( J_i = 1 \) the block group total equals the known census tract total, hence no estimation is necessary.

The resulting collection of estimated block group means \( E(Y_{ij}|y) \) are shown in Table 1, along with the actual case counts \( y_{ij} \). (The occasional noninteger values of \( y_{ij} \) in the table are not errors, but arise from a few cases in which the precise block group of occurrence is unknown, resulting in fractional counts being allocated to several block groups.) Note that, like other interpolation methods, the sum of the estimated cases in each census tract is the same as the corresponding sum for the actual case counts. Additional comparison of these two columns is facilitated by the graphical displays in Figure 9(a) and (b). This figure reveals that our method performed well, especially in Ithaca and the village of Groton (the area in the northeast corner near the rural waste site). Only the surprisingly high disease count in block group #2 in the town of Caroline (southeast corner) is not anticipated by our model, which instead spreads the total over the four block groups in the Caroline-Danby tract.

“Bubble plots” of the type shown in Figure 9 are effective ways of displaying spatial information,
Figure 9: Schematic maps of interpolated and actual block group disease counts, Tompkins County. Horizontal and vertical coordinates are in kilometers from a central New York reference point. A circle with diameter proportional to the corresponding disease count is plotted at each block group centroid.

and are easily created in S-Plus. However, they omit subregional boundaries, and represent each subregion’s value by a symbol plotted at its centroid. A more complete picture of this same information can be seen in the GIS plots shown in Figure 10. While the GIS used to create these plots, ARC/INFO, is difficult to learn, in this case at least the effort seems to have been well-spent. For instance, the precise locations and natures of the two pockets of elevated disease counts (in the villages of Cayuga Heights and Groton) are much more easily seen in Figure 10(a) than Figure 9(a).

For a precise quantification of the quality of fit of our model, we might use the sum of absolute prediction errors,

$$S\text{AE} = \sum_{i=1}^{I} \sum_{j=1}^{J_i} |y_{ij} - \hat{y}_{ij}|,$$

where for our Bayesian method we take $\hat{y}_{ij} = E(Y_{ij} | y)$. The resulting SAE value is 22.46. This compares favorably with the SAE’s obtained using $\hat{y}_{ij}$ values produced by ARC/INFO using ei-
ther population-based (SAE = 22.66) or area-based (SAE = 25.79) interpolation, the only two approaches currently feasible within the ARCGIS/INFO package. The similarity of the SAE values for the Bayes and population-based interpolation methods arises because the two methods produce different \( \hat{y}_{kj} \) values only for the four census tracts (numbers 1, 6, 9, and 10) in which the component block groups do not all share the same covariate values.

To get an idea of the variability inherent in the posterior surface, we now consider mapping the estimated posterior variances of our interpolated counts. Since the block group-level vari-
Figure 11: Schematic map of estimated variability in the interpolated block group disease counts, Tompkins County. Horizontal and vertical coordinates are in kilometers from a central New York reference point. A circle with diameter proportional to the estimated standard deviation (i.e., with area proportional to the estimated variance) of the corresponding disease count is plotted at each block group centroid.

\[ \text{Var}(Y_{ij} | \mathbf{y}) = E(Y_{ij}^2 | \mathbf{y}) - [E(Y_{ij} | \mathbf{y})]^2, \]

where \( E(Y_{ij}^2 | \mathbf{y}) \) are the estimated means (already calculated), and

\[
E(Y_{ij}^2 | \mathbf{y}) = E[E(Y_{ij}^2 | \mathbf{m}, \mathbf{y})] = E[y_i p_{ij}(1 - p_{ij}) + y_i^2 (p_{ij})^2]
\approx \frac{1}{G} \sum_{g=1}^{G} \left[ y_i p_{ij}^{(g)} (1 - p_{ij}^{(g)}) + y_i^2 (p_{ij}^{(g)})^2 \right],
\]

where \( p_{ij} \) is again the appropriate binomial probability for block group \((i, j)\). Substituting in our Monte Carlo samples \( \mathbf{m}^{(g)} \), Figure 11 offers a bubble plot of the resulting block group variability estimates. Note again that since block group totals are known exactly in census tracts where \( J_i = 1 \), the corresponding estimated variability is 0; these block groups are indicated as the tiniest circles in the figure. Variability is generally higher in census tracts having larger numbers of block groups or
larger total counts – an outcome of our Poisson model specification, for which the variance increases with the mean.

We remark that most of the census tracts are composed of homogeneous block groups (e.g., all rural with no waste site nearby); in these instances the resulting binomial probability for each block group is free of \( m \). In such cases, posterior means and variances are readily available without any need for mixing over the Metropolis samples, as in equations (9) and (10).

### 3.2 Aggregate estimation: Block groups near the Ithaca waste site

In order to assess the number of leukemia cases we expect in those block groups within 2 km of the Ithaca waste site, analogously to the approach of Figure 4 we can sample binomial distributions that correspond to the various regions within the 2-km radii, sum the results, and draw a histogram of these sums. From Table 1, we see that twelve block groups in five census tracts fall within these 2-km radii: all of the block groups in census tracts 11, 12, and 13, plus two of the three (block groups 2 and 3) in tract 6 and three of the four (block groups 2, 3, and 4) in tract 10. Since the totals in census tracts 11, 12, and 13 are known to our analysis, we need only sample from two binomial distributions, one each for the conglomerations of near-waste site block groups within tracts 6 and 10. Defining the sum over the twelve block groups as \( Z \), we have

\[
Z^{(s)} = Y_{6,(2,3)}^{(s)} + Y_{10,(2,3,4)}^{(s)} + y_{11.} + y_{12.} + y_{13.}.
\]

A histogram of these values is shown in Figure 12. The estimated median value of 10 happens to be exactly equal to the true value of 10 cases in this area. The sample mean, 9.43, is also an excellent estimate. Note that the minimum and maximum values in Figure 12, \( Z = 7 \) and \( Z = 11 \), are imposed by the data structure: there must be at least as many cases as the total known to have occurred in census tracts 11, 12, and 13 (which is 7), and there can be no more than the total
number known to have occurred in tracts 6, 10, 11, 12, and 13 (which is 11).

Finally, we may again compare our results to those produced by ARC/INFO under either area-based or population-based interpolation. The former produces a mean estimate of 9.28, while the latter gives 9.59. These are close to the Bayesian mean 9.43, but neither approach produces an associated confidence interval, much less a full graphical display of the sort given in Figure 12. The Flowerdew and Green (1994) EM approach could be extended (e.g., in the manner of Louis, 1982) to produce variance estimates associated with the imputed $\hat{y}_{ij}$ values, but since these estimates are clearly not statistically independent it is not clear how they should be combined to obtain a corresponding variance estimate for $Z$. 
4 Discussion and future directions

The method presented in this paper extends that of Flowerdew and Green (1989), who approached the problem of population interpolation from a classical perspective. Our hierarchical Bayesian method appears to offer several advantages. First, we obtain not just point estimates for target zone populations, but also estimates of the entire population distribution. One might argue that even in Flowerdew and Green (1989) the target population distribution ends up being known (i.e., Poisson with an estimated rate parameter). However, our method is more comprehensive in that it does not rely on point estimates for the rate parameters \( \mathbf{m} \), but instead simulates the entire predictive distribution of \( Y \) by conditioning on a MCMC sample from the posterior distribution of these rate parameters.

A second advantage of our method is its flexibility in the use of prior information. While we chose fairly vague priors for \( \mathbf{m} \), in many situations arising in the study of environmental justice or epidemiology there will be reliable data-based prior information available that could be incorporated into these priors. Alternatively, recalling equation (8), the prior could be specified on the \( \theta \) scale if this were more convenient or more appropriate. In the case of Tompkins County dataset, however, placing the prior and performing the sampling on the \( \theta \) scale had a negative effect on our results. Specifically, assuming the \( \theta_i \)'s to be Normal(\( \mu = 3, \sigma^2 = 1 \)) a priori (a distribution roughly equivalent to the log of our original \( G(2,10) \) prior for the \( m_i \)), we obtained chains with slightly higher sample autocorrelations and Gelman and Rubin convergence diagnostics than when running on the \( \mathbf{m} \) scale. Moreover, the questionable assumption of prior independence of the \( \theta_i \)'s (which correspond to regression effects, instead of the \( m_i \)'s, which correspond to covariate-induced population subgroups) leads to confounding which eliminates the “marginal significance” of the waste site effect \( \theta_2 \) in the fitted posterior. As a result, we did not pursue this option further in the
An alternative to our Metropolis sampling implementation would be a data augmentation approach, wherein besides sampling over the parameter space we also sample missing data values \( \{y_{ij}, i = 1, \ldots, I, j = 1, \ldots, J_i\} \). Under our Poisson-gamma model, this restores conjugacy between likelihood and prior, with the full conditional distributions for the \( m_k \) and \( y_i = (y_{i1}, \ldots, y_{iJ_i})' \) emerging as gamma and multinomial, respectively. The MCMC algorithm could therefore be implemented using a series of Gibbs steps, rather than Metropolis steps. While the associated algorithm would be straightforward to program and require no tuning, the substantial increase in the parameter burden would likely slow convergence, due to the greater amount of generation required and the loss of the beneficial effect of \textit{collapsing} offered by our lower-dimensional algorithm (Liu, 1994).

Our method allows for more than just a single dichotomous choice of either \( m_1 \) or \( m_2 \) as the rate parameter. There is no inherent limit to the number of covariates we might use; the choice is limited only by the application and the data at hand. In fact, in settings where the independent Poisson likelihood assumption is not appropriate (e.g., where the data are dependent and/or continuous), our hierarchical Bayesian MCMC approach may well be the only feasible alternative.

The Tompkins County data were not “inherently misaligned;” the misalignment arose by declaring a target zone (e.g., a 2-km radius of a waste site) that was not aligned with census tract boundaries. The choice of block groups as a mutual refinement of both zoning systems (source and target) was both natural and convenient, creating a “nested” setting in which covariates at the refined level were readily available from census data. The location of the population-weighted centroid was a convenient way of determining the exposure covariate. But there are certain weaknesses with using centroids, and so there may well be situations where such an approach is undesirable. Moreover, in many data settings the covariate information may be collected over a completely separate geographical grid (say, zip codes instead of census tracts), creating a “nonnested” model and
a correspondingly greater challenge in managing the data within the GIS, performing the MCMC sampling, and displaying the final results. Nevertheless, our hierarchical modeling approach is still applicable, and pays the same dividends as in the nested case illustrated above. Indeed, our approach is useful in any environmental or biometric setting where the researcher must make inferences from incomplete data collected over incompatible zones.

Finally, a logical generalization of our method is its use in spatial smoothing of the underlying log-relative risks $\hat{h}_{ij}$, instead of imputing missing data values $Y_{ij}$ (Clayton and Kaldor, 1987). This could involve simply plotting the fitted values from (8), but under our current model and binary covariate structure this would produce a map with only four levels of shading. Instead, we might augment our log-relative risk model (8) with a spatial smoothing term $\phi_{ij}$, assigned a spatially autoregressive (SAR) or conditionally autoregressive (CAR) prior distribution (Besag, 1974). The smoothing could even be done over time, leading into a method for handling spatio-temporally misaligned data. In any of these high-dimensional model settings, MCMC methods again offer the only feasible computational approach. We hope to investigate these and other related issues in a future article.

Acknowledgements

The research of both authors was supported in part by National Institute of Environmental Health Sciences (NIEHS) Grant 1-R01-ES07750. In addition, the research of the second author was also supported in part by National Institute of Allergy and Infectious Diseases (NIAID) FIRST Award 1-R29-AI33466. The contents of this paper are solely the responsibility of the authors and do not necessarily represent the official views of the NIEHS, NIAID, or NIH. The authors are grateful to Prof. Lance Waller for providing the Tompkins County data set, as well as substantial related
analytic advice, to Ms. Erin Conlon and Ms. Li Zhu for invaluable assistance with GIS plotting and database issues, and to the associate editor and a referee for helpful advice which led to substantial improvements in the presentation.

References


<table>
<thead>
<tr>
<th>census tract (i)</th>
<th>block group (j)</th>
<th>( n_{ij} )</th>
<th>( w_{ij} )</th>
<th>( w_{ij} )</th>
<th>( y_{ij} )</th>
<th>location name</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>1617</td>
<td>-24.9</td>
<td>-13.9</td>
<td>0</td>
<td>0.5826     0.3</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>1526</td>
<td>-30.0</td>
<td>-13.6</td>
<td>0</td>
<td>1.668      0.8</td>
</tr>
<tr>
<td>1</td>
<td>3</td>
<td>1368</td>
<td>-32.1</td>
<td>-13.8</td>
<td>0</td>
<td>1.4959     1.8</td>
</tr>
<tr>
<td>1</td>
<td>4</td>
<td>702</td>
<td>-36.1</td>
<td>-12.6</td>
<td>0</td>
<td>0.2529     1.1</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>1546</td>
<td>-39.8</td>
<td>-16.1</td>
<td>0</td>
<td>0.3946     0.0</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>962</td>
<td>-43.9</td>
<td>-15.7</td>
<td>0</td>
<td>0.2457     1.0</td>
</tr>
<tr>
<td>2</td>
<td>3</td>
<td>1408</td>
<td>-48.7</td>
<td>-13.8</td>
<td>0</td>
<td>0.3596     0.0</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>887</td>
<td>-40.5</td>
<td>-22.9</td>
<td>0</td>
<td>0          0.0</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>475</td>
<td>-43.5</td>
<td>-26.2</td>
<td>0</td>
<td>0          0.0</td>
</tr>
<tr>
<td>5</td>
<td>1</td>
<td>1335</td>
<td>-48.3</td>
<td>-24.4</td>
<td>0</td>
<td>0.3792     0.0</td>
</tr>
<tr>
<td>5</td>
<td>2</td>
<td>2065</td>
<td>-55.0</td>
<td>-19.1</td>
<td>0</td>
<td>0.5866     1.0</td>
</tr>
<tr>
<td>5</td>
<td>3</td>
<td>1269</td>
<td>-50.1</td>
<td>-24.6</td>
<td>0</td>
<td>0.3605     0.0</td>
</tr>
<tr>
<td>5</td>
<td>4</td>
<td>782</td>
<td>-55.1</td>
<td>-30.1</td>
<td>0</td>
<td>0.2221     0.0</td>
</tr>
<tr>
<td>5</td>
<td>5</td>
<td>1590</td>
<td>-51.1</td>
<td>-29.1</td>
<td>0</td>
<td>0.4516     1.0</td>
</tr>
<tr>
<td>6</td>
<td>1</td>
<td>1343</td>
<td>-46.1</td>
<td>-27.4</td>
<td>1</td>
<td>0.4257     1.0</td>
</tr>
<tr>
<td>6</td>
<td>2</td>
<td>767</td>
<td>-42.2</td>
<td>-29.2</td>
<td>1</td>
<td>0.2665     0.0</td>
</tr>
<tr>
<td>6</td>
<td>3</td>
<td>1366</td>
<td>-42.4</td>
<td>-29.6</td>
<td>1</td>
<td>0.3678     0.0</td>
</tr>
<tr>
<td>7</td>
<td>1</td>
<td>2010</td>
<td>-46.7</td>
<td>-26.9</td>
<td>0</td>
<td>3.5429     3.0</td>
</tr>
<tr>
<td>7</td>
<td>2</td>
<td>1394</td>
<td>-46.4</td>
<td>-28.3</td>
<td>0</td>
<td>2.4571     3.0</td>
</tr>
<tr>
<td>8</td>
<td>1</td>
<td>2630</td>
<td>-38.6</td>
<td>-26.8</td>
<td>0</td>
<td>3           3.0</td>
</tr>
<tr>
<td>9</td>
<td>1</td>
<td>1678</td>
<td>-39.1</td>
<td>-28.3</td>
<td>1</td>
<td>0.6192     0.0</td>
</tr>
<tr>
<td>9</td>
<td>2</td>
<td>699</td>
<td>-39.5</td>
<td>-30.0</td>
<td>0</td>
<td>0.3808     1.0</td>
</tr>
<tr>
<td>10</td>
<td>1</td>
<td>4436</td>
<td>-41.0</td>
<td>-28.7</td>
<td>1</td>
<td>1.1537     0.0</td>
</tr>
<tr>
<td>10</td>
<td>2</td>
<td>1856</td>
<td>-41.9</td>
<td>-29.3</td>
<td>1</td>
<td>0.4018     2.0</td>
</tr>
<tr>
<td>10</td>
<td>3</td>
<td>4250</td>
<td>-40.9</td>
<td>-29.5</td>
<td>1</td>
<td>0.92       1.0</td>
</tr>
<tr>
<td>10</td>
<td>4</td>
<td>2473</td>
<td>-41.7</td>
<td>-30.1</td>
<td>1</td>
<td>0.5354     0.0</td>
</tr>
<tr>
<td>11</td>
<td>1</td>
<td>3075</td>
<td>-40.7</td>
<td>-30.4</td>
<td>1</td>
<td>2.1168     3.0</td>
</tr>
<tr>
<td>11</td>
<td>2</td>
<td>1283</td>
<td>-42.3</td>
<td>-30.8</td>
<td>1</td>
<td>0.8832     0.0</td>
</tr>
<tr>
<td>12</td>
<td>1</td>
<td>2687</td>
<td>-41.6</td>
<td>-31.1</td>
<td>1</td>
<td>2.0776     2.0</td>
</tr>
<tr>
<td>12</td>
<td>2</td>
<td>1193</td>
<td>-42.6</td>
<td>-31.3</td>
<td>1</td>
<td>0.9224     1.0</td>
</tr>
<tr>
<td>13</td>
<td>1</td>
<td>1870</td>
<td>-42.3</td>
<td>-30.2</td>
<td>1</td>
<td>0.3231     0.5</td>
</tr>
<tr>
<td>13</td>
<td>2</td>
<td>1071</td>
<td>-42.7</td>
<td>-30.1</td>
<td>1</td>
<td>0.1902     0.5</td>
</tr>
<tr>
<td>13</td>
<td>3</td>
<td>2689</td>
<td>-43.7</td>
<td>-30.5</td>
<td>1</td>
<td>0.4756     0.0</td>
</tr>
<tr>
<td>14</td>
<td>1</td>
<td>971</td>
<td>-47.5</td>
<td>-30.8</td>
<td>0</td>
<td>0          0.0</td>
</tr>
<tr>
<td>15</td>
<td>1</td>
<td>5613</td>
<td>-40.5</td>
<td>-32.9</td>
<td>0</td>
<td>1          1.0</td>
</tr>
<tr>
<td>16</td>
<td>1</td>
<td>524</td>
<td>-36.4</td>
<td>-33.9</td>
<td>0</td>
<td>0          0.0</td>
</tr>
<tr>
<td>17</td>
<td>1</td>
<td>884</td>
<td>-37.1</td>
<td>-29.9</td>
<td>0</td>
<td>1          1.0</td>
</tr>
<tr>
<td>18</td>
<td>1</td>
<td>3625</td>
<td>-34.1</td>
<td>-27.8</td>
<td>0</td>
<td>0          0.0</td>
</tr>
<tr>
<td>19</td>
<td>1</td>
<td>1527</td>
<td>-33.4</td>
<td>-21.5</td>
<td>0</td>
<td>0.1977     0.0</td>
</tr>
<tr>
<td>19</td>
<td>2</td>
<td>427</td>
<td>-29.0</td>
<td>-22.5</td>
<td>0</td>
<td>0.0553     0.0</td>
</tr>
<tr>
<td>19</td>
<td>3</td>
<td>1509</td>
<td>-25.8</td>
<td>-26.9</td>
<td>0</td>
<td>0.1954     0.0</td>
</tr>
<tr>
<td>19</td>
<td>4</td>
<td>1925</td>
<td>-28.6</td>
<td>-22.6</td>
<td>0</td>
<td>0.2493     1.0</td>
</tr>
<tr>
<td>19</td>
<td>5</td>
<td>2335</td>
<td>-27.1</td>
<td>-28.1</td>
<td>0</td>
<td>0.3623     0.0</td>
</tr>
<tr>
<td>20</td>
<td>1</td>
<td>1242</td>
<td>-26.5</td>
<td>-37.6</td>
<td>0</td>
<td>0.7161     0.0</td>
</tr>
<tr>
<td>20</td>
<td>2</td>
<td>1550</td>
<td>-32.3</td>
<td>-39.2</td>
<td>0</td>
<td>0.8937     3.0</td>
</tr>
<tr>
<td>20</td>
<td>3</td>
<td>1635</td>
<td>-37.7</td>
<td>-38.4</td>
<td>0</td>
<td>0.5968     0.0</td>
</tr>
<tr>
<td>20</td>
<td>4</td>
<td>1376</td>
<td>-42.1</td>
<td>-42.2</td>
<td>0</td>
<td>0.7934     0.0</td>
</tr>
<tr>
<td>21</td>
<td>1</td>
<td>2713</td>
<td>-48.1</td>
<td>-42.0</td>
<td>0</td>
<td>0.6165     0.0</td>
</tr>
<tr>
<td>21</td>
<td>2</td>
<td>1688</td>
<td>-54.2</td>
<td>-42.2</td>
<td>0</td>
<td>0.3835     1.0</td>
</tr>
<tr>
<td>22</td>
<td>1</td>
<td>1610</td>
<td>-42.0</td>
<td>-24.8</td>
<td>0</td>
<td>0          0.0</td>
</tr>
<tr>
<td>23</td>
<td>1</td>
<td>2629</td>
<td>-42.7</td>
<td>-27.3</td>
<td>0</td>
<td>2          2.0</td>
</tr>
</tbody>
</table>

Table 1: Tompkins County data set (see text for definition of symbols)