Fully Model Based Approaches for Spatially Misaligned Data

By Andrew S. Mugglin, Bradley P. Carlin, and Alan E. Gelfand

May 20, 1999
Fully Model Based Approaches forSpatially Misaligned Data

Abstract

In this paper we consider inference using multivariate data that are spatially misaligned, i.e., involving variables (typically counts or rates) which are aggregated over differing sets of regional boundaries. Geographic information systems (GISs) enable the simultaneous display of such data sets, but their current capabilities are essentially only descriptive, not inferential. We describe a hierarchical modeling approach which provides a natural solution to this problem through its ability to sensibly combine information from several sources of data and available prior information. Illustrating in the context of counts, allocation under non-nested regional grids is handled using conditionally independent Poisson-multinomial models. Explanatory covariates and multilevel responses are also easily accommodated, with spatial correlation modeled using a conditionally autoregressive (CAR) prior structure. Methods for dealing with missing values in spatial “edge zones” are also discussed. Like many recent hierarchical Bayesian applications, computing is implemented via a carefully tailored Metropolis-Hastings algorithm. We illustrate our method with a complex data set involving inhalation exposure to radon emanating from a depleted uranium fuel processing plant in southwestern Ohio. Structure counts (obtained from U.S. Geological Survey topographical maps) are used to realign sex- and age group-specific U.S. Census block group population counts onto a 160-cell circular “windrose” centered at the plant.

Key Words: Areal interpolation; Bayesian methods; Environmental risk analysis; Geographic Information System (GIS); Markov chain Monte Carlo.
1 Introduction

Geographically referenced data are becoming more and more readily available, but for reasons of confidentiality, convenience, or feasibility they are often reported in aggregate over regions that partition a parcel of land. Different reporting agencies often use different partitions of the same parcel, however, and related data can thus be “misaligned.” For example, hospital admissions may be reported by zip code, while disease incidence is available at the county level, socioeconomic factors available by federal census tract, and health hazard assessments by regions constructed to be within certain distances or directions of a toxic waste source. Geographic information systems (GISs) provide useful software for the display and database management of such disparate sources of data, but their capabilities are descriptive rather than inferential.

To associate measurements observed on misaligned regions requires estimating the values of variables on regions over which they were not measured. This process is known as areal interpolation. Following the geography literature, we refer to regions on which data are available as “source” zones, and regions for which data are to be imputed as “target” zones.

Areal interpolation is similar to kriging, which is spatial prediction given responses observed at a set of locations. While the literature on kriging is substantial (see e.g. Cressie 1993 for a review), until recently the areal interpolation problem has not been widely addressed in the statistical literature, with the notable exception of Tobler (1979). However, the geography literature has many references to the problem, including the papers by Goodchild and Lam (1980), Lam (1983), Flowerdew and Green (1989, 1991, 1992, 1994), Flowerdew et al. (1991), Langford et al. (1991), Goodchild et al. (1993), and Fisher and Langford (1995).

Of the previous approaches to the problem, simple areal interpolation is the easiest to implement. This is the process by which source values are imputed to target zones proportionally to the areas
of the zones. This method is widely available in many GISs, but it obviously depends on the unrealistic assumption that the source values are uniformly distributed over the entire area in question. Rogers and Killough (1997) employed a modification wherein population is imputed to target zones proportionally to the number of buildings in the zones, where the constant of proportionality could be one of four values, depending on the location of the target zone. Flowerdew and Green (1989) employed an iterative Poisson regression using the EM algorithm to estimate target zone characteristics. In their setting the variable of interest was count data, and they used information in the form of a binary covariate (available on the target zones) to estimate the target count values. They later extended their approach (Flowerdew and Green 1992) to include continuous (typically normally distributed) outcome variables. However, none of these methods reflects a fully inferential approach to the problem of areal interpolation.

Mugglin and Carlin (1998) and Mugglin et al. (1999) present a hierarchical Bayesian method for interpolation and smoothing of Poisson responses with covariates, where the principal advantage is full inference (e.g., enabling interval estimates) for the distributions of target zone populations. In their setting the source zones are U.S. census tracts, while the target zones (and the zones on which covariate data are available) are U.S. census block groups. That is, the target zonation is merely a refinement of the source zonation, a situation they term nested misalignment. Best, Ickstadt and Wolpert (1998) propose a method of Bayesian estimation for Gamma/Poisson random field models, in which the data are misaligned and not nested. In their setting the response regions are small enough that, together with the covariate values, they can be regarded as a marked point process. Unfortunately this is not feasible for many data applications (especially those involving U.S. public health records).

Here, we develop a framework for hierarchical Bayesian interpolation, estimation, and spatial smoothing over non-nested misaligned data grids in settings where the regions are too large to be
regarded as marked point processes. In Section 2 we summarize a dataset collected in response to possible contamination resulting from the former Feed Materials Production Center (FMPC) in southwestern Ohio with the foregoing analytic goals. In Section 3 we develop the theory of our modeling approach in a general framework. Section 4 discusses our Markov chain Monte Carlo approach and several particular challenges that arise in implementing our method for the FMPC data. In Section 5 we set forth the conclusions resulting from our analysis of the FMPC data, and finally in Section 6 we assess our findings and suggest avenues for further research.

2 Motivating Dataset

Risk-based decision making is often used for prioritizing cleanup efforts at U.S. Superfund sites. Often these decisions will be based on estimates of the past, present, and future potential health impacts. These impact assessments usually rely on estimation of the number of outcomes, and the accuracy of these estimates will depend heavily on the ability to estimate the number of individuals at risk. Our motivating dataset is connected with just this sort of risk assessment.

In the years 1951-1988 near the town of Ross in southwestern Ohio, the former Feed Materials Production Center (FMPC) processed uranium for weapons production. Draft results of the Fernald Dosimetry Reconstruction Project, sponsored by the Centers for Disease Control and Prevention (CDC), indicated that during production years the FMPC released radioactive materials (primarily radon and its decay products and, to a lesser extent, uranium and thorium) from the site. Although radioactive liquid wastes were released, the primary exposure to residents of the surrounding community resulted from breathing radon decay products. The potential for increased risk of lung cancer is thus the focus of intense local public interest and ongoing public health studies (see Devine et al., 1998).
Estimating the number of adverse health outcomes in the population (or in subsets thereof) requires estimation of the number of individuals at risk. Population counts, broken down by age and sex, are available from the U.S. Census Bureau according to federal census block groups, while the areas of exposure interest are dictated by both direction and distance from the plant. Rogers and Killough (1997) construct an exposure “windrose,” which consists of 10 concentric circular bands at 1-kilometer radial increments divided into 16 compass sectors (N, NNW, NW, WNW, W, etc.). Through the overlay of such a windrose onto U.S. Geological Survey (USGS) maps, they provide counts of the numbers of “structures” (residential buildings, office buildings, industrial building complexes, warehouses, barns, and garages) within each subdivision (cell) of the windrose.

Figure 1 shows the windrose centered at the FMPC. We assign numbers to the windrose cells, with 1-10 indexing the cells starting at the plant and running due north, then 11-20 running from the plant to the north-northwest, and so on. Structure counts are known for each cell; the hatching pattern in the figure indicates the areal density (structures per square kilometer) in each cell.

Also shown in Figure 1 are the boundaries of 39 Census Bureau block groups, for which 1990 population counts are known. These are the source zones for our interpolation problem. Shading intensity indicates the population density (persons per square kilometer) for each block group. The intersection of the two (non-nested) zonation systems results in 389 regions we call atoms, which can be aggregated appropriately to form either cells or block groups.

The plant was in operation for 38 years, raising concern about the potential health risks it has caused—a question that has been under active investigation by the CDC for some time. Present efforts to assess the impact of the FMPC on cancer morbidity and mortality require the analysis of this misaligned dataset; in particular, it is necessary to interpolate gender- and age group-specific population counts to the windrose exposure cells. These numbers of persons at risk could then be combined with cell-specific dose estimates obtained by Killough et al. (1996) and estimates
Figure 1: Census block groups and 10-km windrose near the FMPC site, with 1990 population density by block group and 1980 structure density by cell (both in counts per km^2).

of the cancer risk per unit dose to obtain expected numbers of excess cancer cases by cell. The resulting expected counts would not only suggest the overall level of risk to the community, but also indicate whether sufficient power existed to justify the initiation of a full epidemiological study of the observed cancers near the plant.

In fact, such an expected death calculation has already been made by Devine et al. (1998), using traditional life table methods operating on the Rogers and Killough (1997) cell-level population estimates (which were in turn derived simply as proportional to the structure counts). However, these estimates were only for the total population in each cell; sex- and age group-specific counts
were obtained by “breaking out” the totals into subcategories using a standard table (i.e., the same table in each cell, regardless of its true demographic makeup). In addition, the uncertainty associated with the cell-specific population estimates was quantified in a rather ad-hoc way through distributional assumptions on the true number of structures in each cell and the true occupancy rate per structure, with this uncertainty propagated through to the estimated numbers of lung cancer deaths. We instead adopt a Bayesian approach that allows not only formal prior-to-posterior updating of point estimates and standard errors, but full posterior distributions for the imputed population counts in the 160 target cells in the windrose. We are also able to incorporate other available prior and covariate information, and to benefit from the implicit spatial smoothing.

In what follows we create a “snapshot” of the population for the year 1990. This particular year was convenient for us because of the ready availability of census data and block group boundary locations from CD-ROMs and the World Wide Web. Multiyear modeling was beyond the scope of our project, since the Census Bureau typically changes the boundaries of its block groups with every census, and earlier boundaries are not available in computerized form. Nevertheless, it is possible to extend our methods to the temporal case.

While our population data and boundaries are from 1990, the structure counts we use are actually somewhat older. This is because Rogers and Killough (1997) provide structure counts for only two years (1950 and 1980) based on corresponding USGS maps made at those times. For our purposes we chose to consider 1980 structure counts and 1990 population counts as being essentially concurrent; inflating all structures by some constant growth factor would have made no difference in our interpolated population values. Rogers and Killough (1997) suggest using the 1950 and 1980 datapoints to fit separate exponential growth models for the structure counts in each cell and then extrapolating to 1990. We find this unattractive since it amounts to fitting 160 two-parameter models, each to only two datapoints.
3 Model Development

We confine our model development to the case of two misaligned spatial grids. Given this development, the extension to more than two grids will be conceptually apparent. The additional computational complexity and bookkeeping detail will also be evident.

Let the first grid have regions indexed by $i = 1, ..., I$, denoted by $B_i$, and let $S_B = \bigcup_i B_i$. Similarly, for the second grid we have regions $C_j$, $j = 1, ..., J$ with $S_C = \bigcup_j C_j$. In some applications $S_B = S_C$, i.e., the $B$-cells and the $C$-cells offer different partitions of a common region. A special case includes the situation where one partition is a refinement of the other, e.g., the case where each $C_j$ is contained entirely in one and only one $B_i$ (as considered by Mugglin and Carlin 1998). Another possibility is that one data grid contains the other; say, $S_B \subset S_C$. In this case, there will exist some $C$ cells for which a portion lies outside of $S_B$. In the most general case, there is no containment and there will exist $B$-cells for which a portion lies outside of $S_C$ and $C$-cells for which a portion lies outside of $S_B$. Figure 2 illustrates this most general situation.

Atoms are created by intersecting the two grids. For a given $B_i$, each $C$-cell which intersects $B_i$ creates an atom (which possibly could be a union of disjoint regions). There may also be a portion
of $B_i$ which does not intersect with any $C_j$. We refer to this portion as the edge atom associated with $B_i$, i.e., a $B$-edge atom. In Figure 2, atoms $B_{11}$ and $B_{21}$ are $B$-edge atoms. Similarly, for a given $C_j$, each $B$-cell which intersects with $C_j$ creates an atom, and we analogously determine $C$-edge atoms (atoms $C_{11}$ and $C_{22}$ in Figure 2). It is crucial to note that each non-edge atom can be referenced relative to an appropriate $B$-cell, say $B_i$, and denoted as $B_{ik}$. It also can be referenced relative to an appropriate $C$ cell, say $C_j$, and denoted by $C_{jk}$. Hence, there is a one-to-one mapping within $S_B \cap S_C$ between the set of $ik$'s and the set of $jl$'s, as shown in Figure 2 (which also illustrates our convention of indexing atoms by area, in descending order). Formally we can define the function $c$ on non-edge $B$-atoms such that $c(B_{ik}) = C_{jk}$, and the inverse function $b$ on $C$-atoms such that $b(C_{jk}) = B_{ik}$. For computational purposes we suggest creation of “look-up” tables to specify these functions. (Note that the possible presence of both types of edge cell precludes a single “ij” atom numbering system, since such a system could index cells on either $S_B$ or $S_C$, but not their union.)

Without loss of generality we refer to the first grid as the response grid, that is, at each $B_i$ we observe a response $Y_i$. We seek to explain $Y_i$ using a variety of covariates. Some of these covariates may, in fact, be observed on the response grid; we denote the value of this vector for $B_i$ by $W_i$. But also, some covariates are observed on the second or explanatory grid. We denote the value of this vector for $C_j$ by $X_j$.

We seek to explain the observed $Y$’s through both $X$ and $W$. The misalignment between the $X$’s and $Y$’s is the obstacle to standard regression methods. What levels of $X$ should be assigned to $Y_i$? We propose a fully model-based approach in the case where the $Y$’s and $X$’s are aggregated measurements. The advantage of a model-based approach implemented within a Bayesian framework is full inference both with regard to estimation of model parameters and prediction using the model.

The assumption that the $Y$’s are aggregated measurements means $Y_i$ can be envisioned as
\[ \sum_k Y_{ik}, \text{where the } Y_{ik} \text{ are unobserved or latent and the summation is over all atoms (including perhaps an edge atom) associated with } B_i. \] To simplify, we assume that the \( X \)'s are also scalar aggregated measurements, i.e., \( X_j = \sum_l X_{jl} \) where the summation is over all atoms associated with \( C_j. \) As for the \( W \)'s, we assume that each component is either an aggregated measurement or an \textit{inheritable} measurement. For component \( r, \) in the former case \( W_i^{(r)} = \sum_k W_{ik}^{(r)} \) as with \( Y_i; \) in the latter case \( W_{ik}^{(r)} = W_i^{(r)}. \)

In addition to (or perhaps in place of) the \( W_i \) we will introduce \( B \)-cell random effects \( \mu_i, \) \( i = 1, \ldots, I. \) These effects are employed to capture spatial association among the \( Y_i \)'s. The \( \mu_i \) can be given a spatial prior specification. A Markov random field form (Besag 1974, Bernardinelli and Montomoli 1992), as described below, is convenient. Similarly we will introduce \( C \)-cell random effects \( \omega_j, j = 1, \ldots, J \) to capture spatial association among the \( X_j \)'s. It is assumed that the latent \( Y_{ik} \) inherit the effect \( \mu_i \) and that the latent \( X_{j\ell} \) inherit the effect \( \omega_j. \)

For aggregated measurements which are counts, we assume the latent variables are conditionally independent Poissons. As a result, the observed measurements are Poissons as well and the conditional distribution of the latent variables given the observed is a product multinomial. For aggregated measurements which are continuous, a convenient distributional assumption is conditionally independent Gammas whence the latent variables would be rescaled to product Dirichlet. An alternative choice is the normal, whence the latent variables would have a distribution which is a product of conditional multivariate normals. In the sequel we detail the Poisson case.

As mentioned above, area traditionally plays an important role in allocation of spatial measurements. Letting \(|A|\) denote the area of region \( A, \) if we apply the standard assumption of allocation proportional to area to the \( X_{j\ell} \) in a stochastic fashion, we would obtain

\[ X_{j\ell} \mid \omega_j \sim P \mathcal{d}(e^\omega_j | C_{j\ell}), \] (1)
assumed independent for $\ell = 1, 2, ..., L_j$. Then $X_j \mid \omega_j \sim P_0(\omega_j \mid C_j)$ and $(X_{j1}, X_{j2}, ..., X_{jL_j} \mid X_j, \omega_j) \sim \text{Mult}(X_j; q_{j1}, ..., q_{jL_j})$ where $q_{j\ell} = |C_{j\ell}|/|C_j|$. 

Such strictly area-based modeling cannot be applied to the $Y_{ik}$'s since it fails to connect the $Y$'s with the $X$'s (as well as the $W$'s). To do so we again begin at the atom level. For non-edge atoms we use the previously mentioned look-up table to find the $X_{j\ell}$ to associate with a given $Y_{ik}$. It is convenient to denote this $X_{j\ell}$ as $X'_{ik}$. Ignoring the $W_i$ for the moment, we assume

$$Y_{ik} \mid \mu_i, \theta_{ik} \sim P_0(e^{\mu_i} | B_{ik} \mid h(X'_{ik}/|B_{ik}|; \theta_{ik}))$$

independent for $k = 1, 2, ..., K_i$. Here $h$ is a preselected parametric function, the part of the model specification which adjusts an expected proportional-to-area allocation according to $X'_{ik}$. Since (1) models expectation for $X_{j\ell}$ proportional to $|C_{j\ell}|$, it is natural to use the standardized form $X'_{ik}/|B_{ik}|$ in (2). Particular choices of $h$ include $h(z; \theta_{ik}) = z$ yielding $Y_{ik} \mid \mu_i \sim P_0(e^{\mu_i} X'_{ik})$, which would be appropriate if we chose not to use $|B_{ik}|$ explicitly in modeling $E(Y_{ik})$. In our FMPC implementation, we actually select $h(z; \theta_{ik}) = z + \theta_{ik}$ where $\theta_{ik} = \theta/(K_i |B_{ik}|)$ and $\theta > 0$; see equation (8) below and the associated discussion.

If $B_i$ has no associated edge atom, then

$$Y_i \mid \mu_i, \theta, \{X_{j\ell}\} \sim P_0 \left( e^{\mu_i} \sum_k |B_{ik}| h(X'_{ik}/|B_{ik}|; \theta_{ik}) \right).$$

If $B_i$ has an edge atom, say $B_{iE}$, since there is no corresponding $C_{j\ell}$ there is no corresponding $X'_{iE}$. Hence, we introduce a latent $X'_{iE}$ whose distribution is determined by the non-edge atoms which are neighbors of $B_{iE}$. Paralleling equation (1), we model $X'_{iE}$ as

$$X'_{iE} \mid \omega_i^* \sim P_0(e^{\omega_i*} | B_{iE})$$
thus introducing an additional set of spatial random effects \( \{ \omega^*_i \} \) to the existing set \( \{ \omega_j \} \). These two sets then jointly receive a Markov random field prior specification. An alternative is to model 
\[
X'_{iE} \sim Po \left( |B_{iE}| \left( \sum_{N(B_{iE})} X'_{i} / \sum_{N(B_{iE})} |B_i| \right) \right),
\]
where \( N(B_{iE}) \) is the set of neighbors of \( B_{iE} \) and \( t \) indexes this set. Effectively, we multiply \( |B_{iE}| \) by the overall count per unit area in the neighboring non-edge atoms. While this model is somewhat more data-dependent than the (more model-dependent) one given in (4), we remark that it can actually lead to better MCMC convergence due to the improved identifiability in its parameter space: the spatial similarity of the structures in the edge zones is being modeled directly, rather than indirectly via the similarity of the \( \omega^*_i \) and the \( \omega_j \).

Now, with an \( X'_{ik} \) for all \( ik \), (2) is extended to all \( B \)-atoms and the conditional distribution of \( Y_i \) is determined for all \( i \) as in (3). But also \( Y_{i1}, ..., Y_{ii} \mid Y_i, \mu_i, \theta_{ik} \sim Mult(Y_i; q_{i1}, ..., q_{ik}) \) where 
\[
q_{ik} = \frac{|B_{ik}| h(X'_{ik} / |B_{ik}| ; \theta_{ik})}{\sum_k |B_{ik}| h(X'_{ik} / |B_{ik}| ; \theta_{ik})}.
\]

To capture the spatial nature of the \( B_i \) we may adopt a Markov random field prior for the \( \mu_i \)'s. Following Bernardinelli and Montomoli (1992) we assume that 
\[
f(\mu_i \mid \mu_{i' \neq i}) = N \left( \sum_{i'} u_{i,i'} \mu_{i'} / u_{ii}, 1/(\lambda \mu u_{ii}) \right)
\]
where \( u_{ii} = 0, u_{i'i'} = u_{i'i} \) and \( u_{ii} = \sum_{i'} u_{i'i} \). The joint distribution associated with these conditional distributions is uniquely determined but improper. Spatial heterogeneity assumes that the physical location of the units determine the \( u_{ii} \). If we put \( u_{i'i} = 1 \) for \( B_{i'} \) a neighbor of \( B_i \) and \( u_{i'i} = 0 \) otherwise, we obtain the standard “adjacency” form of the conditionally autoregressive (CAR) prior (Besag et al., 1995). If, for instance, all \( u_{i'i} = 1 \) we obtain an exchangeable prior (EXCH) which captures heterogeneity but not local clustering.
Similarly we assume that

$$f(\omega_j \mid \omega_{j',j' \neq j}) = N\left( \sum_{j'} v_{jj'} \omega_{j'} / v_{j'}, 1/(\lambda_\omega v_{j'}) \right).$$

We adopt a proper Gamma prior for $\lambda_\mu$ and also for $\lambda_\omega$. When $\theta$ is present we require a prior which we denote by $f(\theta)$. The choice of $f(\theta)$ will likely be vague but its form depends upon the adopted parametric form of $h$.

The entire specification can be given a representation as a graphical model, as in Figure 3. In this model the arrow from $\{X_{ji}\} \rightarrow \{X'_{ik}\}$ indicates the inversion of the $\{X_{ji}\}$ to $\{X'_{ik}\}$, augmented by any required edge atom values $X'_{iE}$. The $\{\omega^*_i\}$ would be generated if the $X'_{iE}$ are modeled using (4). Since the $\{Y_{ik}\}$ are not observed, but are distributed as multinomial given the fixed block group totals $\{Y_i\}$, this is a predictive step in our model, as indicated by the arrow from $\{Y_i\}$ to $\{Y_{ik}\}$ in the figure. In fact, as mentioned above the further predictive step to impute $Y'_{j}$, the $Y$ total associated with $X_j$ in the $j^{th}$ target zone, is of key interest. If there are edge atoms $C_{jE}$, this will require a model for the associated $Y'_{jE}$. Since there is no corresponding $B$-atom for $C_{jE}$. 

Figure 3: Graphical version of the model, with variables as described in the text. Boxes indicate data nodes, while circles indicate unknowns.
a specification such as (2) is not appropriate. Rather, we can imitate the above modeling for \(X_i\) using (4) by introducing \(\{\mu_j^*\}\) which along with the \(\mu_i\) follow the prior in (5). The \(\{\mu_j^*\}\) and \(\{Y_j\}\) would add two consecutive nodes to the right side of Figure 3, connecting from \(\lambda\) to \(\{Y_j\}\).

The entire distributional specification overlaid on this graphical model has been supplied in the foregoing discussion and (in the absence of \(C_j\) edge atoms, as in Figure 1) takes the form

\[
\begin{align*}
P_i \cdot f(Y_{i1}, \ldots, Y_{ik} | Y_i, \theta) &= \prod_i f(Y_i | \mu_i, \theta, \{X_i^T\}) \cdot f(\{X_i^T\} | \omega_i, \{X_j\}) \\
&\times \prod_j f(X_{j1}, \ldots, X_{jk} | X_j) \cdot f(X_j | \omega_j) \\
&\times f(\{\mu_i\} | \lambda) \cdot f(\lambda) \cdot f(\omega) \cdot f(\theta).
\end{align*}
\]

(7)

Bringing in the \(W_i\) merely revises the exponential term in (2) from \(\exp(\mu_i)\) to \(\exp(\mu_i + W_{ik}^T \beta)\).

Again, for an inherited component of \(W_i\), say \(W_i^{(r)}\), the resulting \(W_i^{(r)} = W_i^{(r)}\). For an aggregated component of \(W_i\), again say \(W_i^{(r)}\), we imitate (1) assuming \(W_i^{(r)} \sim \text{Po}(e^{\mu_i^{(r)} | B_{ik}})\), independent for \(k = 1, \ldots, K\). A spatial prior on the \(\mu_i^{(r)}\) and a Gaussian (or perhaps flat) prior on \(\beta\) completes the model specification.

Finally, on the response grid, for each \(B_i\) rather than observing a single \(Y_i\) we may observe \(Y_{im}\), where \(m = 1, 2, \ldots, M\) indexes levels of factors such as sex, race or age group. Here we seek to use these factors, in an ANOVA fashion, along with the \(X_j\) (and \(W_i\)) to explain the \(Y_{im}\). Ignoring \(W_i\), the resultant change in (2) is that \(Y_{im}\) will be Poisson with \(\mu_i\) replaced by \(\mu_{im}\), where \(\mu_{im}\) has an appropriate ANOVA form. For example, in the case of sex and age classes, we might have a sex main effect, an age main effect, and a sex-age interaction effect. In our application these effects are not nested within \(i\); we include only a spatial overall mean effect indexed by \(i\). We have not explored nested effects nor have we considered ANOVA forms for the \(\theta\)s (perhaps with an appropriate link function). Modification to incorporate the \(W_i\) is straightforward.
4 Implementation Issues

We elucidate the notation of the previous section for our Section 2 dataset as follows. The responses $Y$ are population counts, while the misaligned covariates $X$ are structure counts; our model has no additional covariates $W_i$. In the context of Figure 1, the $B_i$ are the block groups and the $C_j$ are the windrose cells. The union of the census blocks contains the windrose ($S_C \subseteq S_B$), so only $B$-edge atoms are present. Finally, henceforth we emphasize the summing over atoms within a block group or cell by replacing the appropriate subscript with a dot (e.g., $X_j = \sum_{i=1}^{L_j} X_{ji}$).

The modeling ideas in Section 3 are implemented for the FMPC dataset with a Markov chain Monte Carlo approach, implemented with Fortran and S-PLUS routines. The basic structure is that of a Gibbs sampler, frequently employing Metropolis and Hastings substeps (see e.g. Carlin and Louis 1996, Sec. 5.4.4). We run five independent chains for all parameters, using trace plots, lag 1 sample autocorrelations, and Gelman and Rubin (1992) diagnostic statistics to assess convergence of the algorithm. As mentioned in Section 3 the $Y_i \rightarrow \{Y_{ik}\}$ step is predictive, which allows us to implement it after the model has been fitted. That is, the $\{Y_{ik}\}$ may be analytically integrated out of the full model specification (7), the sampler run on the resulting lower-dimensional parameter space, and the required $\{Y_{ik}\}$ samples drawn from the appropriate multinomial at the very end.

We faced several challenges in implementing our model on the FMPC dataset. The first was the amount of bookkeeping in the “lookup” tables that allowed for atoms $p$ to be referenced in any of three systems: $p$, $(ik)$, and $(jl)$. Introduction of a CAR prior on the $\omega_j$ and $\omega^*_{i}$ requires yet more lookup tables, as it is necessary to keep track of which regions are neighbors (the customary CAR model adjacency matrix). Making the assumption that adopting a CAR prior for the structure spatial effects will result in sufficient smoothing of the fitting maps, we simply placed an iid (not a CAR) prior on the $\{\mu_k\}$. Details concerning the number and types of necessary lookup tables are
given in Mugglin (1999).

A second difficulty stemmed from the discrete multivariate (and sometimes zero-valued) nature of the multinomial distribution in the MCMC setting. This presented several obstacles, the first being the choice of candidate density for the Metropolis-Hastings rejection step. Initially we chose to propose structures in each cell as a multinomial value, allocating structures to atoms proportionally to area (i.e., equating the candidate and the prior distributions). This was convenient computationally, but did not produce acceptable convergence because in many cells the proposed values were not similar enough to the posteriors. A first-order Maclaurin approximation to provide multinomial candidates that were closer to the posterior also proved ineffective. We finally adopted a multinomial proposal based only on the present state of the chain. That is, if this state is \((X_{j1}, \ldots, X_{jL_j})\), we propose a multinomial value drawn from a \(\text{Mult}(X_j; q_{j1}, \ldots, q_{jL_j})\) where \(q_{jl} = X_{jl}/X_j\). A further correction is required since whenever a present \(X_{jl}\) becomes 0, under this proposal it will be 0 in all subsequent iterations as well. Hence we modify \(q_{jl}\) to \(q_{jl} = (X_{jl} + 1)/(X_j + L_j)\). We emphasize that this modification affects only the Metropolis-Hastings proposal density, and not the model itself.

A somewhat different problem required a similar solution. We originally adopted the identity function for \(h\) in (2), producing the model \(Y_{ik} \sim P_0(\mu_i(X'_{ik}))\), which in turn implies \(Y_i \sim P_0(\mu_i(X'_{ik}))\). Suppose however that \(Y_i > 0\) for a particular block group \(i\), but in some MCMC iteration no structures are allocated to any of the atoms of the block group. The result is a flawed probabilistic specification. To ensure \(h > 0\) even when \(z = 0\), we revised our model to \(h(z; \theta_{ik}) = z + \theta_{ik}\) where \(\theta_{ik} = \theta/(K_i|B_{ik}|)\) with \(\theta > 0\), resulting in

\[
Y_{ik} \sim P_0\left(\mu_i\left(X'_{ik} + \frac{\theta}{K_i}\right)\right).
\] (8)
This adjustment eliminates the possibility of a zero-valued Poisson parameter, but does allow for the possibility of a nonzero population count in a region where there are no structures observed. When conditioned on \( Y_i \), we find \( (Y_{i1}, \ldots, Y_{iK_i} \mid Y_i) \sim \text{Mult}(Y_i; p_{i1}, \ldots, p_{iK_i}) \), where

\[
p_{ik} = \frac{X_{ik}^l + \theta / K_i}{X_i^l + \theta}, \quad \text{and} \quad Y_i \sim \text{Po}(e^{\eta_i} (X_i^l + \theta)).
\] (9)

Our basic model then consists of (8) – (9) together with the following:

\[
\mu_i \sim \text{iid } N(\eta_i, 1/\tau_i), \quad X_{jl} \sim \text{Po}(e^{\omega_j} | C_{jl}|) \Rightarrow X_{jl} \sim \text{Po}(e^{\omega_j} | C_{jl}|),
\]

\[
(X_{j1}, \ldots, X_{jL_j} \mid X_{jl}) \sim \text{Mult}(X_{jl}; q_{j1}, \ldots, q_{jL_j}), \quad \text{where} \quad q_{ji} = |C_{jl}| / |C_j|,
\] (10)

\[
X_{iE}^l \sim \text{Po}(e^{\omega_i^*} | R_{iE}|), \quad \text{and} \quad (\omega_j, \omega_i^*) \sim \text{CAR}(\lambda_\omega),
\]

where \( X_{iE}^l \) and \( \omega_i^* \) refer to edge atom structure counts and log relative risk parameters, respectively.

While \( \theta \) could be estimated from the data, in our implementation we simply set \( \theta = 1 \); Section 6 discusses the impact of alternate selections.

A third implementation challenge we faced was the high autocorrelations in the sampled chains. We implemented a standard Metropolis-Hastings algorithm with five independent parallel chains. With the structures missing outside the windrose and highly constrained (by their cell totals) within the windrose, the algorithm appears to traverse the parameter space slowly: the median of the lag-1 autocorrelations over the 1006 parameters is 0.76; 8% of them are larger than 0.99. Wherever possible we tuned candidate densities so that proposed values were accepted at rates between 35% and 50%, as recommended by Gelman et al. (1996), as well as years of Metropolis “folklore.” It should be noted, however, that the structures (proposed as multinomials in the windrose cells and as Poissons in the edge atoms) could not be tuned as with usual Gaussian or \( t \) proposals, because the variances could not be specified independently of the means. Fortunately in all cases we were
still able to obtain a reasonable minimum level of candidate acceptance.

Despite the high autocorrelations, Gelman and Rubin convergence diagnostic statistics maintained acceptable levels (95.5% of the 1006 parameters had G&R estimates of at most 1.1), and the five independent chains, when viewed graphically, appeared to mix well and traverse the stationary distribution adequately. Thus, more sophisticated techniques such as blocking (Liu, Wong, and Kong 1994) and tempering (Geyer and Thompson 1995) were not implemented.

5 Data Analysis

We turn now to the particulars of the FMPC data analysis, examining three different models in the context of the misaligned data as described in Section 2. In the first case we take up the problem of total population interpolation, while in the second and third cases we consider age- and sex-specific population interpolation.

5.1 Total population interpolation model

We begin by taking $\eta_\mu = 1.1$ and $\tau_\mu = 0.5$ in (10). The choice of mean value reflects the work of Rogers and Killough (1997), who found population per household (PPH) estimates for four of the seven townships in which the windrose lies. Their estimates ranged in value from 2.9 to 3.2, hence our choice of $\eta_\mu = 1.1 \approx \log(3)$. The value $\tau_\mu = 0.5$ is sufficiently small to make the prior for $\mu_i$ large enough to support all feasible values of $\mu_i$ (two prior standard deviations in either direction would enable PPH values of 0.18 to 50.8).

For $\omega = \{\omega_j, \omega^*_i\}$ we adopted a CAR prior and fixed $\lambda_\omega = 10$. We did not impose any centering of the elements of $\omega$ around 0, allowing them to determine their own mean level in the MCMC algorithm. Since most cells have four neighbors, the value $\lambda_\omega = 10$ translates into a conditional prior standard deviation for the $\omega$’s of $\sqrt{1/(10\cdot 4)} = .158$, hence a marginal prior standard deviation.
of roughly $1.58/7 \approx 0.23$ (Bernardinelli et al. 1995). In any case, we found $\lambda_\omega < 10$ too vague to allow MCMC convergence. Typical posterior medians for the $\omega$'s ranged from 2.2 to 3.3 for the windrose $\omega_j$'s and from 3.3 to 4.5 for the edge $\omega_i$'s.

Running 5 parallel sampling chains, acceptable convergence obtains for all parameters within 1500 iterations. We discarded this initial sample and then continued the chains for an additional 5000 iterations each, obtaining a final posterior sample of size 25,000. From the resulting samples, we can examine the posterior distributions of any parameters we wish. It is instructive first to examine the distributions of the imputed structure counts $X_{ji}$. For example, consider Figure 4, which shows the posterior distributions of the structure counts in cell 106 (the sixth one from the windrose center in the SE direction), for which $L_j = 4$. The known cell total $X_{106}$ is 55. Note that the structure values indicated in the histograms are integers. The vertical bars in each histogram indicate how the 55 structures would be allocated if imputed proportionally to area. In this cell we observe good general agreement between these naively imputed values and our histograms, but the advantage of assessing variability from the full distributional estimates is immediately apparent.

Population estimates per cell for cells 105 through 110 (again in the SE direction, from the
Figure 5: Posterior distributions of populations in cells 105-110. Vertical bars represent estimates formed by multiplying structures per cell by a constant population per household (PPH) of 3.0. Middle to outer edge of the windrose are indicated in Figure 5. Vertical bars here represent estimates calculated by multiplying the number of structures in the cell by a fixed (map-wide) constant representing population per household (PPH), a method roughly equivalent to that employed by Rogers and Killough (1997), who as mentioned above actually used four different PPH values. Our reference lines use a constant value of 3 (the analogue of our prior mean). While cells 105 and 106 indicate good general agreement in these estimates, cells 107 through 110 display markedly different population estimates, where our estimates are substantially higher than the constant-PPH estimates. This is typical of cells toward the outer edge of the southeast portion of the windrose,
since the suburbs of Cincinnati encroach on this region. We have population data only (no structures) in the southeastern edge atoms, so our model must estimate both the structures and the population in these regions. The resulting PPH is higher than a mapwide value of 3 (one would expect suburban PPH to be greater than rural PPH) and so the CAR model placed on the \( \{\omega_j, \omega_i^*\} \) parameters induces a spatial similarity that can be observed in Figure 5.

We next implement the \( \{Y_{i*}\} \rightarrow \{Y_{i*}\} \) step. From the resulting \( \{Y_{i*}\} \) come the \( \{Y_{*j}\} \) cell totals by appropriate reaggregation. Figure 6 shows the population densities by atom \( (Y_{i*}/|B_{i*}|) \), calculated by taking the posterior medians of the population distributions for each atom and dividing by atom area in square kilometers. This figure clearly shows the encroachment by suburban Cincinnati.
on the southeast side of our map, with some spatial smoothing between the edge cells and the outer windrose cells. Finally, Figure 7 shows population densities by cell \( \frac{Y_j}{|C_j|} \), where the atom-level populations have been aggregated to cells before calculating densities. Posterior standard deviations, though not shown, are also available for each cell. While this figure, by definition, provides less detail than Figure 6, it provides information at the scale appropriate for combination with the exposure values of Killough et al. (1996). Moreover, the scale of aggregation is still fine enough to permit identification of the locations of Cincinnati suburban sprawl, as well as the communities of Ross (contained in cells ENE 4-5 and NE 4), Shandon (NW 4-5), New Haven (WSW 5-6), and New Baltimore (SSE 4-5).
5.2 Age and sex effects

Recall from Section 2 that we seek population counts not only by cell but also by sex and age group. This is because the dose resulting from a given exposure will likely differ depending on gender and age, and because the risk resulting from that dose can also be affected by these factors. Again we provide results only for the year 1990; the extension to other timepoints would of course be similar. Population counts at the block group level by sex and age group are provided by the U.S. Census Bureau. Specifically, age is recorded as counts in 18 quinquennial (5-year) intervals: 0-4, 5-9, ..., 80-84, and 85+. We consider two extensions of our basic model (8) – (10) to the sex- and age group-specific case.

5.2.1 Additive Model

Here we start with the assumption that the population counts in atom $k$ of block group $i$ for gender $g$ at age group $a$ is Poisson-distributed as

$$Y_{ikg} \sim Po\left(e^{\delta_{iga}} \left(X_{ik} + \frac{\theta_i}{K_i}\right)\right), \text{ where } \delta_{iga} = \mu_i + g \, \alpha + \sum_{a=1}^{17} \beta_a I_a$$

where $g=0$ for males and 1 for females and $I_a$ is a 0–1 indicator for age group $a$ ($a = 1$ for ages 5-9, $a = 2$ for 10-14, etc.). The $\mu_i$ are block group-specific baselines (in our parametrization, they are the logs of the fitted numbers of males in the 0-4 age bracket), and $\alpha$ and the $\{\beta_a\}$ function as main effects for sex and age group, respectively. Note the $\alpha$ and $\{\beta_a\}$ parameters are not specific to any one block group, but rather apply to all 39 block groups in the map.

With each $\mu_i$ now corresponding only to the number of baby boys (not the whole population) in block group $i$, we expect its value to be decreased accordingly. Because there are 36 age-sex divisions, we modified the prior mean $\eta_\mu$ to $-2.5 \approx \log(3/36)$. We placed vague independent
Table 1: Quantiles and significance of gender and age effects for the age-sex additive model.

<table>
<thead>
<tr>
<th>effect</th>
<th>parameter</th>
<th>median</th>
<th>2.5%</th>
<th>97.5%</th>
</tr>
</thead>
<tbody>
<tr>
<td>gender</td>
<td>$\alpha$</td>
<td>0.005</td>
<td>-0.012</td>
<td>0.021</td>
</tr>
<tr>
<td>ages 5-9</td>
<td>$\beta_1$</td>
<td>0.073</td>
<td>0.033</td>
<td>0.116</td>
</tr>
<tr>
<td>ages 10-14</td>
<td>$\beta_2$</td>
<td>0.062</td>
<td>0.021</td>
<td>0.106</td>
</tr>
<tr>
<td>ages 15-19</td>
<td>$\beta_3$</td>
<td>-0.003</td>
<td>-0.043</td>
<td>0.041</td>
</tr>
<tr>
<td>ages 20-24</td>
<td>$\beta_4$</td>
<td>-0.223</td>
<td>-0.268</td>
<td>-0.177</td>
</tr>
<tr>
<td>ages 25-29</td>
<td>$\beta_5$</td>
<td>-0.021</td>
<td>-0.063</td>
<td>0.024</td>
</tr>
<tr>
<td>ages 30-34</td>
<td>$\beta_6$</td>
<td>0.137</td>
<td>0.095</td>
<td>0.178</td>
</tr>
<tr>
<td>ages 35-39</td>
<td>$\beta_7$</td>
<td>0.118</td>
<td>0.077</td>
<td>0.16</td>
</tr>
<tr>
<td>ages 40-44</td>
<td>$\beta_8$</td>
<td>0.044</td>
<td>0.001</td>
<td>0.088</td>
</tr>
<tr>
<td>ages 45-49</td>
<td>$\beta_9$</td>
<td>-0.224</td>
<td>-0.27</td>
<td>-0.18</td>
</tr>
<tr>
<td>ages 50-54</td>
<td>$\beta_{10}$</td>
<td>-0.404</td>
<td>-0.448</td>
<td>-0.357</td>
</tr>
<tr>
<td>ages 55-59</td>
<td>$\beta_{11}$</td>
<td>-0.558</td>
<td>-0.609</td>
<td>-0.507</td>
</tr>
<tr>
<td>ages 60-64</td>
<td>$\beta_{12}$</td>
<td>-0.627</td>
<td>-0.677</td>
<td>-0.576</td>
</tr>
<tr>
<td>ages 65-69</td>
<td>$\beta_{13}$</td>
<td>-0.896</td>
<td>-0.951</td>
<td>-0.839</td>
</tr>
<tr>
<td>ages 70-74</td>
<td>$\beta_{14}$</td>
<td>-1.32</td>
<td>-1.386</td>
<td>-1.255</td>
</tr>
<tr>
<td>ages 75-79</td>
<td>$\beta_{15}$</td>
<td>-1.72</td>
<td>-1.797</td>
<td>-1.643</td>
</tr>
<tr>
<td>ages 80-84</td>
<td>$\beta_{16}$</td>
<td>-2.32</td>
<td>-2.424</td>
<td>-2.224</td>
</tr>
<tr>
<td>ages 85+</td>
<td>$\beta_{17}$</td>
<td>-2.836</td>
<td>-2.969</td>
<td>-2.714</td>
</tr>
</tbody>
</table>

Table 1: Quantiles and significance of gender and age effects for the age-sex additive model.

$N(0,10^2)$ priors on $\alpha$ and the $\beta$s, and kept all other prior values the same as in Section 5.1. Convergence of the MCMC algorithm obtains in about 1500 iterations. (The slowest parameters to converge are those pertaining to the edge atoms, where we have no structure data. Some parameters converge much faster: the $\alpha$ and $\beta_a$ parameters, for example, converge by about 500 iterations.) We then ran 5000 iterations for each of 5 chains, resulting in a final sample of 25,000.

Population interpolation results are quite similar to those outlined in Section 5.1, except that population distributions are available for each cell at any combination of age and sex. While we do not show these results here, we do include a summary of the main effects for age and sex. Table 1 shows the posterior medians and 2.5% and 97.5% quantiles for the $\alpha$ and $\beta_a$ parameters. Among the $\beta_a$ parameters, we see a significant negative value of $\beta_4$ (ages 20-24), reflecting a relatively small group of college-aged residents in this area. After a slight increase in the age distribution for ages 30-44, we observe increasingly negative values as $a$ increases, indicating the expected decrease in
population with advancing age.

### 5.2.2 Additive-plus-interaction model

A slightly more complicated model allows for sex-age interaction:

\[ Y_{ikga} \sim Po \left( e^{\delta_{iga}} \left( X'_{ik} + \frac{\theta}{K_i} \right) \right), \quad \text{where} \quad \delta_{iga} = \mu_i + g \alpha + \sum_{a=1}^{17} \beta_a I_a + g \sum_{a=1}^{17} \gamma_a. \]

Here the \( \mu_i \) retain their interpretation as log-baby boys per structure in block group \( i \), \( \alpha \) is now an adjustment for baby girls, the \( \beta_a \) can now be interpreted as the effect of age on males, and the \( \gamma_a \) are an adjustment for the effect of age on females. As with \( \alpha \) and \( \{\beta_a\} \), we placed independent \( N(0, 10^2) \) priors on the \( \{\gamma_a\}_{a=1}^{17} \), and kept all other priors the same as in Section 5.2.1. Convergence obtains once again by about 1500 iterations, with some parameters converging notably faster: the \( \alpha \), \( \beta \), and \( \gamma \) parameters converge by about 500 iterations. Once again we used 5000 samples from each of 5 independent chains collected after this initial sampling period.

Table 2 shows the posterior medians and 2.5% and 97.5% quantiles for each of the \( \alpha \), \( \beta \), and \( \gamma \) parameters. The most apparent feature is that, for older age groups, the \( \beta \) and \( \gamma \) parameters are trending away from 0: as in the additive model, the \( \beta \)'s are decreasing, indicating a general decrease in the prevalence of older males, but the \( \gamma \)'s are increasing, indicating a progressive relative increase in the numbers of older women over older men, in agreement with intuition. The \( \alpha \) parameter is now slightly negative, reflecting the fact that the number of baby boys in the dataset (2229) exceeds that of baby girls (2104). The “bump” in the mortality curve for men in their early 20s, familiar to actuaries, is also apparent in our data.

Finally, population interpolation results can again be found and mapped for any level of age and sex; in the interest of space we suppress their display here. We also experimented with somewhat larger values of the smoothing parameter \( \lambda_{\omega} \) (50 and 500), obtaining fitted maps qualitatively
Table 2: Quantiles and significance of gender, age, and interaction effects for additive-plus-interaction model.

similar to those in Figures 6 and 7.

6 Summary and Discussion

In this paper we have outlined a method for handling spatially misaligned data which allows the customary Bayesian borrowing of strength across units (perhaps via a spatial smoothing prior), and enables full posterior inference for imputed counts at any desired level of aggregation. The approach appears to be an attractive alternative to more ad hoc interpolation and estimation methods currently implemented in most GISs; in particular, it offers the potential for specifically estimating the combination of sampling variance and uncertainty inherent in population estimates, and using these components to make risk assessment inferences in a comprehensive manner.

As mentioned below equation (10), our results all assume $\theta$ in (8)-(9) is a tuning constant taking
a fixed value of 1. Since this is an admittedly rather ad-hoc model modification, we undertook a brief study of the effect of the choices \( \theta \in \{.01, 1, 10, 100\} \) on the resulting posterior summaries of interest. We found minimal impact on the posterior medians for the 39 \( \mu_i \) and the 160 imputed \( Y_{ij}^r \) cell totals for every choice except \( \theta = 100 \). Posterior medians for the \( X_{ij}^r \) ranged from 5 to 685 with a median of 292, and exhibited little sensitivity to the choice of \( \theta \). Equation (9) shows that large choices of \( \theta \) will have an appreciable effect when the \( X_{ij}^r \) tend to be small. MCMC convergence summaries for the \( \mu_i \) were acceptable for all of the \( \theta \) values listed above. Values of \( \theta < .01 \) still work well for our dataset, though they might not in settings where sample values of \( X_{ij}^r = 0 \) are more common.

There are several topics for future investigation in this area. In our FMPC data analysis, we may wish to compare various prior or model specifications using predictive methods (Gelfand and Ghosh 1998) or penalized likelihood criteria designed to account for the hierarchical structure in our models (Spiegelhalter et al. 1998). We might also extend the scope of the project by combining our population and structure count information with the exposure summaries by Killough et al. (1996) and cancer incidence data from the Ohio Department of Health, in order to investigate the link between radon exposure and excess lung cancer incidence in the Fernald assessment domain. While our methods are in principle able to handle this setting, there are several practical impediments to obtaining accurate results. These include the relatively small number of observed lung cancer cases during the observation period, the lack of an effective control variable (or surrogate marker) for smoking behavior, and the difficulty in assessing the impact of migration in and out of the study area (e.g., it may be that many persons retired and moved out of the study area before their radon-induced lung cancer was diagnosed). However, incorporation of cancer death information (painstakingly collected from death certificates at area hospitals) into the model-based risk assessment of Devine et al. (1998) is the subject of ongoing work at the CDC.
Regarding more general methodological development, several more issues come to mind. The application of our method to continuous data, following the modeling suggestions preceding equation (1), should be explored in detail. One might also attempt an extension of our models to the spatio-temporal case. Here the misalignment can arise from a single set of grid boundaries that changes over time (e.g. zip codes or census tracts, which are frequently updated to reflect changing population), or as in our work from multiple grid boundaries, possibly with temporal change as well (Zhu et al., 1999). Another important extension is the combination of misaligned areal data of the sort we have described here with point process data, such as that analyzed in Wakefield and Morris (1997). Finally, one might investigate the feasibility of our methods under higher-dimensional (e.g. three) misaligned grids, or in problems involving a multivariate response.

References


31