

# Bayesian Multivariate Areal Wombling for Multiple Disease Boundary Analysis

HAIJUN MA AND BRADLEY P. CARLIN<sup>1</sup>

*MMC 303, School of Public Health, University of Minnesota,  
Minneapolis, Minnesota 55455-0392, U.S.A.*

Correspondence author: Bradley P. Carlin

telephone: (612) 624-6646

fax: (612) 626-0660

email: brad@biostat.umn.edu

December 21, 2006

---

<sup>1</sup>Haijun Ma is Senior Biostatistician, Amgen Corporation, Thousand Oaks, CA, 91360. Bradley P. Carlin is Mayo Professor in Public Health in the Division of Biostatistics, School of Public Health, University of Minnesota, Minneapolis, MN, 55455. The work of both authors was supported in part by NIH grant 1-R01-CA95955-01.

# Bayesian Multivariate Areal Wombling for Multiple Disease Boundary Analysis

## Abstract

Multivariate data summarized over areal units (counties, zip codes, etc.) are common in the field of public health. Estimation or testing of geographic boundaries for such data may have varied goals. For example, for data on multiple disease outcomes, we may be interested in a single set of “composite” boundaries for all diseases, separate boundaries for each disease, or both. Different areal wombling (boundary analysis) techniques are needed to meet these different requirements. But in any case, the underlying statistical model needs to account for correlations across both diseases and locations. Utilizing recent developments in multivariate conditionally autoregressive (MCAR) distributions and spatial structural equation modeling, we suggest a variety of Bayesian hierarchical models for multivariate areal boundary analysis, including some that incorporate random neighborhood structure. Many of our models can be implemented via standard software, namely WinBUGS for posterior sampling and R for summarization and plotting. We illustrate our methods using Minnesota county-level esophagus, larynx, and lung cancer data, comparing models that account for both, only one, or neither of the aforementioned correlations. We identify both composite and cancer-specific boundaries, selecting the best statistical model using the DIC criterion. Our results indicate primary boundaries in both the composite and cancer-specific response surface separating the mining- and tourism-oriented northeast counties from the remainder of the state, as well as secondary (residual) boundaries in the Twin Cities metro area.

**KEY WORDS:** Areal data; Cancer; Multivariate conditionally autoregressive (MCAR) model; Surveillance, Epidemiology and End Results (SEER) data.

# 1 Introduction

Recently, there has been increasing interest in the spatial problem of detecting barriers separating regions of high and low response for certain quantities of interest. This area is often referred to as *boundary analysis* or *wombling*, the latter name paying tribute to an early important paper in the area (Womble, 1951). In public health, wombling is useful for detecting regions of significantly different disease mortality or incidence, thus improving decision-making regarding disease prevention and control, allocation of resources, and so on.

Multivariate areal data are common in public health studies. Much of this data is *areal* (aggregated at a certain regional level) to protect subjects' privacy. Since geographic location is often a surrogate for a mix of lifestyle, environmental, and possibly genetic factors that may underlie geographical differences (Elliot and Best, 1998), observations collected over a map are often spatially correlated. At the same time, the multiple variables under study (e.g., diseases sharing etiologic or other risk factors) are often correlated too.

The Surveillance, Epidemiology and End Results (SEER; <http://seer.cancer.gov>) database provides the numbers of deaths and corresponding numbers of person-years at risk in quinquennial age brackets for each county in various states and each of several cancer sites. Here we will study SEER data on the numbers of deaths due to cancers of the lung, larynx and esophagus in the years from 1990 to 2000 at the county level in the U.S. state of Minnesota. These three cancer sites are all part of the upper aerodigestive tract, and hence closely related anatomically. Epidemiological evidence also shows a strong and consistent relationship between exposure to alcohol and tobacco and the risk of cancer at these sites (Baron et al., 1993). In particular, the link between tobacco use and lung cancer, the leading cause of cancer death in the U.S., is by now well-established (<http://www.lungcancer.org>). Moreover, a new report (Institute of Medicine, 2006) finds sufficient evidence for a causal link between asbestos exposure (a known cause of lung cancer) and laryngeal cancer.

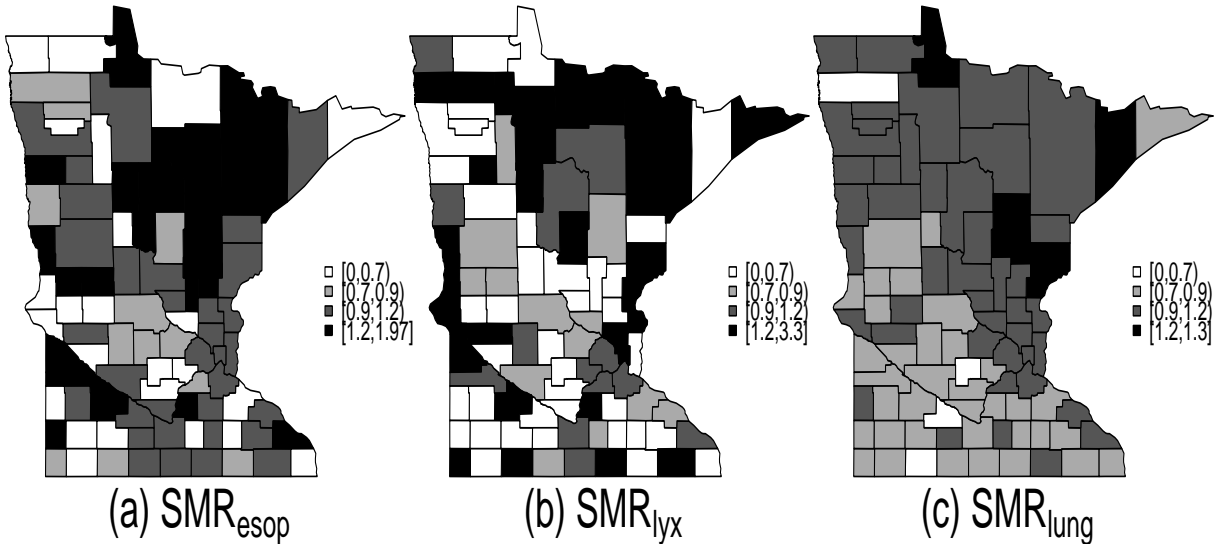


Figure 1: Maps of age-adjusted standardized mortality ratios: (a) esophagus cancer; (b) larynx cancer; (c) lung cancer.

Both larynx cancer and esophagus cancer are far rarer than lung cancer. For our dataset, the raw mortality rates are less than 1 per 100,000 person for larynx cancer, about 4 per 100,000 for esophagus cancer, and 45 per 100,000 for lung cancer. These rates are roughly comparable to those seen in the U.S. generally, though the Minnesota larynx rate remains small by comparison.

Figure 1 gives the raw county-level standardized mortality ratios (SMRs) based on our data, calculated as  $Y_{ik}/E_{ik}$  for  $i = 1, \dots, n$  and  $k = 1, 2, 3$ , where  $Y_{ik}$  is the cancer death count and  $E_{ik}$  is the age-adjusted expected count for cancer  $k$  in county  $i$ . The SMR maps for all three cancers indicate a pattern of decrease from northeast to southwest, suggesting positive association among them. The pattern is strongest for lung cancer, which also has less variable SMRs due to the higher case counts. To further investigate this pattern, we performed a preliminary regression analysis using overall (all three cancers) county-level age-adjusted SMR as the response variable, and the sum of the  $x$  and  $y$  coordinates of each county's centroid as the predictor variable. This variable's coefficient emerges as positive and statistically significant, indicating the increase from southwest to northeast is important (consistent with the visual impression from the maps), and motivating a full hierarchical

spatial analysis. Such an analysis' primary substantive goal would be to identify significant boundaries on these maps for the cancers individually, as well as for any underlying spatial common factor, all while accounting for correlation across both counties and cancers. Boundaries are important here for public health professionals tasked with identifying geographic regions in need of certain cancer-related intervention efforts, e.g., a cancer education or screening campaign focused at a few shopping malls or other public locations. They are also useful for identifying regions of rapid change in the fitted cancer surface, so that these areas can be studied in more detail for clues (say, missing covariates) that might explain why cancer mortality differs across the identified boundaries.

For correlated areal variables, the most popular modeling approach has been through the conditionally autoregressive (CAR) distribution (Besag, 1974) and its variants. Mardia (1988) described the theoretical background for multivariate CAR (MCAR) specifications using Gaussian Markov random fields (MRFs). Many generalizations of MCAR allow more flexible modeling of associations between different variables and areal units; see e.g. Kim et al. (2001), Carlin and Banerjee (2003), and Gelfand and Vounatsou (2003). Recent development in Bayesian statistical computing and software has made hierarchical MCAR modeling available to practitioners. For example, an important special case of the MCAR, the multivariate intrinsic autoregressive (MIAR) distribution, is included in the WinBUGS package (<http://www.mrc-bsu.cam.ac.uk/bugs/>) under the name `mv.car`. Also, Jin et al. (2005) specified the MCAR conditionally, allowing more versatile yet conceptually straightforward (and OpenBUGS-implementable) modeling of the intervariable correlations.

Another line of research for multivariate area data has arisen from structural equation modeling (SEM); see e.g. Wang and Wall (2003) and Liu et al. (2005). The “shared component” model of Knorr-Held and Best (2001) is also a special kind of SEM. In this approach, the correlations among the spatially-referenced variables are modeled through some collec-

tion of shared latent variables and their corresponding residual variations.

The rest of this article is organized as follows. Section 2 is devoted to providing background information on boundary analysis and multivariate areal modeling via MCAR and SEM. Section 3 then describes our proposed multivariate areal wombling techniques. The methods are illustrated using our SEER cancer data in Section 4, where models are compared using the Deviance Information Criterion (DIC; Spiegelhalter et al., 2002). Both composite and disease-specific boundaries are constructed and evaluated for the DIC-best models. Finally, Section 5 concludes and mentions directions for further work in this area.

## 2 Methodological background

### 2.1 Existing methods for univariate areal boundary analysis

Boundary analysis typically involves estimation or testing of “lines” on a continuous surface. In the areal case, however, defensible boundaries can only be a subset of the borders that determine the areal units, since we lack within-county information. Early studies in this area were purely algorithmic in nature; see e.g. Csillag et al. (2001), Jacquez and Greiling (2003), and the GIS software package `BoundarySEER`.

Lu and Carlin (2005; henceforth LC) embedded areal boundary analysis in a Bayesian hierarchical modeling framework. For counts of a single disease  $Y_i$ , they assume

$$Y_i | \boldsymbol{\beta}, \phi_i \sim \text{Poisson}(\mu_i) \text{ with } \log \mu_i = \log E_i + \mathbf{x}'_i \boldsymbol{\beta} + \phi_i, \quad i = 1, \dots, n, \quad (1)$$

$$\text{and } \boldsymbol{\phi} | W, \tau_\phi \propto \exp \left\{ -\frac{\tau_\phi}{2} \sum_{i \sim j} w_{ij} (\phi_i - \phi_j)^2 \right\}, \quad (2)$$

where (2) is the CAR prior for the areal random effects  $\boldsymbol{\phi} = (\phi_1, \dots, \phi_n)'$ ,  $\tau_\phi$  is a positive scale parameter, and  $i \sim j$  indicates that areas  $i$  and  $j$  are neighbors. The CAR distribution is an MRF formed from consideration of all *pairs* of neighbors. More specifically, the neighborhood

structure is specified in an  $n \times n$  *proximity matrix*,  $W$ , whose elements  $w_{ij}$  measure “closeness” of each pair of areas  $(i, j)$ . The most common choice is a “0/1” proximity matrix, wherein  $w_{ij} = 1$  if areas  $i$  and  $j$  are neighbors (spatially adjacent), and 0 otherwise. More general classes of weights (say, using intercentroidal distances) are also possible; see e.g. Cressie (1993, p.385) or Banerjee et al. (2004, pp.70-71). In our Section 4 data analysis we focus on the 0/1 case, but not before considering alternatives.

Let  $D = \text{Diag}(w_{i+})$ , where  $w_{i+} = \sum_j w_{ij}$  is the sum of  $W$ 's entries in row  $i$ . Then (2) can be rewritten as  $\exp\left(-\frac{\tau_\phi}{2}\phi'(D - W)\phi\right)$ . Due to the singularity of  $D - W$ , this CAR distribution is improper, and sometimes called an *intrinsically autoregressive* (IAR) distribution. A commonly used proper CAR is obtained by instead taking the joint distribution of  $\phi$  to be  $p(\phi|W, \rho) \propto \exp\left[-\frac{\tau_\phi}{2}\phi'(D_w - \rho W)\phi\right]$ , where  $\rho$  is chosen to make  $(D_w - \rho W)$  nonsingular (Cressie, 1993, Sec. 6.3; Banerjee et al., 2004, pp.80-81).

In areal wombling, changes over county boundaries are quantified by boundary likelihood values (BLVs). Lu and Carlin (2005) determine BLVs using posterior summaries of changes in the corresponding fitted mean structures. For example, the BLV can be the posterior mean of  $\Delta_{|\eta|,ij} = |\eta_i - \eta_j|$  for all adjacent  $i$  and  $j$ , where  $\eta_i = \mu_i/E_i$  measures the true underlying relative risk in area  $i$ . An edge element  $ij$  can then be thought of as part of the boundary if  $E(\Delta_{|\eta|,ij}|\mathbf{y}) > c$ , where  $c > 0$  is some predefined constant.

The boundary segments identified using this method are often disconnected, even though boundaries formed as series of connected segments may be preferred. Also note that in this model, the CAR smooths over all neighbors regardless of auxiliary physical (say, mountain range) or sociodemographic (say, racial) information that may be relevant. Lu et al. (2006) allow the areal adjacency weights  $w_{ij}$  to be random and subject to the influence of such covariate information. Ma et al. (2006) further extend this idea, including both areal and edge random effects in the modeling. Following edge detection techniques in the image

restoration literature (e.g. Jeng and Woods, 1991, Dass and Nair, 2003), this paper proposed a “site-edge” (SE) approach that jointly models two types of random effects. First, a set of site-level (areal) random effects  $\boldsymbol{\phi}^S = \{\phi_i^S\}$  are given a CAR prior to account for spatial association among the areas. Then a second set of *edge-level* random effects  $\boldsymbol{\phi}^E = \{\phi_{ij}^E\}$  are included and their distribution modeled using edge adjacency information from the map, enabling smoother, more connected boundaries. Continuing to assume that the observations can be modeled as Poisson counts, the SE model replaces (2) with

$$\boldsymbol{\phi}^S \mid \boldsymbol{\phi}^E, \tau_\phi \propto \exp \left\{ -\frac{\tau_\phi}{2} \sum_{i \sim j} (1 - \phi_{ij}^E) (\phi_i^S - \phi_j^S)^2 \right\}, \quad (3)$$

$$\text{and } \boldsymbol{\phi}^E \propto \exp \left\{ -\nu \sum_{ij \sim kl} \phi_{ij}^E \phi_{kl}^E \right\}, \quad (4)$$

where  $\phi_i^S \in \mathfrak{R}$  and  $\phi_{ij}^E \in \{0, 1\}$ . The conditional distribution in (3) is IAR, with the  $(1 - \phi_{ij}^E)$  playing the roles of the  $w_{ij}$  in (2). That is,  $\phi_{ij}^E = 1$  if edge  $(i, j)$  is a boundary element, and 0 otherwise, so smoothing of neighboring site effects  $\phi_i^S$  and  $\phi_j^S$  is only encouraged if there is no boundary between them. The prior for  $\boldsymbol{\phi}^E$  in (4) is an *Ising* model with “binding strength” parameter  $\nu$  (Geman and Geman, 1984, p.725). Smaller (or even negative) values of  $\nu$  lead to more connected boundary elements, hence more separated areal units. Finally, edges  $ij$  and  $kl$  are considered adjacent ( $ij \sim kl$ ) if they are connected on the map. Ma et al. (2006) use this approach to find Medicare service area boundaries for two competing hospice systems headquartered in Duluth, Minnesota, but their work did not address potential dependence between the two hospices.

## 2.2 MCAR models for multivariate areal data

The primary challenge in multivariate areal data modeling is formulating versatile and practical probability distributions that sensibly model correlations both between and within

areas. Because of the relatively complex model and limited information provided by data, various simplifying assumptions (e.g., separability of the two types of correlation) are often made. In this subsection, we give a brief review of two MCAR models that are readily fit in WinBUGS, illustrating in the case of  $p = 2$  diseases for simplicity.

The MIAR distribution is a direct multivariate extension of the IAR model. As in the univariate case, the MIAR is improper and thus can only be used as a random effect distribution, not a likelihood. Let areal random effects corresponding to the two diseases be  $\Phi = (\phi'_1, \phi'_2)$ , where  $\phi'_1 = (\phi_{11}, \dots, \phi_{n1})$ ,  $\phi'_2 = (\phi_{12}, \dots, \phi_{n2})$ , and  $n$  is again the number of areal units. Under the MIAR model, the multivariate joint distribution is defined as  $p(\Phi) \propto \exp \{-1/2\Phi'[\Lambda \otimes (D - W)]\Phi\}$ , where  $\Lambda$  is  $2 \times 2$  and positive definite, and  $\otimes$  denotes the Kronecker product. This corresponds to the conditional distribution

$$\begin{pmatrix} \phi_{i1} \\ \phi_{i2} \end{pmatrix} \Big| \phi_{-(i1,i2)} \sim N \left( \begin{pmatrix} \bar{\phi}_{i1} \\ \bar{\phi}_{i2} \end{pmatrix}, (w_{i+}\Lambda)^{-1} \right), \quad (5)$$

where  $\phi_{-(i1,i2)}$  stands for the collection of all  $\phi_{ij}$  except  $\phi_{i1}$  and  $\phi_{i2}$ . Let  $\bar{\phi}_{i1} = \sum_j w_{ij}\phi_{j1}/w_{i+}$  and  $\bar{\phi}_{i2} = \sum_j w_{ij}\phi_{j2}/w_{i+}$ , the averages of the random effects for area  $i$ 's neighbors specific to variables 1 and 2, respectively. It can be seen that  $\Lambda$  serves as scaled conditional precision for  $(\phi_{i1}, \phi_{i2})$ , where  $w_{i+}$  is a scale parameter. Areas with more neighbors have higher precision. Since  $\Lambda$  is common for all areas  $i = 1, \dots, n$ , it controls the conditional precision for each pair of variables at the same site averaged over all areas. Letting  $\Sigma = \Lambda^{-1}$ ,  $\frac{1}{w_{i+}}\Sigma$  is the conditional covariance matrix with  $\rho_{12} = \sigma_{12}/\sqrt{\sigma_{11}\sigma_{22}}$  as the conditional correlation between  $\phi_{i1}$  and  $\phi_{i2}$ ,  $i = 1, \dots, n$ . The implementation of the same conditional correlation across all areas units facilitates computation, but can be too restrictive; we may want to borrow more information (enforce stronger correlations) across areas where we have less data.

Jin et al. (2005) propose a generalized MCAR (GMCAR) model that formulates the joint

distribution for a multivariate MRF by specifying simpler conditional and marginal models. The GMCAR is constructed as  $p(\boldsymbol{\phi}_1, \boldsymbol{\phi}_2) = p(\boldsymbol{\phi}_1|\boldsymbol{\phi}_2)p(\boldsymbol{\phi}_2)$ , where  $\boldsymbol{\phi}_1|\boldsymbol{\phi}_2 \sim N(A\boldsymbol{\phi}_2, [(D - \alpha_1 W)\tau_1]^{-1})$  and  $\boldsymbol{\phi}_2 \sim N(\mathbf{0}, [(D - \alpha_2 W)\tau_2]^{-1})$  are both proper CAR with  $0 < \alpha_1 < 1$  and  $0 < \alpha_2 < 1$ . As  $E(\boldsymbol{\phi}_1|\boldsymbol{\phi}_2) = A\boldsymbol{\phi}_2$ , the  $A$  matrix is defined such that  $E(\boldsymbol{\phi}_1|\boldsymbol{\phi}_2) = (\eta_0 I + \eta_1 W)\boldsymbol{\phi}_2$ , where  $\eta_0$  is a “bridging parameter” associating  $\phi_{i1}$  with  $\phi_{i2}$ , and  $\eta_1$  is that associating  $\phi_{i1}$  with  $\phi_{j2}$ ,  $i \sim j$ . Since the conditional and marginal distributions are both proper CAR, this model can be fit in WinBUGS as well using the `car.proper` function. However, choosing the order in which the random effects for different variables enter the model is a thorny practical issue. Note also that  $\alpha_1, \tau_1$  are *conditional* spatial autocorrelation and precision parameters, while  $\alpha_2, \tau_2$  instead pertain to a *marginal* distribution, complicating hyperprior selection for these parameters. Another drawback of this formulation is that extension to  $p > 2$  variables is increasingly awkward, since there are  $p!$  potential orderings.

### 2.3 Spatial factor analysis

Often it is scientifically interesting to find whether multiple diseases share common underlying factors, which are often understood as a mixture of shared risk factors, socioeconomic status, and so on. An example is the shared component model of Knorr-Held and Best (2001), illustrated in the WinBUGS user manual (2004). The model partitions the geographical variation in two diseases into a common (shared) component  $\boldsymbol{\theta}$ , and two disease-specific (residual) components  $\boldsymbol{\psi}_1$  and  $\boldsymbol{\psi}_2$ . Again assuming the death counts  $Y_{ik}$  for disease  $k$  in area  $i$  to be independent Poisson variables with mean  $\mu_{ik}$ ,  $k = 1, 2$ , they model

$$\log(\mu_{ik}) = \log E_{ik} + \mathbf{x}'_{ik} \boldsymbol{\beta}_k + \delta_k \theta_i + \psi_{ik} , \quad (6)$$

where  $E_{ik}$  are the expected counts, and the scaling parameters  $\delta_k$  allow different “risk gradients” for different diseases. The covariates  $\mathbf{x}_{ik}$  are the multiple disease analog of those in (1); if unavailable, the  $\beta_k$  become univariate intercept parameters  $\beta_k$ . Convolution priors (Besag et al., 1991) are given to the three components  $\theta$ ,  $\psi_1$ , and  $\psi_2$ , expressing each areal random effect as a sum of a spatially structured random effect (usually CAR) and a non-spatial *i.i.d.* white noise random effect. Best et al. (2005) compared a shared component model with convolution priors and a multivariate convolution model using an MIAR prior for the random effects, and showed the former delivered superior simulated DIC performance.

The advantage of the shared component model is that the common spatial pattern of the two diseases can be extracted, while disease-specific spatial residuals are still available for further investigation. This model described above is a special case of *spatial factor analysis*. Wang and Wall (2003) proposed a generalized spatial factor model for the relationship between several observed variables and a single spatially distributed latent factor. Hogan and Tchernis (2004) used essentially this model to quantify material deprivation in Rhode Island census tracts. Liu et al. (2005) offered a generalized SEM framework for regression analysis using latent and manifest (observed) variables while accounting for spatial correlation.

### 3 Multivariate boundary analysis

In this section, we develop several multivariate boundary analysis methods using the building blocks discussed in the previous section. We start with a multivariate extension of the LC method. Assuming  $p$  rare diseases, we again use a conditionally independent Poisson likelihood to model  $Y_{ik}$ , the observed count for disease  $k$  in area  $i$ , where  $i = 1, \dots, n$  and  $k = 1, \dots, p$ . Again  $E_{ik}$  denotes the expected count, assumed fixed and known and typically internally standardized (Banerjee et al., 2004, p.158) by disease. We arrange the

areal random effects as  $\Phi = (\phi_{1,1}, \dots, \phi_{n,1}, \phi_{1,2}, \dots, \phi_{n,2}, \phi_{1,p}, \dots, \phi_{n,p})'$ , and write

$$Y_{ik} | \mu_{ik} \sim \text{Poisson}(\mu_{ik}), \text{ with } \log \mu_{ik} = \log E_{ik} + \mathbf{x}'_{ik} \boldsymbol{\beta}_k + \phi_{ik}, \quad (7)$$

$$\text{and } \Phi | W, \Lambda \propto \exp \left\{ -\frac{1}{2} \Phi' \Lambda \otimes (D - W) \Phi \right\}, \quad (8)$$

where  $\Lambda$  is  $p \times p$  and positive definite. The spatial and inter-variable correlations are captured by the MIAR distribution in (8). Following Ma et al. (2006), the BLVs for disease  $k$  can be defined as  $\Delta_{\eta,ij,k} = \eta_{ik} - \eta_{jk}$ , the signed relative risk difference. Disease-specific boundaries can thus be constructed based on posterior summaries of the  $\Delta_{\eta,ij,k}$ , as in the univariate case. On the other hand, if we are interested in a single set of composite boundaries, we need to define a set of BLVs that represent the overall change for all diseases under study. One possible choice is  $\Delta_{\bar{\eta},ij} = \frac{1}{p} \sum_{k=1}^p (\eta_{ik} - \eta_{jk})$ , i.e., the composite BLV is defined as the average of the disease-specific BLVs. Other definitions may be appropriate depending on the scientific problem under study. The model is completed by choosing vague priors for the hyperparameters; for example, a Wishart prior is usually chosen for  $\Lambda$ , independent vague Gaussian (or possibly flat) priors are used for the  $\boldsymbol{\beta}_k$ , and so on.

The GMCAR model adopts the approach described in Section 2.2, instead of directly specifying the joint prior distribution of the multivariate areal random effects  $\Phi$ . Specifically, we define the prior distribution for the three sets of areal random effects in our dataset by conditioning sequentially as  $p(\boldsymbol{\phi}_1, \boldsymbol{\phi}_2, \boldsymbol{\phi}_3) = p(\boldsymbol{\phi}_2 | \boldsymbol{\phi}_1, \boldsymbol{\phi}_3) p(\boldsymbol{\phi}_1 | \boldsymbol{\phi}_3) p(\boldsymbol{\phi}_3)$ . In the context of our Figure 1 data, this conditioning order (3, then 1, then 2) translates to modeling lung cancer, followed by esophagus cancer given lung cancer, and finally larynx cancer given the other two. This particular order is chosen so that  $\boldsymbol{\phi}$  for diseases with less information are

conditioned on those with more information, and produces

$$\begin{aligned}
p(\boldsymbol{\phi}_3) &\sim N(\mathbf{0}, [(D - \alpha_3 W)\tau_3]^{-1}), \\
p(\boldsymbol{\phi}_1|\boldsymbol{\phi}_3) &\sim N(A\boldsymbol{\phi}_3, [(D - \alpha_1 W)\tau_1]^{-1}), \\
\text{and } p(\boldsymbol{\phi}_2|\boldsymbol{\phi}_1, \boldsymbol{\phi}_3) &\sim N\left(B\begin{pmatrix} \boldsymbol{\phi}_1 \\ \boldsymbol{\phi}_3 \end{pmatrix}, [(D - \alpha_2 W)\tau_2]^{-1}\right),
\end{aligned} \tag{9}$$

where  $|\alpha_1| < 1$ ,  $|\alpha_2| < 1$ , and  $|\alpha_3| < 1$ . Here,  $A$  is  $n \times n$  while  $B$  is  $n \times 2n$ . Following Jin et al. (2005), we set  $A = \xi_1 I + \xi_2 W$  and  $B = [\xi_3 I + \xi_4 W, \xi_5 I + \xi_6 W]$ , where  $I$  is the  $n \times n$  identity matrix. Thus  $\boldsymbol{\xi} = (\xi_1, \dots, \xi_6)'$  are ‘‘bridging’’ parameters, where  $\xi_1$ ,  $\xi_3$ , and  $\xi_5$  associate pairs of areal random effects defined on the same area unit, while  $\xi_2$ ,  $\xi_4$ , and  $\xi_6$  associate areal random effects among neighboring units. Thus  $\xi_1$  and  $\xi_2$  pertain to diseases 1 and 3 (esophagus and lung),  $\xi_3$  and  $\xi_4$  to diseases 1 and 2 (esophagus and larynx), and finally  $\xi_5$  and  $\xi_6$  to diseases 2 and 3 (larynx and lung). It is important to remember that, like the  $\alpha$  and  $\tau$  parameters, the  $\boldsymbol{\xi}$  are defined on different conditioning levels, and thus cannot be directly compared. Note that the three equations given in (9) are proper CAR models, and hence the joint prior distribution is valid as long as all three of the distributions in (9) are valid. Boundary analysis may then proceed as before, with individual boundaries based on the posterior distributions of the  $\Delta_{\eta,ij,k}$ , and composite boundaries perhaps based on the  $\Delta_{\bar{\eta},ij}$  posteriors.

Turning to shared-component boundary analysis, we use the above Poisson likelihood and the link function given in (6), but slightly different priors than in Subsection 2.3, namely

$$\boldsymbol{\theta}|\tau_s \sim CAR(\tau_s, W), \tag{10}$$

$$\text{and } \psi_{ik}|\tau_{\psi_k} \stackrel{iid}{\sim} N(0, 1/\tau_{\psi_k}). \tag{11}$$

Boundaries for the latent factor  $\boldsymbol{\theta}$  identify edges of abrupt change that are common to all of the diseases. Note that (11) assumes the residuals  $\psi_{ik}$  are independent across both regions and diseases. We could easily generalize to independent CAR models (i.e.,  $\boldsymbol{\psi}_k | \tau_{\psi_k}, W \overset{ind}{\sim} CAR(\tau_{\psi_k}, W)$ ) or even use an MCAR for  $\boldsymbol{\psi} = (\boldsymbol{\psi}'_1, \dots, \boldsymbol{\psi}'_k)'$ , although our inclusion of the shared component  $\boldsymbol{\theta}$  in (10) may well make such complexity unnecessary. Defining the “shared” BLV for edge  $ij$  as  $\Delta_{\theta,ij} = \theta_i - \theta_j$ , boundary elements can be identified based on posterior summaries of  $\Delta_{\theta,ij}$  (say,  $E(\Delta_{\theta,ij} | \mathbf{y})$  or their magnitudes). Disease-specific residual boundaries (separating areas that differ due to factors other than the spatial common factor) can analogously be based on  $\Delta_{\psi,ij,k} = \psi_{ik} - \psi_{jk}$  for each disease  $k$ . We hasten to add that, as in all factor analysis modeling, constraints need to be imposed on  $\delta$  to avoid identifiability problems. For  $p = 2$ , Knorr-Held and Best (2001) set  $\delta_1 = 1/\delta_2$  and place a lognormal prior on  $\delta_1$ ; Held et al. (2005) generalize this approach for  $p > 2$ . We instead follow Liu et al. (2005) and fix  $\delta_p = 1$ , but leave the remaining  $\delta_k$  unconstrained.

To avoid the oversmoothing often associated with the CAR model in boundary analysis, we can allow the areal adjacency matrix for the common factor to be random. Boundary analysis is straightforward using this approach: we simply replace the prior in (10) with the random adjacency CAR given in (3) and (4). That is, we assume  $\boldsymbol{\theta} | \boldsymbol{\phi}^E, \tau_s \sim CAR(\tau_s, W)$ , where the  $\phi_{ij}^E = 1 - w_{ij}$  have the Ising prior given in (4). In other words,

$$\boldsymbol{\phi}^E \propto \exp\{-\nu \boldsymbol{\phi}^{E'} W^* \boldsymbol{\phi}^E\}, \quad (12)$$

where  $W^*$  is the edge-space adjacency matrix (i.e.,  $W_{mm'}^* = 1$  if edge  $m \equiv ij$  is connected to edge  $m' \equiv kl$ ). We refer to this model as SESHARED. Alternatively, allowing the adjacency matrix in an MIAR distribution to be random can be achieved by assigning the Ising prior (12) to the  $\phi_{ij}^E$  in (8). We refer to this as the SEMIAR model. Note that the overall boundary

map could now be based on posterior summaries (say, posterior means) of the edge random effects  $\phi^E$  themselves, if desired.

It is computationally difficult to specify a hyperprior for  $\nu$  and estimate it from the data. This is because its full conditional under (12) requires evaluation of  $\sum \exp(-\nu \sum_{ij \sim kl} \phi_{ij}^E \phi_{kl}^E)$ , where the first summation is over all possible 0-1 statuses of the adjacent edges ( $2^{216}$  possibilities in our 216-edge setting). For a given  $\nu$ , there will be at most  $m + 1$  unique values for this sum, where  $m$  is the number of edge adjacencies (406 in our map). Evaluation of this sum may be possible for a smaller map, but even then, acceptable convergence of (say) Metropolis steps for  $\nu$  is very much in doubt. As a result, we follow previous convention and think of  $\nu$  as a tuning constant, sometimes comparing results under a few plausible values before making a final decision. Even with  $\nu$  fixed, evaluation of this model is complicated by the need to include the proper normalizing constant in the full conditional for  $\theta$  (since  $W$  in this distribution depends on  $\phi^E$ , which is random). In this paper, we follow Ma et al. (2006) and include  $\tau_s^{G/2} |D - W|_+$  as this normalizing constant, where  $G$  is the rank of  $(D - W)$  and  $|D - W|_+$  is the product of its nonnegative eigenvalues.

Finally, we note an interesting connection between the GMCAR and shared component models. The latter’s link function (6) can be written as  $\log(\mu_k) = \log \mathbf{E}_k + \mathbf{X}_k \beta_k + \delta_k' \theta + \psi_k$ , where  $\delta_k = \delta_k \mathbf{1}_{n \times 1}$ . On the other hand, from the conditional specification of GMCAR given in (9), we have  $\phi_1 = A \phi_3 + \epsilon_1$ , where  $\epsilon_1 \sim N(\mathbf{0}, [(D - \alpha_1 W) \tau_1]^{-1})$  and  $\phi_2 = B \begin{pmatrix} \phi_1 \\ \phi_3 \end{pmatrix} + \epsilon_2$ , where  $\epsilon_2 \sim N(\mathbf{0}, [(D - \alpha_2 W) \tau_2]^{-1})$ . Plugging these into (7), we get  $\log \mu_k = \log \mathbf{E}_k + \mathbf{X}_k \beta_k + C_k \mathbf{f} + \epsilon_k$ , where  $C_1 = [\mathbf{0}_{n \times n}, A]$ ,  $C_2 = B$  and  $C_3 = [\mathbf{0}_{n \times n}, I_{n \times n}]$  are all  $n \times 2n$  matrices, for  $\mathbf{f} = \begin{pmatrix} \phi_1 \\ \phi_3 \end{pmatrix}$  and  $\epsilon_3 = \mathbf{0}_{n \times 1}$ . Thus  $\mathbf{f} = \begin{pmatrix} \phi_1 \\ \phi_3 \end{pmatrix}$  is analogous to the latent factor vector  $\theta$  in the shared component model, while the  $C_k$  are playing the role of the factor loadings  $\delta_k$ . Of course, there are differences; the GMCAR has fewer areal random effects ( $\epsilon_1, \epsilon_2, \phi_3$ ) than the shared component ( $\psi_1, \psi_2, \psi_3, \theta$ ), and six “factor loading” parameters ( $\xi_1, \dots, \xi_6$ ) instead of just

two  $(\delta_1, \delta_2)$ . The use of proper CAR specifications within either or both of these frameworks would also alter their properties, and thus how similarly they behave.

## 4 Data analysis

### 4.1 Multivariate areal wombling model selection

In this section, we illustrate and compare our various multivariate boundary analysis techniques in the context of the SEER three cancer dataset introduced in Section 1 and mapped in Figure 1. We begin by selecting the best model. Here one often thinks of cross-validatory or other out-of-sample forecasting methods, but they are less natural for areal data due to the interconnected and often irregular nature of the spatial lattice. The underlying map with its adjacency structure breaks down as soon as we decide to leave out one or more regions. Leaving regions out will not only result in a loss of information, but more importantly may lead to “islands” or other components in the map for which implementing the model itself will be difficult.

As such, we instead turn to DIC, an extension of the Akaike information criterion (AIC) that reflects both goodness of fit and complexity of hierarchical models. This criterion is based upon the *deviance* statistic,  $D(\boldsymbol{\theta}) = -2 \log f(\mathbf{y}|\boldsymbol{\theta}) + 2 \log h(\mathbf{y})$ , where  $\boldsymbol{\theta}$  is the collection of parameters in the model,  $f(\mathbf{y}|\boldsymbol{\theta})$  is the likelihood, and  $h(\mathbf{y})$  is any standardizing function of the data alone. The DIC is then defined as  $DIC = \bar{D} + p_D$ , where  $\bar{D} = E(D(\boldsymbol{\theta})|\mathbf{y})$  is the posterior mean deviance, and  $p_D$  is the effective number of parameters in the model (i.e., a count that is typically less than the actual number of parameters due to the shrinkage of random effects toward their own grand mean). Using an asymptotic normal approximation to the posterior, Spiegelhalter et al. (2002) show that  $p_D$  is sensibly defined as  $p_D = \bar{D} - D(\hat{\boldsymbol{\theta}})$ , where  $\hat{\boldsymbol{\theta}}$  is a suitable “plug-in” estimate of  $\boldsymbol{\theta}$  (say, the posterior mean). DIC can be interpreted approximately as the expected posterior loss in prediction when adopting

a particular model. Thus using DIC to compare the various boundary analysis models is reasonable, since we are interested in selecting the model that leads to the best prediction of the boundary elements in the study region. Models with smaller DIC values are preferred.

We begin with a comparison of several neighborhood structures for our data. Table 1 gives  $p_D$  and DIC values for several of our simpler models using different forms of the weight matrix  $W$ . All models in this table were fit using WinBUGS. The first column of the table indicates the spatial model. IndCAR refers to the model that fits three independent CAR models; i.e., it assumes independence between the diseases, but spatial association within each disease. Next, MIAR refers to the multivariate IAR model given in (7) and (8). SHARED-G is described in (10) and (11), while SHARED-C replaces (11) with independent CAR priors for the  $\psi_k$ , i.e.,  $\psi_k | \tau_{\psi_k}, W \stackrel{ind}{\sim} CAR(\tau_{\psi_k}, W)$ . We adopt independent gamma priors with mean 1 and variance 100 for all the precision parameters, flat priors for the intercepts  $\beta$ , and  $N(0, 10^2)$  priors for the loading factors  $\delta_1$  and  $\delta_2$  (recall  $\delta_3$  is fixed at 1).

The second column of Table 1 indicates the proximity matrix  $W$  used in each CAR model. Here we compare five possibilities, the first of which is the usual 0/1 adjacency structure. Next are two structures for which we let  $w_{ij} = \exp(-\omega d_{ij})$  if areas  $i$  and  $j$  share a common boundary, and 0 otherwise. Here  $d_{ij}$  is the distance between the centroids of the two counties, and  $\omega$  is a spatial decay parameter. This neighborhood structure may be more appropriate if neighboring areas are sometimes of very different sizes, and we want to downweight the association of adjacent but large areas. Choosing small  $\omega > 0$  leads to a proximity matrix similar to the 0/1 matrix. Since WinBUGS will not permit unknown  $\omega$ , we show results for two fixed choices of  $\omega$ , 0.3 and 1. Finally, we consider  $w_{ij} = \exp(-\omega d_{ij})$  for *all* county pairs. This neighborhood structure allows areas that are not physically connected to be neighbors, while still yielding  $W$  elements very close to 0 for widely separated county pairs. Table 1 indicates this assumption is not well supported by our data. Combining the adjacency and

model	$W$ matrix	$\overline{D}(\boldsymbol{\theta})$	$D(\widehat{\boldsymbol{\theta}})$	$p_D$	DIC
IndCAR	0/1	1472.35	1398.53	73.82	1546.16
	$0/e^{-0.3d_{ij}}$	1472.56	1398.55	74.01	1546.57
	$0/e^{-d_{ij}}$	1470.69	1395.00	75.70	1546.37
	$e^{-0.3d_{ij}}$	1472.45	1375.28	97.17	1569.62
	$e^{-d_{ij}}$	1467.49	1372.80	94.70	1562.19
MIAR	0/1	1461.21	1385.06	76.16	1537.37
	$0/e^{-0.3d_{ij}}$	1481.88	1426.39	55.49	1537.36
	$0/e^{-d_{ij}}$	1505.15	1463.36	41.79	1546.94
	$e^{-0.3d_{ij}}$	1769.84	1759.01	10.83	1780.66
	$e^{-d_{ij}}$	1756.40	1746.61	9.785	1766.18
SHARED-G	0/1	1454.43	1373.86	80.57	1535.00
	$0/e^{-0.3d_{ij}}$	1455.16	1375.35	79.81	1534.97
	$0/e^{-d_{ij}}$	1454.15	1373.32	80.83	1534.98
	$e^{-0.3d_{ij}}$	1453.11	1353.90	99.21	1552.32
	$e^{-d_{ij}}$	1454.19	1361.68	92.51	1546.70
SHARED-C	0/1	1464.77	1393.52	71.25	1536.03
	$0/e^{-0.3d_{ij}}$	1463.30	1390.26	73.04	1536.34
	$0/e^{-d_{ij}}$	1462.48	1387.37	75.11	1537.59
	$e^{-0.3d_{ij}}$	1465.16	1378.02	87.15	1552.31
	$e^{-d_{ij}}$	1462.30	1376.16	86.14	1548.43

Table 1: Comparison of  $p_D$  and DIC values for multivariate boundary analysis models with different neighborhood structures, SEER esophagus, larynx and lung cancer data.

distance information (as the  $0/e^{-\omega d_{ij}}$  matrices do) leads to better results, but the simple 0/1 choice typically performs quite competitively. As such, we adopt the usual 0/1 proximity matrix for all subsequent models.

Next, we investigate the impact of different “binding strength” parameters  $\nu$  on our SE models. Table 2 gives DIC and related summaries for three models and three choices of  $\nu$ . Here, IndSE denotes the model that fits separate SE models for every variable. The SEMIAR (site-edge MIAR) model is the one described near equation (12), while the SESHARED model adds an extra layer of random edge effects  $\boldsymbol{\phi}^E$  onto SHARED-G by replacing (10) with (3) and (4). This table indicates different binding strengths do not significantly affect the trade-offs between goodness of fit (expressed as  $\overline{D}(\boldsymbol{\theta})$ ) and model complexity (expressed as  $p_D$ ). However, there is some preference for smaller  $\nu$  in the shared component model,

model	binding strength	$\overline{D}(\boldsymbol{\theta})$	$D(\widehat{\boldsymbol{\theta}})$	$p_D$	DIC
IndSE	$\nu = 0.1$	1522.50	1481.36	41.14	1563.64
	$\nu = 1$	1488.95	1433.36	55.59	1544.54
	$\nu = 100$	1478.47	1414.30	64.17	1542.64
SEMIAR	$\nu = 0.1$	1460.93	1380.47	80.46	1541.39
	$\nu = 1$	1464.10	1392.27	71.84	1535.94
	$\nu = 100$	1460.51	1386.07	74.44	1534.94
SESHARED	$\nu = 0.1$	1447.75	1362.89	84.87	1532.62
	$\nu = 1$	1449.22	1364.50	84.72	1533.94
	$\nu = 100$	1449.88	1361.40	88.47	1538.35

Table 2: Comparison of  $p_D$  and DIC values for multivariate boundary analysis models with different binding strength parameters  $\nu$  for Ising prior, SEER esophagus, larynx and lung cancer data.

while larger  $\nu$  fare better in the other two models. This difference is not surprising, since in the former case  $\nu$  directly affects only the underlying factor, whereas in the latter two cases it affects all of the disease-specific spatial random effects.

Table 3 gives a summarizing DIC comparison for our various areal wobbling models. All models used the conditionally independent Poisson likelihood, 0/1 adjacency matrix, and feature only an intercept  $\beta_k$  (no covariates  $\mathbf{x}_{ik}$ ). The models in the first two blocks of rows in the table follow the log link function given in (7). The models in the third block of rows instead use the shared component link function (6), which requires estimation of the unknown factor loadings  $\boldsymbol{\delta}$ . The remaining differences across models lie mainly in their random effect specifications, as we now describe.

The first block of rows in the table shows DIC values for models that do not account for spatial and inter-variable correlations simultaneously. For example, in the IID model, the areal random effects  $\phi_{ik}^S$  are assumed to follow an *i.i.d.* Gaussian distribution, which models neither the spatial nor the inter-variable correlations. In the MultiIID model, inter-variable correlations are modeled by assigning a multivariate normal prior to the random effects at the same areal unit, i.e., replacing (8) by  $\boldsymbol{\phi}_i^S \stackrel{iid}{\sim} N(\mathbf{0}, \Lambda^{-1})$ , where  $\boldsymbol{\phi}_i^S = (\phi_{i1}^S, \dots, \phi_{ip}^S)'$  is the set of areal random effects for different variables over areal unit  $i$ . Spatial correlations are thus

model	$\overline{D}(\boldsymbol{\theta})$	$D(\widehat{\boldsymbol{\theta}})$	$p_D$	DIC
IID	1464.45	1365.64	98.81	1563.26
MultIID	1445.12	1338.09	107.13	1552.36
IndCAR	1472.35	1398.53	73.82	1546.16
IndSE	1478.47	1414.30	64.17	1542.64
MIAR	1461.21	1385.06	76.16	1537.37
GMCAR	1463.10	1392.20	70.74	1533.84
SEMIAR	1460.51	1386.07	74.44	1534.94
SHARED-G	1454.43	1373.86	80.57	1535.00
SHARED-C	1467.77	1393.52	71.25	1536.03
SESHARED	1447.75	1362.89	84.87	1532.62

Table 3: Comparison of  $p_D$  and DIC values for multivariate boundary analysis models, SEER esophagus, larynx and lung cancer data.

ignored by this model. Conversely, the IndCAR model ignores inter-variable (in our case, cross-cancer) correlations, assigning independent CAR priors to the areal random effects for different variables. That is, we let  $\phi_k^S \overset{ind}{\sim} CAR(\tau_{\phi_k}, W)$ , where  $\phi_k^S \equiv (\phi_{1k}^S, \dots, \phi_{nk}^S)'$  is the set of areal random effects for variable  $k$ . In a similar vein, the IndSE model fits separate SE models for every variable; this is the approach of Ma et al. (2006).

For the hyperparameters of the models in the first row block, vague conjugate prior distributions are adopted, including flat priors for the intercepts, Gamma(mean 1, variance 100) priors for the precision parameters in the IID and IndCAR models, and a Wishart( $R, r$ ) prior for the  $p$ -dimensional precision matrix  $\Lambda$  in the MultiIID model. We chose  $r = p$  and selected  $R$  to be diagonal (cancer counts at same locations are independent a priori), where the prior means for the diagonal entries of  $\Lambda$  were taken as 15 with variance 150.

For the models in the second block of Table 3's rows, the MIAR is the model given in (7) and (8), while the SEMIAR (site-edge MIAR) model is described near equation (12). For both models, we use the same Wishart prior for  $\Lambda$  as used above, and set the binding strength parameter  $\nu = 1$  in SEMIAR, based on the results of Table 2. The GMCAR model adopts the generalized MCAR approach using the structure and disease conditioning order

given in (9). For this model, we assign  $N(0, 10^2I)$  priors to  $\xi$ , flat priors to the intercepts  $\beta$ , independent gamma priors with mean 1 and variance 100 to  $\tau_1, \tau_2$ , and  $\tau_3$ , and independent  $Uniform(0, 0.99)$  priors to  $\alpha_1, \alpha_2$ , and  $\alpha_3$ . Finally, the models in the third row block of Table 3 are the shared component (spatial factor analysis) models SHARED-G, SHARED-C, and SESHARED. In the latter case, we alter the binding strength parameter in (4) to  $\nu = 0.1$ , as suggested by Table 2; all priors remain as specified earlier.

All of the models except IndSE, SEMIAR and SESHARED can be fit in WinBUGS or OpenBUGS. Comparing the DIC values for IID, MultiIID and IndCAR in Table 3, we see that adding spatial correlation improved DIC score more than adding inter-variable correlations. IndSE performs better than the other ‘‘Block 1’’ models, but only marginally better than IndCAR and less well than any of the multivariate spatial models. The models accounting for both spatial and inter-variable correlations perform the best. The MIAR, SEMIAR, SHARED-G and SHARED-C have very similar DIC values, while SESHARED and GMCAR (which we have seen is related to the SHARED models) have the smallest DICs of all.

The SESHARED model is relatively robust to prior changes, and its interpretation is more straightforward than that of GMCAR. Moreover, the conditioning order used by our GMCAR is partially data-based; other orders either do not perform as well or lead to poor MCMC convergence. As such, we adopt the SESHARED model for our subsequent choropleth areal and wombled boundary maps.

## 4.2 Choropleth and wombled maps for the SEER data

Previous epidemiological studies and basic anatomical relations suggest that esophagus, larynx and lung cancer are related. As we noticed in Figure 1, the raw SMRs for the three cancers show similar patterns of decrease from northeast to southwest. Thus it seems plausible to model an underlying common factor connecting the three diseases. Panel (a)

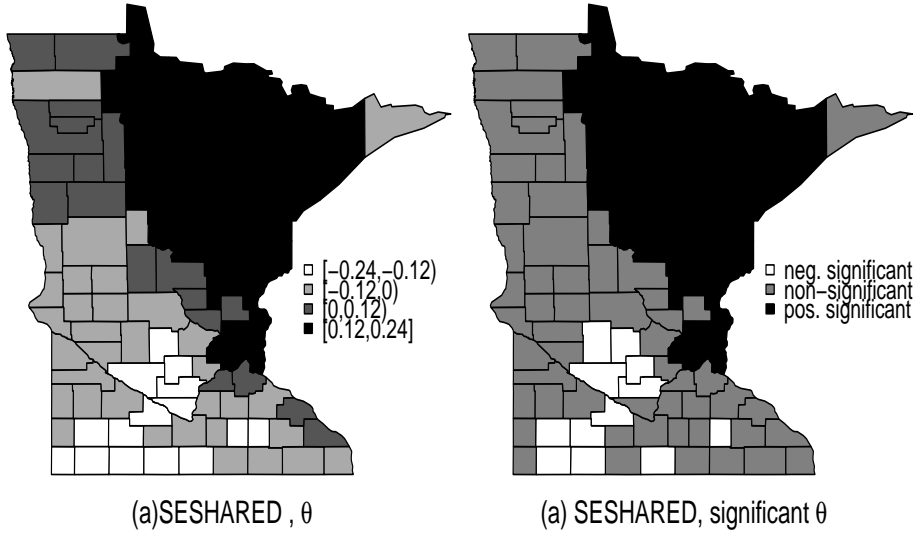


Figure 2: Maps of SESHARED model, Minnesota SEER cancer mortality data: (a) posterior medians of the spatial shared component; (b) significant/nonsignificant shared components.

of Figure 2 plots the posterior medians of the underlying shared component  $\theta$ . Panel (b) identifies the counties with significantly positive or negative  $\theta$ , i.e., those whose posterior 95% CI contains only positive or negative values. The remaining counties (those whose 95% CI includes zero) are labeled non-significant. The shared component exhibits a strong spatial pattern, with counties in northeastern Minnesota subject to higher cancer risk, and the risk decreasing as we move to the southwest. However, in the presence of the spatially structured shared component, the disease-specific residual effects  $\psi_{ik}$  are all non-significant.

The  $(\delta_1, \delta_2, \delta_3)'$  in (6) are interpreted as “log-relative risk gradients” for the different diseases. In the case of no disease-specific effects  $\psi_{ik}$ , we have  $\log(\mu_{ik}) = \log E_{ik} + \beta_k + \delta_k \theta_i$ . Since  $\delta_3 = 1$ , this means that  $\theta_i = \log(\eta_{i3}) - \beta_3$ . Thus the shared component is the log-relative risk of lung cancer, “shifted” by  $\beta_3$ . Since we internally standardize for each disease, we expect  $\beta_3 \approx 0$ , so this shift should be minimal. The posterior distributions of  $\delta_1$  and  $\delta_2$  are fairly symmetric around 0.98 and 0.92, respectively, indicating little adjustment of the shared component’s scale is needed for the other two cancers. The posterior 95% CI for  $\delta_1$  does not include 0 while that for  $\delta_2$  does, reflecting the paucity of information regarding larynx cancer in the dataset.

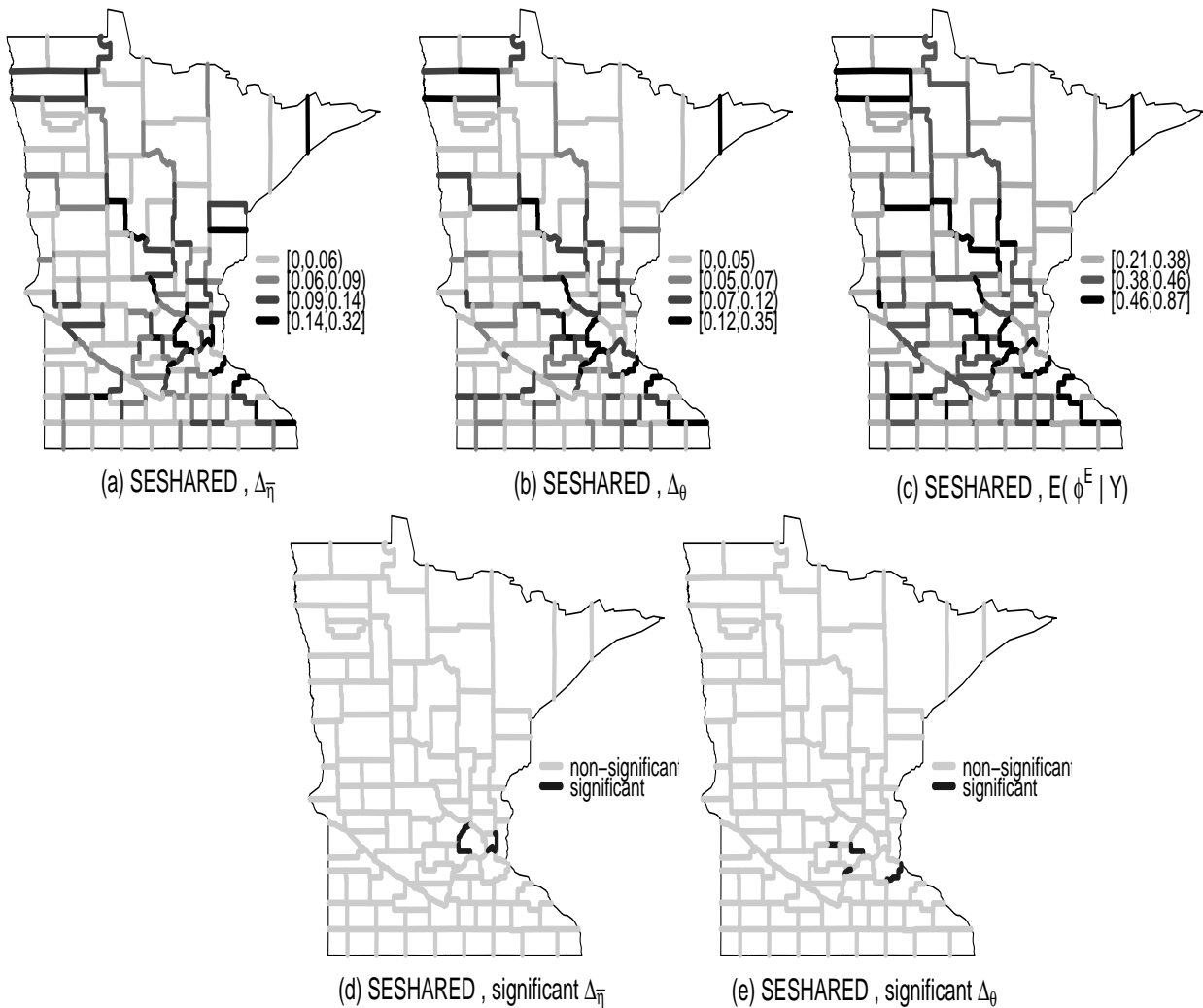


Figure 3: Composite boundary maps under the SESHARED model, Minnesota SEER data: (a) posterior medians of  $\Delta_{\bar{\eta},ij}$ ; (b) posterior medians of  $\Delta_{\theta,ij}$ ; (c) posterior means of  $\phi_{ij}^E$ ; (d) significant/nonsignificant  $\Delta_{\bar{\eta},ij}$ ; (e) significant/nonsignificant  $\Delta_{\theta,ij}$ .

The composite boundary map can be constructed based on either the shared component, or some summary of the disease-specific relative risks (which incorporate both shared component and disease-specific effect information). Panels (a), (b), and (c) of Figure 3 give composite boundaries based on  $\Delta_{\bar{\eta},ij}$ ,  $\Delta_{\theta,ij}$ , and  $\phi_{ij}^E$ , respectively. The mean-based boundaries in panels (a) and (b) are quite similar, whereas the variance-based boundaries in panel (c) show different patterns that seem less helpful for boundary detection. Panels (d) and (e) identify mean-based edges whose posterior 95% CI for the corresponding parameter excludes 0. There are not many, but both panels identify boundaries separating portions of the Twin Cities metro area from the exurban and rural regions nearby.

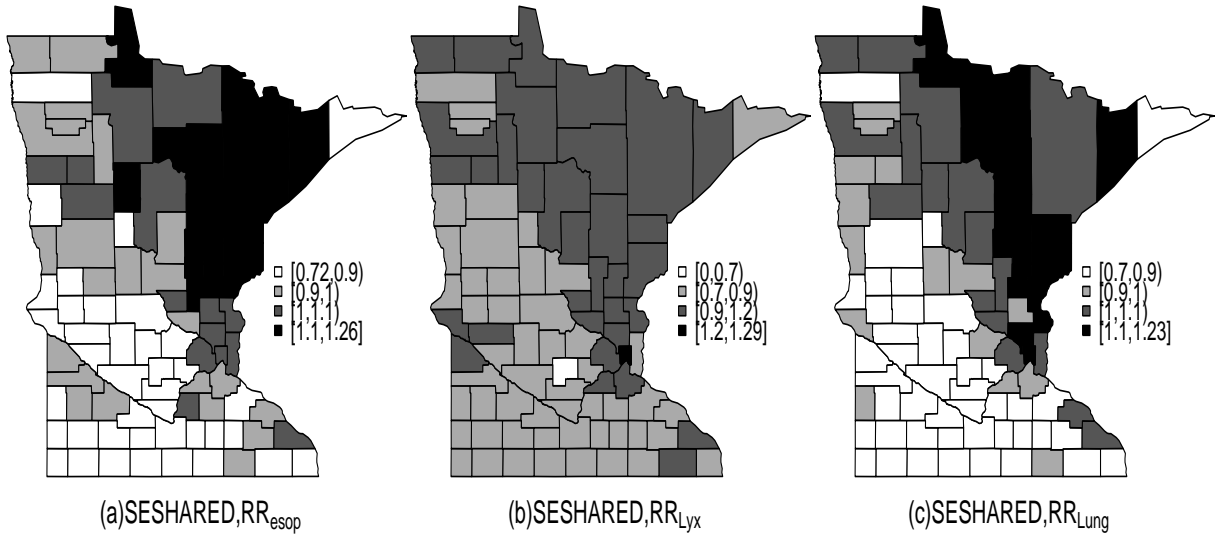


Figure 4: Disease-specific fitted relative risks  $\eta_{ik}$ , Minnesota SEER data: (a) esophagus cancer; (b) larynx cancer; (c) lung cancer.

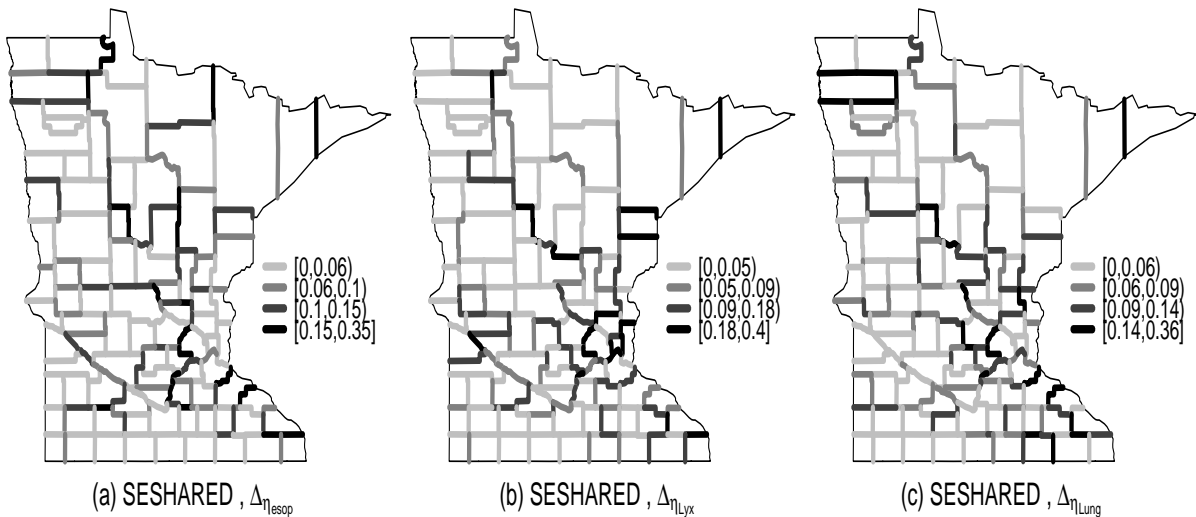


Figure 5: Disease-specific boundary maps based on posterior medians of  $\Delta_{\eta_{i,j,k}}$ , Minnesota SEER data: (a) esophagus cancer; (b) larynx cancer; (c) lung cancer.

Next, Figure 4 shows maps of fitted relative risks  $\eta_{ik} = \mu_{ik}/E_{ik}$  for the three cancers. Comparing them to the raw SMR plots shown in Figure 1, we notice the characteristic shrinkage and spatial smoothing in the Bayesian estimates. Larynx cancer is a fairly rare disease, so with so little data about it we would like to borrow information from the other two diseases, as well as across space. Figure 4(b) reflects this desired smoothing.

Using the usual BLV idea, disease-specific boundary maps can be constructed based on  $\Delta_{\eta_{i,j,k}}$  posterior summaries. These boundaries are shown in Figure 5. The boundary maps indicate similar patterns across disease, but subtle differences (say, east of the Twin Cities

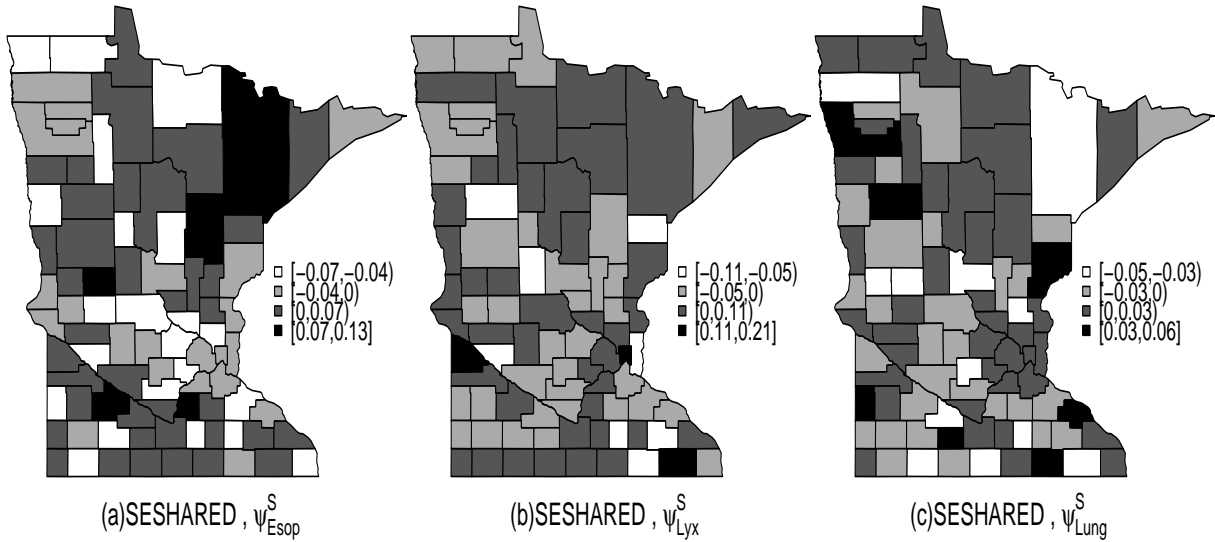


Figure 6: Disease-specific fitted residual effects  $\psi_{ik}$ , Minnesota SEER data: (a) esophagus cancer; (b) larynx cancer; (c) lung cancer.

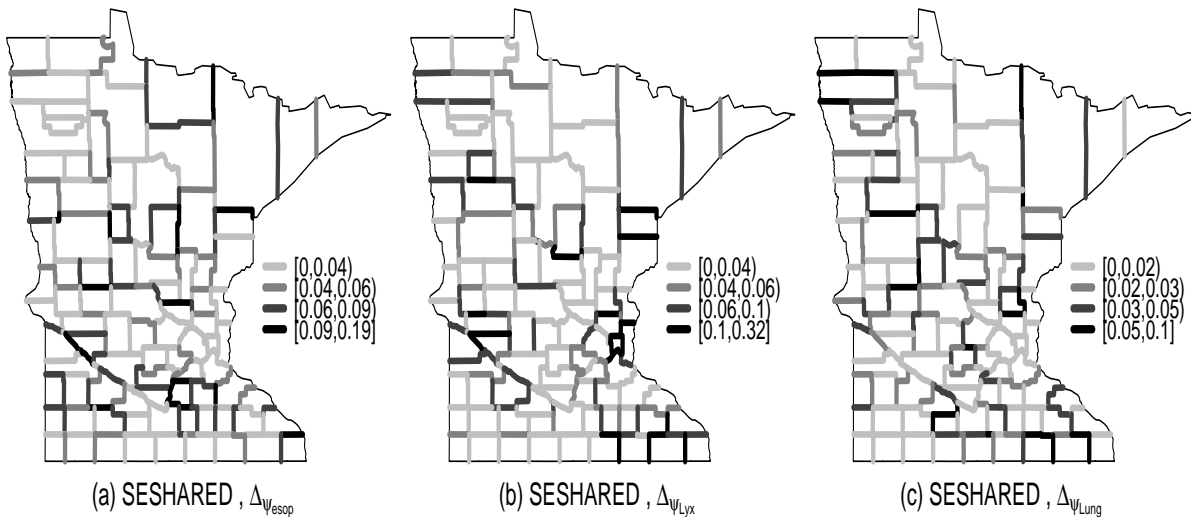


Figure 7: Disease-specific boundary maps based on posterior medians of  $\Delta_{\psi,ij,k}$ , Minnesota SEER data: (a) esophagus cancer; (b) larynx cancer; (c) lung cancer.

metro area in panel (b), the larynx map) can be appreciated.

Finally, Figure 6 shows choropleth posterior median maps of the disease-specific residuals  $\psi_{ik}$ , while Figure 7 gives corresponding mean-based boundaries. Recalling these random effects are for the most part not significantly different from 0, we regard these last plots as primarily exploratory tools. Still, Ramsey County (darkest shaded region in Figure 4(b); contains the city of St. Paul) appears to show an unusually high rate of larynx cancer. This is reflected in its Figure 6(b) shading, as well as its separation from all of its neighbors in Figure 7(b). The lung cancer rate in Hennepin County (the slightly larger county just west

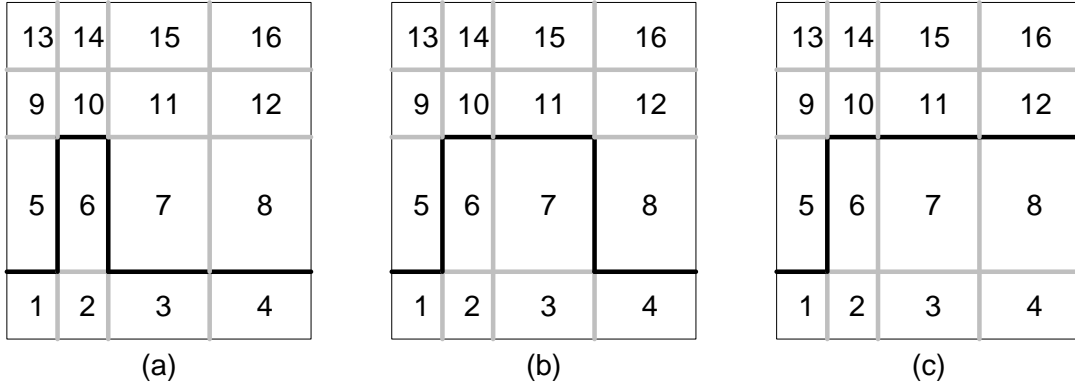


Figure 8: Areal template used for small simulation study, with assumed 0/1 adjacency structure indicated for three cancers.

of Ramsey; contains the city of Minneapolis) also appears elevated in Figure 6(c), but this appears to have less impact on the corresponding boundaries in Figure 7(c), perhaps since the data are more conclusive regarding lung cancer.

## 5 Discussion and future work

In this paper, we have proposed several models that can be adopted to carry out multivariate boundary analysis. Accounting for correlations across both space and variables can improve the modeling, both in terms of improved DIC scores and more meaningful and easily interpretable maps and other summaries of the corresponding enhanced model parameters. The SESHARED model emerged as best for our Minnesota cancer dataset, but many of the models we considered improved on the IndCAR and IndSE, two models that carefully account for spatial association but ignore correlation among the cancers.

In order to investigate the performance of our preferred SESHARED model more broadly, we carried out a brief simulation study. To control the size of the computation, we replace our Minnesota county-level grid by an idealized  $4 \times 4$  template. We generate independent Poisson counts with one mean for a set of contiguous “lower” regions, and another mean for the remaining, “upper” regions. We then create “true” boundaries that are correlated

model	$P(CS)$	$\overline{p_D}$	$\overline{DIC}$
IndCAR	.752	32.93	227.81
MultIID	.798	24.20	219.54
SESHARED	.842	23.45	215.40

Table 4: Simulated probabilities of correct selection (CS) and average  $p_D$  and DIC scores, 1000 datasets drawn over the  $4 \times 4$  regular template.

but not identical across the cancers as shown in Figure 8: we set the “lower” units to be Regions 1, 2, 3, 4, and 6 for Cancer 1, but add Region 7 to this list for Cancer 2, and further add Region 8 to this list for Cancer 3. We generate 1000 sets of artificial data, setting the two means equal to 10 and 2. We also assume equal populations in all areas, so that the internally standardized expected counts  $E_{ik}$  for disease  $k$  are all equal to the average count  $\bar{Y}_k$  over all areas.

Table 4 gives the empirical probabilities of correct selection (averaged over the 1000 simulated datasets) for the IndCAR, MultIID, and SESHARED models. (Note the first two of these models account for either spatial or inter-variable correlations, but not both.) The correct selection probability for each cancer is computed as  $m/M$ , where  $m$  is the number of the  $M$  true boundary segments identified by the model (i.e., amongst those having the top 6 BLVs), and  $M$  equals 6 for Cancers 1 and 2 but equals 5 for Cancer 3. We do this based on the absolute values of the posterior medians of the  $\Delta_{\eta,ij,k}$ , as in Section 4. The table shows SESHARED to have the largest empirical probability of correct selection averaged over the three cancers. The table also compares model fit and complexity by showing the  $p_D$  and DIC scores averaged over the 1000 simulations. The SESHARED model again emerges as best of this class, with modest complexity and the smallest average DIC score.

The southwest-to-northeast pattern evident in Figure 1 motivates a search for a suitable spatially-oriented covariate to include in our models. Sadly, while a thorough search of the U.S. Census Bureau site `quickfacts.census.gov` yielded several county-level income,

poverty, and occupational covariates, none emerged as worthy of inclusion in any of our spatial models. The proportion of each county's business establishments classified as forestry, fishing, hunting, mining, or agriculture support was significantly associated with lung cancer for some models, but even this did not lead to worthwhile improvements in DIC score.

Other specifications could also be adopted for the shared-component and the disease-specific random effects in (10) and (11). For example, we might decompose the common factor  $\theta_i$  in (6) as  $\pi\phi_i + (1 - \pi)\epsilon_i$ , where  $\phi \sim CAR$ ,  $\epsilon_i \stackrel{iid}{\sim} N(0, \sigma^2)$ , and  $\pi \sim Bernoulli(p)$  with a vague prior for  $p$ . This is a convolution model encouraged by Lawson and Clark (2002). As this model is highly overparameterized, model selection could be applied to see if modeling both spatially structured and unstructured random effects is warranted.

Our SEMIAR model assumed that the all disease-specific areal random effects share a common underlying neighborhood structure. A more general framework would be to allow each variable to have its own neighborhood structure, i.e.,

$$\begin{aligned} \Phi | \Phi^E, \Lambda &\propto \exp \left\{ -\frac{1}{2} \Phi' R(\Lambda, \Phi^E) \Phi \right\}, \\ \text{and } \phi_k^E &\propto \exp \{ -\nu_k \phi_k^{E'} W_k^* \phi_k^E \}, \quad k = 1, \dots, p, \end{aligned} \tag{13}$$

where  $\Phi^E = (\phi_1^{E'}, \dots, \phi_p^{E'})'$ ,  $\phi_{ij,k}^E = 1 - w_{ij,k}$ , and  $R(\Lambda, \Phi^E)$  denotes any suitable (typically proper) inverse covariance matrix. Here  $w_{ij,k}$  is the  $ij$  entry of  $W_k$ , the proximity matrix for variable  $k$ .

We note the IndSE model also uses a different  $W$  matrix for each disease, but does not allow modeling of any association among these matrices. Such inter-variable association can be incorporated through specification of the covariance structure of  $\Phi$ ,  $\Sigma = R^{-1}$ . Carlin and Banerjee (2003) and Gelfand and Vounatsou (2003) proposed useful techniques, but used a single, fixed neighborhood structure for all variables. In our SEMIAR setting, we could

generalize their approach, decomposing  $\Sigma$  in the 3-variable case as

$$\begin{pmatrix} R'_1 R_1 \lambda_{11} & R'_1 R_2 \lambda_{12} & R'_1 R_3 \lambda_{13} \\ R'_2 R_1 \lambda_{12} & R'_2 R_2 \lambda_{22} & R'_2 R_3 \lambda_{23} \\ R'_3 R_1 \lambda_{13} & R'_3 R_2 \lambda_{23} & R'_3 R_3 \lambda_{33} \end{pmatrix} = \begin{pmatrix} R'_1 & 0 & 0 \\ 0 & R'_2 & 0 \\ 0 & 0 & R'_3 \end{pmatrix} (\Lambda \otimes I_{n \times n}) \begin{pmatrix} R_1 & 0 & 0 \\ 0 & R_2 & 0 \\ 0 & 0 & R_3 \end{pmatrix},$$

where  $R'_k R_k = D - \alpha_k W_k$ ,  $k = 1, 2, 3$ , i.e.,  $R_k$  is the upper-triangular Cholesky decomposition of  $D - \alpha_k W_k$ . Gelfand and Vounatsou (2003) instead recommend a spectral decomposition,  $R_k = \text{Diag}(1 - \alpha_k \omega_i)^{\frac{1}{2}} P' D^{\frac{1}{2}} P$ , where the  $\omega_i$  are the eigenvalues of  $D^{-\frac{1}{2}} W_k D^{-\frac{1}{2}}$  and  $P$  is an orthogonal matrix with the corresponding eigenvectors as its columns.

Finally, multivariate areal wombling as described herein leads naturally to methods of *spatiotemporal* areal wombling, as needed to track changes in areal boundaries over time (say, for annual cancer surveillance purposes). Here the temporal units play the role of the  $p$  variables above. Space-time separability (analogous to our use of the Kronecker product in (8)) would be a natural assumption to ease conceptual and computational difficulties, but may or may not be justified for any given dataset.

## References

- Banerjee, S., Carlin, B.P., Gelfand, A.E. (2004). *Hierarchical Modeling and Analysis for Spatial Data*. Boca Raton, FL: Chapman and Hall/CRC Press.
- Baron, A.E., Franceschi, S. Barra, S. Talamini, R. and La Vecchia, C. (1993). Comparison of the joint effect of alcohol and smoking on the risk of cancer across sites in the upper aerodigestive tract. *Cancer Epidemiology Biomarkers and Prevention*, **2**, 519–523.
- Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems (with discussion). *J. Roy. Statist. Soc., Ser. B*, **36**, 192–236.

- Besag, J. and Higdon, D. (1999). Bayesian analysis of agricultural field experiments (with discussion). *J. Roy. Statist. Soc., Ser. B*, **61**, 691–746.
- Besag, J., York, J. and Mollie, A. (1991). Bayesian image restoration, with two applications in spatial statistics (with discussion). *Annals of the Institute of Statistical Mathematics*, **43**, 1–59.
- Best, N.G., Richardson, S. and Thomson, A. (2005). A comparison of Bayesian spatial models for disease mapping. *Statistical Methods in Medical Research*, **14**, 35–59.
- Carlin, B.P. and Banerjee, S. (2003). Hierarchical multivariate CAR models for spatio-temporally correlated survival data (with discussion). In *Bayesian Statistics 7*, eds. J.M. Bernardo, M.J. Bayarri, J.O. Berger, A.P. Dawid, D. Heckerman, A.F.M. Smith, and M. West, Oxford: Oxford University Press, pp. 45–63.
- Cressie, N.A.C. (1993). *Statistics for Spatial Data*, 2nd ed. New York: Wiley.
- Csillag, F., Boots, B. Fortin, M-J., Lowell, K. and Potvin, F. (2001). Multiscale characterization of boundaries and landscape ecological patterns. *Geomatica*, **55**, 291–307.
- Dass, S.C. and Nair, V.N. (2003). Edge detection, spatial smoothing, and image restoration with partially observed multivariate data. *J. Amer. Statist. Assoc.*, **98**, 77–89.
- Elliot, P. and Best, N.G. (1998). Geographical patterns of disease. *Encyclopedia of Biostatistics*, (eds. P. Armitage and T. Colton). London: Wiley.
- Gelfand, A.E. and Vounatsou, P. (2003). Proper multivariate conditional autoregressive models for spatial data analysis. *Biostatistics*, **4**, 11–25.
- Geman, S. and Geman, D. (1984). Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **6**, 721–742.

- Held, L., Natário, I., Fenton, S.E., Rue, H., and Becker, N. (2005). Towards joint disease mapping. *Statistical Methods in Medical Research*, **14**, 61–82.
- Hogan, J.W. and Tchernis, R. (2004). Bayesian factor analysis for spatially correlated data, with application to summerizing area-level material deprivation from census data. *J. Amer. Statist. Assoc.*, **99**, 314–324.
- Institute of Medicine (2006). Asbestos: selected cancers. Technical report, National Academy of Science, Washington, DC; report released June 6, 2006.
- Jacquez, G.M. and Greiling, D.A. (2003) Geographic boundaries in breast, lung and colorectal cancers in relation to exposure to air toxics in Long Island, New York. *International Journal of Health Geographics*, **2:4**.
- Jeng, F.C., and Woods, J.W. (1991) Compound Gauss-Markov random fields for image estimation. *IEEE Transactions in Signal Processing*, **39**, 683-691.
- Jin, X., Carlin, B.P., and Banerjee, S. (2005). Generalized hierarchical multivariate CAR models for areal data. *Biometrics*, **61**, 950–961.
- Kim, H., Sun, D. and Tsutakawa, R.K. (2001). A bivariate Bayes method for improving the estimates of mortality rates with a twofold conditional autoregressive model. *J. Amer. Statist. Assoc.*, **96**, 1506–1521.
- Knorr-Held, L. and Best, N.G. (2001). A shared component model for detecting joint and selective clustering of two diseases. *J. Roy. Statist. Soc. Ser. A*, **164**, 73–85.
- Lawson, A.B. and Clark, A. (2002). Spatial mixture relative risk models applied to disease mapping. *Statistics in Medicine*, **21**, 359–370.
- Liu, X., Wall, M. and Hodges, J. (2005). Generalized spatial structural equation models. *Biostatistics*, **6**, 539–557.

- Lu, H. and Carlin, B.P. (2005). Bayesian areal wombling for geographical boundary analysis. *Geographical Analysis*, **37**, 265–285.
- Lu, H., Reilly, C., Banerjee, S., and Carlin, B.P. (2006). Bayesian areal wombling via adjacency modeling. To appear *Environmental and Ecological Statistics*.
- Ma, H., Carlin, B.P., and Banerjee, S. (2006). Hierarchical and joint site-edge methods for Medicare hospice service region boundary analysis. Research Report 2006–010, Division of Biostatistics, University of Minnesota.
- Mardia, K.V. (1988). Multi-dimensional multivariate Gaussian Markov random fields with application to image processing. *Journal of Multivariate Analysis*, **24**, 265–284.
- Spiegelhalter, D.J., Best, N.G., Carlin, B.P., and van der Linde, A. (2002). Bayesian measures of model complexity and fit (with discussion). *J. Roy. Statist. Soc., Ser. B*, **64**, 583–639.
- Wang, F. and Wall, M. (2003). Generalized common spatial factor model. *Biostatistics*, **4**, 569–582.
- Womble, W.H. (1951). Differential systematics. *Science*, **114**, 315–322.