Hierarchical Models for Spatio-Temporally Correlated Public Health Data

Bradley P. Carlin

Division of Biostatistics
School of Public Health
University of Minnesota, USA

email: brad@biostat.umn.edu
web: www.biostat.umn.edu/~brad/
Overview

Biostatisticians are increasingly faced with the task of analyzing data that are:

- *highly multivariate*, with many important predictors and response variables
- *geographically referenced*, and often presented as *maps*
- *temporally correlated*, as in longitudinal or other time series structures

**Example:** Lung, breast, colorectal, and cervical cancer rates by county and year in a particular state, with smoking, mammography, and other important screening and staging information also available at some level.

Public health professionals who collect such data are charged not only with surveillance, but also *statistical inference* tasks:

♠ *modeling* of trends and correlation structures
♠ *estimation* of underlying model parameters
♠ *hypothesis testing*, or *comparison* of competing models
♠ *prediction* of observations at unobserved times or locations
Possible solution: **Hierarchical Bayesian Modeling**!

Made feasible (finally!) in the last decade by two important computing developments:

- *Geographic Information Systems* (GISs) for the simultaneous graphical display and summary of the data (common software: *ArcView*).

- *Markov chain Monte Carlo* (MCMC) methods for the estimation of relevant posterior quantities (common software: *WinBUGS*).

---

![Map](image)

**Figure 1**: *ArcView* map of percent of surveyed population with household income below 200% of the federal poverty limit, SHAPE survey units in Hennepin County, MN.
Types of Spatial Data

(1) Geostatistical Model (Point Source Data Model): For a univariate response $Y(s)$ indexed \textit{continuously} by $s$ throughout a space $D$. Given data $\mathbf{Y} \equiv \{Y(s_i)\}$ at known \textit{source} locations $s_i$, $i = 1, \ldots, I$, we might assume

$$\mathbf{Y} | \mu, \theta \sim N_I(\mu, H(\theta)) ,$$

where

- $N_I$ denotes the $I$-dimensional normal distribution
- $\mu$ is the (stationary) mean level
- $(H(\theta))_{ii'}$ gives the covariance between $Y(s_i)$ and $Y(s_{i'})$

Simplest form for $H$: \textit{isotropic}, where spatial correlation depends only on the distance $d_{ii'}$ between $s_i$ and $s_{i'}$, say,

$$(H(\theta))_{ii'} = \sigma^2 \exp(-\phi d_{ii'}), \quad \sigma^2 > 0, \quad \phi > 0 ,$$

an \textit{exponential} form where $\theta = (\sigma^2, \phi)'$. Other isotropic possibilities for $H$ include the powered exponential, spherical, Gaussian, and Matérn (see e.g. Cressie, 1993, or Stein, 1999).

Also might also assume strength of spatial association is higher along one direction than another (\textit{anisotropy}), e.g., if an airborne exposure followed a prevailing wind direction.
Types of Spatial Data (cont’d)

(2) Lattice Model (Regional Summary Data Model): Now define \( Y \) only on \textit{discretely} indexed regions that form a partition of the space \( D \). In this case, the spatial structure is often placed not on the data, but on the \textit{second stage} of the hierarchical model.

Example: the spatial disease mapping model,

\[
Y_i \mid \mu_i \overset{\text{iid}}{\sim} Po (E_i e^{\mu_i}) \, , \text{ where} \\
\begin{align*}
Y_i &= \text{observed disease count}, \\
E_i &= \text{expected count (known), and} \\
\mu_i &= x_i' \beta + \theta_i + \phi_i
\end{align*}
\]

The \( x_i \) are explanatory spatial covariates, having parameter coefficients \( \beta \). The \( \theta_i \) capture regional heterogeneity via

\[
\theta_i \overset{\text{iid}}{\sim} N(0 \, , \, 1/\tau_h) \, ,
\]

while the \( \phi_i \) capture regional clustering via a \textit{conditionally autoregressive} (CAR) prior,

\[
\phi_i \mid \phi_{j \neq i} \sim N(\bar{\phi}_i \, , \, 1/(\tau_c m_i)) \, ,
\]

where \( m_i \) is the number of “neighbors” of region \( i \), and \( \bar{\phi}_i = m_i^{-1} \sum_{j \in \partial i} \phi_j \), the average of the neighboring values.
• CAR prior is a pairwise difference prior (Besag et al., 1995), identified only up to an additive constant. Thus to identify an intercept term \( \beta_0 \) in the log-relative risk, we add the constraint \( \Sigma_{i=1}^{I} \phi_i = 0 \).

• A consequence of our prior specification is that

\[
\phi_i \mid \phi_i \neq i \sim N(\bar{\phi}_i, 1/(\tau_c m_i)),
\]

which gives some intuition about the role of \( \tau_c \).

For either geostatistical or lattice models, Bayesian inference for the parameter vector \( \theta \) proceeds by

1. Picking a prior distribution \( p(\theta) \) for \( \theta \)

2. Basing inference on the posterior distribution,

\[
p(\theta | y) = \frac{p(y | \theta)p(\theta)}{\int p(y | \theta)p(\theta)d\theta} \quad \text{(1)}
\]

or the predictive distribution,

\[
p(y_{I+1} | y) = \int p(y_{I+1} | \theta)p(\theta | y)d\theta. \quad \text{(2)}
\]

Integrals are often intractable, but MCMC enables us to draw samples \( \theta^{(g)} \) and \( y_{I+1}^{(g)} \), \( g = 1, \ldots, G \), from (1) and (2), which can then be summarized or mapped!
Univariate vs. Multivariate Modeling

We can combine CAR models with traditional multivariate models to accommodate a broad range of problems.

**Example:** For breast cancer control model in county $i$, let

\[
Y_{1i} = \text{observed age-adjusted mortality rate per 100,000}
\]

\[
Y_{2i} = \text{observed age-adjusted incidence rate per 100,000}
\]

\[
Y_{3i} = \text{observed percent of late (regional or distant) diagnoses}
\]

\[
Y_{4i} = \% \text{ of surveyed women over the age of 40 who have not had a mammogram in the last two years}
\]

Now assume

\[
Y_{ki} \overset{\text{iid}}{\sim} N(\theta_{ki}, \sigma_{ki}^2), \quad k = 1, \ldots, 4, \quad i = 1, \ldots, I = 87
\]

and set $\sigma_{ki}^2 = \sigma_k^2 / n_{ki}$, where $n_{ki}$ is the number of persons at risk for event $k$ in county $i$. Writing $\theta_k = (\theta_{k1}, \ldots, \theta_{kI})'$, let

\[
\theta_k \sim CAR(\lambda_k)
\]

We then borrow strength across breast cancer indicators by assuming

\[
\lambda_k \overset{\text{iid}}{\sim} G(a, b) \quad \text{and} \quad \sigma_k^2 \overset{\text{iid}}{\sim} IG(a, b)
\]

where $G$ and $IG$ denote the gamma and inverse gamma distributions, respectively.
Parameter of interest: Given a set of weights $\alpha_k$ such that $\sum_{k=1}^{4} \alpha_k = 1$, form the breast cancer “control variable,”

$$\eta_i = \sum_{k=1}^{4} \alpha_k \theta_{ki}.$$ 

A sample from the posterior distribution of $\eta = (\eta_1, \ldots, \eta_I)'$, $p(\eta|y)$, is easily obtained in WinBUGS. Using $\alpha_k = 1/4$ for all $k$, compare the posterior medians of the $\eta_i$ with naive averages:

![Image of maps showing breast cancer control variables](image)

Figure 2: Naive and spatially smoothed breast cancer control variables, Minnesota data, 1993-97.

Note smoothed control variates cover a narrower range, and clarify the overall spatial pattern in the state!

Can extend model to:

- estimate county ranks (instead of medians)
- account for error in mammography survey information ($Y_{4i}$)
- handle multiple cancers (breast, lung, and colorectal): MCAR!
Spatial Misalignment

Many multivariate spatially-referenced datasets needed in the analysis of environmental data suffer from data misalignment:

- differing support: block (areal) summary vs. point-referenced
- collected by differing agencies: EPA, CDC, Census, etc.
- over differing regions: zip codes, census tracts, buffers, etc.

Geographic Information Systems (GISs) are powerful computer tools for analyzing misaligned data, but their current capabilities are primarily descriptive, rather than inferential:

★ When is a “hot spot” truly “hot”?
★ Which predictors are “significant”?
★ How to “realign” data for subsequent analysis?
★ How to interpolate counts in the subregions over a misaligned areal grid?
★ What is the true exposure at a given point in space, or the relative risk of disease in a given region (or subregion)?
★ What will it be next year?
“Change of Support Problem”

- concerned with inference about the values of a variable at points or blocks different from those at which it has been observed. Four cases to consider here:

1. points to points
2. points to blocks
3. blocks to points
4. blocks to blocks

Case 1 has typically been handled via kriging. Substantial literature on Bayesian approaches to this problem: Handcock and Stein (1993), Handcock and Wallis (1994), Ecker and Gelfand (1997), Diggle, Tawn and Moyeed (1998), etc.

Case 4 has been referred to as the modifiable areal unit problem in the geography literature (see e.g. Cressie, 1996). For an extensive variable (where block values are sums of sub-block values), GISs often use areal weighting. For areal allocation using “better” covariates than area, see Flowerdew and Green (1989, 1992, 1994), Mugglin and Carlin (1998), Zhu et al. (2000), and Mugglin, Carlin and Gelfand (2000):
**Block-Block Misalignment:** Population at risk for lung cancer near a putative risk source in SW Ohio

Study area surrounds the Feed Materials Production Center (FMPC), which processed uranium for weapons production.

- Radon gas released into atmosphere
- Uranium contamination of groundwater

Nonnested grid formed by the intersection of US census block group boundaries and an exposure windrose ("bullseye"):

---

Figure 3: 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 km²).
Notation for Nonnested Misaligned Data Model

Suppose 2 partitions of the region, the intersection of which create “atoms”:

- $B_i$, $i = 1, \ldots, I$ (census block groups)
  Then atoms $B_{ij}$ are indexed $j = 1, \ldots, J_i$.
- $C_j$, $j = 1, \ldots, J$ (exposure bullseye cells)
  Then atoms $C_{ji}$ are indexed $i = 1, \ldots, I_j$.

Thus we will require a giant “look-up table,” to match the atom identifiers under the two indexing systems (e.g., $B_{1,1} = C_{47,3}$).

Let $k = 1, 2$ index gender, and $l = 1, \ldots, L$ index age group.

Data available:

<table>
<thead>
<tr>
<th>item</th>
<th>description</th>
<th>source</th>
</tr>
</thead>
<tbody>
<tr>
<td>$n_{i,kl}$</td>
<td>population count</td>
<td>Census Bureau</td>
</tr>
<tr>
<td>$s_{.j}$</td>
<td>structure count</td>
<td>USGS aerial maps</td>
</tr>
</tbody>
</table>

Task: Reallocate the block group-level population counts to the exposure cells, obtaining posterior estimates of the $n_{.jkl}$ for the CDC.
Model for population reallocation problem

Ignoring the age and gender subscripts for now, we let
\[ n_{ij} \sim Po \left( e^{\mu_i} A_{ij} e^{\theta(d_{ij}-c)} \right), \]
where \( A_{ij} \) is the area of atom \( ij \), \( d_{ij} = s_{ij}/A_{ij} \), the structure density, and \( c \) is some ballpark “centering” constant.

Thus the \( e^{\mu_i} \) are area effects (“bodies per unit area”), and \( \theta \) is the structure count effect (a “relative risk”). Remembering that only \( n_i \) is observed, rewrite this model as

\[
\begin{aligned}
(n_{i1}, \ldots, n_{iJ_i} \mid n_i) & \sim \text{Mult}(n_i; p_{i1}, \ldots, p_{iJ_i}) \\
\text{where } p_{ij} & = \frac{A_{ij} e^{\theta(d_{ij}-c)}}{\sum_{j=1}^{J_i} A_{ij} e^{\theta(d_{ij}-c)}} \\
\end{aligned}
\]

AND \( n_{i.} \sim Po \left( e^{\mu_i} \sum_{j=1}^{J_i} A_{ij} e^{\theta(d_{ij}-c)} \right) \)

Priors: \( \theta \sim \text{flat}, \) and \( \mu_i \overset{iid}{\sim} N(\eta_\mu, 1/\tau_\mu) \) or \( CAR(\lambda) \)

Now need models for

• the \( s_{ji} \) (similar to above; “structures per unit area”)

• “edge atoms” \( s_{iE} \) (formed by overlap of some block groups \( i \) outside the bullseye): borrow strength from neighbors!
Full model specification

\[
\begin{align*}
\prod_{i=1}^{I}[n_{i1}, \ldots, n_{iJ_i}, n_i, \theta, \{s'_{ij}\}][n_i|\mu_i, \theta, \{s'_{ij}\}][\mu_i] \ [\theta] \\
\times \left( \prod_{j=1}^{J_i}[s_{j1}, \ldots, s_{ji}] [s_j|\omega_j][\omega_j] \right) \\
\times \left( \prod_{i=1}^{I_E}[s'_{iE}|\omega'_i][\omega'_i] \right)
\end{align*}
\]

MCMC Generation

- \(n_{ij}\): multinomial draws!
- \(\mu_i, \theta, \omega_j, \omega'_i\): Metropolis steps (Gaussian proposals)
- \(s_{ji}\): Hastings steps (multinomial proposals)
- \(s'_{iE}\): Hastings steps (Poisson proposals)

Note that \(n_{ij}\) generation may be pulled out of MCMC loop, and done at the end, along with aggregation to cell totals \(n_{.j}\). 

Concerns: slow MCMC convergence, weak identifiability!

Compare imputed \(n_{.j}\) values to standard \((\hat{n}_{.j} \propto s_{.j})\), using

- 39 bg’s (w/ 29 edge zones), 160 cells, 389 atoms
- moderately informative prior: \(\eta_{\mu} = \log 3 = 1.1\) (Rogers and Killough), \(\eta_{\omega} = 3.167\) (map-based), and
  - iid/iid prior: \(\tau_{\mu} = \tau_{\omega} = 100\) (i.e. \(\sigma_{\mu} = \sigma_{\omega} = 0.1\)).
  - iid/CAR prior: \(\tau_{\mu} = 10, \lambda_{\omega} = 100\) (\(\sigma_{\mu} = \sigma_{\omega} \approx 0.32\)).
Posterior distributions of structure estimates for the four atoms of cell 106 (SE6); \( s_{106} \) is known to be 55. Vertical bars represent structure values if imputed proportionally to area.

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.
Figure 4: Imputed population densities (persons/km²) by atom for the FMPC region
Figure 5: Imputed population densities (persons/km²) by cell for the FMPC windrose
**Point-Block Misalignment**: Ozone levels and pediatric ER visits in metro Atlanta, GA

◊ point data: 1-hour daily max ozone levels at 10 fixed monitors over 92 summer days (June 1 – August 31) in 1993–1995

◊ block data: pediatric asthma ER visit counts and total number of pediatric ER visits by day and zip for the 162 zip codes

◊ Goal #1: Estimate the ozone (exposure potential) levels for the 36 zip codes falling within the city of Atlanta, as well as two selected specific points within a particular zip.

![Map of zip code boundaries in the Atlanta metropolitan area and eight-hour maximum ozone levels (ppm) at the 10 monitoring sites for July 15, 1995.](image_url)

Figure 6: Zip code boundaries in the Atlanta metropolitan area and eight-hour maximum ozone levels (ppm) at the 10 monitoring sites for July 15, 1995.
Model for Static Spatial Case

Suppose an underlying spatial process $X(s)$ for locations $s \in D$. Then point data is $X(s_i)$, while block data arise as

$$X(B) = |B|^{-1} \int_B X(s) ds,$$

where $|B|$ denotes the area of $B$.

Start with a stationary spatial Gaussian process having mean function $\mu(s; \beta)$ and covariance function $c(s - t; \theta)$.

Case 1: point-point: Let $X^T_s = (X(s_1), \ldots, X(s_I))$. Then

$$X_s \mid \beta, \theta \sim N(\mu_s(\beta), H_s(\theta))$$

where $\mu_s(\beta)_i = \mu(s_i; \beta)$ and $(H_s(\theta))_{ii'} = c(s_i - s_{i'}; \theta)$.

To predict at $X^T_{s'} = (X(s'_{1}), \ldots, X(s'_{K}))$, a new set of $K$ target locations, we need

$$f(X_{s'} \mid X_s) = \int f(X_{s'} \mid X_s, \beta, \theta)f(\beta, \theta \mid X_s) d\beta d\theta.$$  \(3\)

But $(X_s, X_{s'} \mid \beta, \theta)$ is $(I + K)$-variate normal, so we have that $p(X_{s'} \mid X_s, \beta, \theta_g)$ is $K$-variate normal by standard formulae. Thus, samples from (3) are easily drawn via composition, i.e.,

$$X_{s', g}^* \sim p(X_{s'} \mid X_s, \beta_g^*, \theta_g^*), \quad \text{where}$$

$$(\beta_g^*, \theta_g^*) \sim p(\beta, \theta \mid X_s), g = 1, \ldots, G,$$

the latter drawn via importance or MCMC sampling.
Case 2: point-block: To predict at target blocks, i.e. at
$X_B^T = (X(B_1), \ldots , X(B_K))^t$, we now require
\[
f(X_B \mid X_s) = \int f(X_B \mid X_s; \beta, \theta) f(\beta, \theta \mid X_s) d\beta d\theta .
\]
Under a Gaussian process, we have
\[
f\left( \begin{pmatrix} X_s \\ X_B \end{pmatrix} \right) \mid \beta, \theta) = N \left( \begin{pmatrix} \mu_s(\beta) \\ \mu_B(\beta) \end{pmatrix} , \begin{pmatrix} H_s(\theta) & H_{s,B}(\theta) \\ H_{s,B}^T(\theta) & H_B(\theta) \end{pmatrix} \right),
\]
where
\[
(\mu_B(\beta))_k = |B_k|^{-1} \int_{B_k} \mu(s; \beta) ds,
\]
\[
(H_B(\theta))_{kk'} = |B_k|^{-1} |B_{k'}|^{-1} \int_{B_k} \int_{B_{k'}} c(s - s'; \theta) ds' ds,
\]
and
\[
(H_{s,B}(\theta))_{ik} = |B_k|^{-1} \int_{B_k} c(s_i - s'; \theta) ds'.
\]
So $X_B \mid X_s, \beta, \theta$ is again distributed as Gaussian, but the mean
and variance components now require integrations!

Monte Carlo approach: Draw an independent set of locations
$s_{k,\ell}, \ell = 1, 2, \ldots , L_k$ uniformly over $B_k$. Replace $(\mu_B(\beta))_k,$
$(H_B(\theta))_{kk'},$ and $(H_{s,B}(\theta))_{ik}$ with
\[
(\mu_B(\beta))_k = L_k^{-1} \sum_\ell \mu(s_{k,\ell}; \beta),
\]
\[
(H_B(\theta))_{kk'} = L_k^{-1} L_{k'}^{-1} \sum_\ell \sum_{\ell'} c(s_{k,\ell} - s_{k',\ell'}; \theta),
\]
and
\[
(H_{s,B}(\theta))_{ik} = L_k^{-1} \sum_\ell c(s_i - s_{k,\ell}; \theta).
\]
Note that the same set of $s_{k,\ell}$’s can be used for each integration
and with each $(\beta^*, \theta^*)$; we need only obtain this set once.

Cases 3 (block-point) and 4 (block-block) follow similarly!
Example 2: Atlanta ozone data, July 15, 1995

Figure 6 shows the 1-hour daily maximum ozone measures at 10 monitoring sites on July 15, 1995. Predict block averages for the 36 city zip codes, and also the specific values at the points A and B.

Let $\mathbf{X}(s)$ be log-ozone exposure, assume a constant mean $\mu$, and let $c(s_i - s_{i' }; \theta) = \sigma^2 e^{-\phi \| s_i - s_{i' } \|}$

Priors: $\mu \sim \text{flat}$, $\sigma^2 \sim IG(3, .05)$, and $\phi \sim G(.03, 100)$.

Algorithm: Generate 3743 random sites $s_{kl}$ over the city, to obtain an average $L_k$ of roughly 104. Run 3 parallel sampling chains for 1000 iterations each.

Next figure maps summaries of the posterior samples for 36 target blocks (city zips) and 2 target points (A and B):

- **posterior medians, $q_{.50}$**: show a clear spatial pattern, reflecting the original data (Figure 6); also note sensibly differing values for points A and B

- **lengths of the 95% equal-tail CIs, $q_{.975} - q_{.025}$**: reflect spatial variability, with lower values occurring in larger areas (which require more averaging) or in areas nearer to monitoring stations (SE, NE, and W city boundaries).
Point-block misaligned regression

Suppose we have area-level observed disease counts $Y_\ell$ available for each area $\ell$, and point-level exposure measurements $X(s_i)$ at $I$ fixed monitoring stations. Defining the expected counts $E_\ell = n_\ell \left( \frac{\sum Y_i}{\sum n_i} \right)$, we assume

$$Y_\ell \sim Po \left( E_\ell \exp (\beta X_\ell + \sum_{k=1}^{K} \alpha_k Z_{k\ell}) \right),$$

where the $Z_{k\ell}$ are region-level sociodemographic covariates. Then using our previous approach to realign the $X$’s, the (MC-approximated) full Bayesian model is

$$\prod_{\ell} f(Y_\ell | \beta, \alpha, X_\ell) \hat{f} \left( \{X_\ell\} | X_s, \mu, \sigma^2, \phi \right) \hat{f} \left( X_s | \mu, \sigma^2, \phi \right) p(\beta, \alpha, \mu, \sigma^2, \phi),$$

where $X_s^T = (X(s_1), \ldots, X(s_I))$. Now we must run a MCMC sampler over the larger space $(\mu, \sigma^2, \phi, \{X_\ell\})$.

---

Application to Atlanta ozone data:

- $Y_\ell =$ pediatric ER visits for asthma from zip $\ell$
- $X(s_i) =$ average ozone (1-hr max) at monitoring station $i$
- $Z_{k\ell} =$ SES and race covariates for zip $\ell$

Spatio-temporal version follows similarly, but many missing $X(s_i, t_j)$ forces use of a 3-stage imputation/MCMC algorithm
Spatial Frailty Models

Frailty models often used for time-to-event data grouped into strata (e.g. counties). For subject \( j \) in stratum \( i \), let

\[
t_{ij} = \text{time to death or censoring}
\]

\[
x_{ij} = \text{vector of individual-specific covariates}
\]

\[
\gamma_{ij} = \text{death indicator (0 if alive, 1 if dead)}
\]

Then under proportional hazards we have

\[
h(t_{ij}; x_{ij}) = h_0(t_{ij}) \exp(\beta^T x_{ij}) , \tag{4}
\]

where \( h_0 \) is the baseline hazard.

In the frailty setting, model (4) is extended to

\[
h(t_{ij}; x_{ij}) = h_0(t_{ij}) \omega_i \exp(\beta^T x_{ij})
\]

\[
= h_0(t_{ij}) \exp(\beta^T x_{ij} + W_i) ,
\]

where \( W_i \equiv \log \omega_i \) is the stratum-specific frailty term, designed to capture differences among the strata.

The spatial frailty model replaces the usual i.i.d. specification for the \( W_i \) with a geostatistical or lattice (CAR) model. For the geostatistical model, the joint posterior distribution is

\[
p(\beta, W, \rho, \theta \mid t, x, \gamma) \propto L(\beta, W, \rho \mid t, x, \gamma) p(W \mid \theta) p(\beta)p(\rho)p(\theta) .
\]
Dataset: Infant Mortality in Minnesota Counties

- 1547 deaths before first birthday
- covariate information: sex, race, birth weight, mother’s age, and mother’s total number of previous births
- spatial information: mother’s county of residence, latitude and longitude of county centroids (for geostat model), and county contiguity map (for CAR model).

Here, $t_{ij}$ is the survival time for those babies who died during their first year; remaining survivors are “censored”!

Under a parametric (Weibull) form for the baseline hazard and vague priors, compare models via effective model size ($p_D$) and the Deviance Information Criterion (DIC; small is good):

<table>
<thead>
<tr>
<th>Model</th>
<th>$p_D$</th>
<th>DIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>No Frailty</td>
<td>8.72</td>
<td>511</td>
</tr>
<tr>
<td>Non-Spatial frailty</td>
<td>39.35</td>
<td>392</td>
</tr>
<tr>
<td>CAR frailty</td>
<td>34.52</td>
<td>371</td>
</tr>
<tr>
<td>Geostat frailty</td>
<td>35.02</td>
<td>360</td>
</tr>
</tbody>
</table>
### Results for CAR frailty model:

<table>
<thead>
<tr>
<th>covariate</th>
<th>2.5%</th>
<th>50%</th>
<th>97.5%</th>
</tr>
</thead>
<tbody>
<tr>
<td>intercept</td>
<td>-2.585</td>
<td>-2.461</td>
<td>-2.405</td>
</tr>
<tr>
<td>sex (boys = 0) girls</td>
<td>-0.224</td>
<td>-0.183</td>
<td>-0.096</td>
</tr>
<tr>
<td>race (white = 0) black</td>
<td>-0.219</td>
<td>-0.148</td>
<td>-0.007</td>
</tr>
<tr>
<td>native American</td>
<td>0.455</td>
<td>0.782</td>
<td>0.975</td>
</tr>
<tr>
<td>unknown</td>
<td>0.351</td>
<td>0.831</td>
<td>1.165</td>
</tr>
<tr>
<td>mother’s age</td>
<td>-0.005</td>
<td>-0.004</td>
<td>-0.003</td>
</tr>
<tr>
<td>birth weight in kg</td>
<td>-1.953</td>
<td>-1.932</td>
<td>-1.898</td>
</tr>
<tr>
<td>total births</td>
<td>0.088</td>
<td>0.119</td>
<td>0.151</td>
</tr>
<tr>
<td>$\rho$</td>
<td>0.470</td>
<td>0.484</td>
<td>0.497</td>
</tr>
<tr>
<td>$\lambda$</td>
<td>12.62</td>
<td>46.07</td>
<td>160.4</td>
</tr>
</tbody>
</table>

- All main effects are “significant”
- $\rho$ summaries suggest decreasing hazard (most deaths occur in the first day)
- Results for geostat model are similar, yet take 10 times longer to produce (matrix inversion)!
Frailties as “spatial residuals”:

Boxplots of the posterior median frailties for the iid and CAR models, with and without covariates:

- tightness of the full CAR boxplot suggests this model is best at reducing the need for the frailty terms.
- all full CAR residuals are in the range (−0.15, 0.10), or (0.86, 1.11) on the hazard scale. This means that any missing spatially-varying covariates have only a modest (10-15%) impact on the hazard

So this model fits quite well!

Maps of these posterior median frailties show overall trends...
Software

• Current: ArcView for database and mapping, but C++ for MCMC generation

• Very Near Future: Read in S–plus county boundary files for any state in the US, and use GeoBUGS to build the adjacency matrix, fit CAR model, and map results!

____________________________________________________

Summary

Hierarchical Bayesian methods, coupled with modern GIS and MCMC software, offer an attractive environment for multivariate spatio-temporal modeling, but more remains to be done:

★ Choice of weights $\alpha_k$ and whether to standardize measures $\theta_{ki}$ in cancer control problem

★ Multivariate response or frailty models, via MCAR (Mardia, 1988) or “2-fold CAR” (Kim et al., 2001)

★ Incorporation of measurement error (in both $X$ and $s$)

★ Incorporating temporal trends and nonparametric baseline hazard functions in spatial frailty setting

____________________________________________________

For more info: http://www.biostat.umn.edu/~brad/