Lecture 5 Hierarchical modeling of spatial data

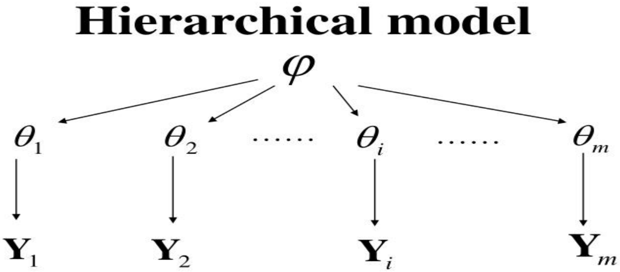
Shiwei Lan[1]

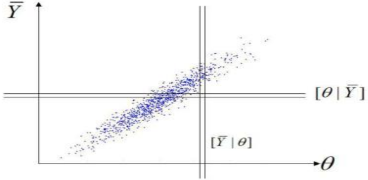
[1]School of Mathematical and Statistical Sciences
Arizona State University

STP598 Spatiotemporal Analysis
Fall 2020

Table of Contents

Bayesian hierarchical modeling





Stationary spatial process models

- The basic model is
$$Y(\mathrm{s}) = \mu(\mathrm{s}) + w(\mathrm{s}) + \epsilon(\mathrm{s}) \tag{1}$$
- The mean structure $\mu(\mathrm{s}) = \mathrm{x}^T \boldsymbol{\beta}$.
- The residual is partitioned into two parts: the spatial part $w(\mathrm{s})$ and the non-spatial errors $\epsilon(\mathrm{s})$.
- Recall that the $w(\mathrm{s})$ introduces the partial sill (σ^2) and the range ϕ; while the $\epsilon(\mathrm{s})$ adds the nugget τ^2.
- Assume stationarity. Pure error $\epsilon(\mathrm{s})$ vs spatial error $w(\mathrm{s})$. We further assume
  - $w(\mathrm{s}+\mathrm{h}) - w(\mathrm{s}) \to 0$ as $\mathrm{h} \to 0$
  - $[w(\mathrm{s}+\mathrm{h}) + \epsilon(\mathrm{s}+\mathrm{h})] - [w(\mathrm{s}) + \epsilon(\mathrm{s})] \not\to 0$ as $\mathrm{h} \to 0$
- *Microscale* view: $\epsilon(\mathrm{s})$ is a spatial process with very rapid decay in association, and only matters at high resolution.

Isotropic models

- Suppose we have data $Y(\mathrm{s}_i)$, $i = 1, \cdots, n$, and let $\mathrm{Y} = (Y(\mathrm{s}_1), \cdots, Y(\mathrm{s}_n))^T$. In the basic Gaussian isotropic kriging model, we assume

$$\Sigma = \sigma^2 H(\phi) + \tau^2 I \tag{2}$$

where H is correlation matrix with $H_{ij} = \rho(\mathrm{s}_i - \mathrm{s}_j; \phi)$ and ρ is a valid isotropic correlation function on $\mathbb{R}^2$, e.g. $\rho(\mathrm{s}_i - \mathrm{s}_j; \phi) = \exp(-\phi \|\mathrm{s}_i - \mathrm{s}_j\|)$

- Collecting all model parameters into a vector $\boldsymbol{\theta} = (\boldsymbol{\beta}, \sigma^2, \tau^2, \phi)$, we have

$$p(\boldsymbol{\theta}|\mathrm{y}) \propto f(\mathrm{y}|\boldsymbol{\theta}) p(\boldsymbol{\theta}) \tag{3}$$

where

$$\mathrm{Y}|\boldsymbol{\theta} \sim N(X\boldsymbol{\beta}, \sigma^2 H(\phi) + \tau^2 I) \tag{4}$$

Prior specification

- Typically, independent priors are chosen for the different parameters

$$p(\boldsymbol{\theta}) = p(\boldsymbol{\beta})p(\sigma^2)p(\tau^2)p(\phi) \tag{5}$$

- We usually adopt the following priors

$$\beta \sim N_{p+1}(\mu_0, \Lambda_0) \tag{6}$$
$$\sigma^2 \sim \Gamma^{-1}(a_1, b_1) \tag{7}$$
$$\tau^2 \sim \Gamma^{-1}(a_2, b_2) \tag{8}$$
$$\phi \sim \Gamma(a_3, b_3) \tag{9}$$

- In Matérn class, the product $\sigma^2\phi^{2\nu}$ can be identified but not the individuals (Zhang, 2004). Often, a very informative prior is imposed for ϕ (e.g. uniform over interval) and a relatively vague prior is used for σ^2.

Prior specification

- When $\tau^2 \equiv 0$, with the exponential covariance, Berger et al. (2001) impose the *objective* priors of the form

$$p(\boldsymbol{\beta}, \sigma^2, \phi) \propto \frac{p(\phi)}{(\sigma^2)^\alpha} \tag{10}$$

- This implies a flat prior for $\boldsymbol{\beta}$. With a uniform prior for ϕ, it can be shown that an improper posterior arises for $\alpha < 2$.
- If $\sigma^2 \sim \Gamma^{-1}(\epsilon, \epsilon)$, then $\alpha = 1 + \epsilon$ getting an improper posterior for small ϵ.
- If $\sigma^2 \sim \Gamma^{-1}(a_1, b_1)$, $a_1 \geq 1$ is recommended $\alpha \geq 2$.
- Closed form of *marginal* posterior of $\boldsymbol{\beta}$ may be available (e.g. with Gaussian likelihood).

Hierarchical modeling

- The hierarchical modeling shares the following generic format

$$[\text{data}|\text{process}, \text{parameters}]\ [\text{process}|\text{parameters}]\ [\text{parameters}] \tag{11}$$

- The previous spatial process model (4) can be recast as a hierarchical model with two stages

$$\mathrm{Y}|\boldsymbol{\theta}, \mathrm{W} \sim N(X\boldsymbol{\beta} + \mathrm{W}, \tau^2 I) \tag{12}$$

$$\mathrm{W}|\sigma^2, \phi \sim N(0, \sigma^2 H(\phi)) \tag{13}$$

where $\mathrm{W} = (w(\mathrm{s}_1), \cdots, w(\mathrm{s}_n))^T$ is the spatial random effect that captures spatial dependence.

- The parameter space is now augmented from $\boldsymbol{\theta}$ to $(\boldsymbol{\theta}, \mathrm{W})$, with the dimension increased by n.

Hierarchical modeling

- The resulting $p(\boldsymbol{\theta}|\mathrm{y})$ is the same, but we have the choice of using MCMC to fit either $f(\mathrm{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})$ or $f(\mathrm{y}|\boldsymbol{\theta},\mathrm{W})p(\mathrm{W}|\boldsymbol{\theta})p(\boldsymbol{\theta})$.
- Interest is often in the spatial surface $\mathrm{W}|\mathrm{y}$ as well as prediction for $W(\mathrm{s}_0)|\mathrm{y}$ for various choices of s_0.
- Note $p(\mathrm{W}|\mathrm{y})$ can be recovered from $p(\boldsymbol{\theta}|\mathrm{y})$:

$$p(\mathrm{W}|\mathrm{y}) = \int p(\mathrm{W}|\boldsymbol{\theta},\mathrm{y})p(\boldsymbol{\theta}|\mathrm{y})d\boldsymbol{\theta} \tag{14}$$

which can be sampled by one for one composition of posterior sampling: $\mathrm{W}^{(g)} \sim p(\mathrm{W}|\boldsymbol{\theta}^{(g)},\mathrm{y})$ with $\boldsymbol{\theta}^{(g)} \sim p(\boldsymbol{\theta}|\mathrm{y})$.

Hierarchical modeling

- Bayesian kriging involves prediction of $Y_0 \equiv Y(\mathrm{s}_0)$ at a new location s_0 with associated covariate $\mathrm{x}(\mathrm{s}_0)$:

$$p(y_0|\mathrm{y}, X, \mathrm{x}_0) = \int p(y_0, \boldsymbol{\theta}|\mathrm{y}, X, \mathrm{x}_0) d\boldsymbol{\theta} = \int p(y_0|\mathrm{y}, \boldsymbol{\theta}, \mathrm{x}_0) p(\boldsymbol{\theta}|\mathrm{y}, X) d\boldsymbol{\theta} \quad (15)$$

- In practice, we use MCMC to obtain posterior samples $\{\boldsymbol{\theta}^{(g)}\}_{g=1}^G$ with $\boldsymbol{\theta}^{(g)} \sim p(\boldsymbol{\theta}|\mathrm{y}, X)$, and approximate the above predictive distribution with

$$\hat{p}(y_0|\mathrm{y}, X, \mathrm{x}_0) = \frac{1}{G}\sum_{g=1}^G p(y_0|\mathrm{y}, \boldsymbol{\theta}^{(g)}, \mathrm{x}_0) \quad (16)$$

- Replicates of prediction $y_0^{(g)}$ can be obtained using the composition sampling.
- Multi-output prediction for $\mathrm{Y}_0 = (Y(\mathrm{s}_{01}), \cdots, Y(\mathrm{s}_{0m}))^T$ is also available

$$p(\mathrm{y}_0|\mathrm{y}, X, \mathrm{x}_0) = \int p(\mathrm{y}_0|\mathrm{y}, \boldsymbol{\theta}, \mathrm{x}_0) p(\boldsymbol{\theta}|\mathrm{y}, X) d\boldsymbol{\theta} \approx \frac{1}{G}\sum_{g=1}^G p(\mathrm{y}_0|\mathrm{y}, \boldsymbol{\theta}^{(g)}, \mathrm{x}_0) \quad (17)$$

Table of Contents

Generalized linear spatial process modeling

- In some point-referenced data sets, measurements $Y(\mathrm{s})$ are not naturally modeled as a normal distribution. e.g. binary data, and counting data.
- The observations $Y(\mathrm{s}_i)$ are modeled independent conditioned on $\boldsymbol{\beta}$ and $w(\mathrm{s}_i)$ within the class of exponential family:

$$p(y(\mathrm{s}_i)|\boldsymbol{\beta}, w(\mathrm{s}_i), \gamma) = h(y(\mathrm{s}_i), \gamma) \exp\{\gamma[y(\mathrm{s}_i)\eta(\mathrm{s}_i) - \psi(\eta(\mathrm{s}_i))]\} \tag{18}$$

  where $g(\eta(\mathrm{s}_i)) = \mathrm{x}^T(\mathrm{s}_i)\boldsymbol{\beta} + w(\mathrm{s}_i)$ for some link function g, and γ is a dispersion parameter.
- We presume the $w(\mathrm{s}_i)$ to be the spatial random effect coming from a Gaussian process, i.e. $\mathrm{W} \sim N(0, \sigma^2 H(\phi))$.

Generalized linear spatial process modeling

- Using conditional independence, we have the joint distribution

$$f(y(\mathrm{s}_1), \cdots, y(\mathrm{s}_n)|\boldsymbol{\beta}, \sigma^2, \phi, \gamma) = \int \left(\prod_{i=1}^n f(y(\mathrm{s}_i)|\boldsymbol{\beta}, w(\mathrm{s}_i), \gamma) \right) p(\mathrm{W}|\sigma^2, \phi) d\mathrm{W} \tag{19}$$

- Note W may not be marginalized out in general.
- Binary data. We model $Y(\mathrm{s})$ through the latent process $Z(\mathrm{s}) = \mathrm{x}(\mathrm{s})^T\boldsymbol{\beta} + w(\mathrm{s}) + \epsilon(\mathrm{s})$:

$$\Pr(Y(\mathrm{s}) = 1) = \Pr(Z(\mathrm{s}) \geq 0) = g^{-1}(\mathrm{x}(\mathrm{s})^T\boldsymbol{\beta} + w(\mathrm{s})) \tag{20}$$

  for some link function $g(\cdot)$ such that g^{-1} takes $p(\mathrm{s}) := \Pr(Y(\mathrm{s}) = 1)$ to $\mathbb{R}^1$.
- Possible choices: the logit $g(x) = \log \frac{x}{1-x}$ and the probit $g(x) = \Phi^{-1}(x)$.

Generalized linear spatial process modeling

- Consider real estate data at 50 locations in Baton Rouge, LA.
- $Y(\mathbf{s}) = 1$ indicates that the price of the property at location $\mathbf{s}$ is "high" (above the median for the region); $Y(\mathbf{s}) = 0$ indicates that the price is "low".
- Covariates: the house's age, total living area, and other area in the property.
- We fit the model with the logit link, assuming vague priors for $\boldsymbol{\beta}$, a $\mathrm{unif}(0, 10)$ prior for ϕ and a $\Gamma^{-1}(0.1, 0.1)$ prior for σ^2.

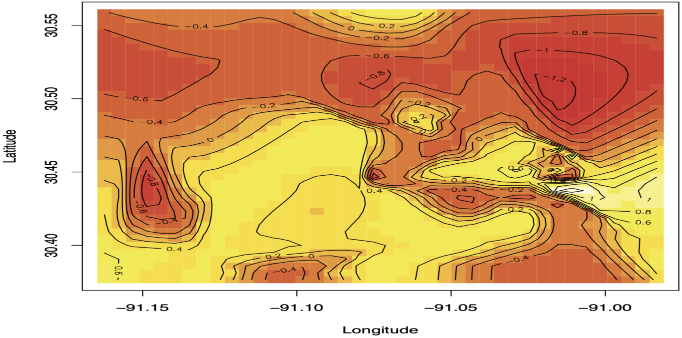


Figure 6.7 Image plot of the posterior median surface of the latent spatial process $w(\mathbf{s})$, binary

Generalized linear spatial process modeling

- Counting data. We model $Y(\mathrm{s})$ through the latent process $Z(\mathrm{s}) = \mathrm{x}(\mathrm{s})^T\boldsymbol{\beta} + w(\mathrm{s}) + \epsilon(\mathrm{s})$:

$$Y(\mathrm{s}) \sim \mathrm{pois}(g^{-1}(Z(\mathrm{s}))) \tag{21}$$

  for some link function $g(\cdot)$, e.g. the canonical link $g(x) = \log(x)$.
- This is related to the log-Gaussian Cox Process (LGCP) model.

Table of Contents

Disease mapping

- An area of strong biostatistical and epidemiological interest is that of *disease mapping*.
- We typically have count data of the following:

$$
\begin{aligned}
Y_i &= \text{observed number of cases of disease in county } i, i = 1, \cdots, I \\
E_i &= \text{expected number of cases of disease in county } i, i = 1, \cdots, I
\end{aligned}
$$

  where E_i are thought of as fixed and known functions of n_i, the number of persons at risk for the disease in county i.
- One can simply assume

$$
E_i = n_i \bar{r} = n_i \frac{\sum_i y_i}{\sum_i n_i} = \sum_i y_i \frac{n_i}{\sum_i n_i} \tag{22}
$$

  i.e. $\bar{r}$ is the overall disease rate in the entire study region.

Disease mapping

- However, such *internal standardization* is "cheating". And one might modify it by age-adjusted rates for the disease

$$E_i = n_{ij} r_j \tag{23}$$

where n_{ij} is the person-years at risk in area i for age group j, and $\bar{r}_j$ is the disease rate in age group j. This process is called *external standardization*.

- The usual model for Y_i is the Poisson model

$$Y_i | \eta_i \sim \mathrm{Pois}(E_i \eta_i) \tag{24}$$

where η_i is the true *relative risk* of disease in region i.

- The maximum likelihood estimate (MLE) of η_i is

$$\hat{\eta}_i = \mathrm{SMR}_i = \frac{Y_i}{E_i} \tag{25}$$

- $\mathrm{Var}(\mathrm{SMR}_i) = \mathrm{Var}(Y_i)/E_i^2 = \eta_i/E_i$. $\widehat{\mathrm{Var}}(\mathrm{SMR}_i) = \hat{\eta}_i/E_i = Y_i/E_i^2$.

Hierarchical Bayesian methods

- For detecting extra-Poisson variability (overdispersion) in the observed rates, we seek *random effects* model for η_i through hierarchical Bayesian modeling.
- Poisson-gamma model

$$Y_i|\eta_i \overset{ind}{\sim} \mathrm{pois}(E_i\eta_i), \quad i = 1, \cdots, I \tag{26}$$

$$\eta_i \overset{iid}{\sim} \Gamma(a, b) \tag{27}$$

- Choose a, b based on $\mu = a/b$ and $\sigma^2 = a/b^2$.
- Due to conjugacy, we have $p(\eta_i|y_i) = \Gamma(a + y_i, b + E_i)$.
- The Bayesian posterior mean (point estimate) of η_i is

$$\mathrm{E}(\eta_i|\mathrm{y}) = \mathrm{E}[\eta_i|y_i] = \frac{y_i + a}{E_i + b} = \frac{y_i + \frac{\mu^2}{\sigma^2}}{E_i + \frac{\mu}{\sigma^2}} = w_i \mathrm{SMR}_i + (1 - w_i)\mu \tag{28}$$

where $w_i = \frac{E_i}{E_i + \frac{\mu}{\sigma^2}}$.

Hierarchical Bayesian methods

- The gamma prior suffers from a serious defect: it fails to allow for spatial correlation among η_i's. We introduce the following model to overcome this issue.
- Poisson-lognormal model. Denote $\psi_i \equiv \log \eta_i$, the log-relative risks.

$$Y_i|\psi_i \overset{ind}{\sim} \mathrm{pois}(E_i e^{\psi_i}), \quad \psi_i = \mathrm{x}_i' \boldsymbol{\beta} + \theta_i + \phi_i \tag{29}$$

$$\theta_i \overset{iid}{\sim} N(0, 1/\tau_h) \tag{30}$$

$$\boldsymbol{\phi}|\mu, \boldsymbol{\lambda} \sim N_I(\mu, H(\boldsymbol{\lambda})), \quad \boldsymbol{\phi} = (\phi_1, \cdots, \phi_I)' \tag{31}$$

- θ_i's capture region-wide *heterogeneity*. These random effects capture extra-Poisson variability in the log-relative risks that varies "globally".
- $\boldsymbol{\phi}$ capture regional *clustering*. They model extra-Poisson variability in the log-relative risks that varies "locally".

Hierarchical Bayesian methods

- In specifying the covariance $H(\boldsymbol{\lambda})$, what are appropriate inter-areal unit distances? Should we use centroid to centroid? Does this make sense with units of quite differing sizes and irregular shapes?
- Big N problem emerges with regard to high-dimensional matrix inversion. We return to CAR (IAR) specifications for $\boldsymbol{\phi}$:

$$\boldsymbol{\phi} \sim \mathrm{CAR}(\tau_c), \quad p(\boldsymbol{\phi}) \propto \exp\left\{-\frac{\tau_c}{2}\sum_{i\neq j} w_{ij}(\phi_i - \phi_j)^2\right\} \tag{32}$$

where w_{ij} is the 0-1 adjacency weights.

- Full conditional of ϕ_i

$$p(\phi_i|\phi_{j\neq i}, \boldsymbol{\theta}, \boldsymbol{\beta}, \mathrm{y}) \propto \mathrm{pois}(y_i|E_i e^{\mathrm{x}_i'\boldsymbol{\beta}+\theta_i+\phi_i}) N(\phi_i|\bar{\phi}_i, 1/(\tau_c m_i)) \tag{33}$$

CAR model

- There are two issues: i) impropriety; ii) selection of τ_h and τ_c.
- Recall that IAR prior is improper due to the singular precision matrix $\Sigma^{-1} = D_w - W$. We could consider either $\Sigma^{-1} = D_w - \rho W$ or $\Sigma^{-1} = M^{-1}(I - \alpha\widetilde{W})$.
- Alternatively, one could ignore the impropriety of the standard CAR model and consider the intrinsic CAR. In practice, we could impose the constraint $\sum_{i=1}^{I} \phi_i = 0$ and numerically implement this centering on the fly.
- There is an identifiability issue of θ_i and ϕ_i in the model. One can choose to fix τ_h and τ_c.
- Alternatively, one could use gamma priors $\tau_h \sim \Gamma(a_h, b_h)$ and $\tau_c \sim \Gamma(a_c, b_c)$ with the following rule of thumb:

$$\mathrm{sd}(\theta_i) = \frac{1}{\sqrt{\tau_h}} \approx \frac{1}{0.7\sqrt{\bar{m}\tau_c}} = \mathrm{sd}(\phi_i) \tag{34}$$

Lip cancer

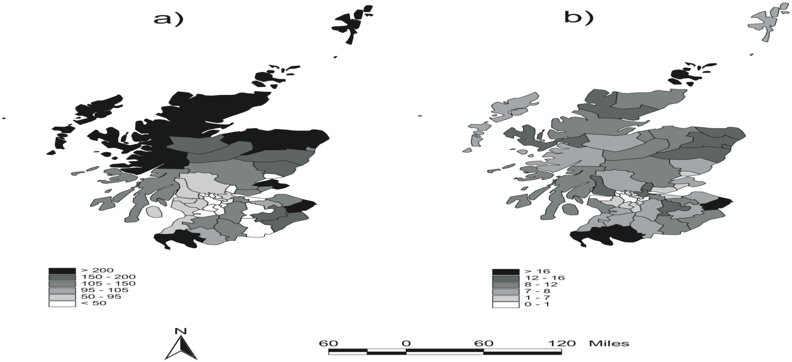


Figure 6.15 *Scotland lip cancer data: (a) crude standardized mortality ratios (observed/expected × 100); (b) AFF covariate values.*

- Consider data from Clayton and Kaldor (1987) are the observed (Y_i) and expected (E_i) cases of lip cancer for the I = 56 districts of Scotland during the period 1975–1980.
- One county-level covariate xi, the percentage of the population engaged in agriculture, fishing or forestry (AFF), is also available.

Lip cancer

- We fit the model of log-relative risk

$$\psi_i = \beta_0 + \beta_1 x_i + \theta_i + \phi_i \tag{35}$$

- Note that y_i cannot inform about θ_i or ϕ_i, but only about their sum $\theta_i + \phi_i$.
- We are more interested in α, the proportion of the variability in the random effects that is due to clustering, and choose priors such that $\alpha \approx 1/2$

$$\alpha = \frac{\mathrm{sd}(\boldsymbol{\phi})}{\mathrm{sd}(\boldsymbol{\theta}) + \mathrm{sd}(\boldsymbol{\phi})} \tag{36}$$

	Posterior for α			Posterior for β		
Priors for τ_c, τ_h	mean	sd	l1acf	mean	sd	l1acf
G(1.0, 1.0), G(3.2761, 1.81)	.57	.058	.80	.43	.17	.94
G(.1, .1), G(.32761, .181)	.65	.073	.89	.41	.14	.92
G(.1, .1), G(.001, .001)	.82	.10	.98	.38	.13	.91
	Posterior for ξ_1			**Posterior for ξ_{56}**		
Priors for τ_c, τ_h	mean	sd	l1acf	mean	sd	l1acf
G(1.0, 1.0), G(3.2761, 1.81)	.92	.40	.33	−.96	.52	.12
G(.1, .1), G(.32761, .181)	.89	.36	.28	−.79	.41	.17
G(.1, .1), G(.001, .001)	.90	.34	.31	−.70	.35	.21

Table 6.6 *Posterior summaries for the spatial model with Gamma hyperpriors for τ_c and τ_h, Scotland lip cancer data; "sd" denotes standard deviation while "l1acf" denotes lag 1 sample autocorrelation.*

Table of Contents

General linear areal data modeling

- Again formulating a hierarchical model, Y_i may be described using a suitable first-stage member of the exponential family.
- Given $\boldsymbol{\beta}$ and ϕ_i, Y_i is modeled conditionally independent:

$$p(y_i|\boldsymbol{\beta}, \phi_i, \gamma) = h(y_i, \gamma) \exp\{\gamma[y_i\eta_i - \psi(\eta_i)]\} \tag{37}$$

  where $g(\eta_i) = \mathrm{x}_i^T \boldsymbol{\beta} + \phi_i$ for some link function g with γ is a dispersion parameter.
- The ϕ_i's will be spatial random effects coming from a CAR model; the pairwise difference, intrinsic (IAR) form is most commonly used.
- In general there is no need to introduce independent heterogeneity effects θ_i's in this generalized linear mixed model with random spatial effects.

Point-referenced models vs areal data models

- Point-referenced data models are defined with regard to an uncountable number of random variables $Y(\mathrm{s})$. For areal units, we envision only a single, finite, n-dimensional distribution for the Y_i, $i = 1, \cdots, I$.
- With point-referenced data $\mathrm{Y} = (Y(\mathrm{s}_1), \cdots, Y(\mathrm{s}_n))'$, we model association directly by covariance Σ_{Y}. With areal data $\mathrm{Y} = (Y_1, \cdots, Y_n)'$ and CAR (or SAR) specifications, we instead model the precision Σ_{Y}^{-1} directly.
- Σ_{Y} encodes *unconditional* association structure; while Σ_{Y}^{-1} provides *conditional* association structure.
- Explanation is a common goal of point-referenced data modeling, but often an even more important goal is spatial prediction or interpolation (i.e., kriging). With areal units, again a goal is explanation, but now often is supplemented by smoothing (MAUP).
- Big N problem for spatial process model due to Σ_{Y}^{-1} and $|\Sigma_{\mathrm{Y}}|$ in the likelihood; not for CAR (full conditionals) or SAR (no matrix inversion in likelihood).