

# Lecture 5 Hierarchical modeling of spatial data

Shiwei Lan<sup>1</sup>

<sup>1</sup>School of Mathematical and Statistical Sciences  
Arizona State University

STP598 Spatiotemporal Analysis  
Fall 2021

Hierarchical  
Model

S.Lan

Stationary  
spatial process  
models

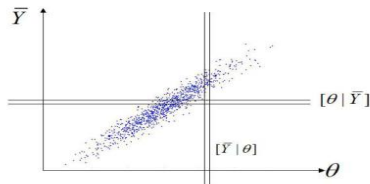
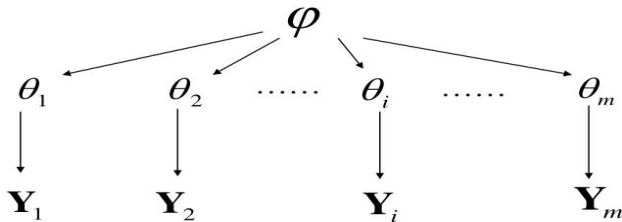
Generalized  
linear spatial  
process  
modeling

Areal data  
models

General linear  
areal data  
modeling

- 1 Stationary spatial process models
- 2 Generalized linear spatial process modeling
- 3 Areal data models
- 4 General linear areal data modeling

## Hierarchical model



- The basic model is

$$Y(\mathbf{s}) = \mu(\mathbf{s}) + w(\mathbf{s}) + \epsilon(\mathbf{s}) \quad (1)$$

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

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

$$\Sigma = \sigma^2 H(\phi) + \tau^2 I \quad (2)$$

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

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

$$p(\theta|\mathbf{y}) \propto f(\mathbf{y}|\theta)p(\theta) \quad (3)$$

where

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

- Typically, independent priors are chosen for the different parameters

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

- We usually adopt the following priors

$$\boldsymbol{\beta} \sim N_{p+1}(\mu_0, \Lambda_0) \quad (6)$$

$$\sigma^2 \sim \Gamma^{-1}(a_1, b_1) \quad (7)$$

$$\tau^2 \sim \Gamma^{-1}(a_2, b_2) \quad (8)$$

$$\phi \sim \Gamma(a_3, b_3) \quad (9)$$

- 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  $\phi$  (e.g. uniform over interval) and a relatively vague prior is used for  $\sigma^2$ .

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

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

- This implies a flat prior for  $\boldsymbol{\beta}$ . With a uniform prior for  $\phi$ , 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  $\epsilon$ .
- 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).

- The hierarchical modeling shares the following generic format

$$[\text{data}|\text{process, parameters}] [\text{process}|\text{parameters}] [\text{parameters}] \quad (11)$$

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

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

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

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

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



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

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

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

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

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

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

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

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

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

Hierarchical  
Model

S.Lan

Stationary  
spatial process  
models

Generalized  
linear spatial  
process  
modeling

Areal data  
models

General linear  
areal data  
modeling

- 1 Stationary spatial process models
- 2 Generalized linear spatial process modeling
- 3 Areal data models
- 4 General linear areal data modeling

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

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

where  $g(\eta(\mathbf{s}_i)) = \mathbf{x}^T(\mathbf{s}_i)\beta + w(\mathbf{s}_i)$  for some link function  $g$ , and  $\gamma$  is a dispersion parameter.

- We presume the  $w(\mathbf{s}_i)$  to be the spatial random effect coming from a Gaussian process, i.e.  $\mathbf{W} \sim N(\mathbf{0}, \sigma^2 H(\phi))$ .

- Using conditional independence, we have the joint distribution

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

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

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

for some link function  $g(\cdot)$  such that  $g^{-1}$  takes  $p(\mathbf{s}) := \Pr(Y(\mathbf{s}) = 1)$  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

Hierarchical  
Model

S.Lan

Stationary  
spatial process  
models

Generalized  
linear spatial  
process  
modeling

Areal data  
models

General linear  
areal data  
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  $\beta$ , a  $\text{unif}(0, 10)$  prior for  $\phi$  and a  $\Gamma^{-1}(0.1, 0.1)$  prior for  $\sigma^2$ .

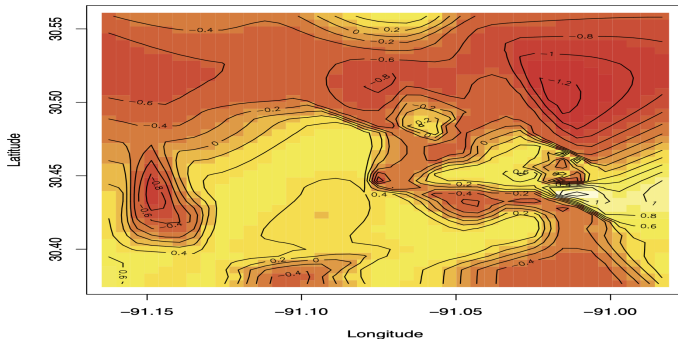


Figure 6.7 Image plot of the posterior median surface of the latent spatial process  $w(\mathbf{s})$ , binary 14 / 27

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

$$Y(\mathbf{s}) \sim \text{pois}(g^{-1}(Z(\mathbf{s}))) \quad (21)$$

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.

Hierarchical  
Model

S.Lan

Stationary  
spatial process  
models

Generalized  
linear spatial  
process  
modeling

Areal data  
models

General linear  
areal data  
modeling

- 1 Stationary spatial process models
- 2 Generalized linear spatial process modeling
- 3 Areal data models
- 4 General linear areal data modeling



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

$Y_i$  = observed number of cases of disease in county  $i, i = 1, \dots, I$

$E_i$  = expected number of cases of disease in county  $i, i = 1, \dots, I$

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

$$E_i = n_i \bar{r} = n_i \frac{\sum_i y_i}{\sum_i n_i} = \sum_i y_i \frac{n_i}{\sum_i n_i} \quad (22)$$

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

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

$$E_i = n_{ij}r_j \quad (23)$$

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

- The usual model for  $Y_i$  is the Poisson model

$$Y_i | \eta_i \sim \text{Pois}(E_i \eta_i) \quad (24)$$

where  $\eta_i$  is the true *relative risk* of disease in region  $i$ .

- The maximum likelihood estimate (MLE) of  $\eta_i$  is the *standard mortality ratio*

$$\hat{\eta}_i = \text{SMR}_i = \frac{Y_i}{E_i} \quad (25)$$

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

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

$$Y_i | \eta_i \stackrel{ind}{\sim} \text{pois}(E_i \eta_i), \quad i = 1, \dots, I \quad (26)$$

$$\eta_i \stackrel{iid}{\sim} \Gamma(a, b) \quad (27)$$

- 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  $\eta_i$  is

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

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

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

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

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

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

- $\theta_i$ '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”.

- In specifying the covariance  $H(\lambda)$ , 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  $\phi$ :

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

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

- Full conditional of  $\phi_i$

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

- There are two issues: i) impropriety; ii) selection of  $\tau_h$  and  $\tau_c$ .
- 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  $\theta_i$  and  $\phi_i$  in the model. One can choose to fix  $\tau_h$  and  $\tau_c$ .
- 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:

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

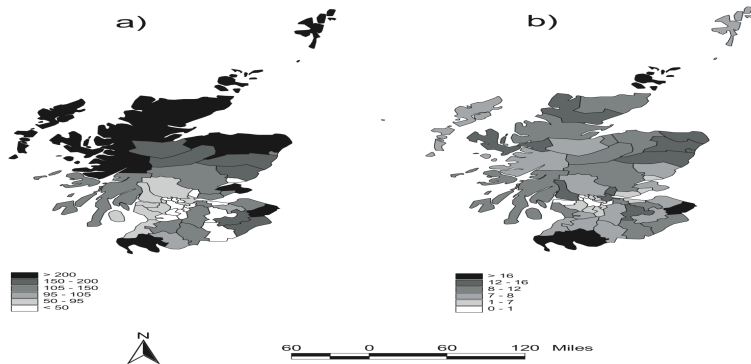


Figure 6.15 Scotland lip cancer data: (a) crude standardized mortality ratios (observed/expected  $\times 100$ ); (b) AFF covariate values.

- 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  $x_i$ , the percentage of the population engaged in agriculture, fishing or forestry (AFF), is also available.

- We fit the model of log-relative risk

$$\psi_i = \beta_0 + \beta_1 x_i + \theta_i + \phi_i \quad (35)$$

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

$$\alpha = \frac{\text{sd}(\phi)}{\text{sd}(\theta) + \text{sd}(\phi)} \quad (36)$$

Priors for $\tau_c, \tau_h$	Posterior for $\alpha$			Posterior for $\beta$		
	mean	sd	llacf	mean	sd	llacf
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
Priors for $\tau_c, \tau_h$	Posterior for $\xi_1$			Posterior for $\xi_{56}$		
	mean	sd	llacf	mean	sd	llacf
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  $\tau_c$  and  $\tau_h$ , Scotland lip cancer data; “sd” denotes standard deviation while “llacf” denotes lag 1 sample autocorrelation.*



Hierarchical  
Model

S.Lan

Stationary  
spatial process  
models

Generalized  
linear spatial  
process  
modeling

Areal data  
models

General linear  
areal data  
modeling

- 1 Stationary spatial process models
- 2 Generalized linear spatial process modeling
- 3 Areal data models
- 4 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  $\beta$  and  $\phi_i$ ,  $Y_i$  is modeled conditionally independent:

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

where  $g(\eta_i) = \mathbf{x}_i^T \beta + \phi_i$  for some link function  $g$  with  $\gamma$  is a dispersion parameter.

- The  $\phi_i$ '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  $\theta_i$ 's in this generalized linear mixed model with random spatial effects.

# Point-referenced models vs areal data models

Hierarchical  
Model

S.Lan

Stationary  
spatial process  
models

Generalized  
linear spatial  
process  
modeling

Areal data  
models

General linear  
areal data  
modeling

- Point-referenced data models are defined with regard to an uncountable number of random variables  $Y(\mathbf{s})$ . For areal units, we envision only a single, finite,  $n$ -dimensional distribution for the  $Y_i$ ,  $i = 1, \dots, I$ .
- With point-referenced data  $\mathbf{Y} = (Y(\mathbf{s}_1), \dots, Y(\mathbf{s}_n))'$ , we model association directly by covariance  $\Sigma_{\mathbf{Y}}$ . With areal data  $\mathbf{Y} = (Y_1, \dots, Y_n)'$  and CAR (or SAR) specifications, we instead model the precision  $\Sigma_{\mathbf{Y}}^{-1}$  directly.
- $\Sigma_{\mathbf{Y}}$  encodes *unconditional* association structure; while  $\Sigma_{\mathbf{Y}}^{-1}$  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  $\Sigma_{\mathbf{Y}}^{-1}$  and  $|\Sigma_{\mathbf{Y}}|$  in the likelihood; not for CAR (full conditionals) or SAR (no matrix inversion in likelihood).