Spatial Autoregressive Models

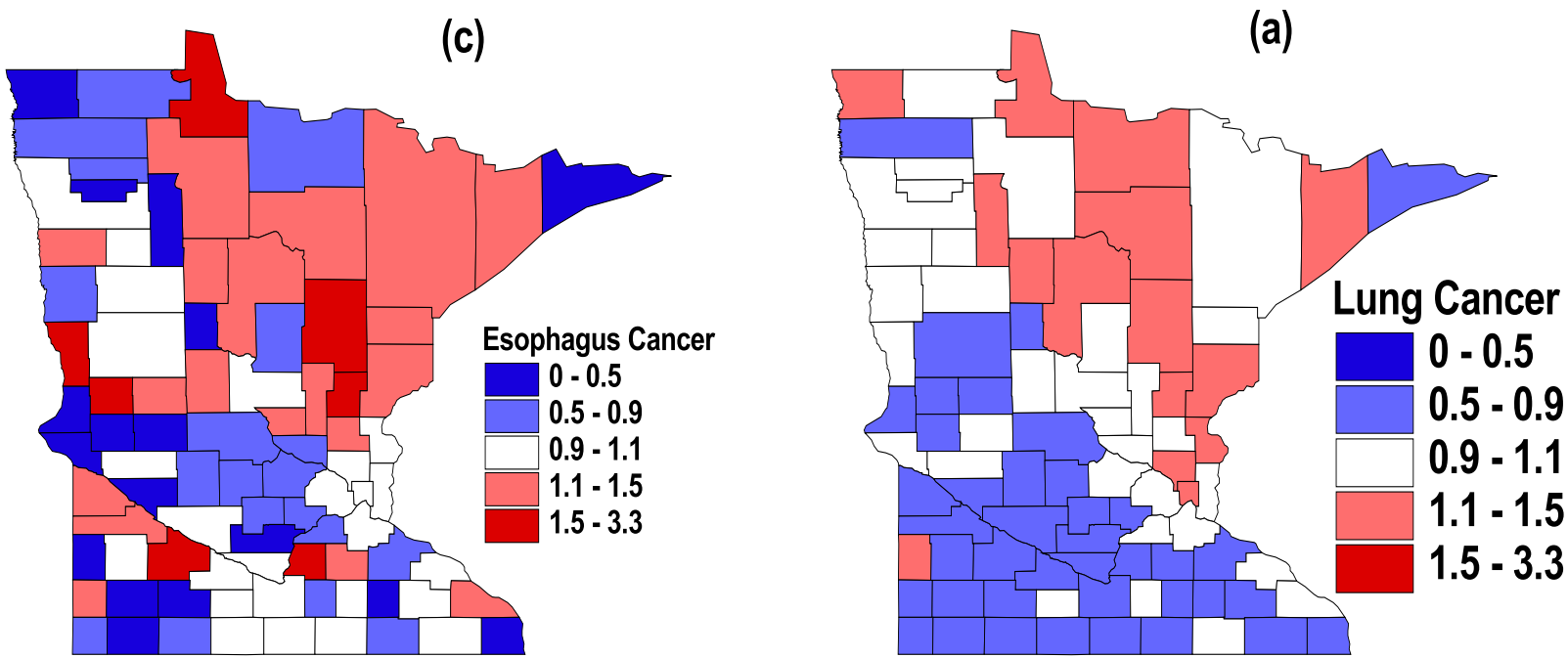
Sudipto Banerjee[1] and Andrew O. Finley[2]

[1] Biostatistics, School of Public Health, University of Minnesota, Minneapolis, Minnesota, U.S.A.

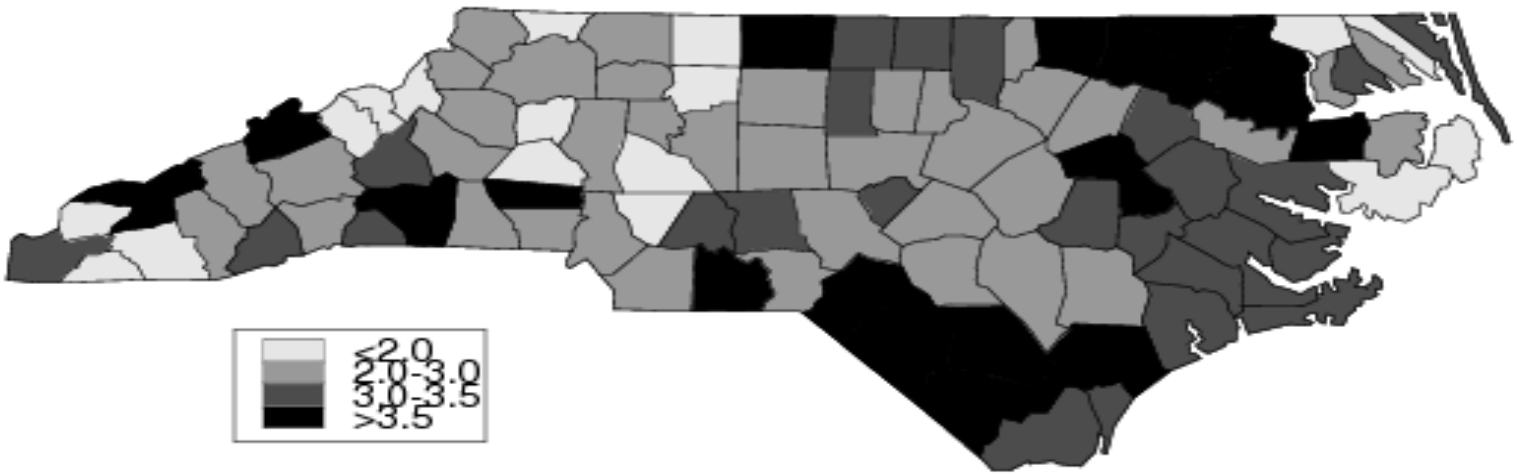
[2] Department of Forestry & Department of Geography, Michigan State University, Lansing Michigan, U.S.A.

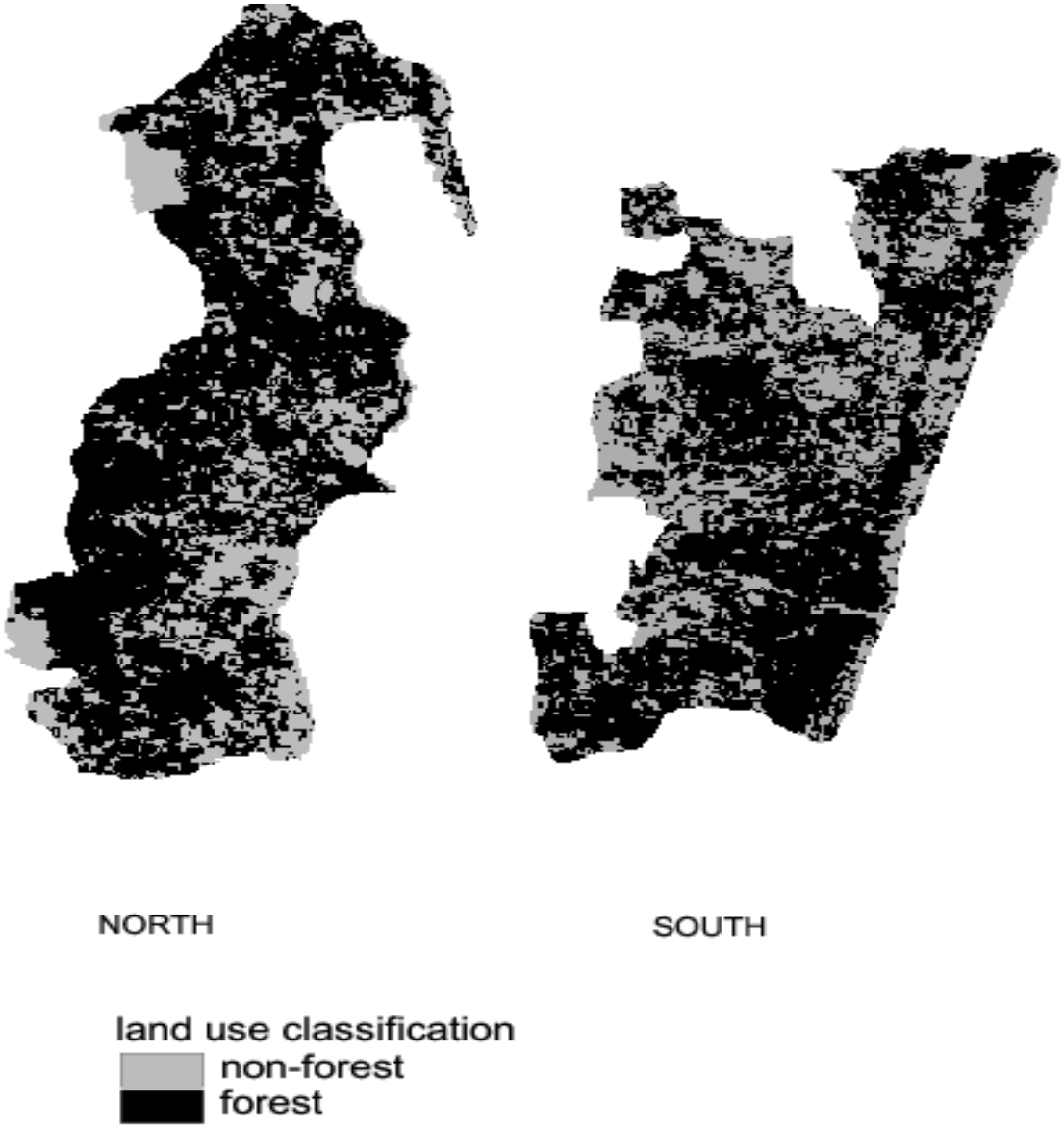
October 15, 2012

Maps of raw standard mortality ratios (SMR) of lung and esophagus cancer between 1991 and 1998 in Minnesota counties



Actual Transformed SIDS Rates




NORTH
SOUTH
land use classification
non-forest
forest

Key Issues

- Is there *spatial* pattern? *Spatial pattern* implies that observations from units closer to each other are more similar than those recorded in units farther away.
- Do we want to *smooth* the data? Perhaps to adjust for low population sizes (or sample sizes) in certain units? How much do we want to smooth?
- Inference for *new* areal units? Is prediction meaningful here? If we modify the areal units to new units (e.g. from zip codes to county values), what can we say about the new counts we expect for the latter given those for the former? This is the Modifiable Areal Unit Problem (MAUP) or Misalignment.

- W, entries w_{ij}, ($w_{ii} = 0$); choices for w_{ij}:
  - $w_{ij} = 1$ if i, j share a common boundary (possibly a common vertex)
  - w_{ij} is an *inverse* distance between units
  - $w_{ij} = 1$ if distance between units is $\leq K$
  - $w_{ij} = 1$ for m nearest neighbors.
- W need not be symmetric.
- $\widetilde{W}$: standardize row i by $w_{i+} = \sum_j w_{ij}$ (row stochastic but need not be symmetric).
- W elements often called “weights”; nicer interpretation?

- Note that proximity matrices are user-defined.
- We can define distance intervals, $(0, d_1]$, $(d_1, d_2]$, and so on.
  - First order neighbours: all units within distance d_1.
  - First order proximity matrix $W^{(1)}$. Analogous to W, $w_{ij}^{(1)} = 1$ if i and j are first order neighbors; 0 otherwise.
  - Second order neighbors: all units within distance d_2, but separated by more than d_1.
  - Second order proximity matrix $W^{(2)}$; $w_{ij}^{(2)} = 1$ if i and j are second order neighbors; 0 otherwise
  - And so on...

- There are analogues for areal data of the empirical correlation function and the variogram.
- Moran's I: analogue of lagged autocorrelation

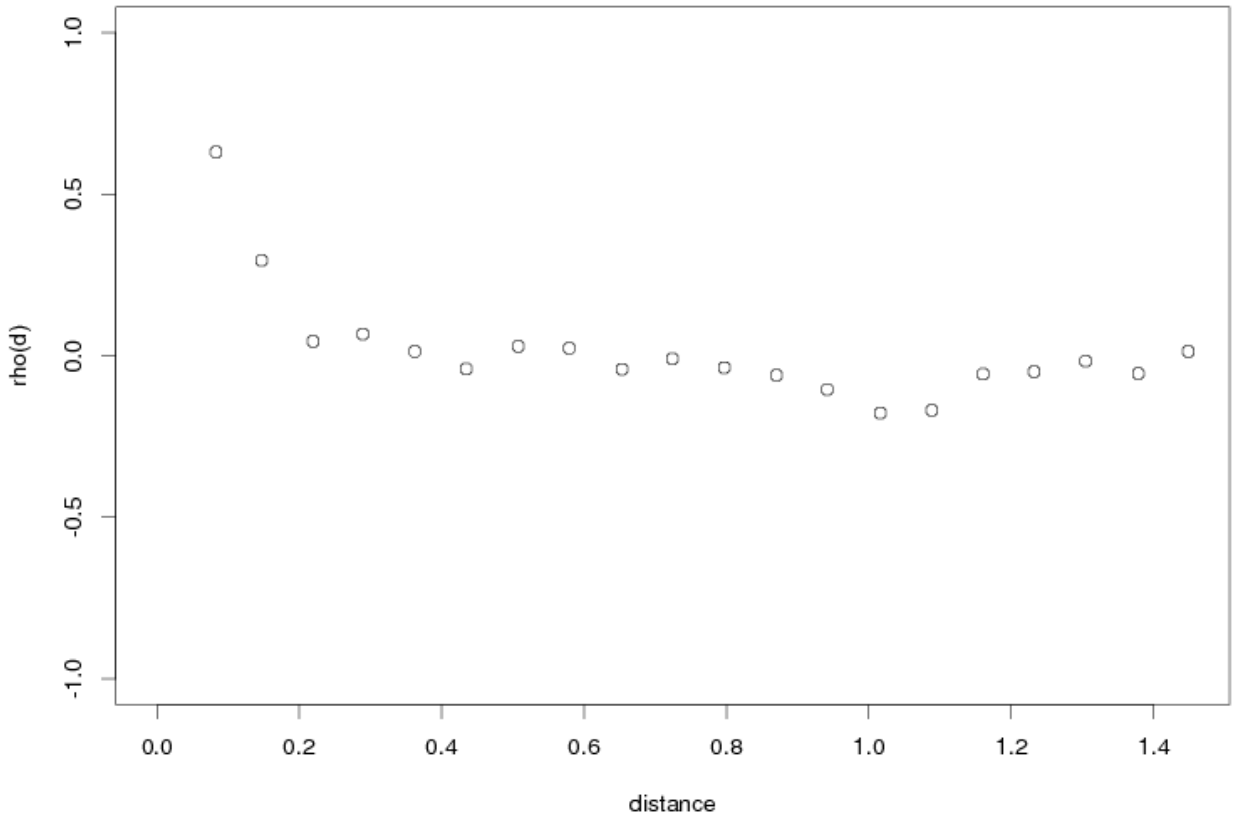
$$I = \frac{n \sum_i \sum_j w_{ij}(Y_i - \bar{Y})(Y_j - \bar{Y})}{(\sum_{i \neq j} w_{ij})(\sum_i (Y_i - \bar{Y})^2}$$

  I is not supported on $[-1, 1]$.
- Geary's C: analogue of Durbin-Watson statistic

$$C = \frac{(n-1) \sum_i \sum_j w_{ij}(Y_i - Y_j)^2}{\sum_{i \neq j} w_{ij}) \sum_i (Y_i - \bar{Y})^2}$$

- Both are asymptotically normal if Y_i are i.i.d., the first with mean $-1/(n-1)$ and the second with mean 1.
- Significance testing using a Monte Carlo test, permutation invariance

- The *areal correlogram* is a useful tool to study spatial association with areal data.
- Working with I, we can replace w_{ij} with $w_{ij}^{(1)}$ taken from $W^{(1)}$ and compute $\rightarrow I^{(1)}$
- Next replace w_{ij} with $w_{ij}^{(2)}$ taken from $W^{(2)}$ and compute $\rightarrow I^{(2)}$, etc.
- Plot $I^{(r)}$ vs. r
- If there is spatial pattern, we expect $I^{(r)}$ to decline in r initially and then vary about 0.


rho(d)
1.0
0.5
0.0
-0.5
-1.0
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4
distance

- To smooth Y_i, replace with $\hat{Y}_i = \frac{\sum_i w_{ij} Y_j}{w_{i+}}$ Note: K-nearest neighbours (KNN) regression falls within this framework.
- More generally,

$$(1-\alpha)Y_i + \alpha\hat{Y}_i$$

  Linear (convex) combination, shrinkage
- Model-based smoothing, e.g., $E(Y_i|\{Y_j,\ j = 1, 2, ..., n\})$

- First, consider $\mathbf{Y} = (y_1, y_2, ..., y_n)$ and consider the set $\{p(y_i \,|\, y_j,\ j \neq i)\}$
- We know $p(y_1, y_2, ...y_n)$ determines $\{p(y_i|y_j, j \neq i)\}$ (full conditional distributions)
- ??? Does $\{p(y_i \,|\, y_j, j \neq i)\}$ determine $p(y_1, y_2, ...y_n)$? If so, we call the joint distribution a *Markov Random Field*.
- In general we cannot write down an arbitrary set of conditionals and assert that they determine the joint distribution. Example:

$$Y_1 \,|\, Y_2 \sim N(\alpha_0 + \alpha_1 Y_2, \sigma_1^2)$$
$$Y_2 \,|\, Y_1 \sim N(\beta_0 + \beta_1 Y_1^3, \sigma_2^2).$$

- The first equation implies that $E[Y_1] = \alpha_0 + \alpha_1 E[Y_2]$, i.e., $E[Y_1]$ is linear in $E[Y_2]$. The second equation implies that $E[Y_2] = \beta_0 + \beta_1 E[Y_1^3]$, i.e. $E[Y_2]$ is linear in $E[Y_1^3]$. Clearly this isn't true in general. Hence no joint distribution.

- Also $p(y_1, \ldots, y_n)$ may be improper even if all the full conditionals are proper.

$$p(y_1, y_2) \propto \exp\{(y_1 - y_2)^2\}$$

But $p(Y_2 \mid Y_1) \propto N(Y_2)$ and $p(Y_1 \mid Y_2) \propto N(Y_2, 1)$. Yet the joint distribution is improper.

- Compatibility: Brook's Lemma. Let $\mathbf{y}_0 = (y_{10}, \ldots, y_{n0})$ be any fixed point in the support of $p(\cdot)$.

$$p(y_1, \ldots, y_n) = \frac{p(y_1 \mid y_2, \ldots, y_n)}{p(y_{10} \mid y_2, \ldots, y_n)} \frac{p(y_2 \mid y_{10}, y_3, \ldots, y_n)}{p(y_{20} \mid y_{10}, y_3, \ldots, y_n)} \cdots \frac{p(y_n \mid y_{10}, \ldots, y_{n-1,0})}{p(y_{n0} \mid y_{10}, \ldots, y_{n-1,0})} p(y_{10}, \ldots, y_{n0}).$$

If LHS is proper, the fact that it integrates to 1 determines the normalizing constant!

- Suppose we want:

$$p(y_i \mid y_j, j \neq i) = p(y_i \mid y_j \in \partial_i)$$

- When does the set $\{p(y_i \mid y_j \in \partial_i)\}$ uniquely determine $p(y_1, y_2, ...y_n)$?
- To answer this question, we need the following important concepts:
  - Clique: A clique is a set of cells such that each element is a neighbor of every other element. We use notation $i \sim j$ if i is a neighbor of j and j is a neighbor of i.
  - Potential: A potential of order k is a function of k arguments that is exchangeable in these arguments. The arguments of the potential would be the values taken by variables associated with the cells for a clique of size k.

- For clique size say 2, $i \sim j$ means $j \sim i$
- For continuous data: $Q(y_i, y_j) = y_i y_j \qquad (\Leftrightarrow (y_i - y_j)^2)$
- For binary data:
  $Q(y_i, y_j) = I(y_i = y_j) = y_i y_j = (1 - y_i)(1 - y_j)$
- Cliques of size 1 $\Leftrightarrow$ independence
- Cliques of size 2 $\Leftrightarrow$ pairwise difference form

$$p(y_1, y_2, \ldots y_n) \propto \exp\left\{ -\frac{1}{2\tau^2} \sum_{i,j} (y_i - y_j)^2 I(i \sim j) \right\}$$

and therefore $p(y_i \,|\, y_j, j \neq i) = N(\sum_{j \in \partial_i} y_i/m_i, \tau^2/m_i)$, where m_i is the number of neighbors of i

- Gibbs distribution: $p(y_1, \ldots, y_n)$ is a Gibbs distribution if it is a function of the y_i's only through potentials on cliques:

$$p(y_1, \ldots, y_n) \propto \exp\{-\gamma \sum_k \sum_{\alpha \in M_k} \phi^{(k)}(y_{\alpha_1}, \ldots, y_{\alpha_k})\},$$

  where $\phi^{(k)}$ is a potential of order k, M_k is the set of all cliques of size k and is indexed by α, and $\gamma > 0$ is a scale parameter.
- Hammersley-Clifford Theorem: If we have a Markov Random Field (i.e., $\{p(y_i \mid y_j \in \partial_i)\}$ uniquely determine $p(y_1, y_2, ...y_n)$), then the latter is a Gibbs distribution
- Geman and Geman (1984) result : If we have a joint Gibbs distribution, then we have a Markov Random Field

Conditionally Auto-Regressive (CAR) models

- Gaussian (autonormal) case

$$p(y_i \mid y_j, j \neq i) = N\left(\sum_j b_{ij} y_j, \tau_i^2\right)$$

- Using Brook's Lemma we can obtain

$$p(y_1, y_2, \ldots y_n) \propto \exp\left\{-\frac{1}{2}\mathbf{y}' D^{-1}(I - B)\mathbf{y}\right\}$$

  where $B = \{b_{ij}\}$ and D is diagonal with $D_{ii} = \tau_i^2$.
- Suggests a multivariate normal distribution with $\mu_y = 0$ and $\Sigma_Y = (I - B)^{-1} D$
- $D^{-1}(I - B)$ symmetric requires

$$\frac{b_{ij}}{\tau_i^2} = \frac{b_{ji}}{\tau_j^2} \text{ for all } i, j$$

- Returning to W, let $b_{ij} = w_{ij}/w_{i+}$ and $\tau_i^2 = \tau^2/w_{i+}$, so

$$p(y_1, y_2, ...y_n) \propto \exp\{-\frac{1}{2\tau^2}\mathbf{y}'(D_w - W)\mathbf{y}\}$$

where D_w is diagonal with $(D_w)_{ii} = w_{i+}$ and thus

$$p(y_1, y_2, ...y_n) \propto \exp\left\{-\frac{1}{2\tau^2}\sum_{i\neq j} w_{ij}(y_i - y_j)^2\right\}$$

- Caution: $(D_w - W)\mathbf{1} = \mathbf{0}$. Intrinsic autoregressive (IAR) model; improper, so requires a constraint (e.g., $\sum_i y_i = 0$)
- Not a valid data model, but only as a random effects model!

- With τ^2 unknown, what should be the power of τ^2? Answer:

$$p(y_1, y_2, ...y_n) \propto (\frac{1}{\tau^2})^{(n-G)/2} \exp\{-\frac{1}{2\tau^2}\mathbf{y}'(D_w - W)\mathbf{y}\},$$

where G is the number of "islands" in the map. In fact, $n - G$ is the rank of $D_w - W$.

- The impropriety can be remedied in an obvious way. Redefine the CAR as:

$$p(y_1, y_2, ...y_n) \propto |D_w - \rho W|^{1/2} \exp\{-\frac{1}{2\tau^2}\mathbf{y}'(D_w - \rho W)\mathbf{y}\},$$

where ρ is chosen to make $D_w - \rho W$ non-singular. This is guaranteed if $\rho \in \left(1/\lambda_{(1)}, 1\right)$, where $\lambda_{(1)}$ is the minimum eigenvalue of $D^{-1/2}WD^{-1/2}$. In practice, the bound $\rho \in (0, 1)$ is often preferred.

To ρ or not to ρ?

- Advantages:
  - makes distribution proper
  - adds parametric flexibility
  - $\rho = 0$ interpretable as independence
- Disadvantages:
  - why should we expect y_i to be a proportion of average of neighbors - sensible spatial interpretation?
  - calibration of ρ as a correlation, e.g.,

$$\rho = 0.80 \text{ yields } 0.1 \leq I \leq 0.15,$$
$$\rho = 0.90 \text{ yields } 0.2 \leq I \leq 0.25,$$
$$\rho = 0.99 \text{ yields } I \leq 0.5$$

  - So, used with random effects, scope of spatial pattern may be limited

Example of a hierarchical model with CAR effects.

- Consider the areal data disease mapping model:

$$
\begin{aligned}
Y_i \mid \mu_i &\overset{ind}{\sim} Po\left(E_i\, e^{\mu_i}\right) \ , \quad \text{where} \\
Y_i &= \text{observed disease count,} \\
E_i &= \text{expected count (known), and} \\
\mu_i &= \mathbf{x}_i'\boldsymbol{\beta} + \phi_i; \ \text{the } \mathbf{x}_i \text{ are explanatory variables}
\end{aligned}
$$

- The ϕ_i capture regional clustering via a conditionally autoregressive (CAR) prior,

$$
\phi_i \mid \phi_{j \neq i} \sim N\left(\bar{\phi}_i \ , \ \frac{\tau^2}{m_i}\right) \ , \ \text{where } \bar{\phi}_i = \frac{1}{m_i} \sum_{j \in \partial_i} \phi_j;
$$

∂_i is the set of "neighbours" of region i, and m_i is the number of these neighbours.

Comments on CAR models

- We specify Σ_y^{-1}, not directly modeling association
- $(\Sigma_y^{-1})_{ii} = 1/\tau_i^2$; $(\Sigma_y^{-1})_{ij} = 0 \Leftrightarrow$ cond'l independence
- Ad hoc prediction: If

$$p(y_0 \,|\, y_1, y_2, ...y_n) = N(\sum_j w_{0j} y_j / w_{0+}, \tau^2 / w_{0+})$$

  then $p(y_0, y_1, ...y_n)$ well-defined but not CAR
- non-Gaussian case, binary data (autologistic)

$$p(y_i \,|\, y_j, j \neq i) \propto \exp\{\phi \sum_j w_{ij} I(y_i = y_j)\}$$

Simultaneous Auto-Regressive (SAR) models

- We may write the CAR model as:

$$\mathbf{y} = B\mathbf{y} + \boldsymbol{\epsilon} \Rightarrow (I - B)\mathbf{y} = \boldsymbol{\epsilon};$$

  Since $\mathbf{y} \sim N(\mathbf{0}, (I - B)^{-1}D)$, we have

$$\boldsymbol{\epsilon} \sim N(\mathbf{0}, D(I - B)').$$

- Instead of letting $\mathbf{y}$ induce the distribution of $\boldsymbol{\epsilon}$, let $\boldsymbol{\epsilon}$ induce a distribution for $\mathbf{y}$. Letting $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \tilde{D})$, where $\tilde{D}$ is diagonal, $\tilde{D}_{ii} = \sigma_i^2$ and let:

$$y_i = \sum_{j=1}^{n} b_{ij} y_j + \epsilon_i.$$

  Assuming $(I - B)^{-1}$ exists, we obtain:

$$\mathbf{y} \sim N\left(\mathbf{0}, (I - B)^{-1}\tilde{D}(I - B)'^{-1}\right).$$

- Often we take $B = \rho W$. If $\rho \in (1/\lambda_{(1)}, 1/\lambda_{(n)})$, where $\lambda_{(1)}$ and $\lambda_{(n)}$ are the minimum and maximum eigenvalues of W. This ensures $(I - \rho W)^{-1}$ exists.
- Alternatively, we can replace W with $\tilde{W} = \{w_{ij}/w_{i+}\}$ where w_{i+} is the sum of the elements in the i-th row of W. Then $|\rho| < 1$ ensures existence of $(I - \rho\tilde{W})^{-1}$.
- Often SAR models are also applied to point-referenced data where W is taken to be the inter-point distance.

- Two variants:
  - The SAR "lag model":

$$\mathbf{y} = B\mathbf{y} + X\boldsymbol{\beta} + \boldsymbol{\epsilon}.$$

  - The SAR "residual" or "error model":

$$(I - B)(\mathbf{y} - X\boldsymbol{\beta}) = \boldsymbol{\epsilon}; \Rightarrow \mathbf{y} = B\mathbf{y} + (I - B)X\boldsymbol{\beta} + \boldsymbol{\epsilon}.$$

- SAR models are well suited to maximum likelihood estimation but not at all for MCMC fitting of Bayesian models. Because it is difficult to introduce SAR random effects (in the CAR framework this is easy because of the hierarchical conditional representation).