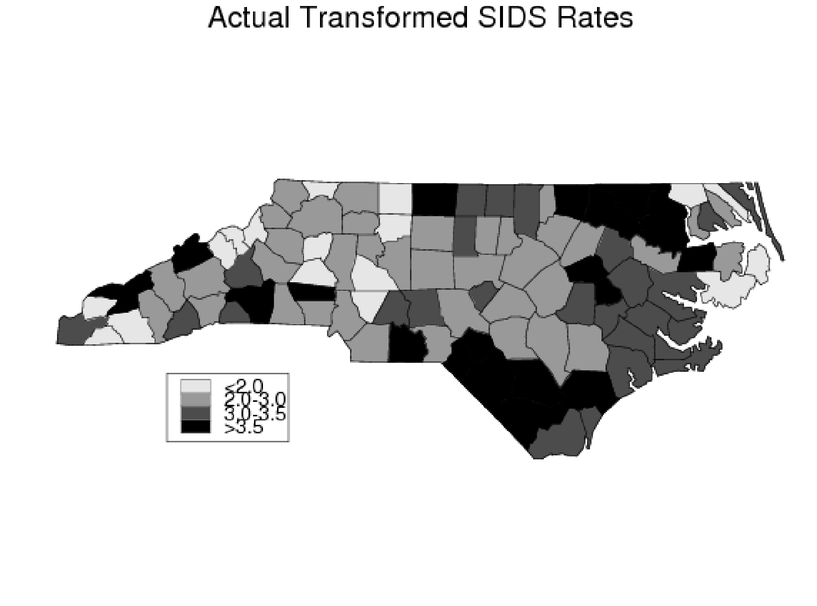
# Basics of areal data models

Dipankar Bandyopadhyay

Div. of Biostatistics, School of Public Health, University of Minnesota, Minneapolis, Minnesota, U.S.A.

## 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. If *no pattern, i.e. i.i.d* data, values are randomly arranged.
- Do we want to *smooth* the data? Perhaps to adjust for low population sizes (or sample sizes) in certain units? A surface of *expected* counts, the higher values are pulled down, and lower up! 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.


Actual Transformed SIDS Rates
<2.0
2.0-3.0
3.0-3.5
>3.5

- $W$ (proximity matrix), 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? Larger $W$ if $j$ is closer to $i$.
- Could also define first-order neighbors $W^{(1)}$, second-order neighbors $W^{(2)}$, etc.

- Note that proximity matrices are user-defined.
- We can define distance intervals, $(0, d_1]$, $(d_1, d_2]$, and so on.
  - First order neighbors: 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 $I$: essentially an "areal covariogram"

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

  $I$ is not supported on $[-1, 1]$.

- Geary's $C$: essentially an "areal variogram"

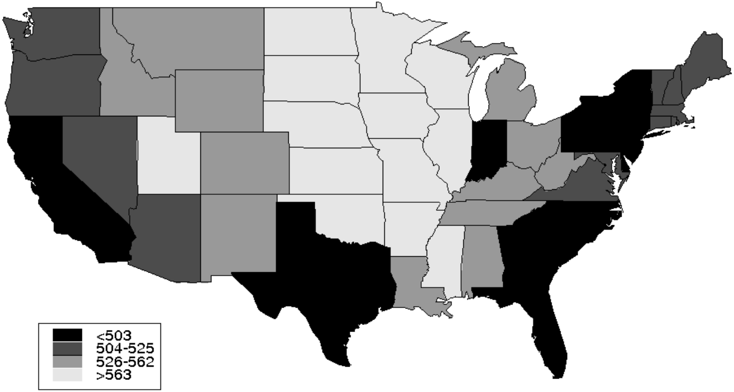
$$C = \frac{(n-1) \sum_i \sum_j w_{ij}(Y_i - Y_j)^2}{(\sum_{i \neq j} w_{ij}) \sum_i (Y_i - \bar{Y})^2}$$

- Both are asymptotically normal if $Y_i$ are i.i.d., the first with mean $-1/(n-1)$ and the second with mean $1$.

- Ratios of quadratic forms

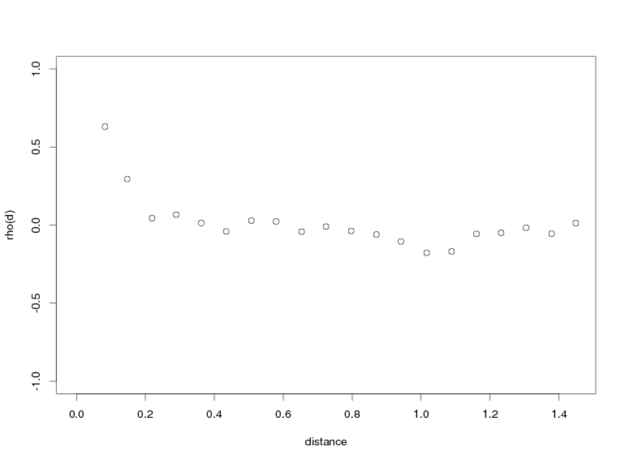
- Relates to ‘Exploratory measure of spatial association’, rather than ‘test of spatial significance’.

# Choropleth map of 1999 average verbal SAT scores across the 48 US (lower) contiguous states.



- For these data, the `spatial.cor` function in `S+SpatialStats` gives a Moran's $I$ of 0.5833, with associated standard error estimate 0.0920 $\Rightarrow$ very strong evidence against $H_0$ : no spatial correlation.
- `spatial.cor` also gives a Geary's $C$ of 0.3775, with associated standard error estimate 0.1008 $\Rightarrow$ again, very strong evidence against $H_0$ (departure from 1)
- Warning: These data have not been adjusted for covariates, such as the proportion of students who take the exam (Midwestern colleges have historically relied on the ACT, not the SAT; only the best and brightest students in these states would bother taking the SAT)
- $\Rightarrow$ the map, $I$, and $C$ all motivate the search for spatial covariates!

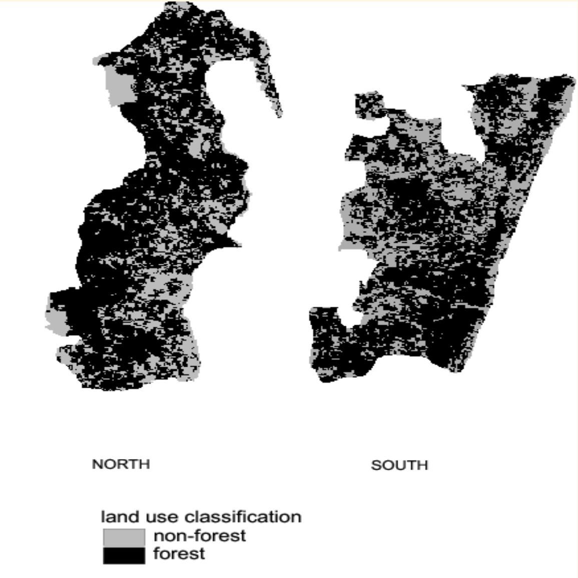
# Correlogram (via Moran's $I$)

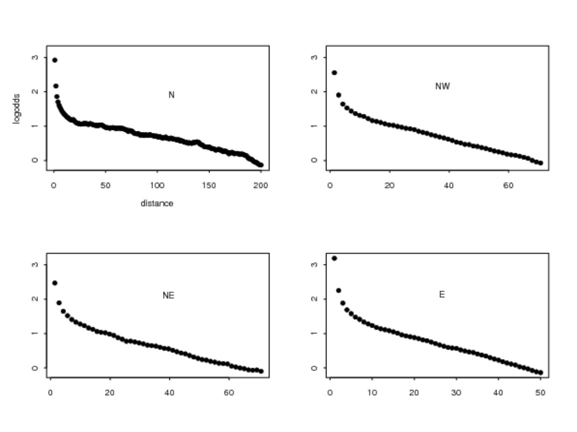


- The *areal correlogram* is a useful tool to study spatial association with areal data.
- Working with $I$, we can replace $w_{ij}$ with $w_{ij}^{(1)}$ taken from $W^{(1)}$ and compute $\rightarrow I^{(1)}$
- Next replace $w_{ij}$ with $w_{ij}^{(2)}$ taken from $W^{(2)}$ and compute $\rightarrow I^{(2)}$, 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.
- spatial analogue of the temporal lag autocorrelation plot

- With large regular grids, we may want to study directional association (say, east-west, north-south, etc).
- Rasterized maps of binary land use in eastern Madagascar, with 25,000 1km X 1km pixels.

# Rasterized binary data map






- A version for a binary map, using two-way tables and log odds ratios at the pixel level.
- Note strongest pattern is to the north (N), but in no direction are the values $\approx 0$ even at 40 km.

## Spatial smoothers

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

- More generally, we can include the actual value $Y_i$

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

  Linear (convex) combination, shrinkage

- Model-based smoothing, i.e. mean of the predictive distribution, 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.

- $f(y_1|y_2)$ and $f(y_2|y_1)$ are incompatible with regards to determining $f(y_1, y_2)$. Hence no joint distribution.
- Also $p(y_1, \ldots, y_n)$ may be improper (does not integrate to 1) even if all the full conditionals are proper.

$$p(y_1, y_2) \propto \exp\left\{-\frac{1}{2}(y_1 - y_2)^2\right\}$$

But $p(Y_2 \,|\, Y_1) \propto N(Y_1, 1)$ and $p(Y_1 \,|\, Y_2) \propto N(Y_2, 1)$. 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 $p(\cdot)$.

$$\begin{aligned} p(y_1, \ldots, y_n) &= \frac{p(y_1 \,|\, y_2, \ldots, y_n)}{p(y_{10} \,|\, y_2, \ldots, y_n)} \frac{p(y_2 \,|\, y_{10}, y_3, \ldots, y_n)}{p(y_{20} \,|\, y_{10}, y_3, \ldots, y_n)} \\ &\quad \cdots \frac{p(y_n \,|\, y_{10}, \ldots, y_{n-1,0})}{p(y_{n0} \,|\, y_{10}, \ldots, y_{n-1,0})} p(y_{10}, \ldots, y_{n0}). \end{aligned}$$

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

Proof: We can rely on the results from conditional probabilities, i.e., $P(A|B)P(B) = P(AB) = P(B|A)P(A)$, and proceed via. induction. Consider the case with 2 arguments. Then, we have:

$$
\begin{aligned}
p(y_1, y_2) &= p(y_1|y_2)p(y_2) = p(y_1|y_2)\frac{p(y_2|y_{1,0})p(y_{1,0})}{p(y_{1,0}|y_2)} \\
&= \frac{p(y_1|y_2)}{p(y_{1,0}|y_2)}\frac{p(y_2|y_{1,0})}{p(y_{2,0}|y_{1,0})}p(y_{2,0}|y_{1,0})p(y_{1,0}) \\
&= \frac{p(y_1|y_2)}{p(y_{1,0}|y_2)}\frac{p(y_2|y_{1,0})}{p(y_{2,0}|y_{1,0})}p(y_{1,0}, y_{2,0})
\end{aligned}
$$

Following this technique, one can prove this for $n$ arguments, given that it holds for $n-1$ arguments.

- Hence, $p(y_1, \ldots, y_n)$ id determined by the full conditional distributions, and the joint density is determined up to a proportionality constant.
- Suppose we want:

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

- When does the set $\{p(y_i \,|\, y_j \in \partial_i)\}$ 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

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

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

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

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

# Conditionally Auto-Regressive (CAR) models

- Gaussian (autonormal) case

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

- Using Brook's Lemma we can obtain

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

  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

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

- Clearly, $B$ is not symmetric.
- Returning to $W$ (assumed to be symmetric), choose $b_{ij} = w_{ij}/w_{i+}$ 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 ,with a little algebra,

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

- 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!

- 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 $\rho$ 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.
- Proper: $D - \rho W \Leftrightarrow y_i \,|\, y_j, j \neq i \sim N\left(\rho \sum_j \frac{w_{ij}}{w_{i+}} y_j, \frac{\tau^2}{w_{i+}}\right)$
- $\rho \sum_j \frac{w_{ij}}{w_{i+}} y_j$ is defined as the 'reaction function', i.e. $\rho$ is the expected proportional 'reaction' of $Y_i$ to $\sum_j \frac{w_{ij}}{w_{i+}} Y_j$.

## To $\rho$ or not to $\rho$?

- **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 $\rho$ 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 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 $\phi_i$ capture regional clustering via a conditionally autoregressive (CAR) prior,

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

$\partial_i$ is the set of “neighbors” of region $i$, and $m_i$ is the number of these neighbors.