Block Referenced Spatial Models

Acknowledgement: Many slides based on / borrowed from Sudipto Banerjee

If you are interested in spatial modeling I recommend: Hierarchical Modeling and Analysis for Spatial Data by Sudipto Banerjee, Alan E. Gelfand, Bradley. P. Carlin (note: text fairly advanced)
Block referenced data

- Data has an location, an attribute and an AREA
- Areas are usually contiguous
- Data often conceived of as being area integrals of some underlying continuous surface

\[
z(B_i) = \frac{1}{|B_i|} \int_{B_i} z(s) \, ds
\]

- Goals
  - Estimate surface \( z(s) \) or new blocks
  - Account for non-independence of adjacent blocks
Figure 1: Choropleth map of 1999 average verbal SAT scores, lower 48 U.S. states.
Proximity matrices

- $W$, entries $w_{ij}$ (with $w_{ii} = 0$). 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.

- $W$ is typically symmetric, but need not be

- $\tilde{W}$: standardize row $i$ by $w_{i+} = \sum_j w_{ij}$

- $W$ elements often called “weights”; interpretation

- Could also define first-order neighbors $W^{(1)}$, second-order neighbors $W^{(2)}$, etc.
Measures of spatial association

- **Moran’s $I$:** essentially an “areal covariogram"

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

- **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.; Moran has mean $-1/(n - 1) \approx 0$, Geary has mean 1
- Significance testing by comparing to a collection of say 1000 random permutations of the $Y_i$
Measures of spatial association (cont’d)

For these data, the Moran’s $I$ is computed as 0.5833, with associated standard error estimate 0.0920 ⇒ very strong evidence against $H_0 : \text{no spatial correlation}$

We obtain a Geary’s $C$ of 0.3775, with associated standard error estimate 0.1008 ⇒ 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)

⇒ the map, $I$, and $C$ all motivate the search for spatial covariates!
Spatial smoothers

- To smooth $Y_i$, replace with $\hat{Y}_i = \frac{\sum_i w_{ij} Y_j}{w_{i+}}$

- More generally, we could include the value actually observed for unit $i$, and revise our smoother to

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

For $0 < \alpha < 1$, this is a linear (convex) combination in “shrinkage” form

Finally, we could try model-based smoothing, i.e., based on $E(Y_i|Data)$, i.e., the mean of the predictive distribution. Smoothers then emerge as byproducts of the hierarchical spatial models we use to explain the $Y_i$’s
Conditional Autoregressive (CAR) Model

\[ y_i = \mu_i + \frac{1}{w_{i+}} \sum_{j \neq i} w_{ij} (y_j - \mu_j) + \epsilon_i \]

- If raster, equivalent to Markov Random Field
- Analogous to AR(1) or our general model for spatial point data

\[ Z(s) = \mu(s|\beta) + \omega(s|\phi) + \epsilon(s) \]

Note: don't naively Google “CAR model”
Conditional Autoregressive (CAR) Model

\[ y_i = \mu_i + \frac{1}{w_{i+}} \sum_{j \neq i} w_{ij} (y_j - \mu_j) + \epsilon_i \]

process model

\[ \text{spatial autocorrelation} \]

\[ \epsilon_i \]

error

\[ \vec{y} \sim N \left( \hat{\mu} \left| (I - \hat{W})^{-1} \sigma^2 I \right. \right) \]

Analogous to time-series

\[ Y_t = \mu + \sum_{i=1}^{p} \rho_i Y_{t-i} + \epsilon_t \rightarrow Y \sim N \left( \mu, \frac{\sigma^2}{1 - \rho^2} R \right) \]
Computation of CAR models

- “GeoBUGS” extension of WinBUGS
Spatial Misalignment Problem

• “Change of support” problem
• Often need to compare / compute / infer spatial data of different types
  – Point – Point (Kriging)
  – Point – Block
  – Block – Point
  – Block - Block
Point to Block

- Collect point data, want to infer the integral of the surface (e.g. county level biomass)
- Traditional approach: sample mean, var
  - Ignores autocorrelation, covariates, etc.
- Recommended Alternative:
  - Bayesian Kriging -> project to a fine grid
  - From each grid, numerically integrate
FIGURE 10.20. Standardized mortality ratios for thirty-nine wards in Birmingham, England, calculated as observed versus expected cases (left), and posterior median relative risk $\gamma(s)$ (right). From Kelsall and Wakefield (2002).
Block-Block Misalignment

Population by census tract; residential structures by “cell”:

- “Areal Allocation”
- Hierarchical Modeling (e.g. CAR)
Bivariate misalignment

Ozone measurements at fixed sites; counts of pediatric asthma cases by zip code in Atlanta, GA:
Bivariate misalignment issues

- When we have two spatially referenced variables, interest often lies in spatial regression.

- But we cannot fit a regression if the two variables are misaligned:
  - X at point level, Y at other points
  - X at point level, Y at block level
  - X at block level, Y at point level
  - X at block level, Y at a different block level

- **Solution:** Bring the X’s to the scale of the Y’s, then fit the model (BCG, Sec 6.4)

- With more than two variables, bring all the variables to a common scale. Highest resolution is obviously preferred, but may be computationally infeasible!