Areal data models

Applied Spatial Statistics

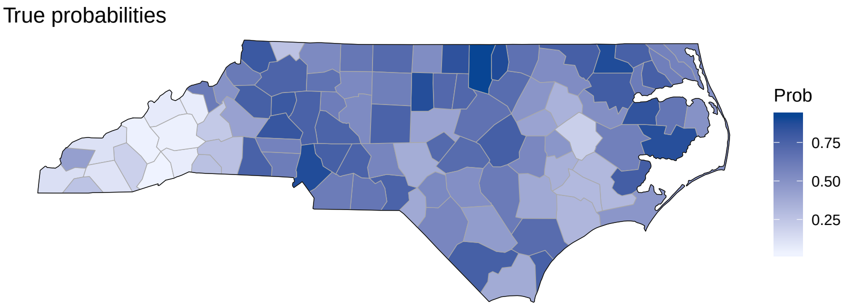
Areal data objectives

- Estimate covariate effects while accounting for dependence
- Borrow strength across space to estimate the true mean each region
- For example, estimating cancer rates in small counties is hard because counts are low
- Averaging across nearby counties can give more precise estimates

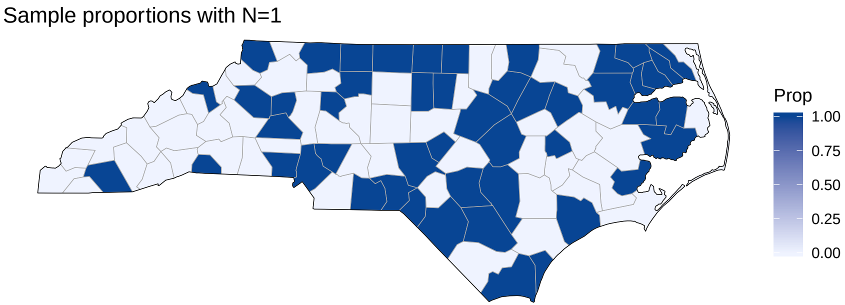
Fake motivating example

- Say the true probability of voting GOP in county i is p_i
- We poll N_i voters in county i and the number that support GOP is $Y_i \sim \text{Binomial}(N_i, p_i)$
- The crude estimate, $\hat{p}_i = Y_i/N_i$, is unstable for counties with small N_i
- Pooling information across neighboring counties might help

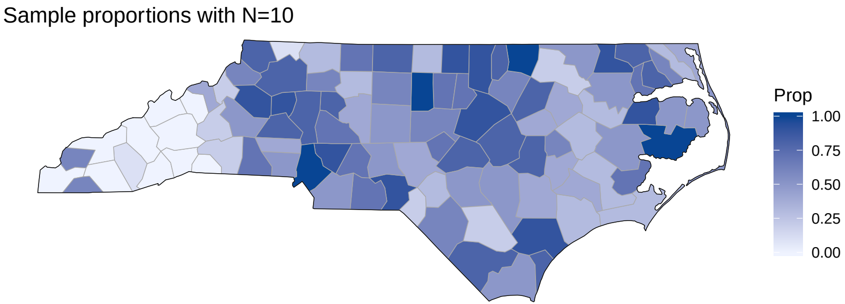
True probability in each county, p_i



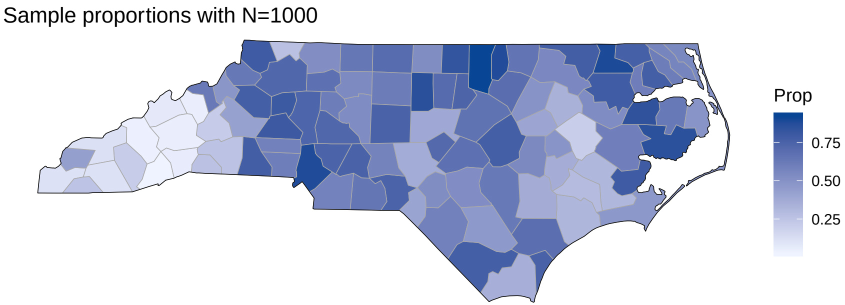
Sample proportions with $N_i = 1$



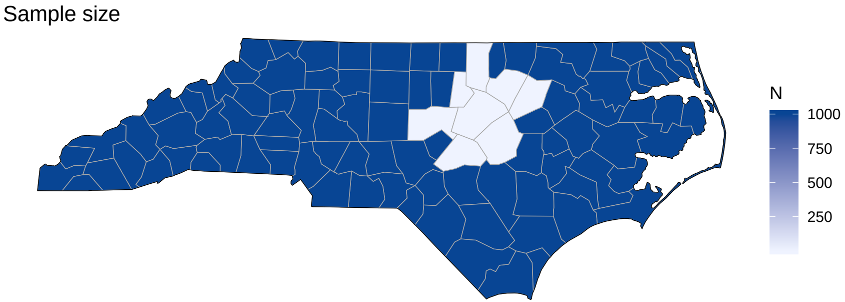
Sample proportions with $N_i = 10$



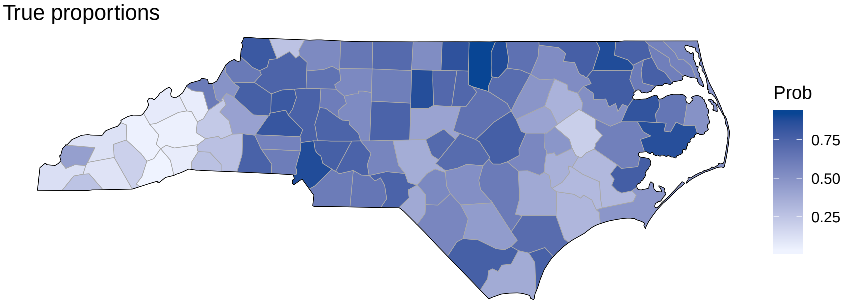
Sample proportions with $N_i = 1000$



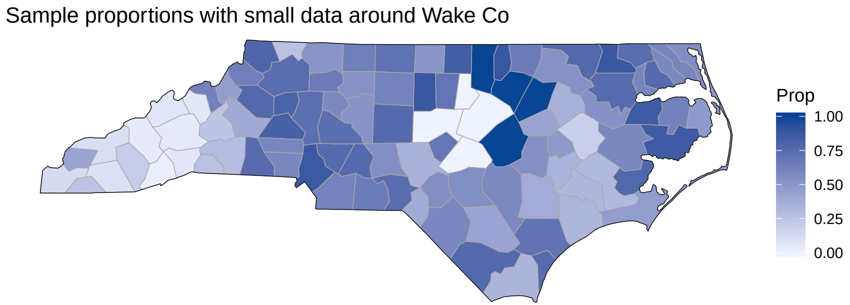
Data with varying N_i



True probability in each county, p_i



Sample proportions



Areal data models

We will start with the familiar model $Y_i = \mathbf{X}_i\beta + Z_i + \varepsilon_i$ for $i \in \{1, ..., n\}$

- The mean $\mathbf{X}_i\beta$ is the same as geostatistical models
- The uncorrelated nugget error is $\varepsilon_i \sim \text{Normal}(0, \tau^2)$
- The spatial term is $\mathbf{Z} = (Z_1, ..., Z_n)^T \sim \text{Normal}(0, \Sigma)$
- As with geostatistics, most of our effort will be dedicated to modeling the $n \times n$ spatial covariance matrix Σ

Applying geostatistical models to areal data

- One option is to assign each region a spatial location and proceed with a geostatistical analysis
- Example: $\mathbf{s}_i$ is the centroid of county i and

$$\Sigma_{ij} = \sigma^2 \exp(-||\mathbf{s}_i - \mathbf{s}_j||/\phi)$$

- This is a valid model in the sense that the covariance is symmetric and positive definite
- It is unsatisfying for irregular regions where distance is difficult to measure

Conditionally autoregressive (CAR) model

- The CAR model is based on adjacency, not distance
- It is defined on the full conditional distributions
- The full conditional distribution is the distribution of Z_i as if all other Z_j are known
- Let Z_{-i} be the collection of the $n-1$ other spatial terms
- Further, define $\bar{Z}_i$ as the mean of Z_j over the m_i regions that neighbor region i

Conditionally autoregressive (CAR) model

- The CAR full conditional distribution of Z_i is

$$Z_i|Z_{-i} \sim \text{Normal}(\rho\bar{Z}_i, \sigma^2/m_i)$$

- Z_i is encouraged to be close it its neighbors, inducing spatial dependence

- The strength of spatial correlation is determined by $\rho \in (0, 1)$

- σ^2 is a variance parameter

- The variance decreases with the number of neighbors

Joint distribution

- The full conditional distributions simultaneously hold for all n regions
- Because the full conditional distribution depends only on neighbors, the model is also called a Gaussian Markov Random Field (GMRF)
- It can be shown that these n full conditional distributions are compatible
- That is, there exists a joint distribution for $\mathbf{Z}$ that gives these full conditionals

Joint distribution

- The joint distribution is MVN with mean zero and

$$\Sigma = (\mathbf{M} - \rho \mathbf{W})^{-1}$$

- $\mathbf{M}$ is the diagonal matrix with diagonal elements $m_1, ..., m_n$
- ρ is the spatial dependence parameter
- $\mathbf{W}$ is the adjacency matrix with elements W_{ij}
- The precision matrix $\Sigma^{-1} = \mathbf{M} - \rho \mathbf{W}$ is sparse

Intrinsic CAR model

- The intrinsic CAR model sets $\rho = 1$ so

$$Z_i | Z_{-i} \sim \text{Normal}(\bar{Z}_i, \sigma^2 / m_i)$$

- This gives one less parameter to estimate

- However, the corresponding covariance matrix is singular

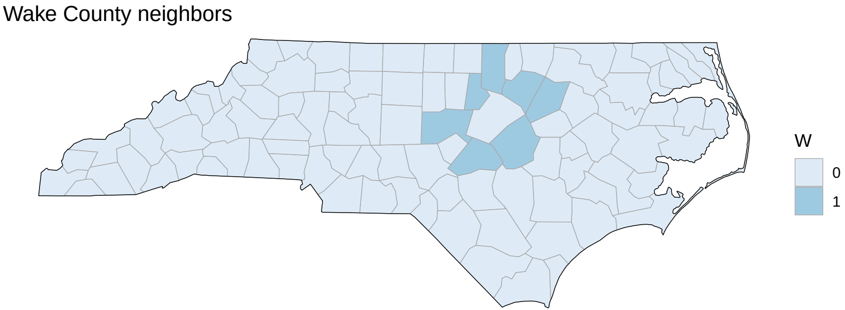
- This implies that the joint MVN distribution is improper, i.e., the PDF does not integrate to one

- This complicates inference, e.g., standard MLE does not apply

Proper CAR model

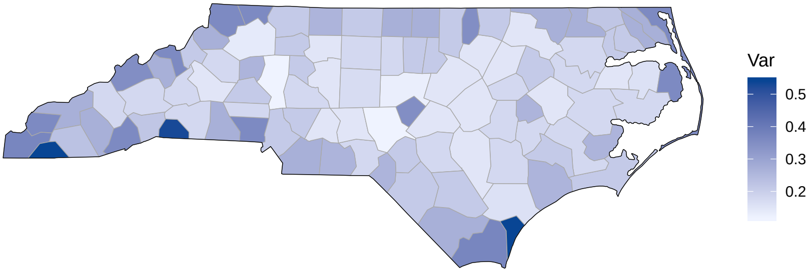
- If $\rho \in [0, 1)$ the MVN distribution is proper and MLE can be used
- Technically, ρ slightly less than zero can also be used
- The lower bound is a complicated function of $\mathbf{W}$
- NOTE: ρ is not the correlation between neighbors
- NOTE: the covariance is non-stationary because $\text{Var}(Z_i) = \Sigma_{ii}$ varies by i

Adjacencies for Wake County

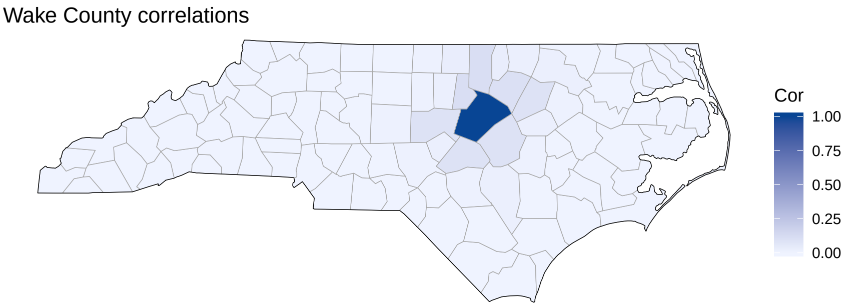


Proper CAR model variance with $\sigma = 1$ and $\rho = 0.5$

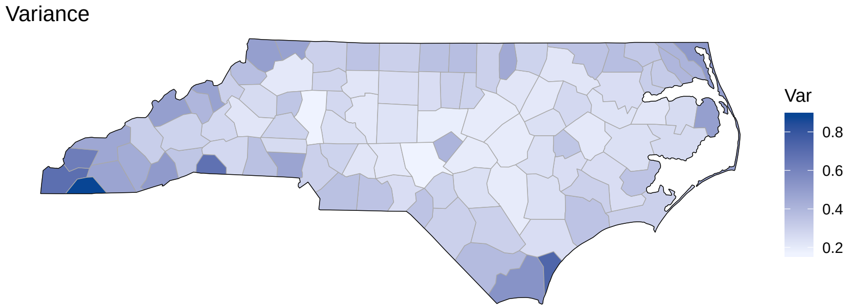
Variance



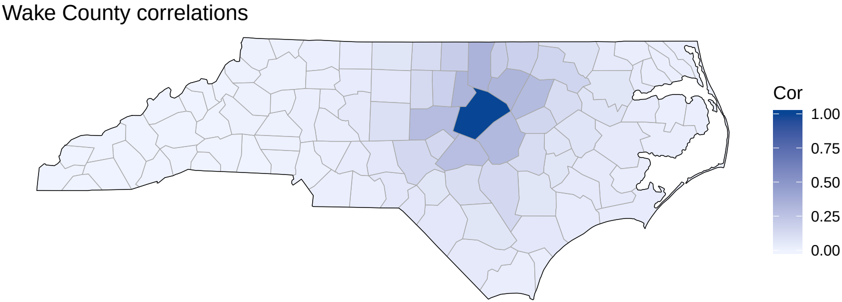
Proper CAR model correlation with $\rho = 0.5$



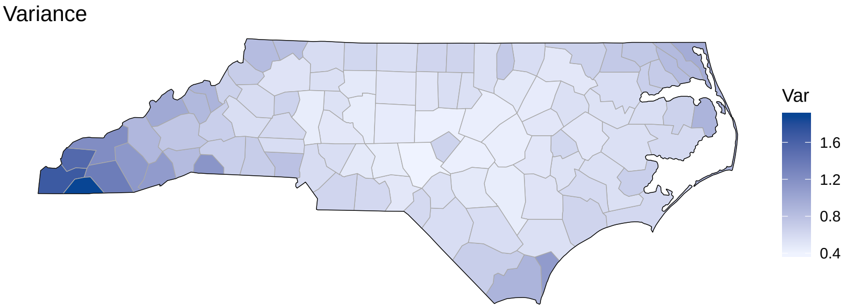
Proper CAR model variance with $\sigma = 1$ and $\rho = 0.9$



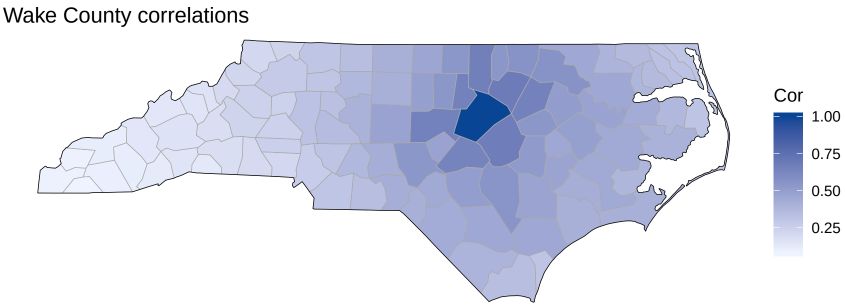
Proper CAR model correlation with $\rho = 0.9$



Proper CAR model variance with $\sigma = 1$ and $\rho = 0.99$



Proper CAR model correlation with $\rho = 0.99$



Other CAR models

- There are other CAR models that use different weights W_{ij}
- You can take the weights to be function of distance between centroids
- When W_{ij} are not binary, set $m_i = \sum_{j=1}^{n} W_{ij}$
- You can take the weights so that the variance is approximately constant across space
- The Leroux parameterization is

$$\Sigma = \sigma^2 \left[(1-\rho)I_n + \rho(\mathbf{M} - \mathbf{W})\right]^{-1}$$

which reduces to an equal-variance model if $\rho = 0$

Simultaneous autoregressive (SAR) model

- The SAR model begins with n simple linear regressions
- For site i, we use the mean of neighbors as the covariate

$$Z_i = \rho \bar{Z}_i + \epsilon_i$$

  where $\epsilon_i \sim \text{Normal}(0, \sigma^2/m_i)$ independent over i

- This is complicated because Z_i appears as a response once and in the covariate m_i times
- As with the CAR model, we must solve for the induced joint distribution

Simultaneous autoregressive (SAR) model

- It can be shown that **Z** is MVN with mean zero and

$$\Sigma = \sigma^2 \left(\mathbf{M} - \rho\mathbf{W}\right)^{-1} \left(\mathbf{M} - \rho\mathbf{W}\right)^{-1}$$

- This is basically the square of the CAR covariance

- The same inferencial methods and choices of weights apply

Software

- There are many packages that can fit these models, but we will use `CARBayes`
- `CARBayes` uses MCMC and is fairly easy to use
- It handles Gaussian and non-Gaussian data
- It can fit the intrinsic (they call it the Besag-York model) and proper (Leroux) models
- It also includes extensions to multivariate data and more sophisticated weighting schemes