Spatial Cluster Detection Using Bayes Factors from Overparameterized Models

Ronald E. Gangnon and Murray K. Clayton

Department of Biostatistics and Medical Informatics

University of Wisconsin – Madison

August 8, 2004

http://www.biostat.wisc.edu/∼ronald

Wisconsin Breast Cancer Data

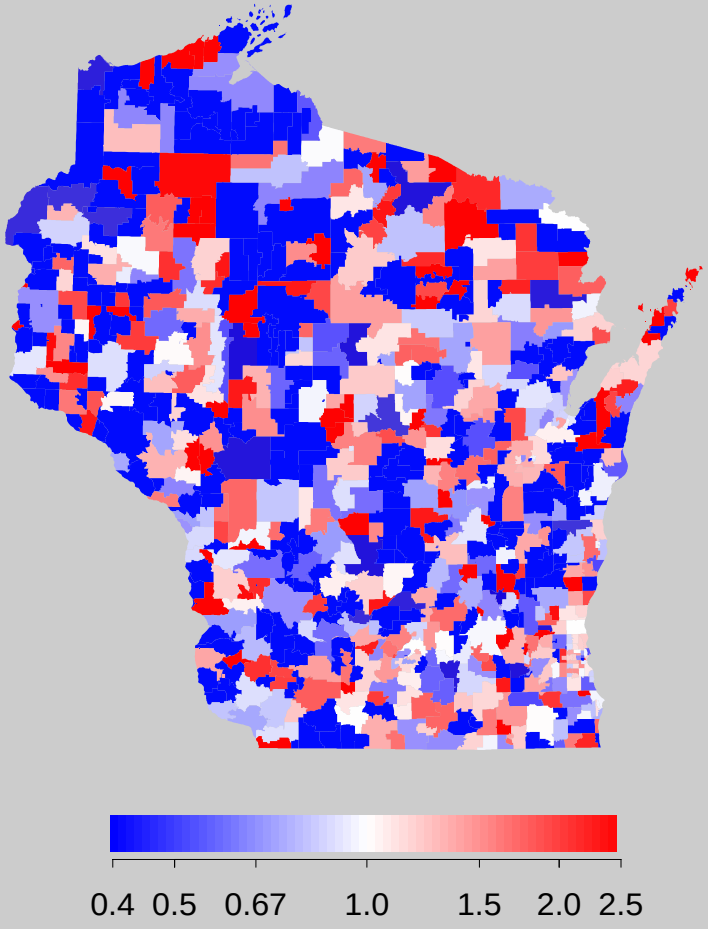
- 1990 zip codes in Wisconsin $(N = 716)$.
- Breast cancer cases from State Cancer Registry, 1990.
- Expected number of cases for zip code.

  Indirect, internal age standardization.

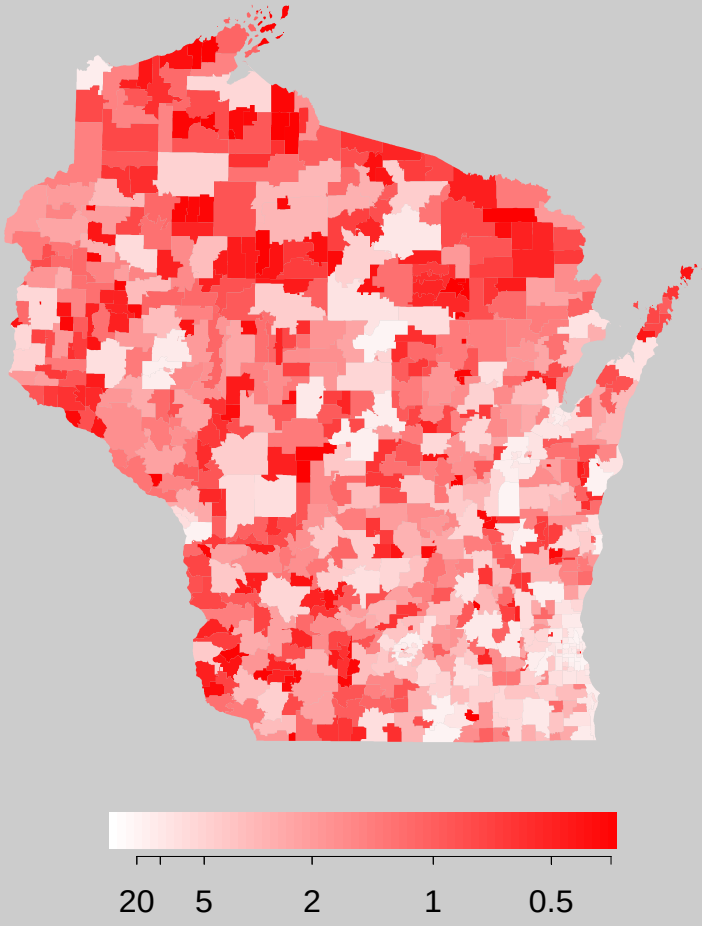
  Age-specific female populations from Census.
- Geographic centroid of zip code.

Wisconsin Breast Cancer Data

Standardized Incidence Ratio



Expected Number of Cases



A General Disease Clustering Model

- First-stage Poisson model:

$$Y_i|\rho_i \overset{\text{ind}}{\sim} \text{Poisson}(\rho_i \cdot E_i)$$

$Y_i =$ number of cases in zip code i.

$E_i =$ expected number of cases in zip code i.

$\rho_i =$ relative risk in zip code i.

- Log-linear model for relative risk:

$$\log(\rho_i) = \alpha + \phi_i + \epsilon_i$$

$\phi_i =$ spatial clustering effect.

$\epsilon_i =$ non-spatial random effect.

A General Disease Clustering Model

- Spatial clustering effect:

$$\phi_i = \sum_{j=1}^{k} \theta_j \mathbf{1}_{\{i \in C_j\}}$$

- $C_1, C_2 \ldots, C_k$ are k clusters.

  Chosen from a restricted set of potential clusters.

  Circles in a relevant metric often convenient.

  Overlapping clusters are OK.

- Choice of clusters reflects prior knowledge, goals of analysis.

Potential Clusters for Wisconsin Data

- Circular clusters centered at zip code centroids.
- Radii ranging from 0 km up to 50 km (31.1 mi).

  Fixed maximum geographic radius.
- Zip code belongs to cluster if centroid falls inside circle.
- Note: Methodology will work for any set of clusters.

  Easily adapted to circles in other (non-Euclidean) metrics.

Prior for Spatial Clustering Effect

- Recall: $\phi_i = \sum_{j=1}^{k} \theta_j \mathbf{1}_{\{i \in C_j\}}$
- Parameters:

  Cluster locations: $C_1, C_2, \ldots, C_k$.

  Cluster risks: $\theta_1, \ldots, \theta_k$.

- First stage: $\theta_1, \theta_2, \ldots, \theta_k | C_1, C_2, \ldots, C_k \overset{\text{iid}}{\sim} N(0, \sigma_\theta^2)$.

  σ_θ^2 must be fixed.

  We use $\sigma_\theta^2 = 0.355$ so that $P(0.25 \leq e^\theta \leq 4) = 0.99$.

- Second stage: $C_1, C_2, \ldots, C_k$ iid with *dartboard prior*.

A *Dartboard Prior* for Clusters

- Select a small area by throwing a dart at study region.
- Center cluster on centroid of that small area.
- Select cluster radius uniformly from available radii.
- Probability of selecting cluster i, j:

$$w_{i,j} = \frac{a_i}{A} \cdot \frac{r_{i,j+1} - r_{i,j}}{r_{max}}$$

- This is a flat prior on center/radius adjusted to avoid empty circles.

What About k?

- Note: if $\phi_i = \sum_{j=1}^{k} \theta_j \mathbf{1}_{\{i \in C_j\}}$, then $\phi_i = \sum_{j=1}^{k+l} \theta_j \mathbf{1}_{\{i \in C_j\}}$.

  $\theta_{k+1} = \theta_{k+2} = \ldots = \theta_{k+l} = 0.$

  $C_{k+1}, C_{k+2}, \ldots, C_{k+l}$ arbitrary.

- Use a fixed, but overly large value for k.

  Posterior will concentrate on needed clusters.

  Locations of excess clusters will follow prior.

  Pro: Don't need to specify prior for k.

  Con: No formal inference about k.

Computation

- Posterior simulation using MCMC algorithm.
- Standard proposal distributions available for

  intercept (α)

  cluster risks ($\theta_1, \theta_2 \ldots, \theta_k$)

  random effects ($\epsilon_1, \epsilon_2, \ldots, \epsilon_N$)

  precision of random effects (τ)

- Proposals for cluster ($C_1, C_2, \ldots, C_k$) from full conditional distribution.

Parameters of Interest

- Primary parameter of interest: ρ_i, $i = 1, 2, \ldots, N$.

  Map of posterior mean for ρ_i.

  Map of posterior SD for $\log(\rho_i)$.

- Secondary parameter of interest: ϕ_i, $i = 1, 2, \ldots, N$.

  Main interest: cluster membership.

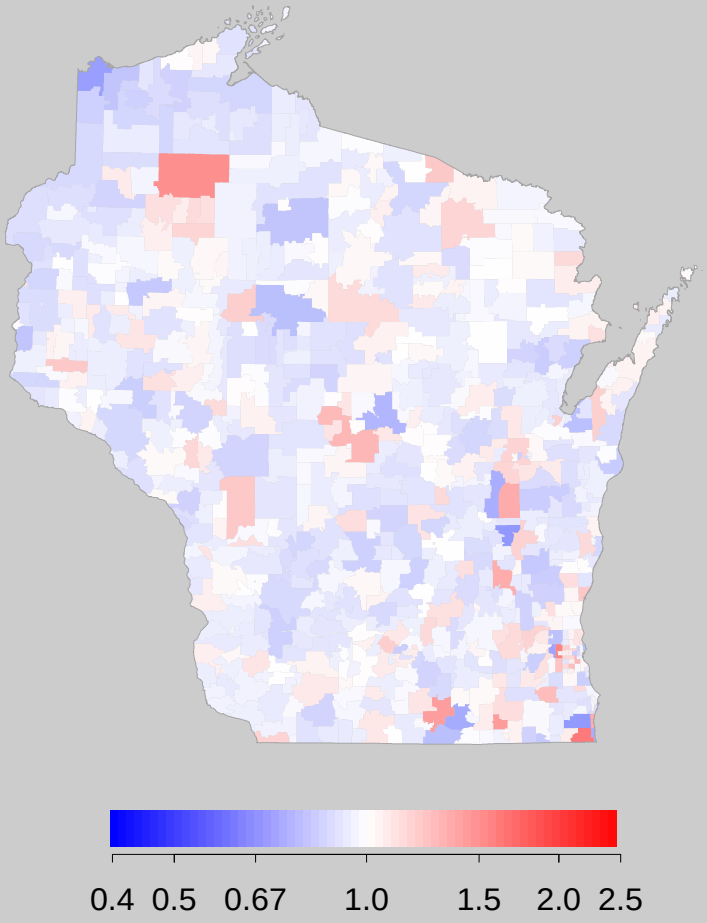
  $\phi_i \neq 0$ iff zip code i belongs to a cluster.

  Map of Bayes factor for $\phi_i \neq 0$.

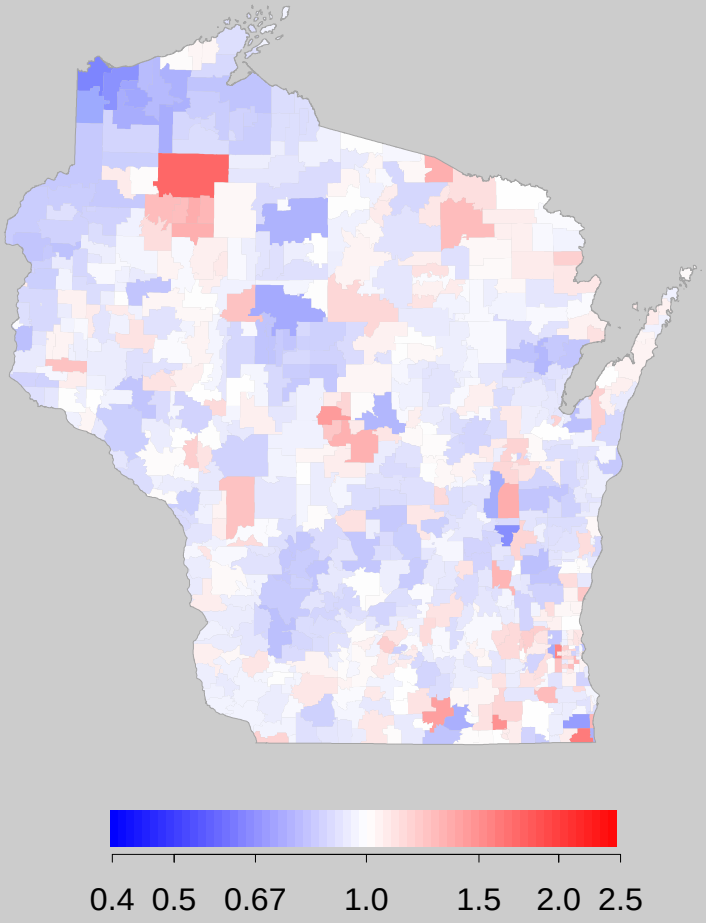
  Possibly separate maps for $\phi_i > 0$ and $\phi_i < 0$.

Results – Posterior Means

10 Cluster Model

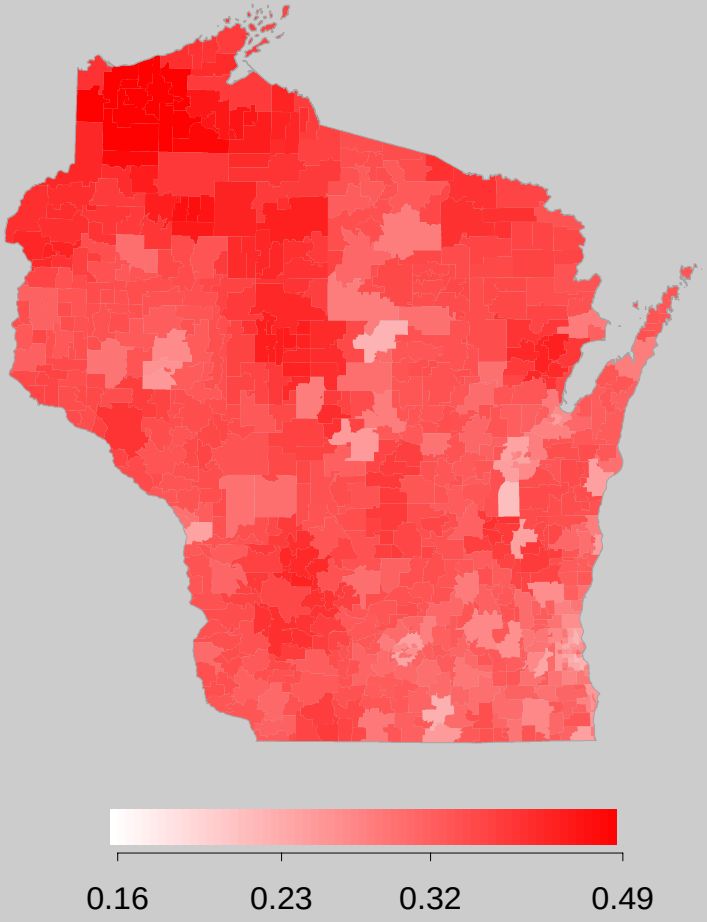


20 Cluster Model

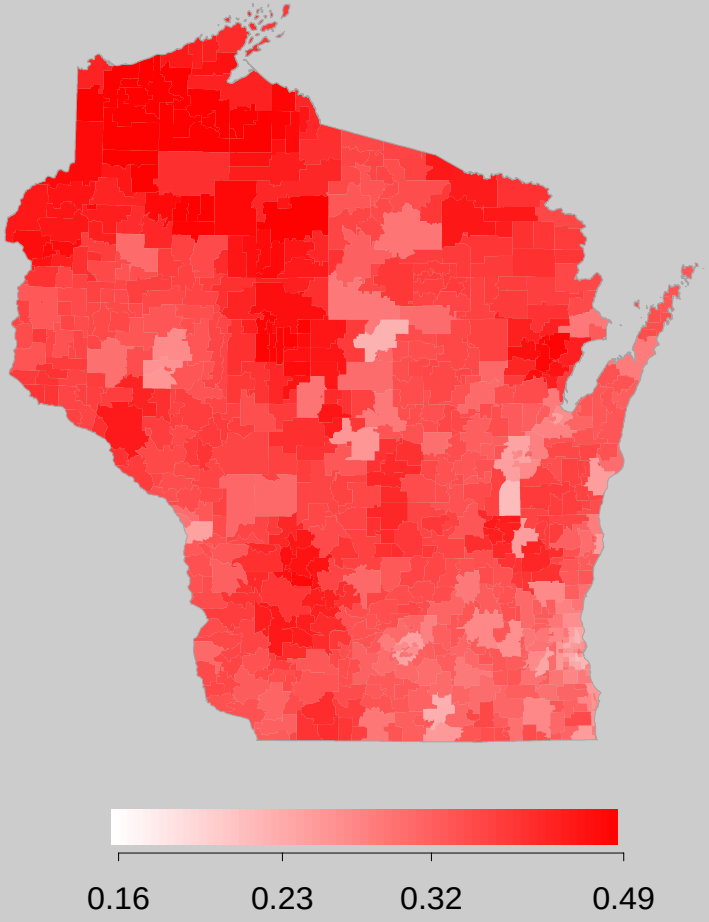


Results – Posterior SDs

10 Cluster Model



20 Cluster Model



Assessing the Evidence for Clustering

- Posterior probability of cluster membership reflects:

  Prior probability (mainly dependent on k) – not of interest.

  Evidence (in the data) for clustering – of interest.

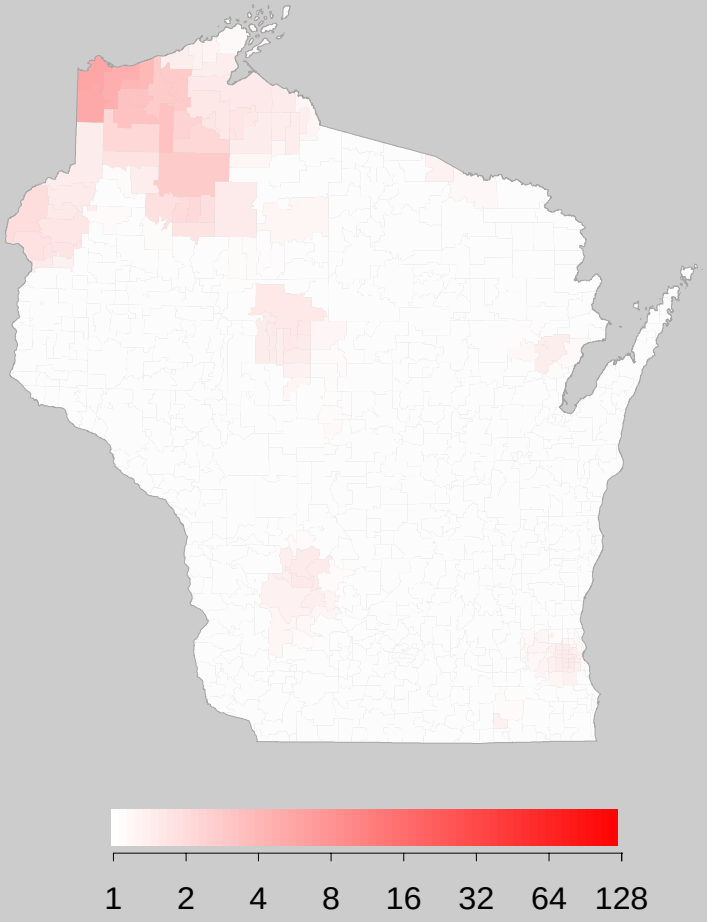
- Better measure of evidence: Bayes factor.

$$\text{BF}_i = \frac{P(\phi_i \neq 0|\mathbf{Y})/P(\phi_i = 0|\mathbf{Y})}{P(\phi_i \neq 0)/P(\phi_i = 0)}$$

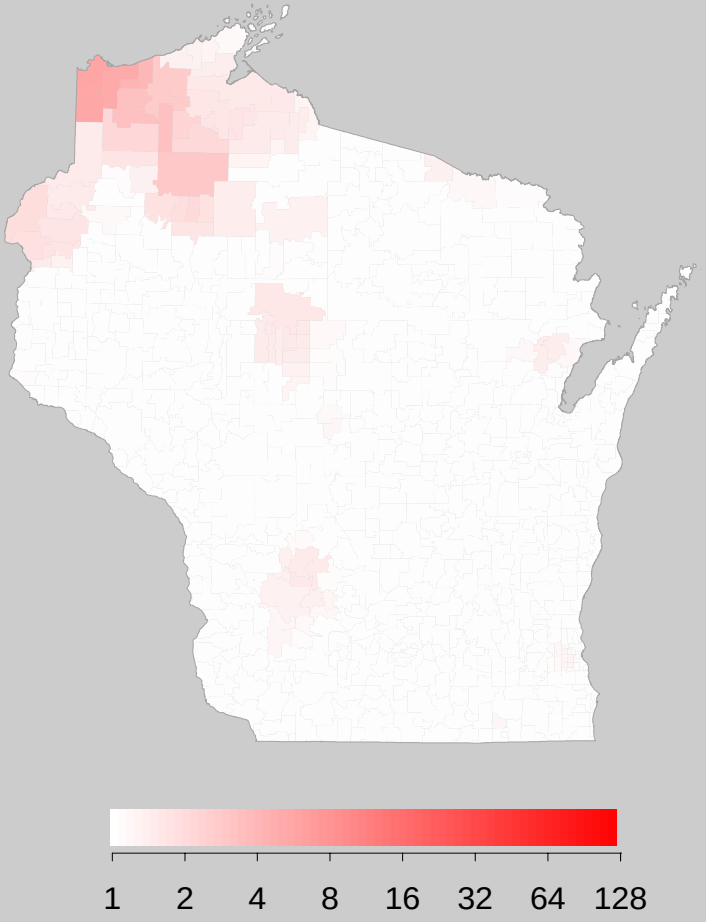
  Minimizes dependence on k.

Results – Bayes Factor

10 Cluster Model



20 Cluster Model



Alternative: Formal Inference for k

- Discrete uniform prior for k.

  Importance sample for other priors.

- Posterior simulation using RJMCMC algorithm.

  Same proposals as fixed k model PLUS:

  ADD/DROP cluster.

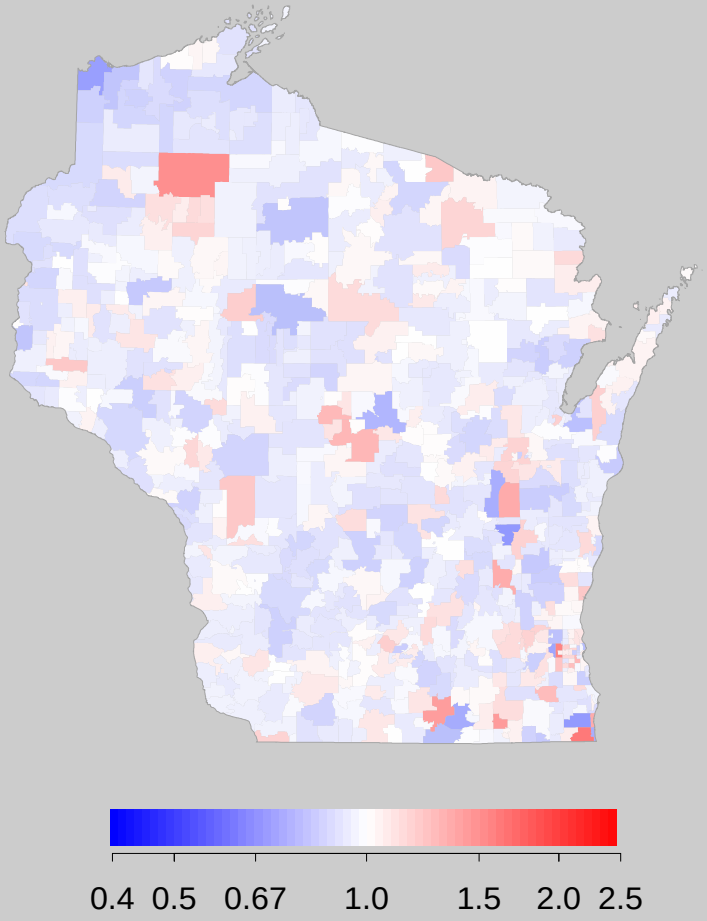
- Parameters of interest:

  Same as fixed k model PLUS:

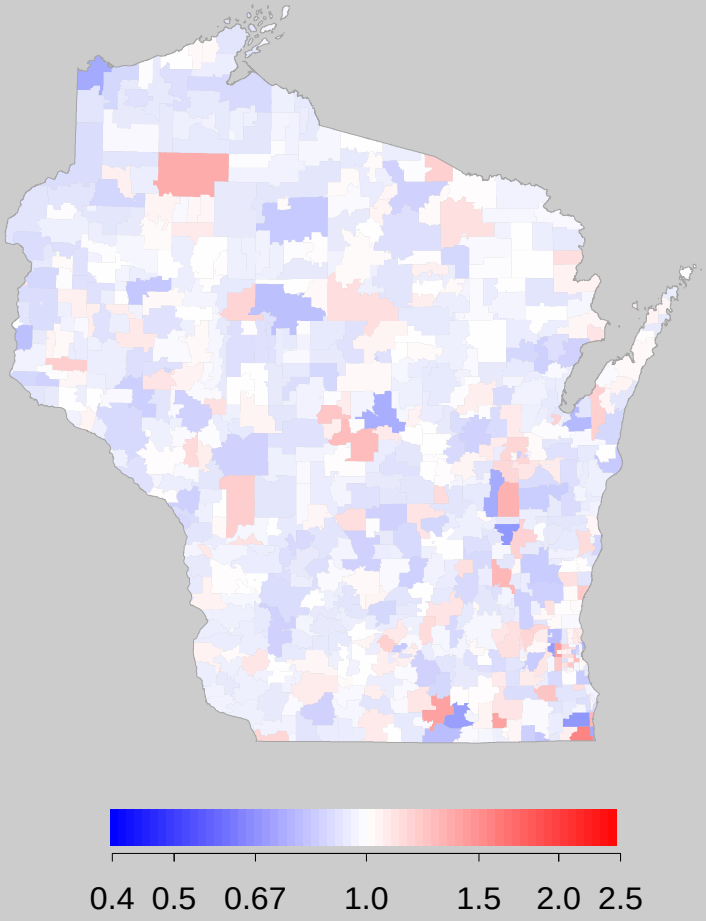
  Number of clusters k.

Comparison of Results – Posterior Mean

10 Cluster Model

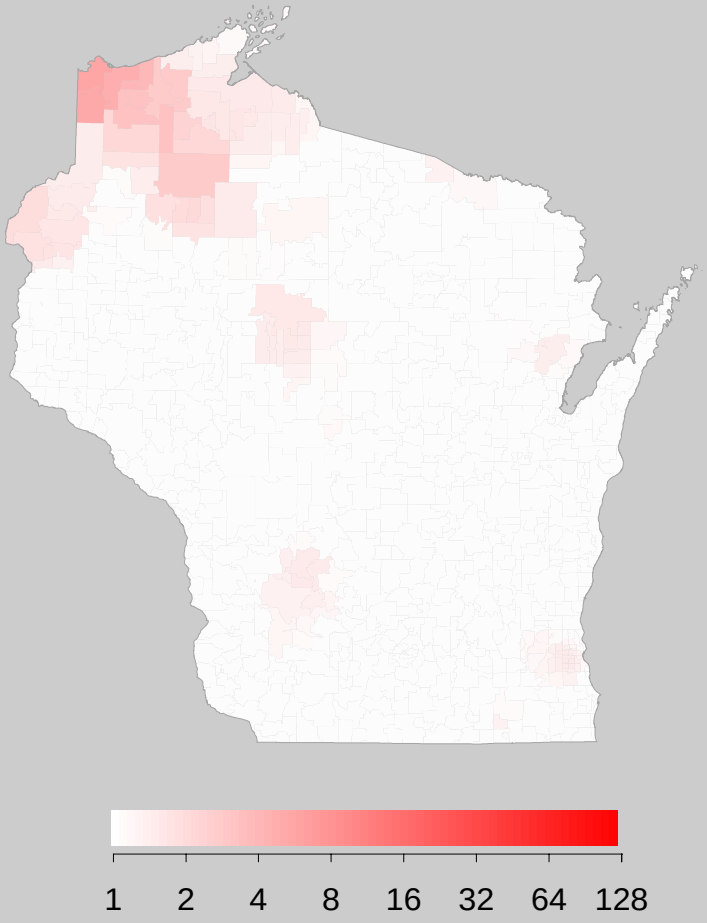


Variable k Model



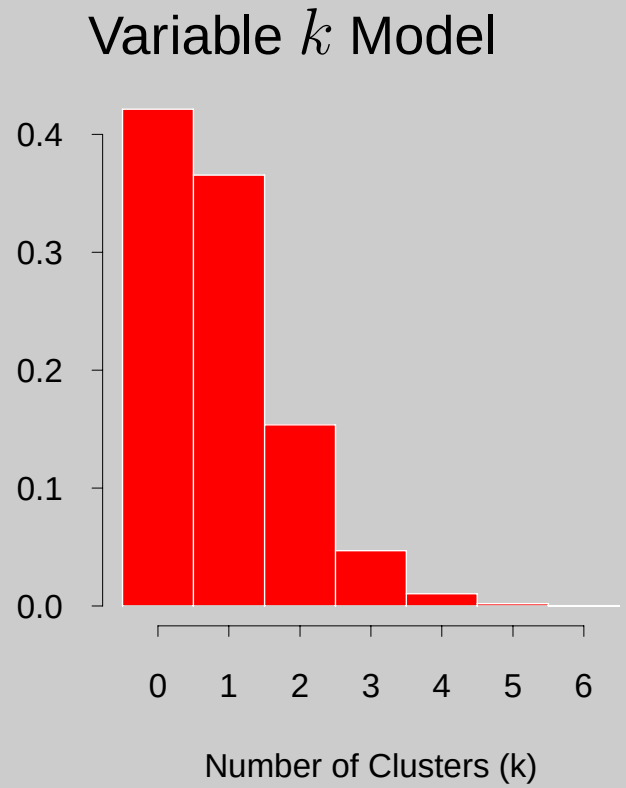
Comparison of Results – Clustering

10 Cluster Model



Variable k Model

Results – Number of Clusters



Example (2): New York Leukemia Data

- Eight counties in Upstate New York.

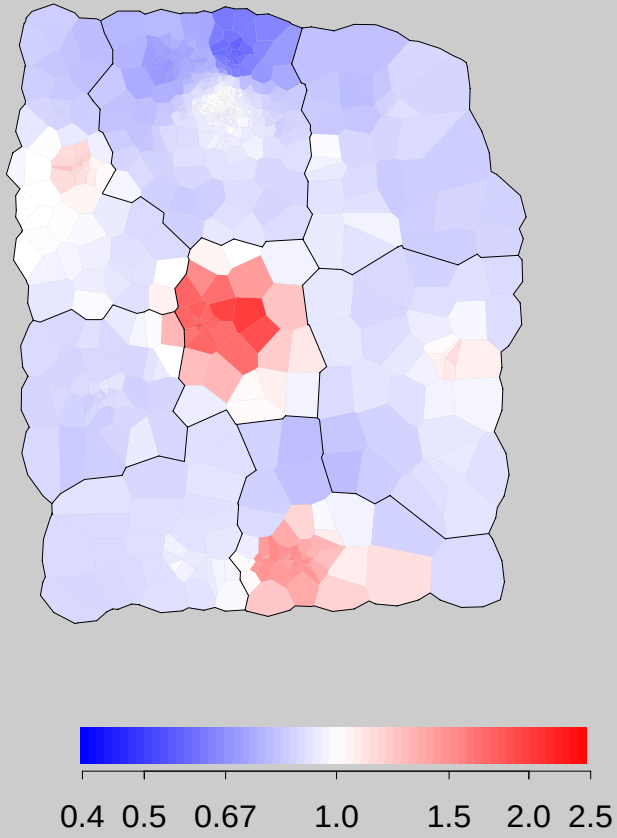
  Census tracts in 7 counties.

  Census blocks in Broome county.

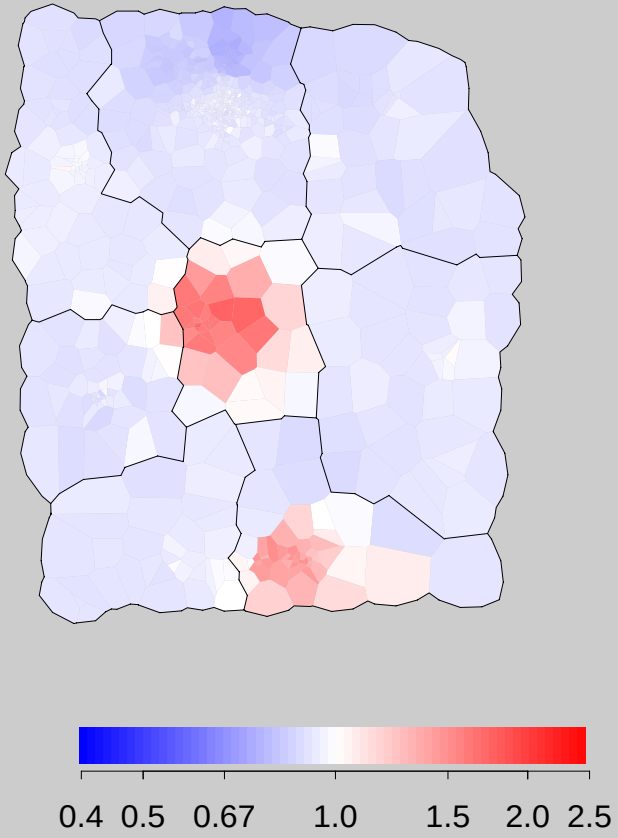
- Leukemia cases from State Cancer Registry, 1978-1982.
- Adult (18+) population from 1980 U.S. Census.
- Geographic centroid of small area from Census Bureau.
- Previous analyses show strong evidence of clustering.

Comparison of Results – Posterior Mean

10 Cluster Model

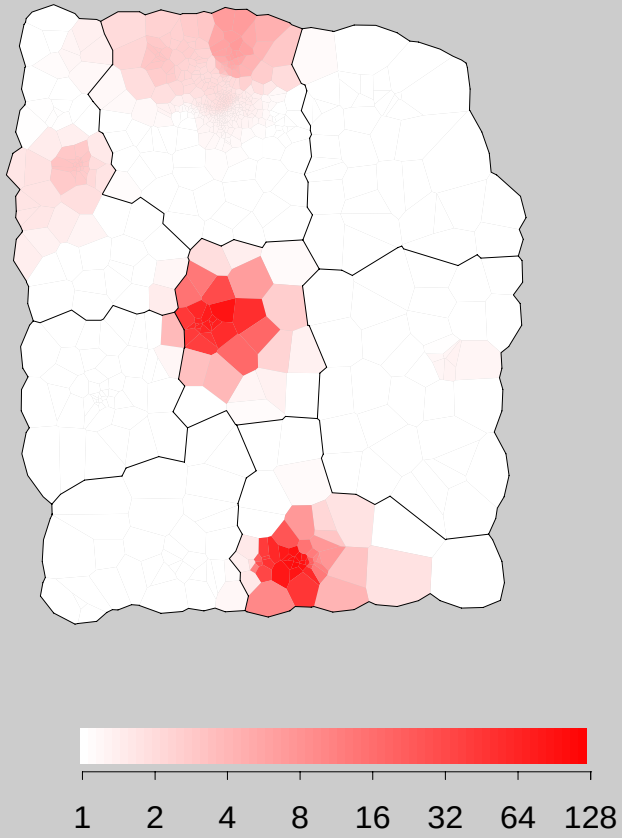


Variable k Model

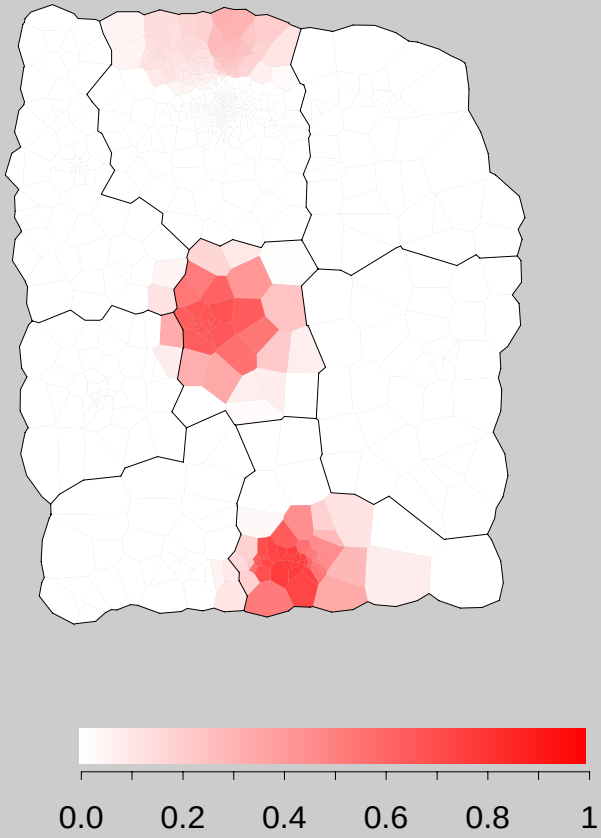


Comparison of Results – Clustering

10 Cluster Model



Variable k Model



Concluding Remarks

- Developed spatial model which includes clustering and heterogeneity components.
- Natural, unbiased prior specification for clustering component of model.
- If primary interest is number of clusters, use RJMCMC.
- If not, we can use fixed, but large k to:

  Estimate disease rates.

  Identify cluster locations using Bayes factors.

  Little dependence on specific k (if large enough).