Bayesian Detection of Spatial Disease Clustering: An Application

Ronald E. Gangnon
and
Murray K. Clayton

Department of Biostatistics and Medical Informatics
Department of Statistics
University of Wisconsin – Madison
August 10, 1999

http://www.biostat.wisc.edu/~ronald
New York Leukemia Data
Goals of Analysis

- Find clusters, i.e., areas with high or low disease rates.

- Estimate cell-specific disease rates.

- Why?
  - Screening.
  - Surveillance.
  - Hypothesis generation.
Basic Statistical Model

- \( i = 1, 2, \ldots, N \) cells.
- \( O_i \) = number of cases in cell \( i \).
- \( n_i \) = population at risk in cell \( i \).
- \( r_i = \mathbb{E}(O_i / n_i) \) = true rate of disease in cell \( i \).
- \( O_i \sim \text{Poisson}(r_i \cdot n_i) \).
Previous Approaches

- Distance-based test statistics, e.g., Whittemore et al. (1987).

- Multiple cluster locations/sizes.
  - All nominally significant circles (Openshaw et al., 1988)
  - Circles of fixed "case radius" (Besag and Newell, 1991)
  - Circles of fixed “population radius” (Turnbull et al., 1990)
  - All circular clusters (Kulldorff and Nagarwalla, 1995)

- Bayesian approaches.
  Point process models (Lawson and Clark, 1999)
  Disease mapping (e.g., Waller et al., 1997)
What’s Missing

• Flexibility
  - Cluster size and shape
  - Lower cluster rates, not just higher

• Multiple clusters

• Estimates
  - Account for uncertainty about true cluster
A Simple Model for Clustering
Prior for Clustering Models

- Markov connected component field (MCCF) 
  (Møller and Waagepetersen, 1997)

Score each cluster based on its properties.

Denote this score by $S_j$.

Calculate prior for model as

$$P(c) = \exp \left\{ - \sum_{j=1}^{k} S_j \right\}$$

- Higher weight on small, circular clusters.

- Define size and shape using perimeter and area.

  $$R_1 = \sqrt{A/\pi}, R_2 = P/(2\pi)$$

  Size = $R_2$, Shape = $R_1/R_2$
Posterior for Clustering Models

- Occam’s window (Madigan and Raftery, 1994)
- Basic idea: Most models have no support in data.
- Only consider good models.

\[
P(c|O) \geq \frac{1}{W} \frac{\max_c P(c|O)}{P(c|O)}
\]

- Randomized search for plausible models.
- Start with saturated model and repeatedly merge adjacent components.
- Merger selection probability roughly proportional to posterior.
Analysis of New York Data

- Prior for rates:

  Background: Gamma(0.739, 1339)

  Cluster(s): Gamma(1.478, 1339)

- Prior for clusters:

- Prior for null model:

  \[ P(H_0) = 0.99. \]
Cluster Membership Probabilities –

Prior 1
Estimated Rates – Replicated
Cluster Membership Probabilities – Replicated

Probability Cell Belongs to Cluster
Analysis of New York Data

• Prior for rates:

  Background: \( \text{Gamma}(0.739, 1339) \)

  Cluster(s): \( \text{Gamma}(1.478, 1339) \)

• Prior for clusters:

  - Prior for null model:

  \[ P(H_0) = 0.99. \]
Estimated Rates – Prior 2

Estimated Rates
Cluster Membership Probabilities –

Prior 2
Analysis of New York Data

- Prior for rates:
  
  Background: Gamma(0.739, 1339)

  Cluster(s): Gamma(1.478, 1339)

- Prior for clusters:

  ![Graph showing prior probabilities for cluster size and shape](image)

- Prior for null model:

  \[ P(H_0) = 0.99. \]
Estimated Rates – Prior 3
Cluster Membership Probabilities –

Prior 3
Analysis of New York Data

- Prior for rates:
  
  Background: Gamma\( (0.739, 1339) \)

  Cluster(s): Gamma\( (1.478, 1339) \)

- Prior for clusters:

- Prior for null model:
  
  \[ P(H_0) = 0.99. \]
Estimated Rates – Prior 4
Cluster Membership Probabilities –

Prior 4
Conclusions

- Bayesian approach to analyzing spatial clustering of disease.

- Flexibility: different cluster sizes and shapes.

- Estimate (not just test for) clustered rates.

- Incorporates uncertainty in cluster memberships.

- Use of multiple priors.