MODULE 9: Spatial Statistics in Epidemiology and Public Health
Lecture 4: Spatial regression

Jon Wakefield and **Lance Waller**

References

- Waller and Gotway (2004, Chapter 9) *Applied Spatial Statistics for Public Health Data*. New York: Wiley.
- Haining, R. (2003). *Spatial Data Analysis: Theory and Practice*. Cambridge: Cambridge University Press.
- Banerjee, S., Carlin, B.P., and Gelfand, A.E. (2014) *Hierarchical Modeling and Analysis for Spatial Data, 2nd Ed.* Boca Raton, FL: CRC/Chapman & Hall.
- Blangiardo, M. and Cameletti, M. (2015) *Spatial and Spatio-temporal Bayesian Models with R-INLA*. Chichester: Wiley.

What do we have so far?

- Tension between *statistical precision* (want large local sample sizes $\rightarrow$ big regions), and *geographic precision* (want small regions for more detail in map).
- *Disease mapping* approaches use *small area estimation* techniques to borrow information from all areas and from neighboring areas to improve local estimation in each area.
- But what about local covariates?
- Can we adjust for those (say, using regression models)?
- And still borrow information?
- With *independent* observations we know how to use *linear* and *generalized linear* models such as linear, Poisson, logistic regression.
- What happens with *dependent* observations?

Caveat

"...all models are wrong. The practical question is how wrong do they have to be to not be useful."
Box and Draper (1987, p. 74)

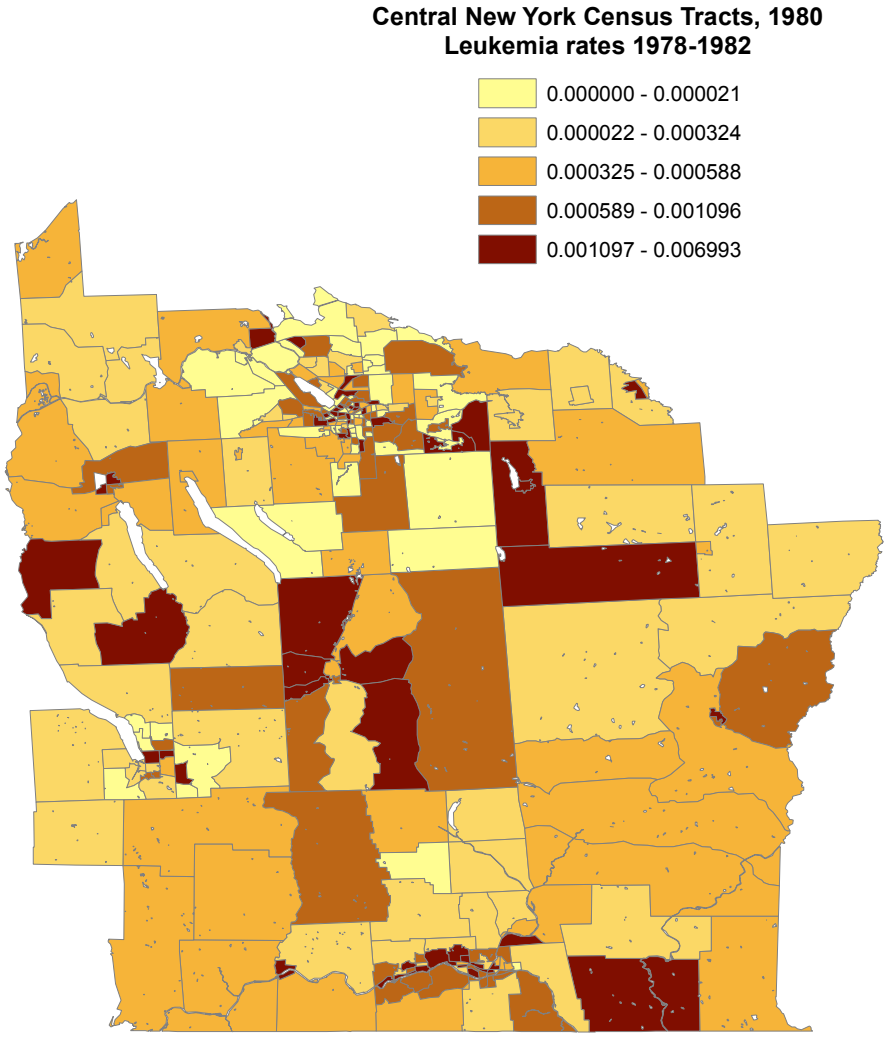
What changes with dependence?

- In statistical modeling, we are often trying to describe the mean of the outcome as a function of covariates, assuming error terms are mutually independent.
- Where do correlated errors come from?
- Perhaps outcomes truly correlated (infectious disease).
- Perhaps we omitted an important variable that has spatial structure itself.

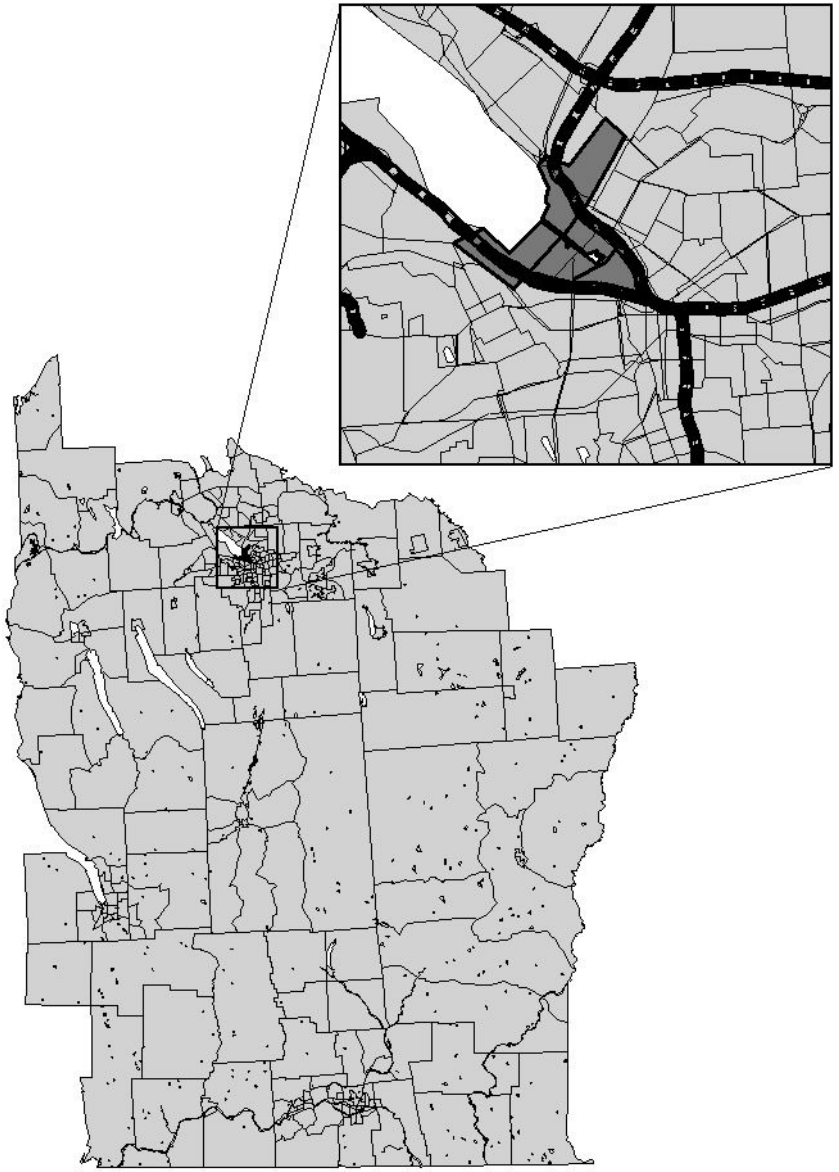
NY leukemia data

- NY leukemia data and covariates (Waller and Gotway, 2004).
- 281 census tracts (1980 Census).
- 8 counties in central New York.
- 592 cases for 1978-1982.
- 1,057,673 people at risk.

Crude Rates (per 100,000)



Outliers, where are the top 3 rates?



Building the model: Poisson regression

- Let Y_i = count for region i.
- Let E_i = *expected* count for region i.
- Let $(x_{i,TCE}, x_{i,65}, x_{i,home})$ be the associated covariate values.
- Poisson regression:

$$Y_i \sim Poisson(E_i \zeta_i)$$

where

$$\log(\zeta_i) = \beta_0 + x_{i,TCE}\beta_{TCE} + x_{i,65}\beta_{65} + x_{i,home}\beta_{home}.$$

Details

- Poisson distribution for counts.
- *Link function*: Natural log of mean of Y_i is a linear function of covariates.
- βs represent multiplicative increases in expected counts, e^{β} a measure of relative risk associated with one unit increase in covariate.
- E_i an *offset*, what we expect if the covariates have no impact.
- Age, race, sex adjustments in either E_i (standardization) or covariates.

Adding spatial correlation: New York data

- Assume E_i known, perhaps age-standardized, or based on global (external or internal) rates.
- Our model is

$$Y_i|\boldsymbol{\beta}, \psi_i \overset{ind}{\sim} \text{Poisson}(E_i \exp(\boldsymbol{x}_i'\boldsymbol{\beta} + \psi_i)),$$

$$\log(\zeta_i) = \beta_0 + x_{i,TCE}\beta_{TCE} + x_{i,65}\beta_{65} + x_{i,home}\beta_{home} + \psi_i.$$

- The ψ_i represent the *random intercepts*.
- Add *overdispersion* via $\boldsymbol{\psi}_i \overset{ind}{\sim} N(0, v_\psi)$.
- Add spatial correlation via

$$\boldsymbol{\psi} \sim MVN(\mathbf{0}, \Sigma).$$

Priors and "shrinkage"

- Overdispersion model (i.i.d. ψ_i) results in each estimate being a compromise between the *local* SMR and the *global average* SMR.
- "Borrows information (strength)" from other observations to improve precision of local estimate.
- "Shrinks" estimate toward global mean. (Note: "shrink" does not mean "reduce", rather means "moves toward").

Local shrinkage

- Spatial model (correlated ψ_i) results in each estimate begin a compromise between the *local* SMR and the *local average* SMR.
- Shrinks each ψ_i toward the average of its *neighbors*.
- Can also include *both* global and local shrinkage (Besag, York, and Mollié 1991).
- How do we fit these models?

Bayesian inference

Bayesian inference regarding model parameters based on *posterior distribution*

$$Pr[\boldsymbol{\beta}, \boldsymbol{\psi} | \boldsymbol{Y}]$$

proportional to the product of the likelihood times the prior

$$Pr[\boldsymbol{Y} | \boldsymbol{\beta}, \boldsymbol{\psi}] Pr[\boldsymbol{\psi}] Pr[\boldsymbol{\beta}].$$

Defers spatial correlation to the prior rather than the likelihood.

Spatial priors

- Could model *joint* distribution

$$\boldsymbol{\psi} \sim MVN(\mathbf{0}, \Sigma).$$

- Could also model *conditional* distribution

$$\psi_i | \psi_{j \neq i} \sim N\left(\frac{\sum_{j \neq i} c_{ij} \psi_j}{\sum_{j \neq i} c_{ij}}, \frac{1}{v_{CAR} \sum_{j \neq i} c_{ij}}\right), i = 1, \ldots, N.$$

  where c_{ij} are *weights* defining the neighbors of region i.
- Adjacency weights: $c_{ij} = 1$ if j is a neighbor of i.

CAR priors

- The conditional specification defines the *conditional autoregressive* (CAR) prior (Besag 1974, Besag et al. 1991).
- Under certain conditions on the c_{ij}, the CAR prior defines a valid multivariate joint Gaussian distribution.
- Variance covariance matrix a function of the *inverse* of the matrix of neighbor weights.

Fitting Bayesian models

- Posterior often difficult to calculate mathematically.
- Markov chain Monte Carlo: Iterative simulation approach to model fitting.
- Given *full conditional* distributions, simulate a new value for each parameter, holding the other parameter values fixed.
- The set of simulated values converges to a sample from the posterior distribution.
- Alternative: *integrated nested Laplace analysis* using the `inla` package (example code).

Complete model specification

$$Y_i|\boldsymbol{\beta}, \psi_i \overset{ind}{\sim} \text{Poisson}(E_i \exp(\boldsymbol{x}_i'\boldsymbol{\beta} + \psi_i)),$$

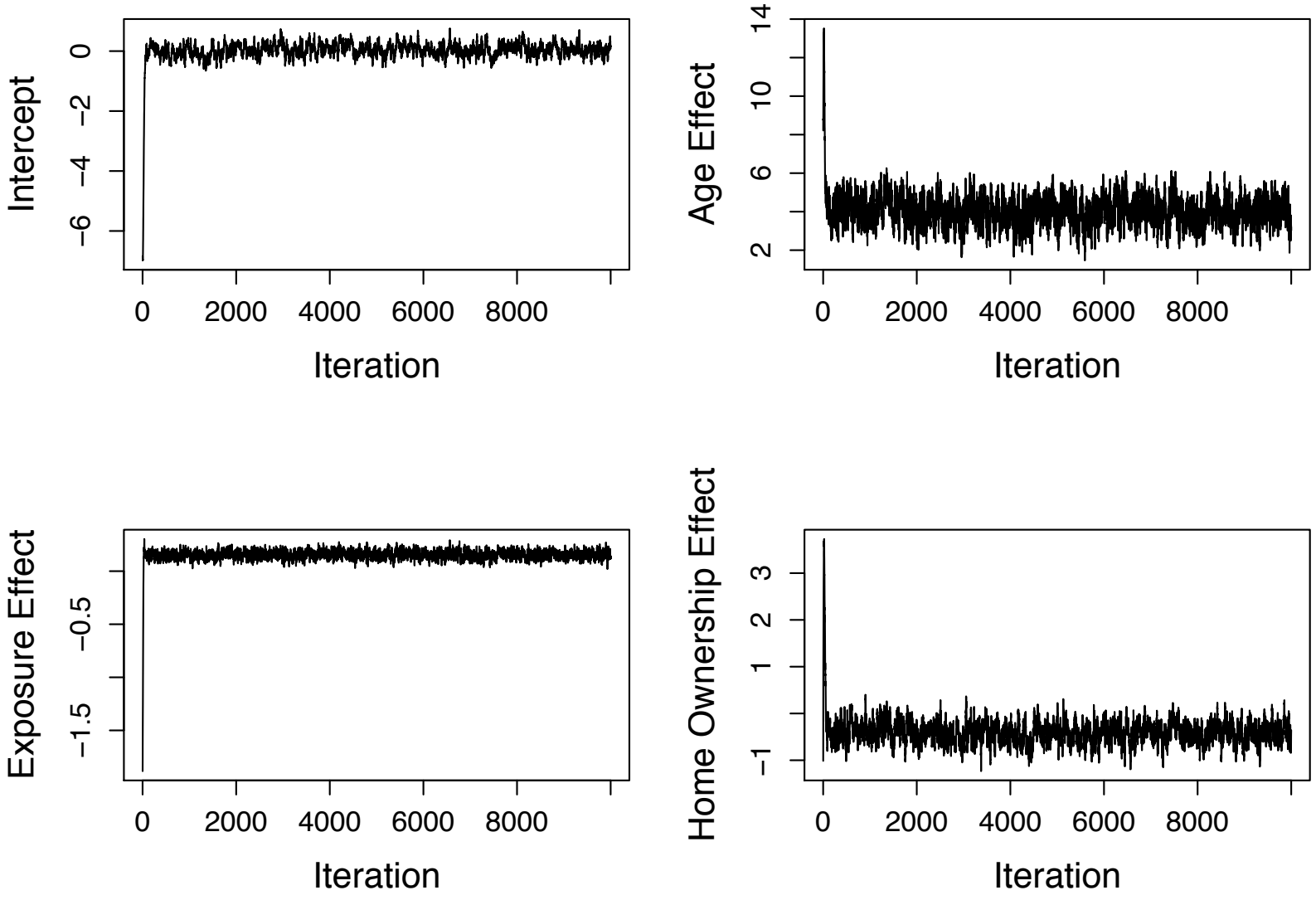
$$\log(\zeta_i) = \beta_0 + x_{i,TCE}\beta_{TCE} + x_{i,65}\beta_{65} + x_{i,home}\beta_{home} + \psi_i.$$

$$\beta_k \sim \text{Uniform}.$$

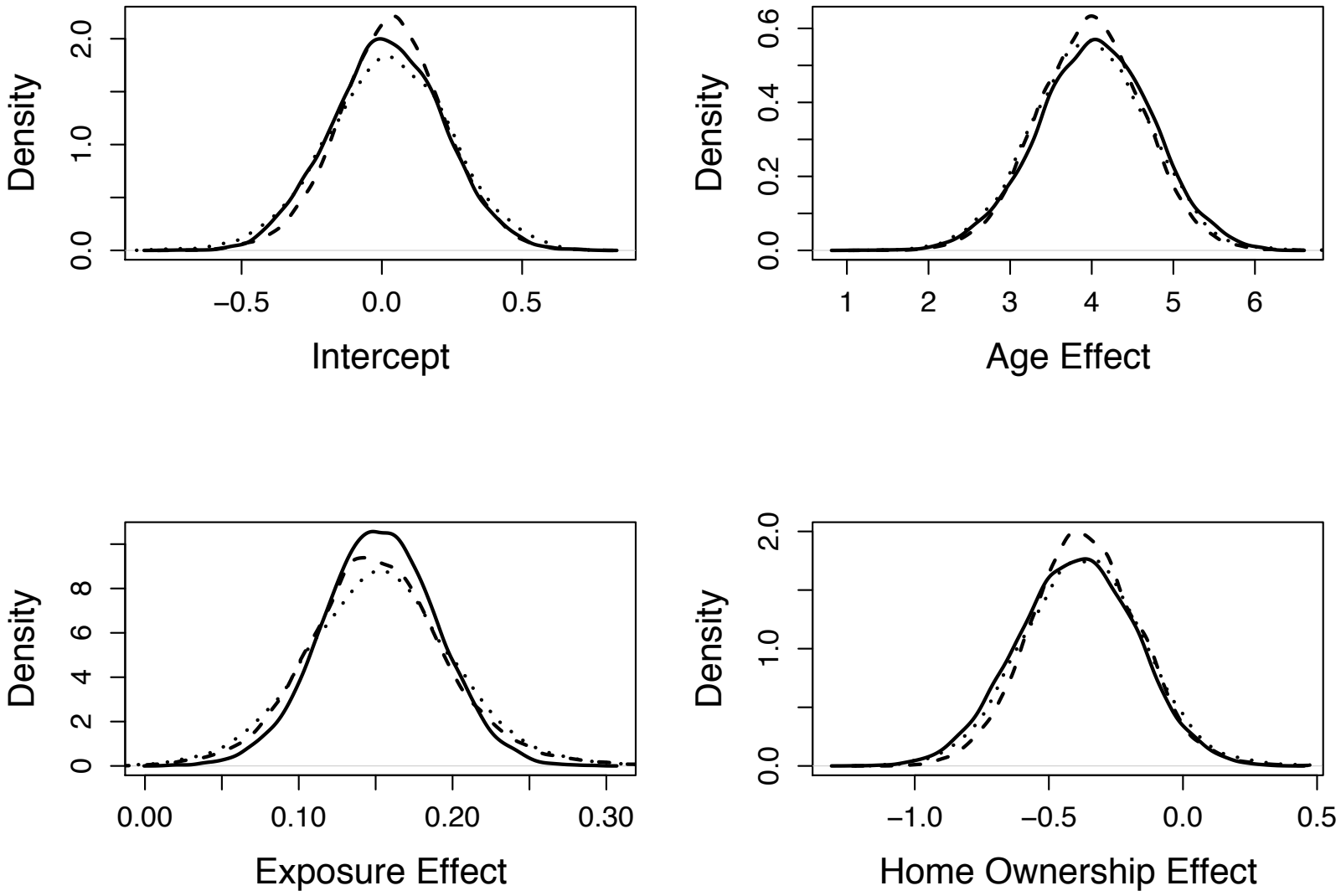
$$\psi_i|\psi_{j\neq i} \sim N\left(\frac{\sum_{j\neq i} c_{ij}\psi_j}{\sum_{j\neq i} c_{ij}}, \frac{1}{v_{CAR}\sum_{j\neq i} c_{ij}}\right), i = 1, \ldots, N.$$

$$\frac{1}{v_{CAR}} \sim \text{Gamma}(0.5, 0.0005).$$

MCMC trace plots



Posterior densities



MCMC posterior estimates

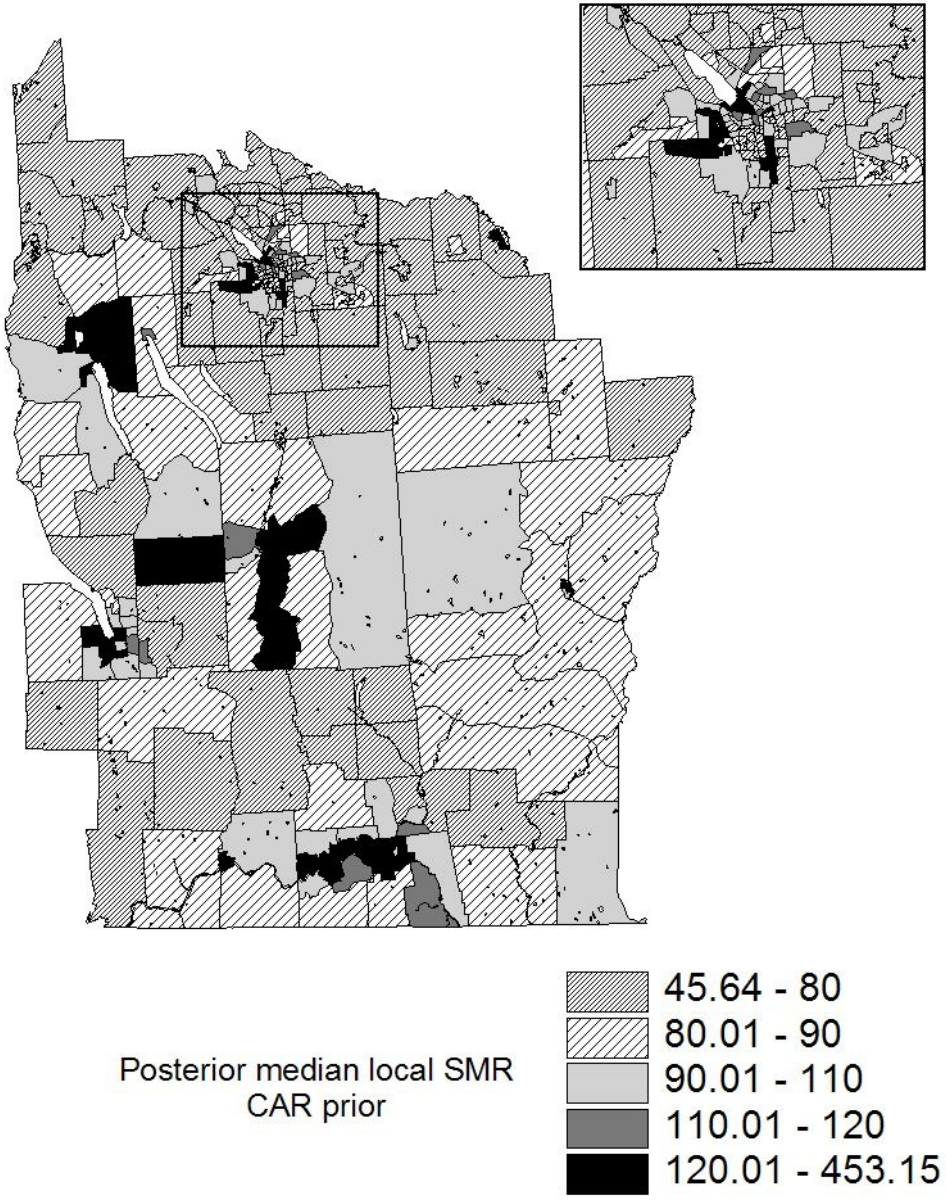
Covariate	Posterior Median	95% Credible Set
β_0	0.048	(-0.355, 0.408)
β_{65}	3.984	(2.736, 5.330)
β_{TCE}	0.152	(0.066, 0.226)
β_{home}	-0.367	(-0.758, 0.049)

But there's more!

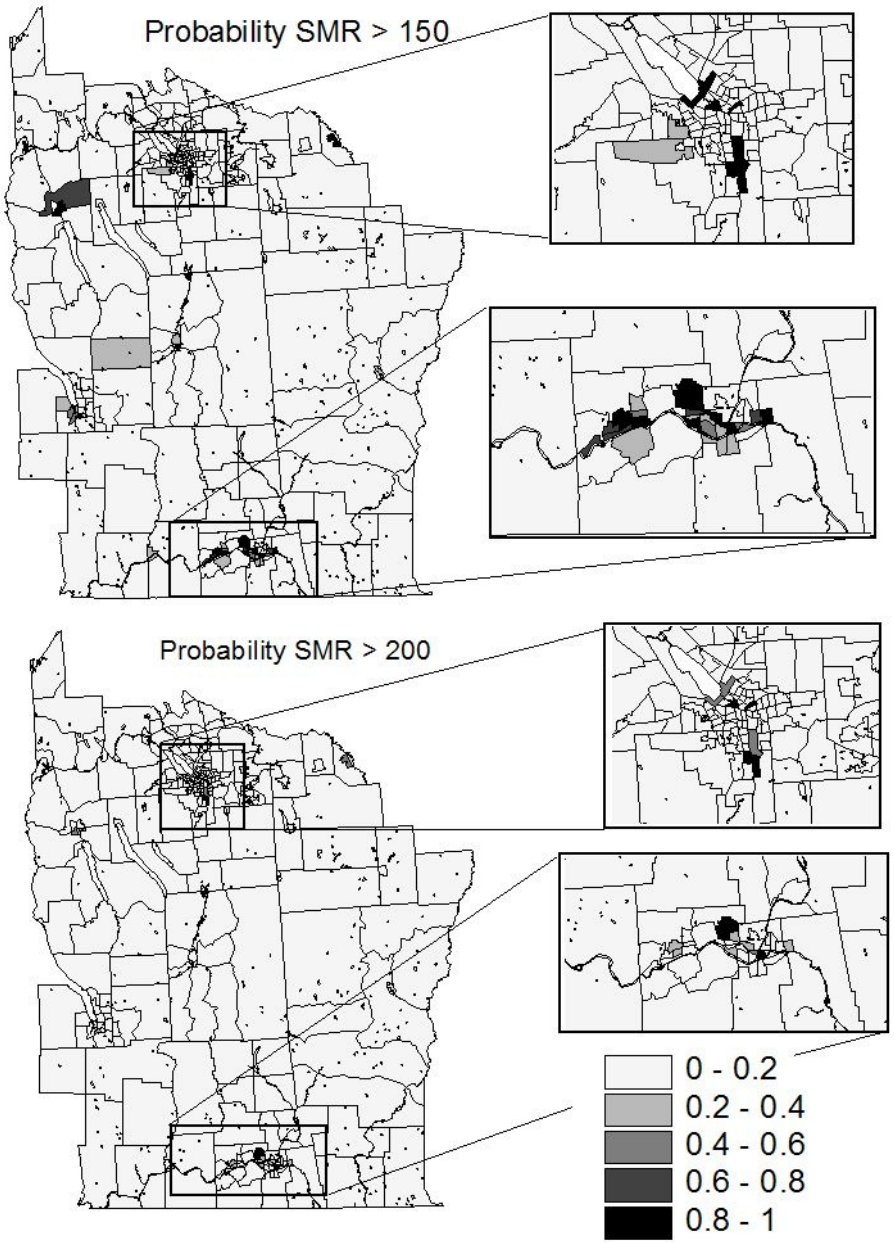
- A nifty thing about MCMC estimates:

  We get posterior samples from any function of model parameters by taking that function of the sampled posterior parameter values.
- Gives us posterior inference for $SMR_i = Y_{i,fit}/E_i$.
- Also can get $Pr[SMR_i > 200|\boldsymbol{Y}]$ and map these *exceedence probabilities*.

Posterior median SMRs



Posterior exceedence probabilities



What do we have?

- Associations between local covariates and local outcomes (counts and rates).
- Spatial correlation between random intercepts (inside the link function).
- (Aside: This is a clever idea since we can use a multivariate Gaussian distribution for correlation...).
- Result: Local rates adjusted for covariates *and* smoothed by borrowing information.
- Many examples in the literature, and many extensions, we'll start with one tomorrow!

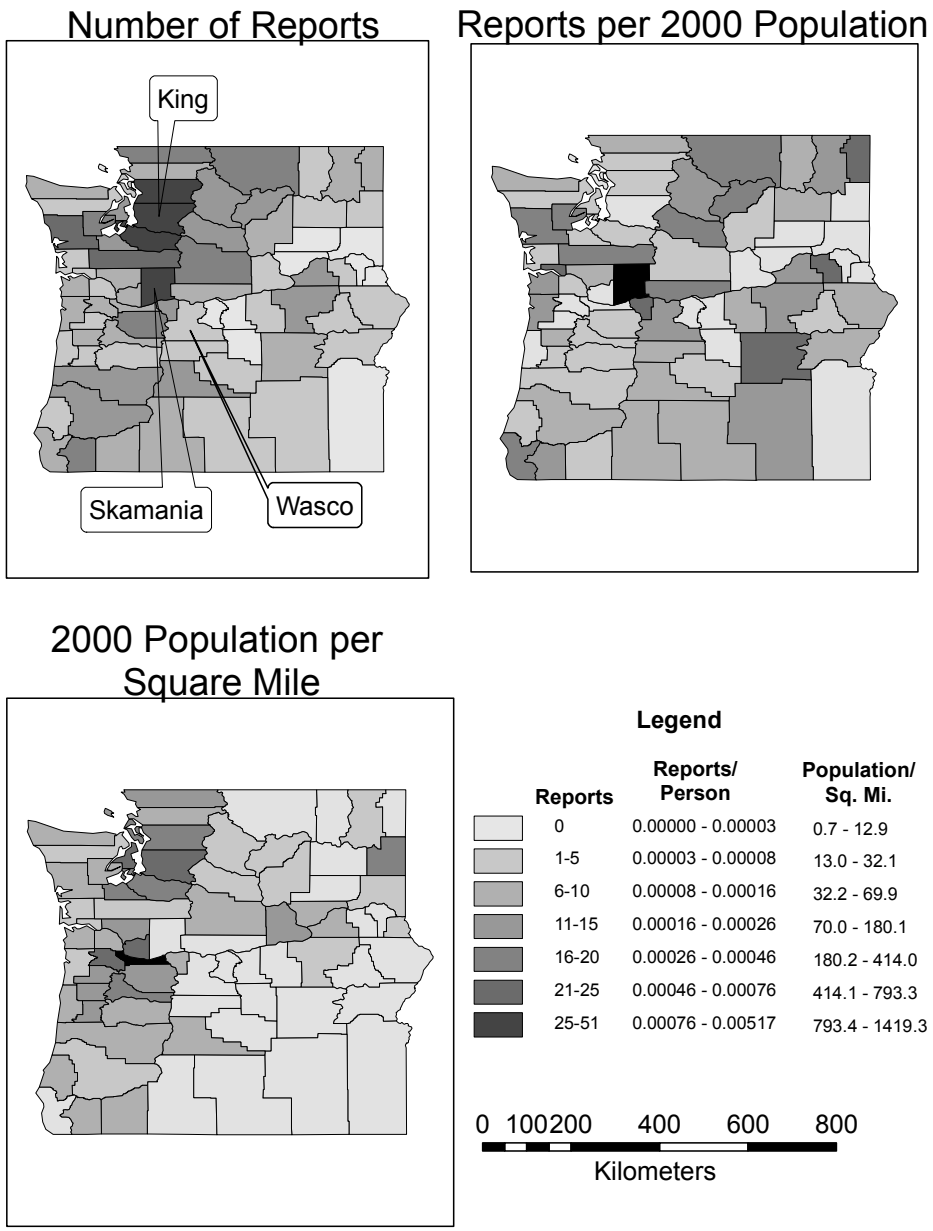
Bonus Example

- Cryptozoology Example: Waller and Carlin (2010) Disease Mapping. In *Handbook of Spatial Statistics*, Gelfand et al. (eds.). Boca Raton: CRC/Chapman and Hall.

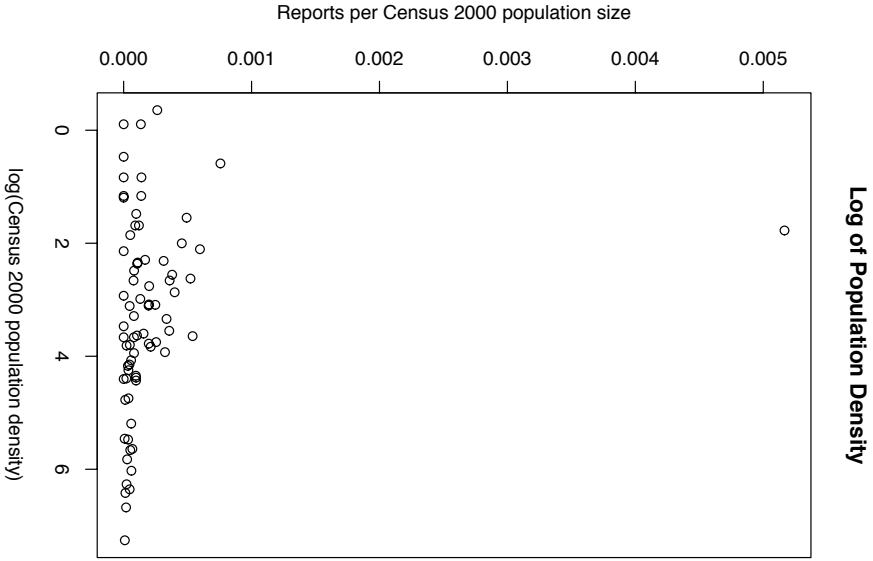
Cryptozoology example

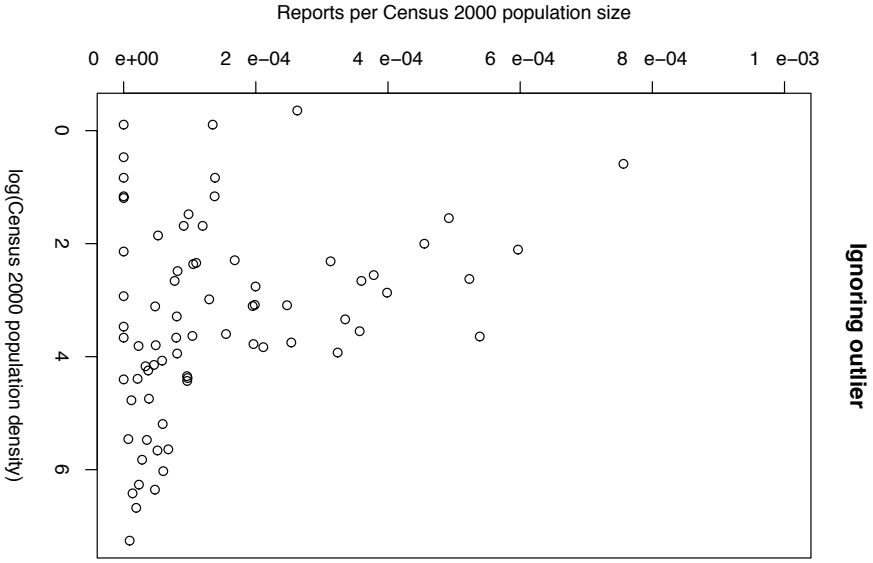
- County-specific reports of encounters with *Sasquatch* (Bigfoot).
- Data downloaded from `www.bfro.net`
- Sightings from counties in Oregon and Washington (Pacific Northwest).
- Probability of report related to population density?
- (Hopefully) rare events in small areas.
- Perhaps spatial smoothing will stabilize local rate estimates.

Sasquatch Data

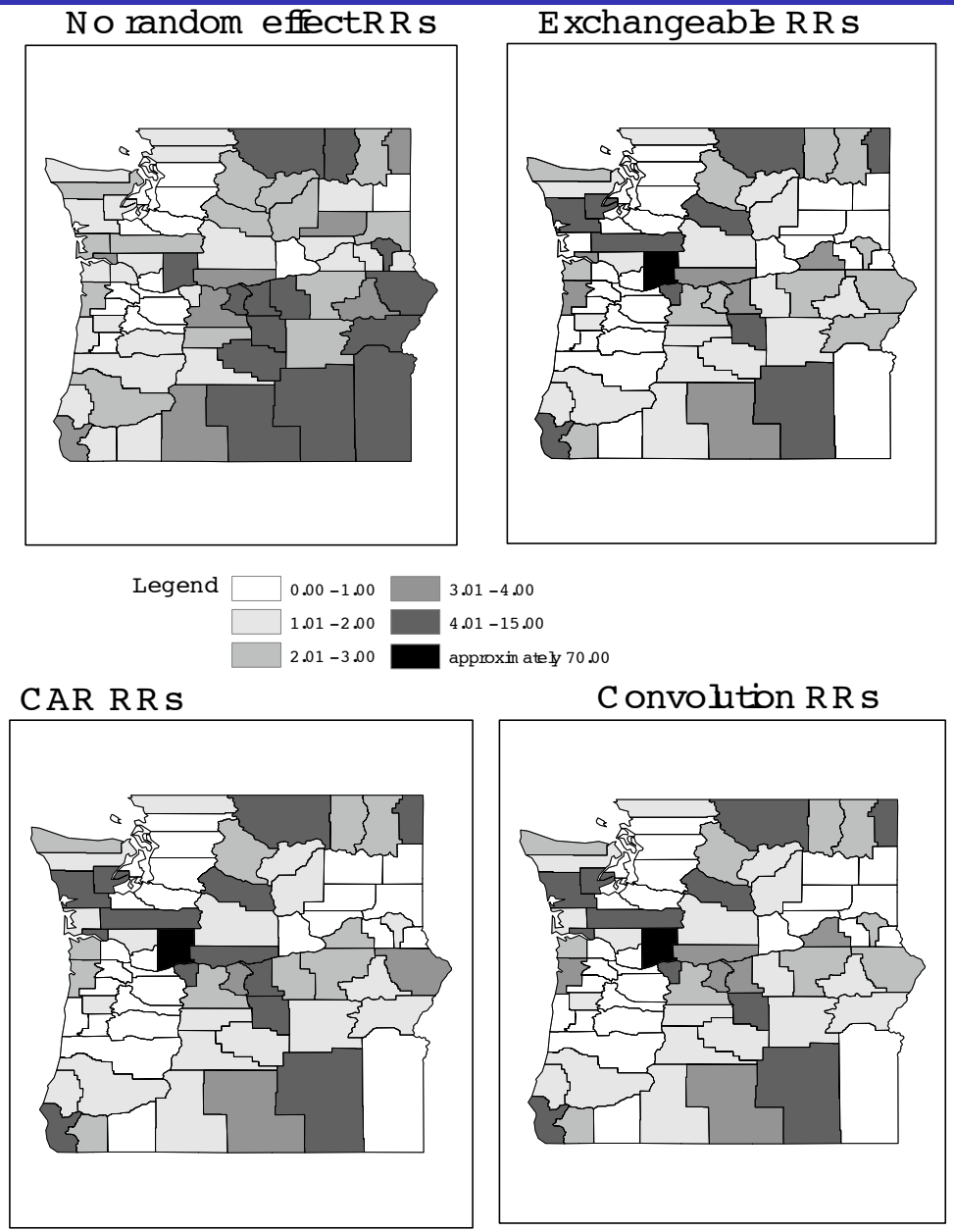


Reports vs. Population Density





Mapped relative risks



Skamania Sasquatch Ordinances

- http://www.skamaniacounty.org/commissioners/homepage/ordinances-2/
- Big Foot Ordinance 69-1: "THEREFORE BE IT RESOLVED that any premeditated, willful and wonton slaying of any such creature shall be deemed a felony punishable by a fine not to exceed Ten Thousand Dollars ($10,000.00) and/or imprisonment in the county jail for a period not to exceed Five (5) years. ADOPTED this 1st day of April, 1969."
- Big Foot Ordinance 1984-2:
  - Repealed felony and jail sentence.
  - Established a Sasquatch Refuge (Skamania County).
  - Clarified penalty (gross misdemeanor vs. misdemeanor) and penalty (fine and jail time), disallowed insanity defense, and clarified distinction between coroner designation of victim as humanoid (murder) or anthropoid (this ordinance).

And...

www.amazon.com/Skamania-County-Washington-Bigfoot-Vintage/dp/B076PWN7ZM

Conclusions

- What method to use depends on what you want data you have and what question you want to answer.
- All methods try to balance trend (fixed effects) with correlation (here, with random effects).
- All models wrong, some models useful.
- Trying more than one approach often sensible.

References

- Waller and Gotway (2004, Chapter 9) *Applied Spatial Statistics for Public Health Data*. New York: Wiley.
- Haining, R. (2003). *Spatial Data Analysis: Theory and Practice*. Cambridge: Cambridge University Press.
- Banerjee, S., Carlin, B.P., and Gelfand, A.E. (2014) *Hierarchical Modeling and Analysis for Spatial Data, 2nd Ed.* Boca Raton, FL: CRC/Chapman & Hall.
- Blangiardo, M. and Cameletti, M. (2015) *Spatial and Spatio-temporal Bayesian Models with R-INLA*. Chichester: Wiley.