MODULE 16: Spatial Statistics in Epidemiology and Public Health
Lecture 8: Disease Ecology

Jon Wakefield and **Lance Waller**

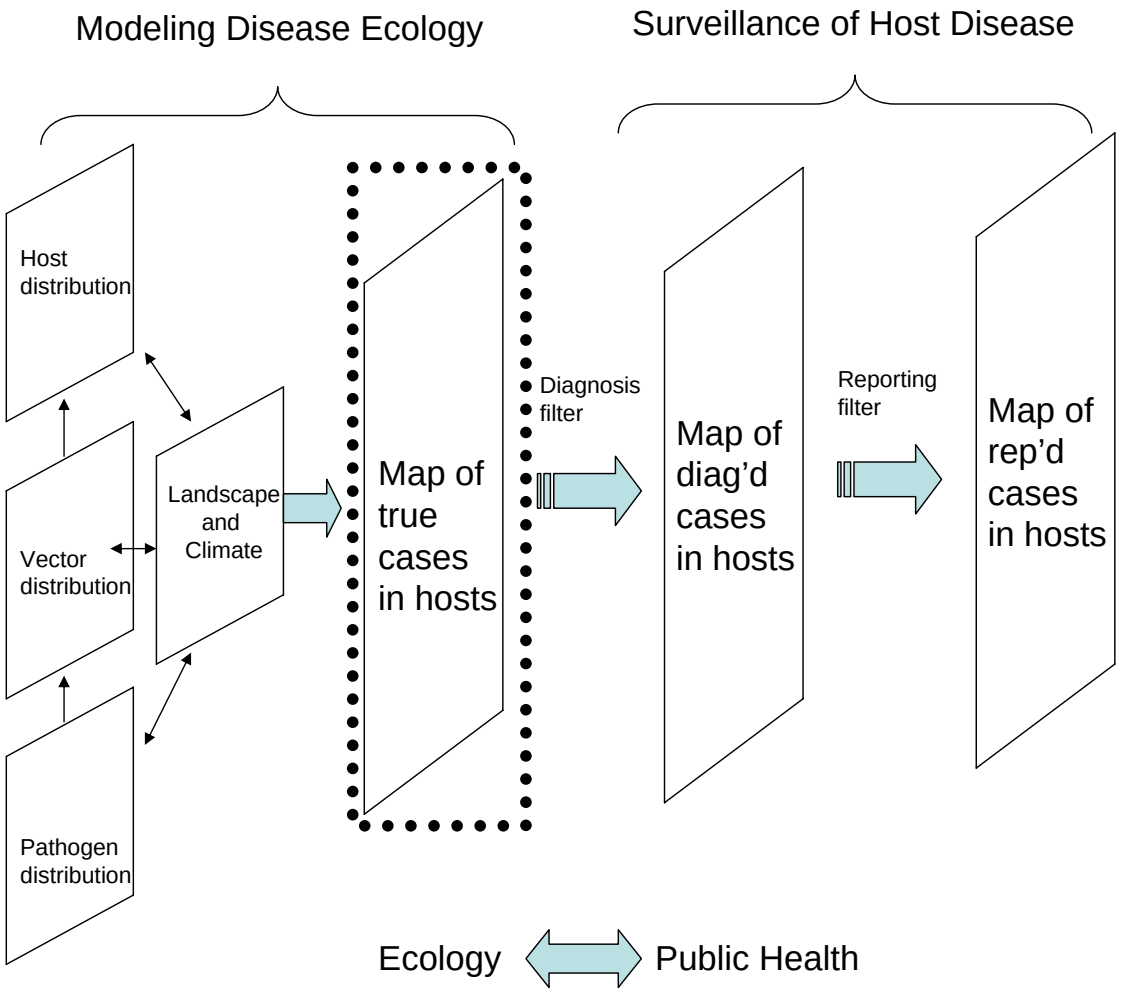
Disease Ecology

- Interactions between virus, host, landscape.
- Landscape epidemiology (Pavlovsky, 1967), landscape ecology (Manel et al. 2003, *TrEE*), spatial epidemiology (Osfeld et al. 2005, *TrEE*), landscape genetics (host and virus) (Biek et al. 2006, *Science*), conservation medicine (Aguirre et al. 2002).
- People, animals, diseases, ecology, environment!
- Spatio-temporal data, mathematical models, genetic sequences, missing data, GIS!

Epizoology and Epidemiology

- Most emergent infectious diseases have animal reservoir (WNV, Ebola, Avian influenza, Monkeypox, SARS, HIV/SIV).
- History of animal/human disease (Torrey and Yolken, 2005, *Beasts of the Earth*).
- Interesting intersection of modelers, ecologists, statisticians, medical geographers, ecological geneticists, public health researchers, epidemiologists.

The "big picture"



Not a new idea (Koch, 2005, *Cartographies of Disease*)

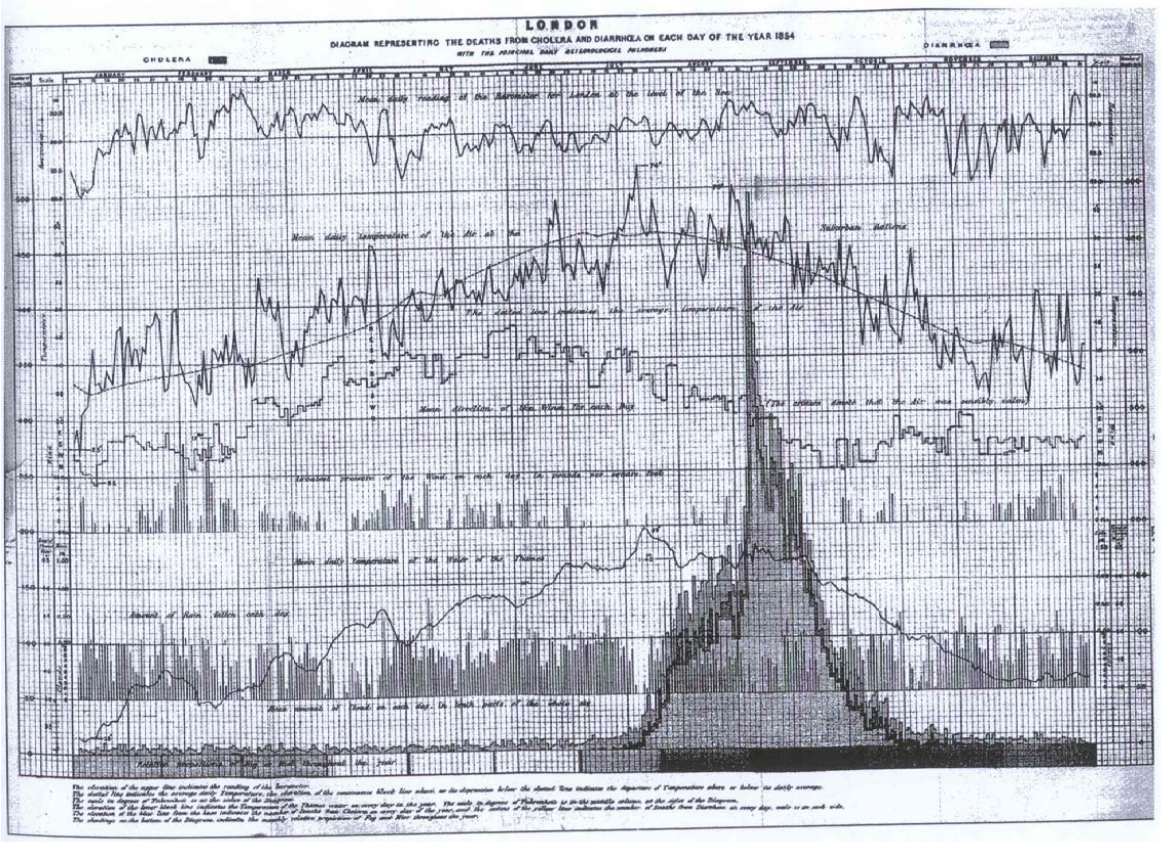


Figure 5.6 A graph of climatic variables joined to incidence of cholera (blue) and chronic diarrhea (yellow) in London, 1854. The map was based on readings from twenty-four urban recording stations in London and prepared by the General Board of Health for a report to both houses of Parliament.

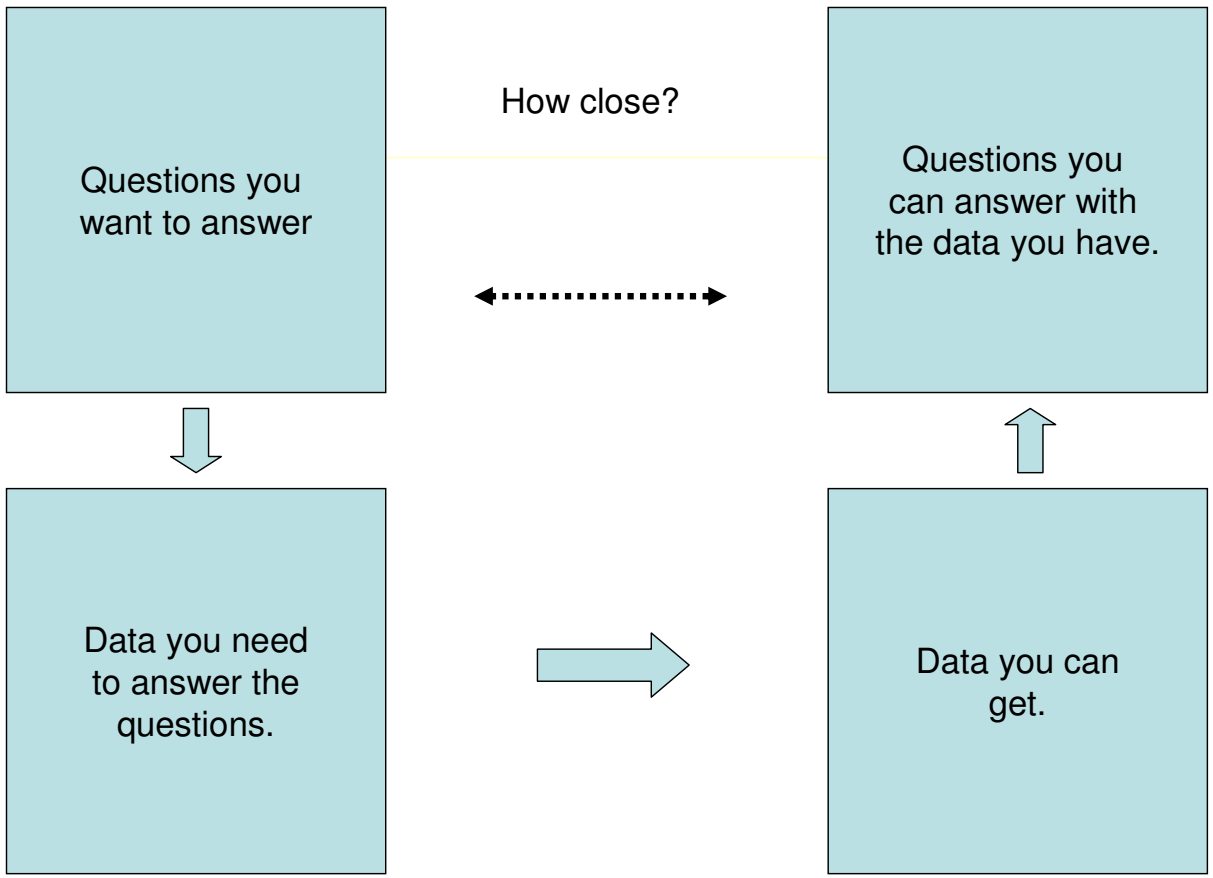
Pattern and Process

- Our ultimate goal is understanding the ecological processes driving the patterns we see in our observations.
- When linking process (model or reality) to pattern (data), typically:
  - Ecology focus: Process to pattern
    - Emphasis on mathematical model, link to available data
  - Statistics: Pattern to process
    - Collected data, hypothesis test or analytic (e.g., regression) model.

Ultimately futile exercise?

- Process may not yield unique pattern (e.g., chaos, stochasticity).
- Pattern may not reveal unique process without additional information (e.g., spatial point patterns, Bartlett (1964)).
- But the real question is, "Can we learn more than we already know?"
- If not, what additional data do we need?

The whirling vortex



1

Exactly wrong or approximately correct?

- John Tukey noted an approximate answer to the right question is better than a precise answer to the wrong question.
- Particularly important here...if available data redefine our answerable questions, we may be changing course without realizing it!
- Let's look at how modelers and statisticians address these questions...

Gaps and Bridges

Conceptual gap

- Mathematical modelers
  - Build assessments using families of models and deriving properties.
  - Data used to calibrate models.
  - Using process (model) to understand pattern (data).
- Statisticians
  - Build inference from probability model defining observations.
  - Data define a likelihood function.
  - Using pattern (data) to understand process (model).

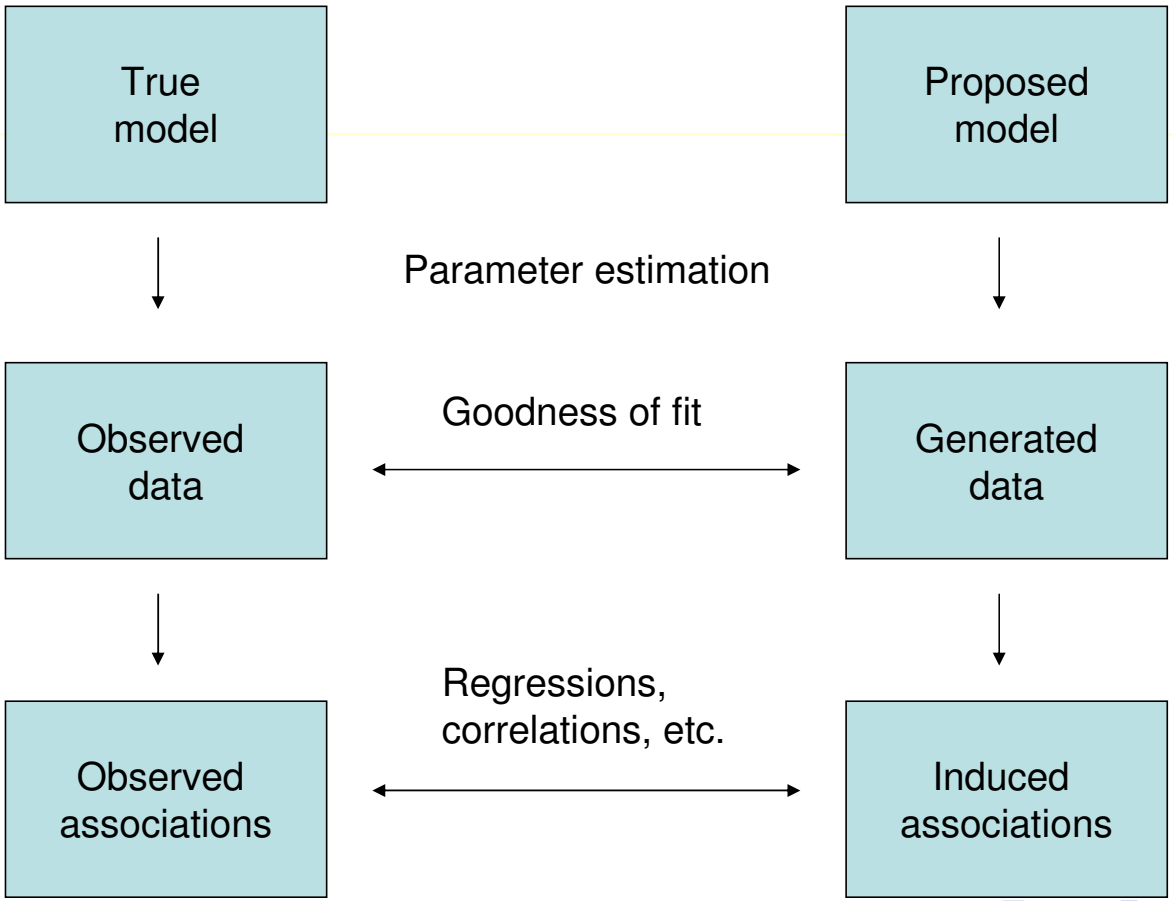
Gaps and Bridges

Training gap

- In current modes of training, mathematical modelers often take one (or fewer) courses in statistics.
- Statisticians often take one (or fewer) courses in mathematical modeling.
- Furthermore, the importance of one area is seldom stressed in the other.
- Few working at the intersection of the two but there is a lot of interesting work to be done!

To see how this might work, consider the following...

How statistics might help...(*Ecology*, 2010)



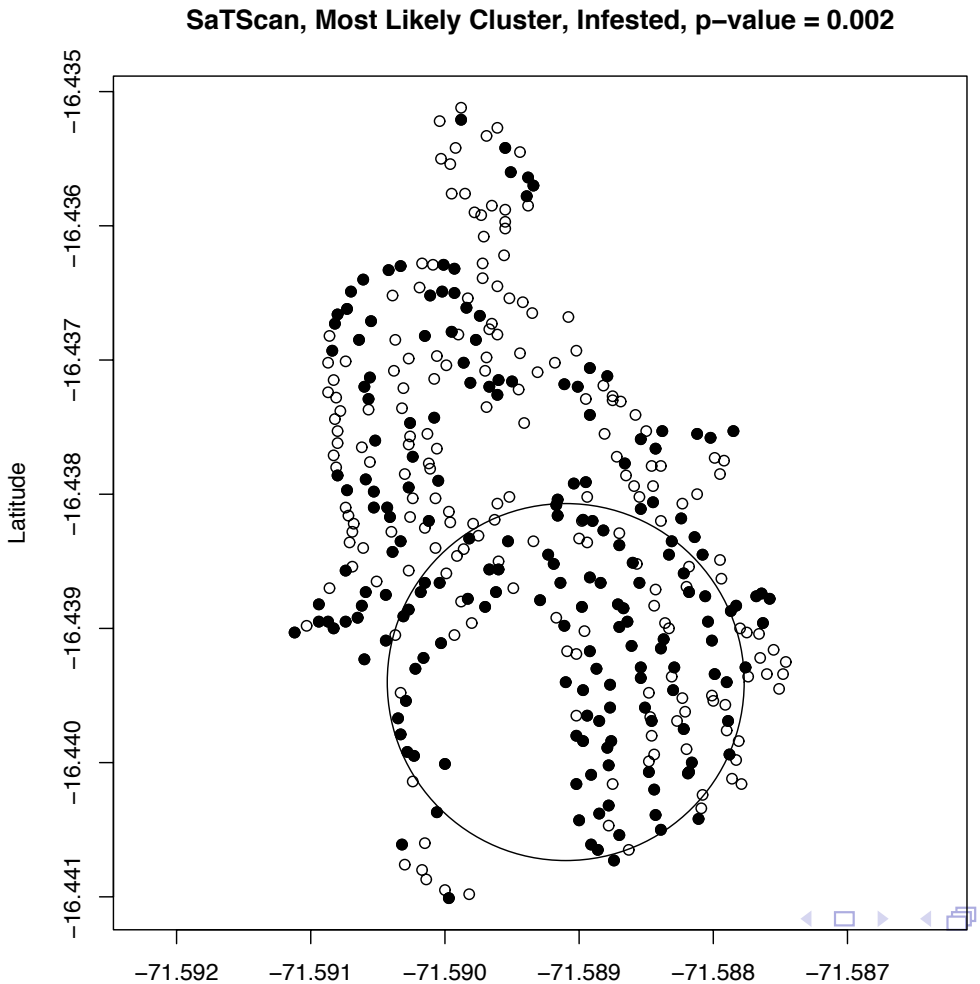
Chagas disease in Peru

- Joint work with Michael Levy (Fogarty International Center, NIH)
- Chagas disease: Vector borne disease (infection with *T. cruzi*).
- Vector (in southern Peru): *Triatoma infestans*.
- Study area: Guadalupe, Peru (peri-urban).
- Fields surrounding rocky hilltops with houses.
- GPS all household locations.
- Spraying campaign, identify house locations, houses with vectors ("infested"), and houses with infected vectors ("infected").

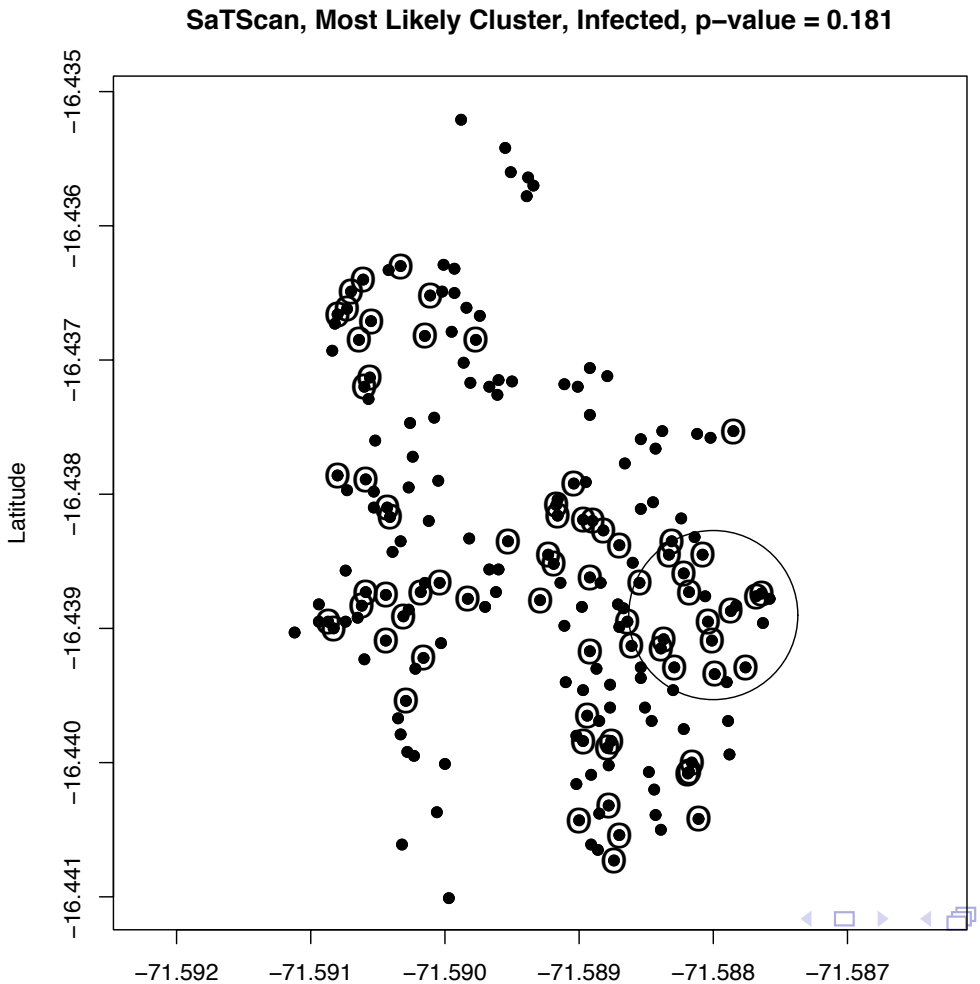
How to find a cluster?

- Consider two approaches: scan statistic and intensity estimators.
- Spatial scan statistic:
  - Define set of potential clusters (elements of scanning window).
  - Assign "score" to each potential cluster.
  - Find "most likely cluster" (MLC) as potential cluster with extreme score.
  - Evaluate significance of most likely cluster via Monte Carlo simulation.
  - Compare observed "score" of MLC to distribution of scores MLCs (regardless of location) under random assignment.
  - SaTScan software (`www.satscan.org`).

SaTScan, Infested among households, Most likely cluster (p=0.002)



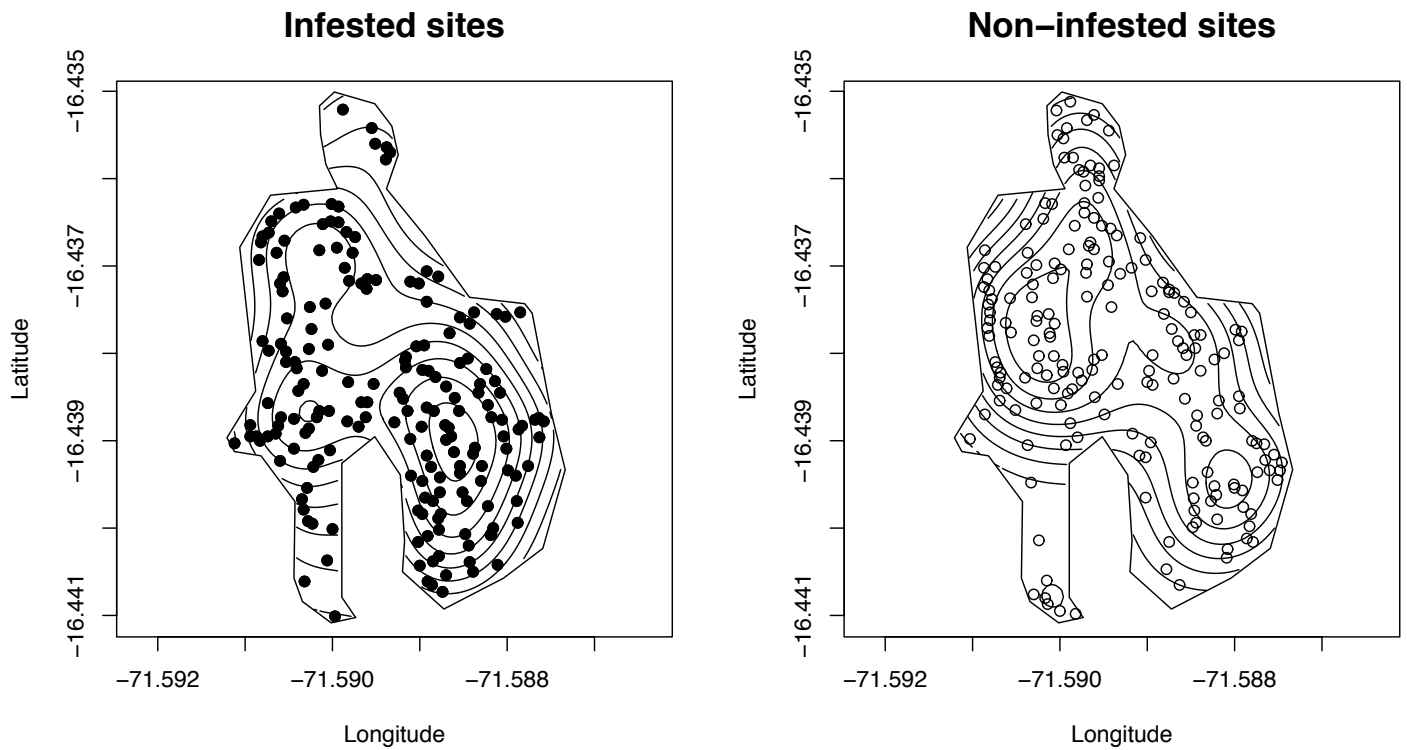
SaTScan, Infected among infested, Most likely cluster (p=0.181)



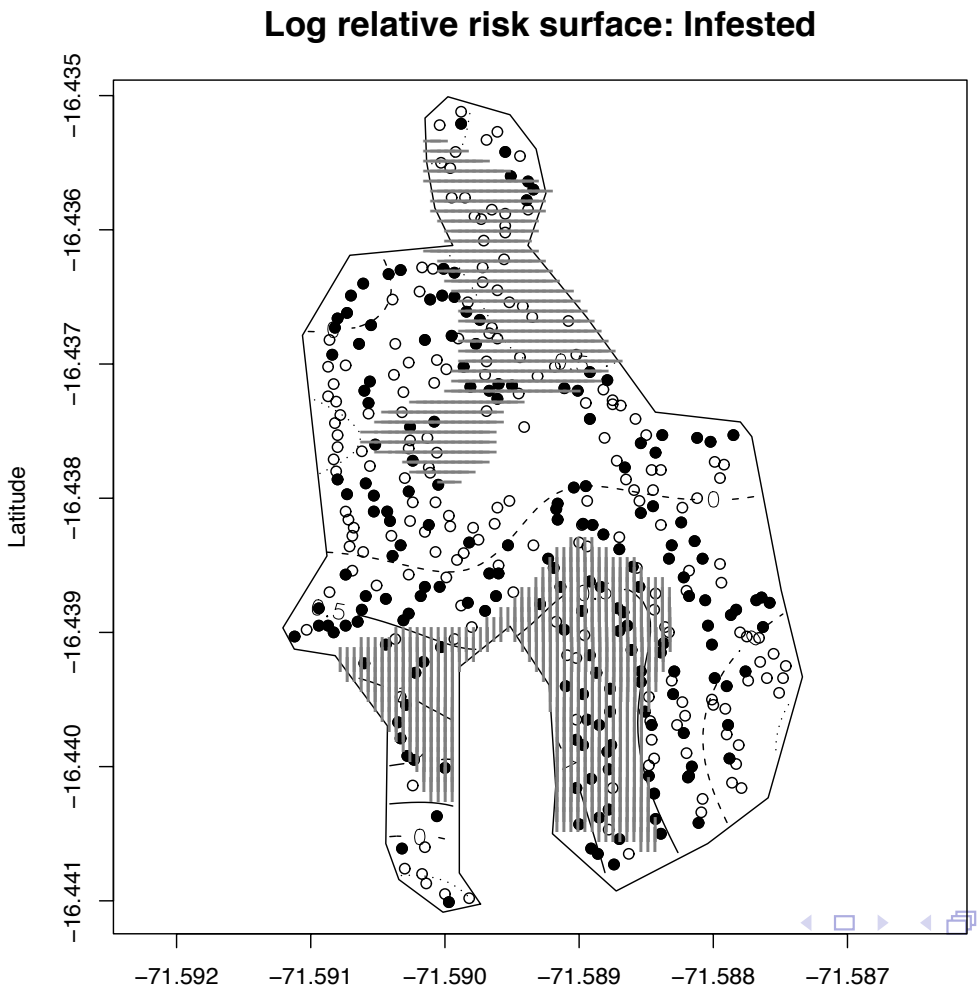
Chagas SaTScan conclusions

- Statistically significant cluster of infested households among all households.
- No statistically significant cluster of infected households among infested households.
- Note circular most likely cluster may include gaps (top of hill).
- What about non-circular clusters?

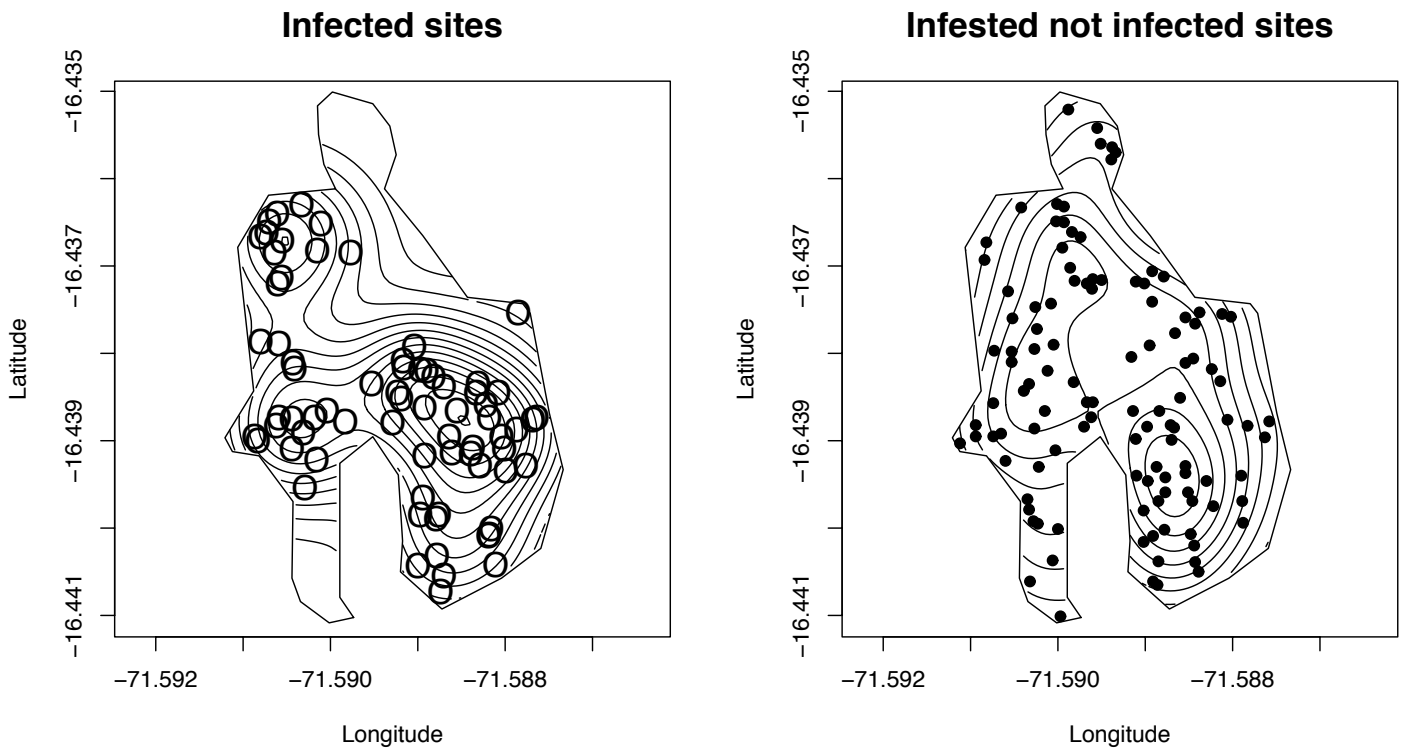
Kernel intensity estimates, infested vs. all households



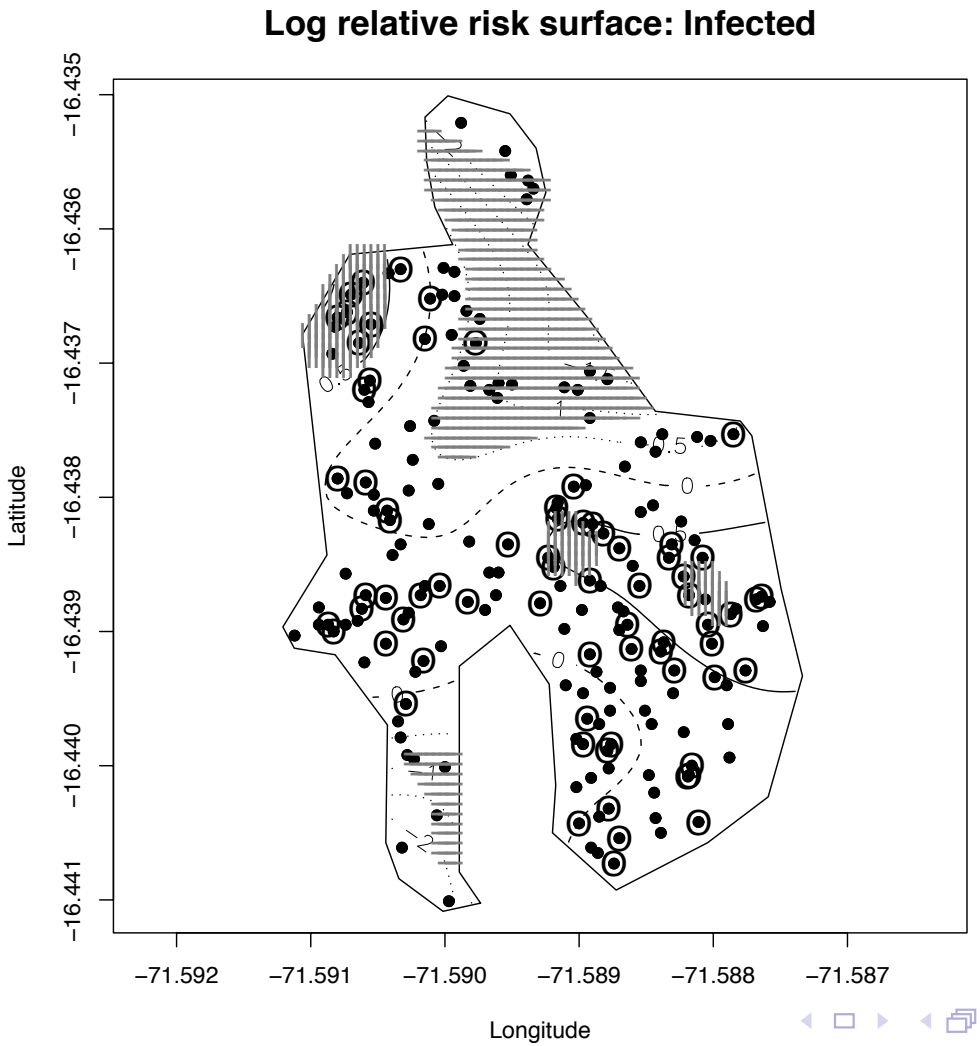
Ratio of kernel intensity estimates, infested vs. all households



Kernel intensity estimates, infected vs. infested households



Ratio of kernel intensity estimates, infected vs. infested



Cluster conclusions

- Relative risk surface adds more geographical precision to patterns initially revealed by SaTScan.
- Large risk of infestation in the south.
- Within this some pockets of increased risk of infection.
- Area of lower risk missed by circular scan statistic, due to its irregular shape.
- Identifies areas for future studies.

Chagas conclusions

- Significant cluster of infested households, but no clusters of infected households (circular clusters).
- Relative risk surface also suggests area of low risk (both infestation *and* infection) in northeast.
- K functions suggest significant *clustering* of *infected* but not *infested* households.
- Taken together, results reveal different aspects of the underlying process.
- A single cluster does not define clustering, nor does clustering imply a single cluster.

Chagas conclusions

- Infestation: pockets of higher and lower relative risk, but level of clustering not different between cases and controls.
- Infection: More clustered at small distances than infestation, but resulting clusters are smaller and more diffuse.
- Scale of clustering different between infestation and infection, and larger than typical range of individual vectors.
- Scale of clustering useful in targeted surveillance for human cases (Levy et al., 2007, *PLoS NTD*).

Raccoon rabies

What is rabies?

- Virus in family of Lyssa ("frenzy") virus.
- Behavioral impact on host.
- Reportable disease.
- Various strains associated with primary host (bat, dog, coyote, fox, skunk, and raccoon).
- Host cross-over, typically transmitted via bite/scratch.
- Most human infection from bat strains.

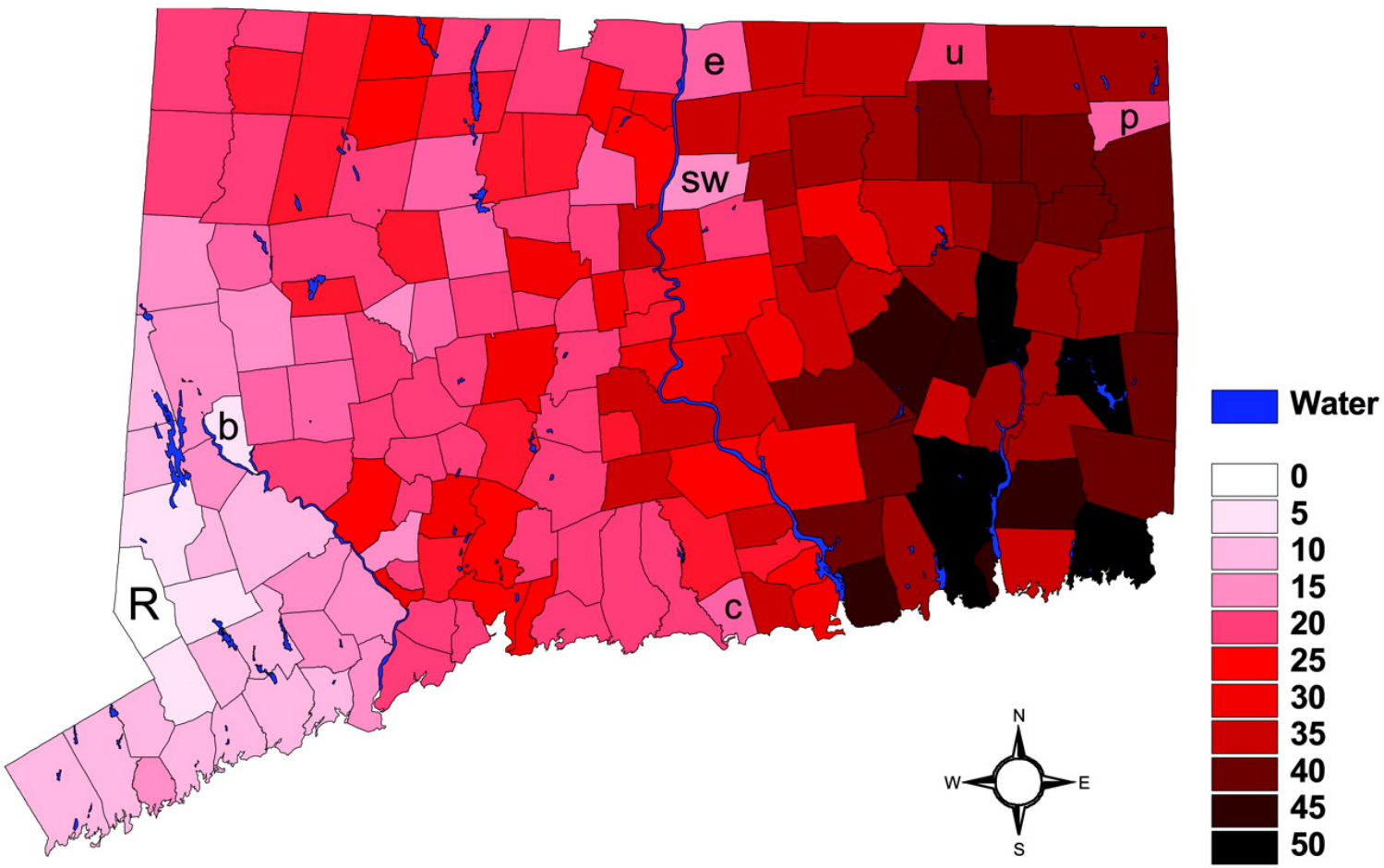
Raccoon rabies

- Endemic in Florida and South Georgia.
- Translocation of rabid animal(s) to VA/WV border circa 1977.
- Wave-like spread since.
- Connecticut first appearance 1991-1996.
- Ohio 2005.
- Joint work with Leslie Real's lab in Population Biology, Evolution, and Ecology (David Smith, Colin Russell, Roman Biek, Scott Duke-Sylvester).

Raccoon rabies in CT

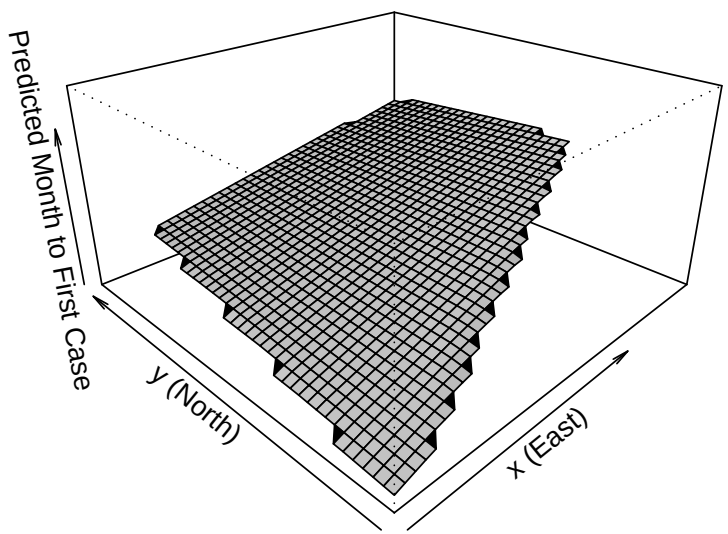
- First appeared in western township in 1991.
- Irregular wave roughly west-to-east.
- Crossed state in ≈ 5 years.
- Features of interest:
  - River effect?
  - Long distance transmittal?
  - Would a *cordon sanitaire* built from vaccinated baits work?

Data: Months to first appearance

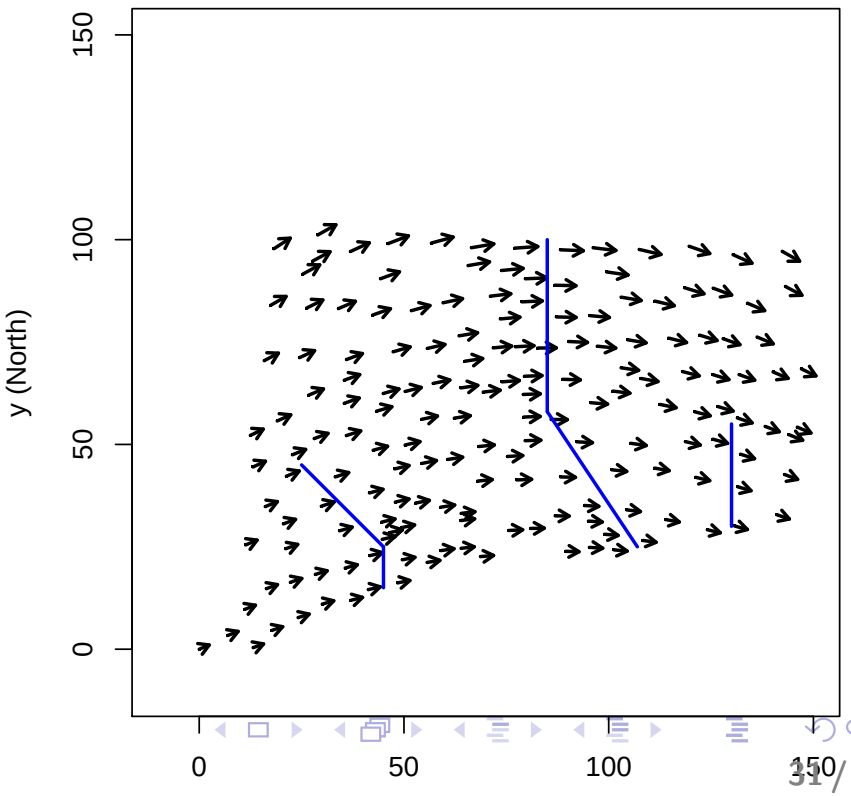


Quadratic Trend Surface

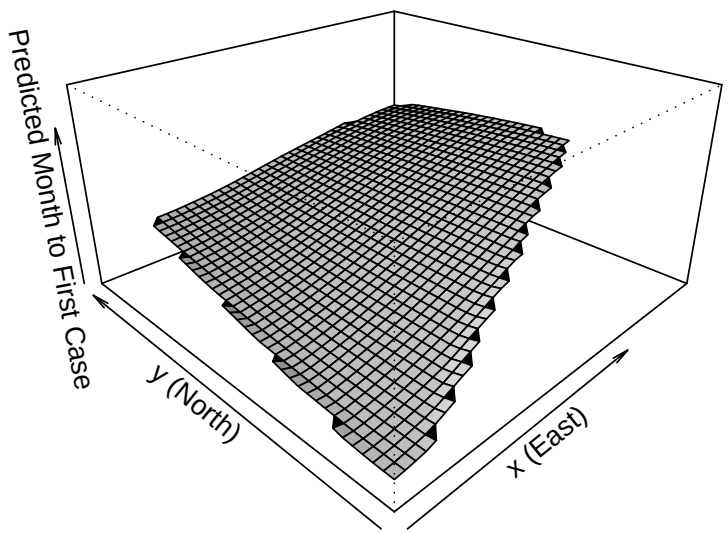
Connecticut Rabies: Best fit quadratic TS

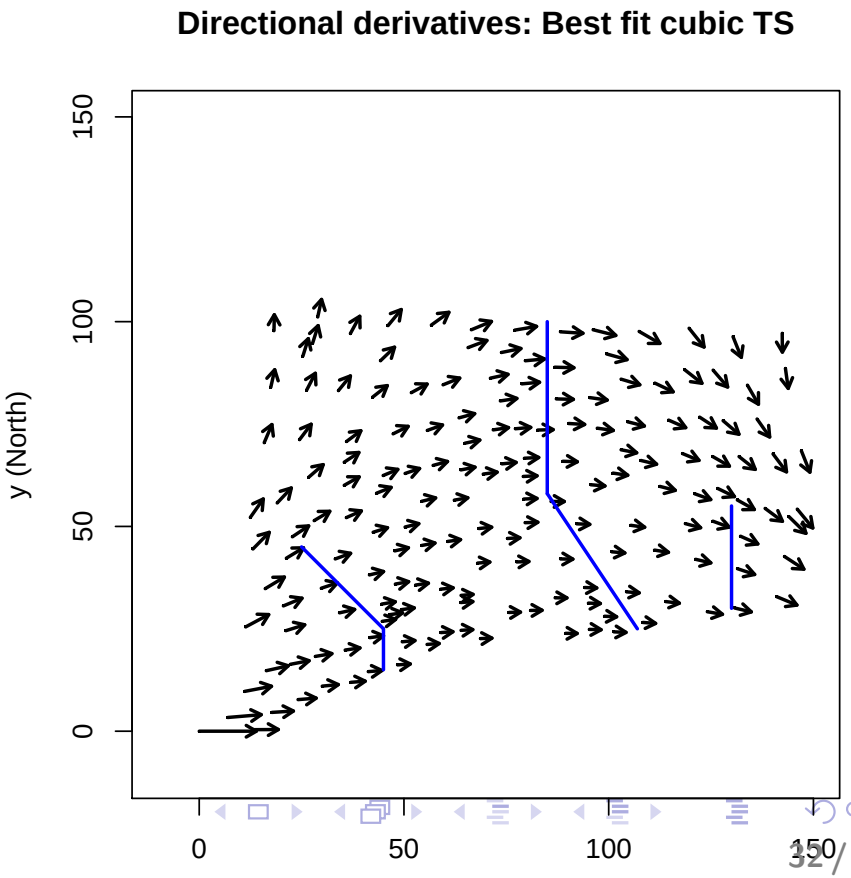


Directional derivatives: Best fit quadratic TS

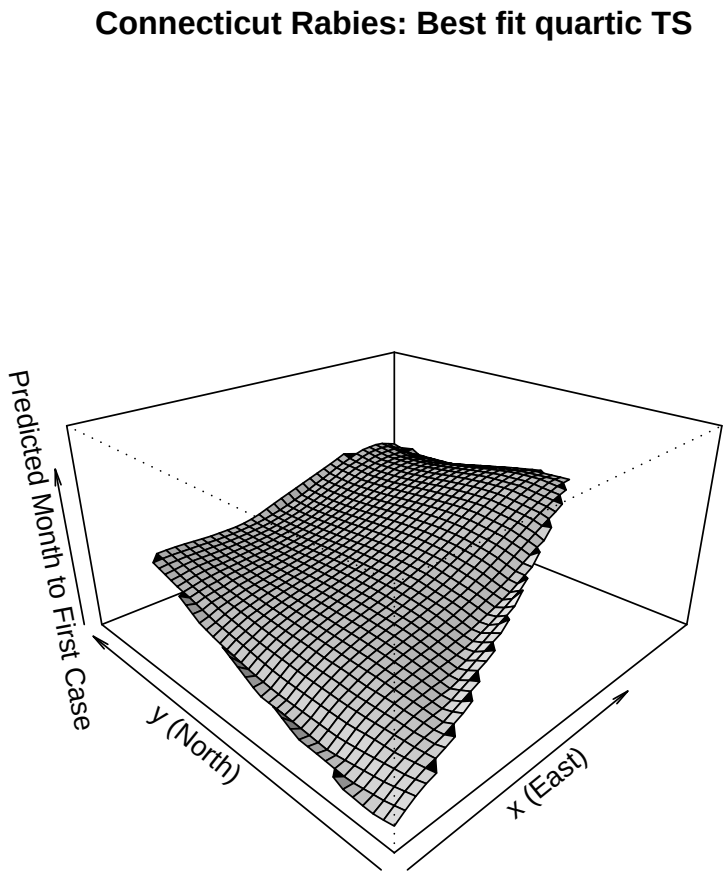


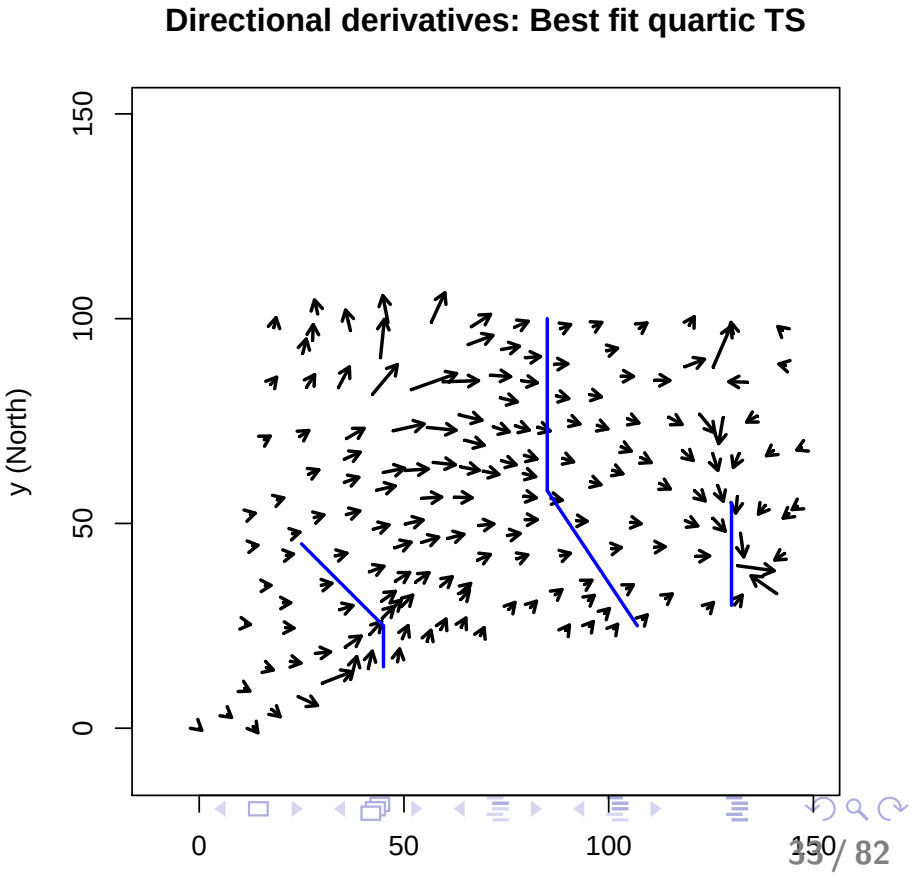
Cubic Trend Surface



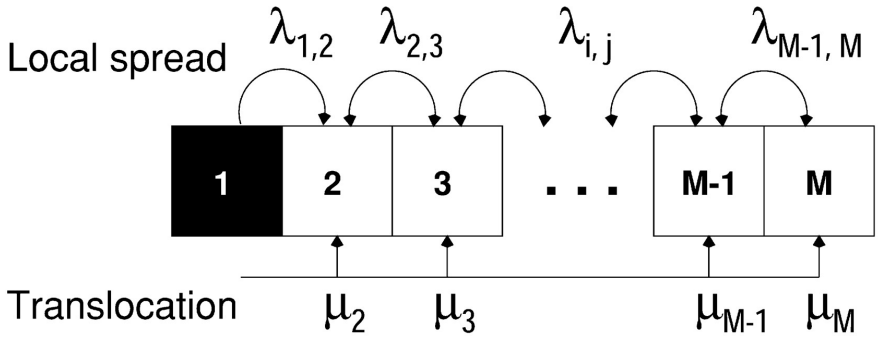


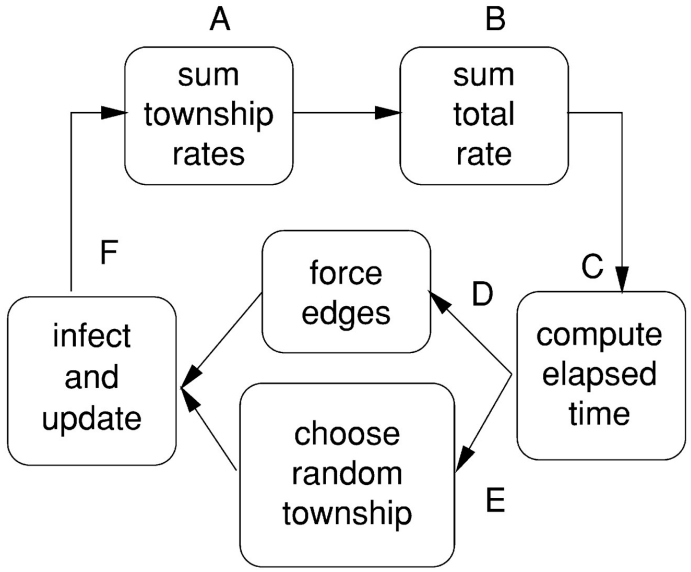
Quartic Trend Surface





Cellular automata stochastic model (David Smith)





Does the model fit the data?

- Smith et al. (2002, *PNAS*), Waller et al. (2003, *Eco Mod*)
- For today: two models of interest:
  1. *Null:* Homogeneous spread ($\lambda_{ij} = \lambda$) + translocation.
  2. *River:* Probability of spread lower across river boundaries (two values for λ_{ij}) + translocation.

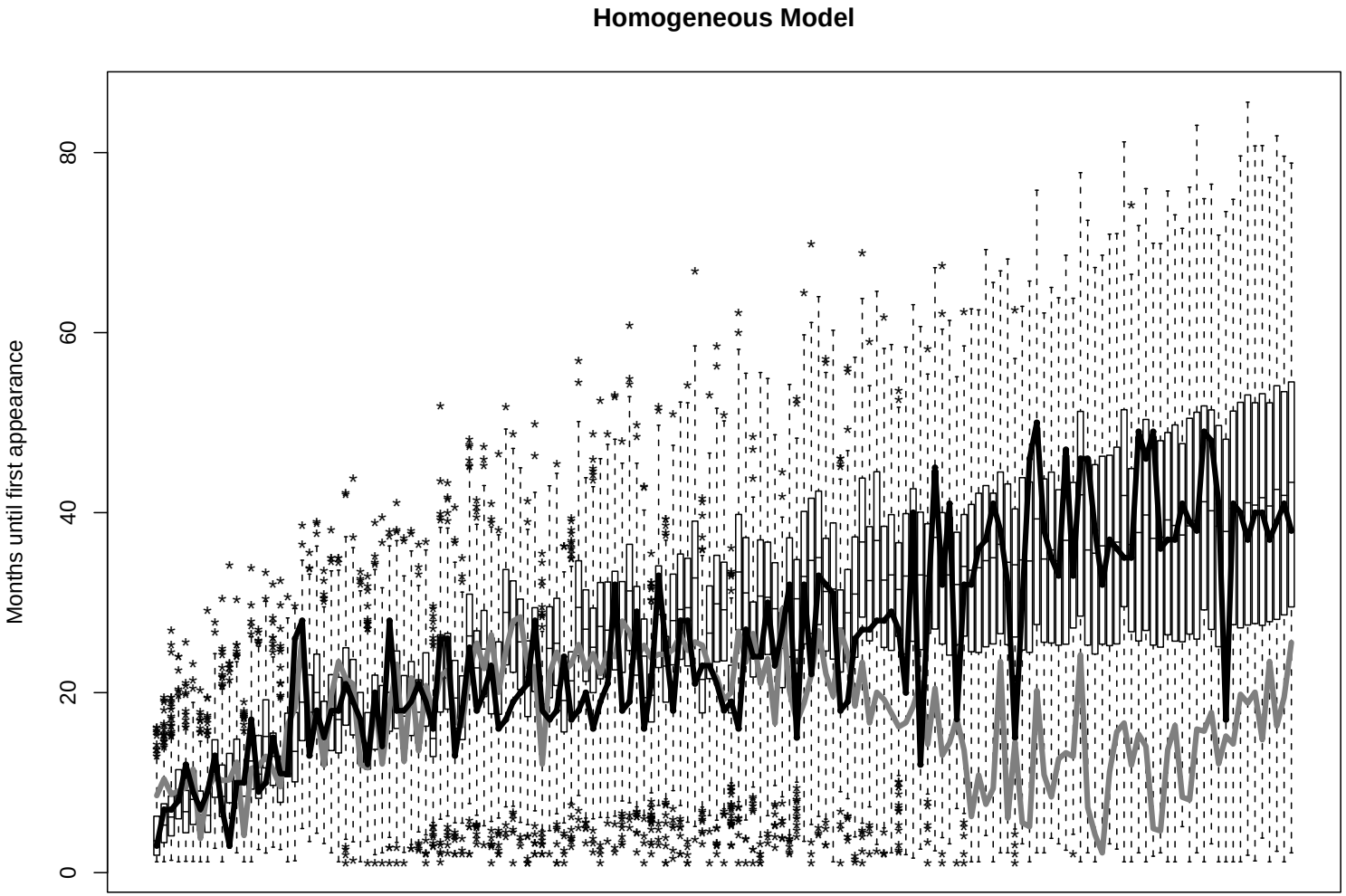
What do we have?

- We have 5,000 independent realizations under the fitted model.
- We have one data realization from the "true" process.
- If we use the data to define a likelihood, we could see if the model seems consistent with the data.
- *OR* we could use the 5,000 realizations and ask "Do the data seem consistent with the model?"
- Do the data look like they could have been a realization of the model?

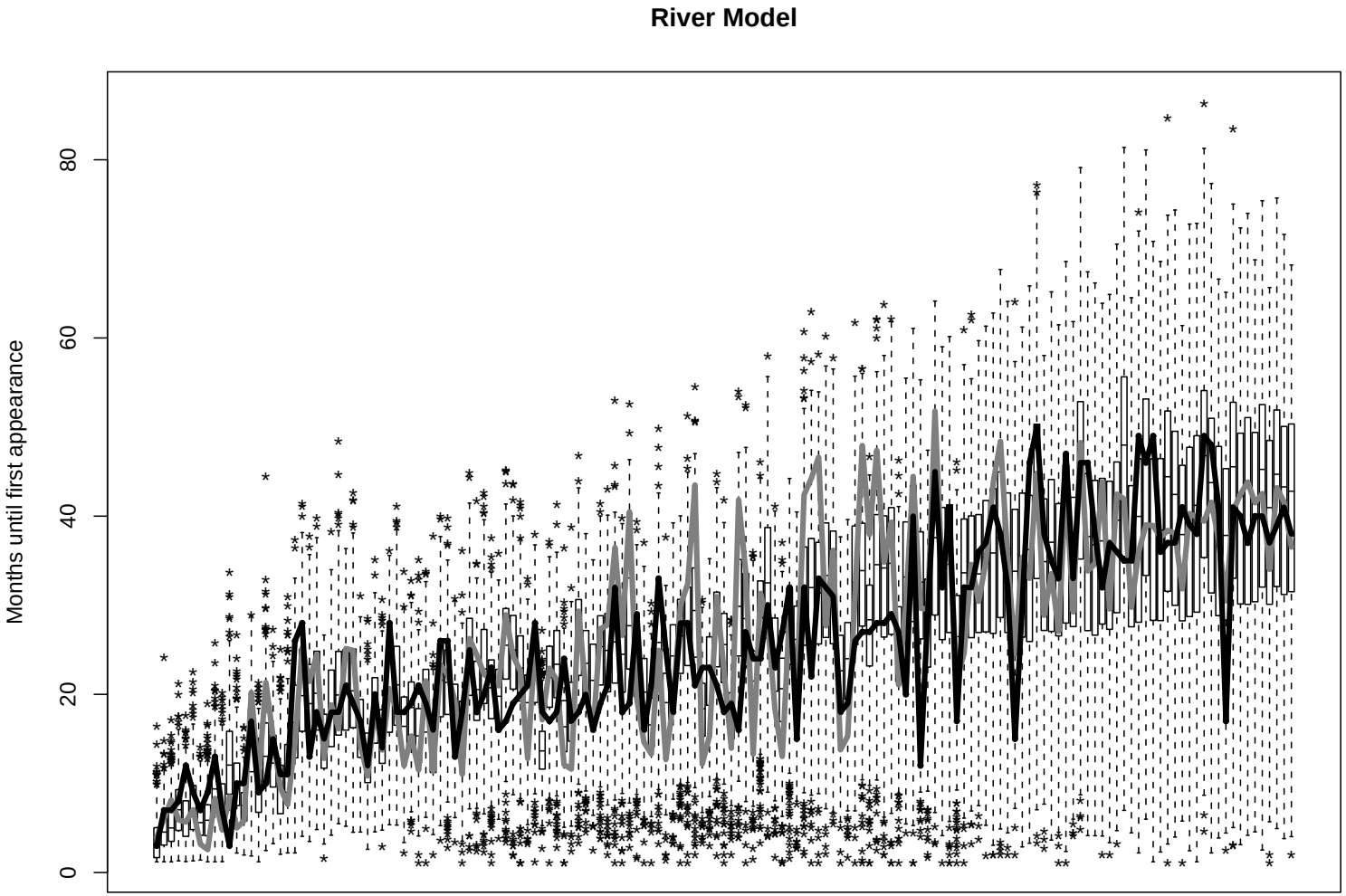
Monte Carlo testing

- Barnard (1963) discussion of Bartlett (1963).
- For a test statistic T, we want the distribution of T under H_0.
- Observe value t^* from the data set.
- p-value $= \Pr[T > t^* | H_0 \text{ true}]$.
- We have 5,000 data sets under H_0 : model is true, calculate T for each of these.
- Histogram of these values approximates distribution of T under H_0.
- Proportion of simulated T's $> t^*$ approximates p-value.

Model realizations: Homogeneous model



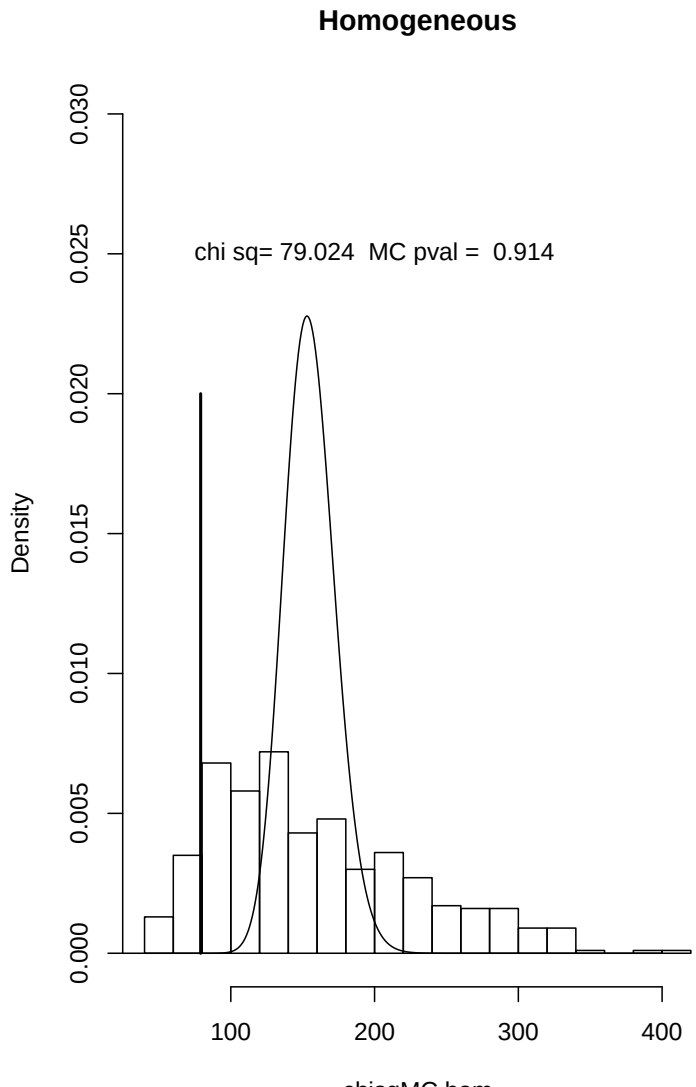
Model realizations: River model

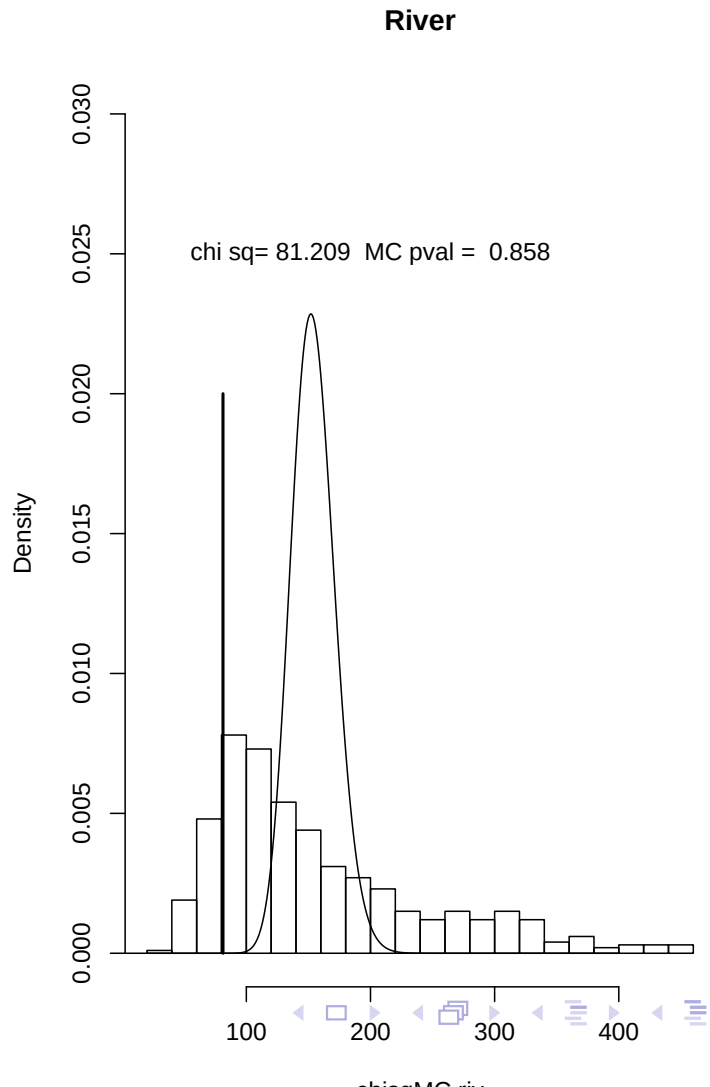


Measuring fit

- Consider $Y^2 = \sum_{i=1}^{n}[(O_i - E_i)^2/V_i]$.
- Sum of squared, standardized residuals.
- Null distribution of Y^2?
- Cross validation approach: Calculate Y^2 for each simulated data set as O_i and other 4,999 defining E_i and V_i.

Adjusted Pearson results

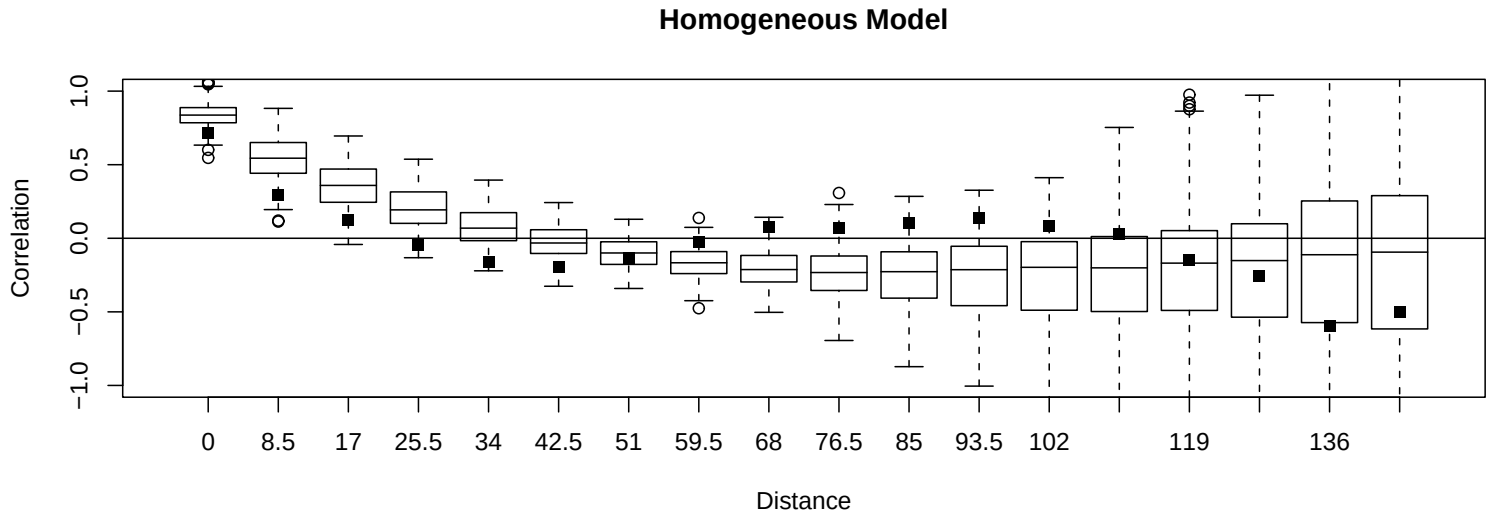


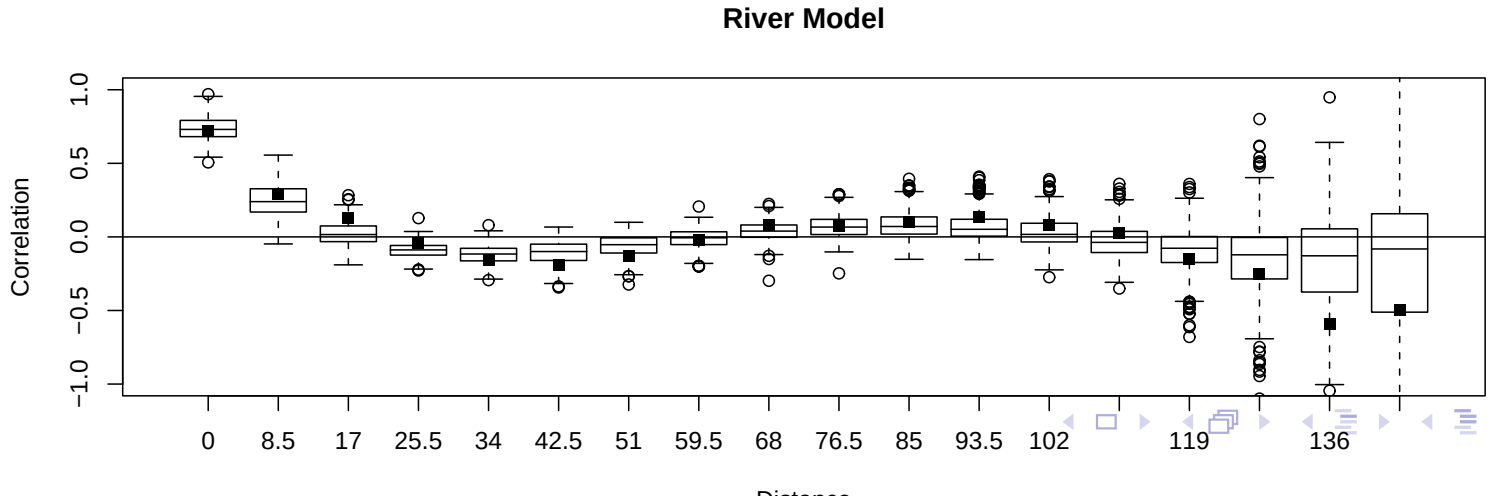


But there's more!

- What about the joint (spatial) fit?
- Models defined by local interactions, induce joint (global) associations.
- Do the models generate spatial patterns similar to the observed pattern?
- Calculate the correlogram (correlation as function of distance) for data and for each realization.

Correlograms





Other measures of fit?

- Mayer and Butler (1993, *Eco Mod*) propose *modelling efficiency*, an R^2 type measure of fit.

$$EF = 1 - \frac{\sum_{i=1}^{n}(O_i - E_i)^2}{\sum_{i=1}^{n}(O_i - \bar{O})^2}$$

  where $\bar{O}$ is the sample mean observed value.
- What fraction of variation around overall mean is captured by variation around model expectations?
- Note: $\bar{O}$ is worst-case regression, not same thing here.

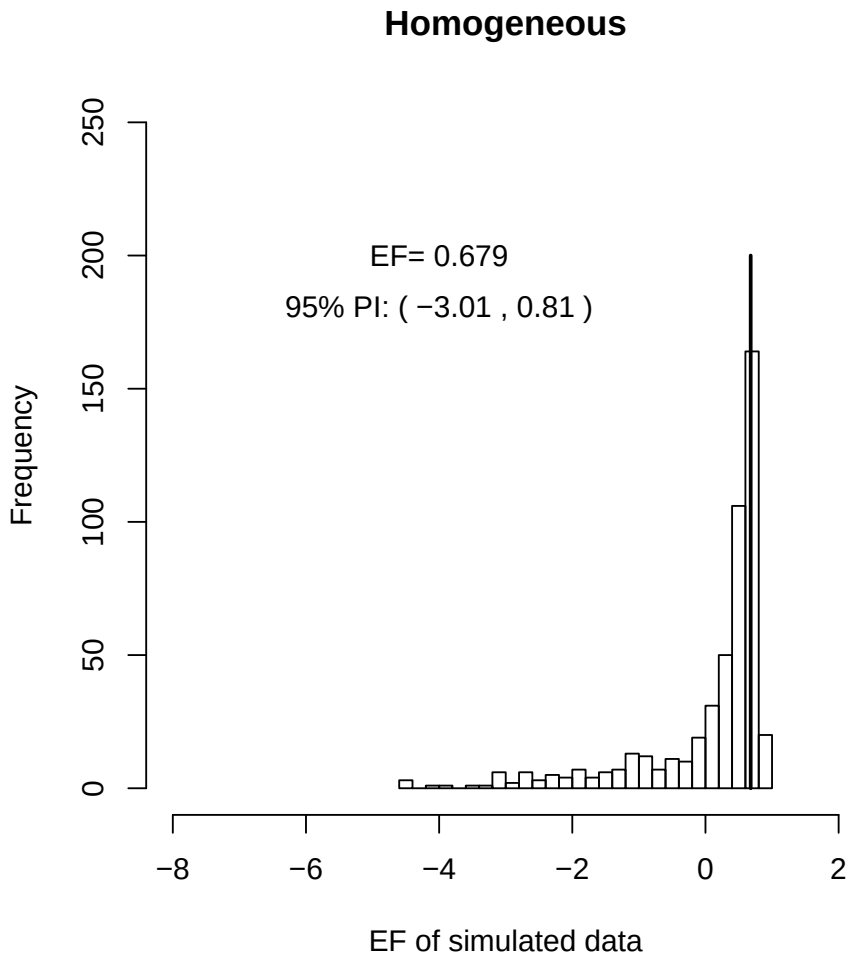
Modelling efficiency

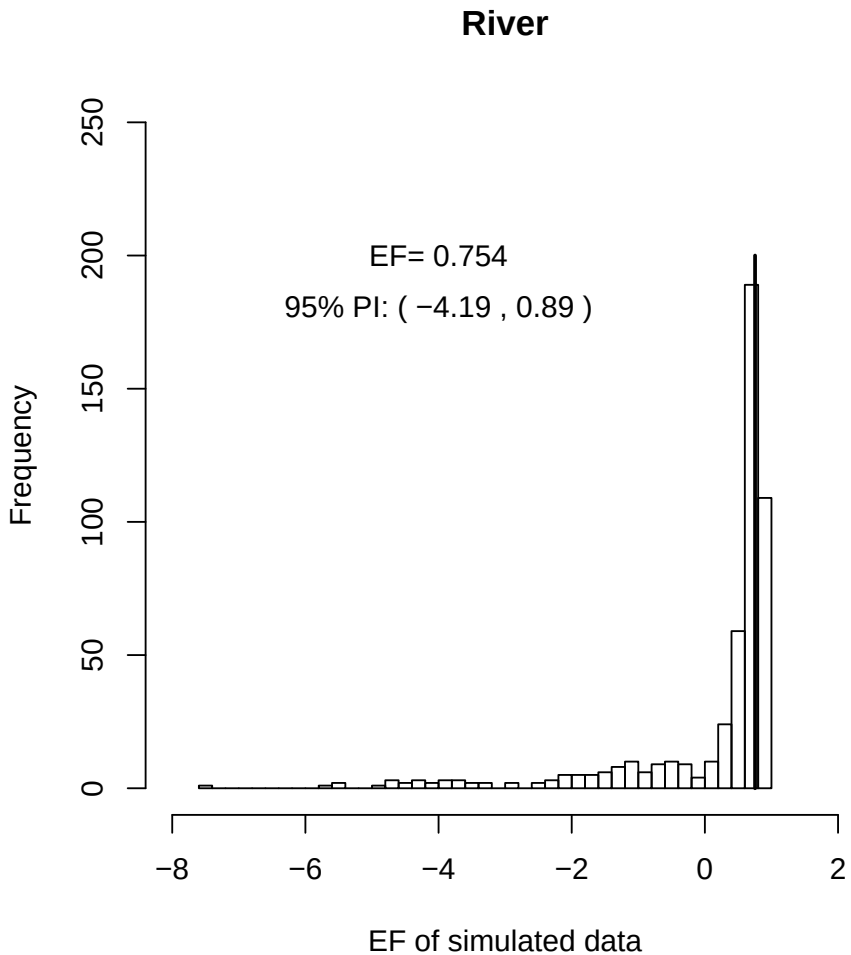
- EF(Homogeneous) = 67.9%, EF(River) = 75.9%
- Variability under H_0, cross-validate again!
- For rth simulation, calculate

$$EF = 1 - \frac{\sum_{i=1}^{n}(O_{r,i} - E_{-r,i})^2}{\sum_{i=1}^{n}(O_{r,i} - \bar{O}_{-r})^2}$$

where subscript r denotes within rth simulation, $-r$ excluding rth simulation.

Modelling efficiency





What we have so far

- Mathematical model of spatio-temporal dynamics of spread on landscape scale.
- Monte Carlo assessments of fit to data.
- Why is it moving faster in Northeast than it did in Southeast?
- Susceptible hosts? Molecular evolution of virus?
- Do all rivers have the same effect?
- Are there other geographical barriers to spread?

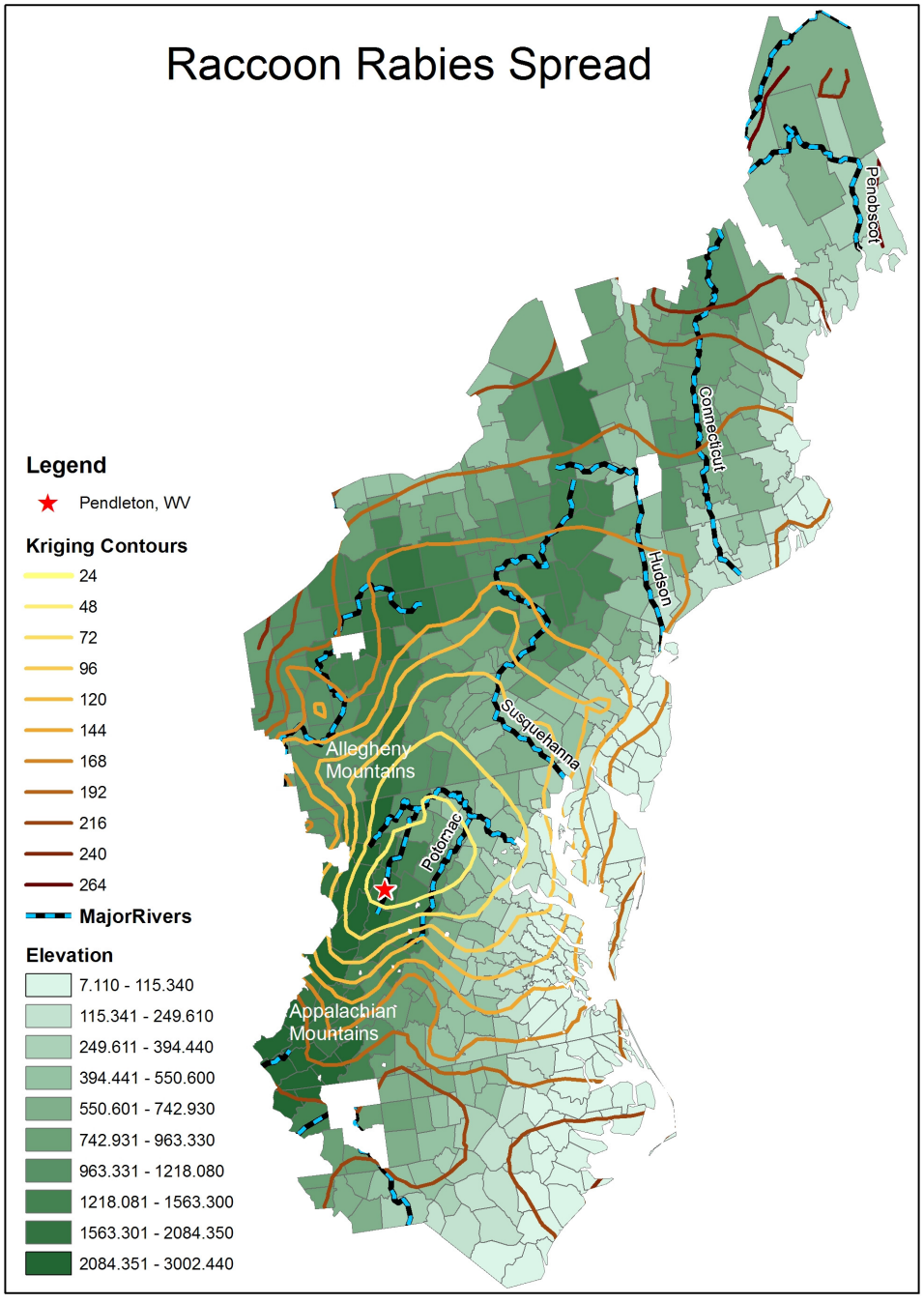
Barrier estimation: What do we want?

- Goal: Measure effect of landscape features, (e.g., mountains and rivers) on the speed of raccoon rabies diffusion.
- Elevation, river or road presence significantly related to raccoon rabies counts (Recuenco et al. 2007) and transmission time (Russell et al. 2004).
- Landscape features may serve as either barriers or gateways to the spread of infectious disease.
- Find and visualize barriers: Do they align with certain landscape features?

Data: What do we have?

- Time in months to first reported raccoon rabies case in 428 contiguous counties in the Eastern US (CDC).
  - 0 for origin county: Pendleton, WV.
- Mean elevation by county (USGS - Geographic Names Information System).
- Indicator for major river presence in county (ESRI data and a geographic information system (GIS)).
- Population density by county (US Census and ESRI).
- Distance between origin county and all counties.

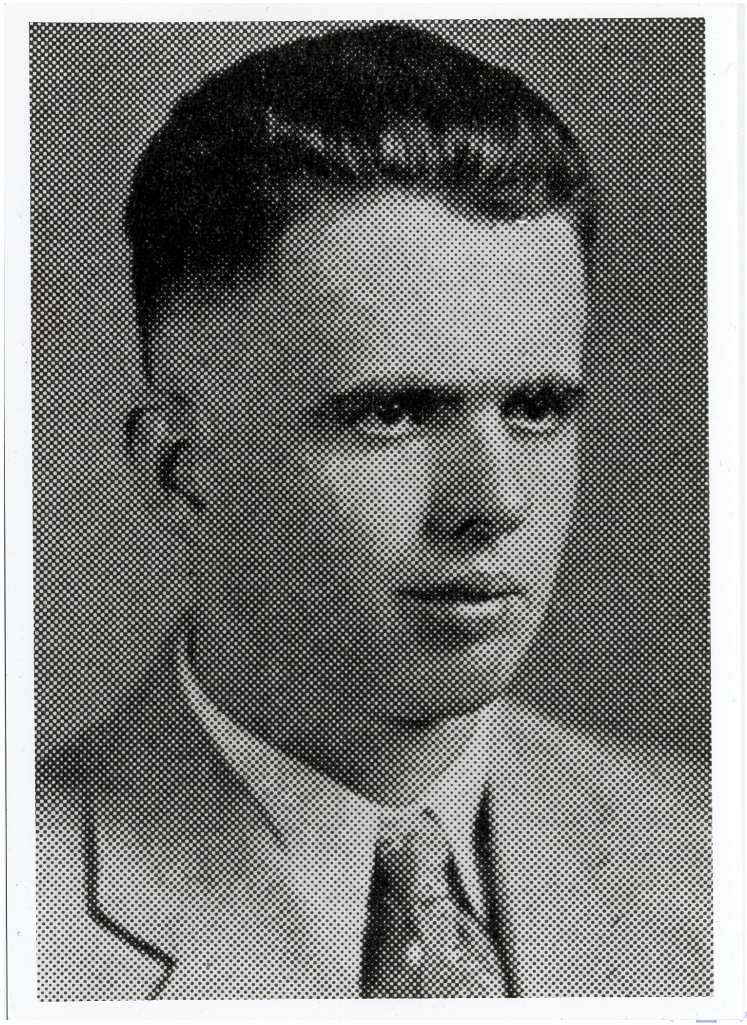
Data


Raccoon Rabies Spread
Legend
Pendleton, WV
Kriging Contours
24
48
72
96
120
144
168
192
216
240
264
MajorRivers
Elevation
7.110 - 115.340
115.341 - 249.610
249.611 - 394.440
394.441 - 550.600
550.601 - 742.930
742.931 - 963.330
963.331 - 1218.080
1218.081 - 1563.300
1563.301 - 2084.350
2084.351 - 3002.440
Penobscot
Connecticut
Hudson
Susquehanna
Allegheny Mountains
Potomac
Appalachian Mountains

Wombling

- Joint work with David Wheeler (Wheeler and Waller, *JABES*, 2008).
- Wombling: determine boundaries on a map by finding where local spread (change) is slower than elsewhere (Womble, 1951 *Science*).
- William H. Womble a bit of an elusive figure...

William H. Womble (?)



Google search: W.H. Womble Professor Robert Stencel

Which leads to...

Are you weady to womble?

- Consider a set of potential boundaries and decide if each is a "real" boundary or not.
- Many algorithmic approaches both deterministic and "fuzzy".
- Adopt a Bayesian hierarchical model for wombling (Lu and Carlin 2005).
- Bayesian approach provides a direct estimate of the probability that a line segment between two adjacent areas is a barrier (fuzzy boundary) in contrast to algorithmic versions based on thresholds, etc.

Bayesian areal wombling

- Model time to first reported raccoon rabies case Y_i:

$$Y_i|\mu_i, \tau \sim N(\mu_i, 1/\tau)$$

where

$$\mu_i = \alpha + \phi_i$$

is the expected value of time to first case per county.

- Spatial random effects follow a conditionally autoregressive (CAR) prior $\phi \sim CAR(\eta)$ with a mean random effect determined by its neighboring values.

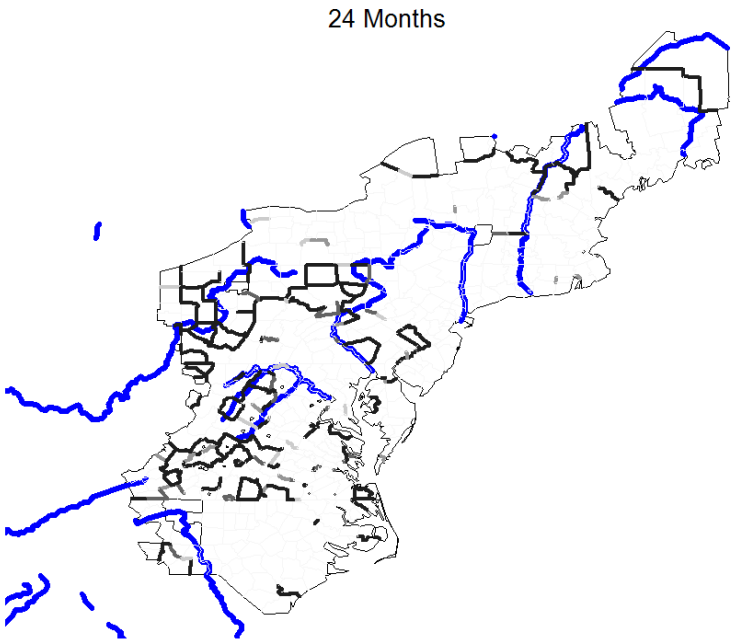
Bayesian areal wombling

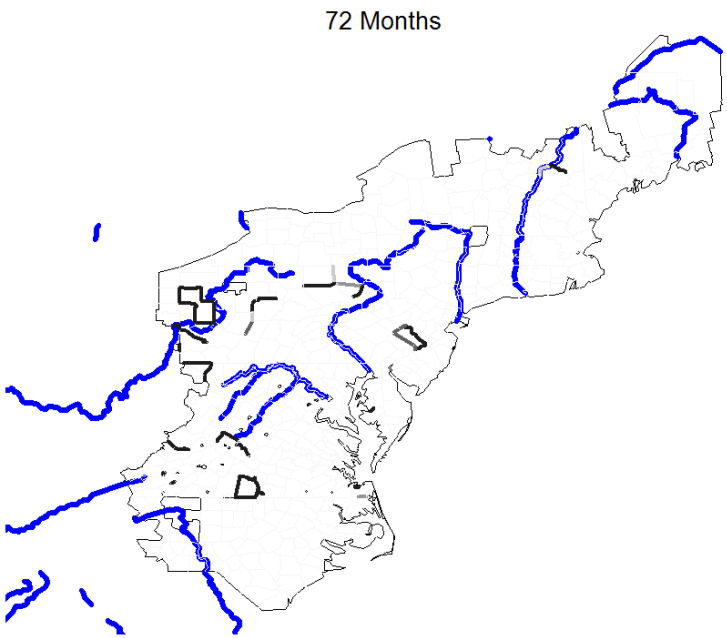
- *Boundary likelihood value* (BLV) assigned to each potential boundary (here, edge between two counties), based on difference in expected (modeled) time to first appearance.

$$\Delta_{ij} = |\mu_i - \mu_j|$$

- Use MCMC to draw sample from posterior $[\Delta_{ij}|\mathbf{y}]$ based on draws from posteriors $[\mu_i|\mathbf{y}]$ and $[\mu_j|\mathbf{y}]$.
- This assigns a posterior probability for each edge, then display edges with with $p(\Delta_{ij} > c|\mathbf{y})$ for some threshold probability c.

Wombling boundaries: $p(\Delta_{ij} > c|\mathbf{y})$





Linking to local covariates

- Bayesian areal wombling provides estimates of barriers but does not allow direct inference regarding the impact of particular landscape barriers on the evidence for barriers.
- We could expand our fixed effect α to $\mathbf{X}'\boldsymbol{\beta}$ to include local covariates (e.g., elevation, boundary based on a river).
- However, what it if the effect of elevation or presence of river varies from place to place?
- Russell et al. (2003, *PNAS*) suggest that river effect depends on direction of movement of the wave (perpendicular? Slower. Parallel? Faster.)

Spatially varying coefficents

- We consider a *spatially varying coefficent* model with CAR priors on the covariate effects $\boldsymbol{\beta}$, i.e.,

$$Y_i|\mu_i, \tau \sim N(\mu_i, 1/\tau)$$

where

$$\mu_i = \mathbf{X}_i'\boldsymbol{\beta}_i + \phi_i$$

- Spatial priors on elements of $\boldsymbol{\beta}_i$.
- More specifically, assign a multivariate CAR prior on the set of $\boldsymbol{\beta}$ (Banerjee et al. 2004).

MultiCAR details

- $\boldsymbol{\beta}_i = (\beta_{i1}, \beta_{i2}, \ldots, \beta_{ip})'$
- $\boldsymbol{\beta}_i | (\boldsymbol{\beta}_{(-i),1}, \boldsymbol{\beta}_{(-i),2}, \ldots, \boldsymbol{\beta}_{(-i),p}) \sim N(\bar{\boldsymbol{\beta}}_i, \Omega/m_i)$ where

$$\bar{\boldsymbol{\beta}}_i = (\bar{\beta}_{i1}, \bar{\beta}_{i2}, \ldots, \bar{\beta}_{ip})'$$

and

$$\bar{\beta}_{i1} = \sum_{k \in \kappa_i} \beta_{k1}/m_i$$

where κ_i = neighbor set for region i, and $|\kappa_i| = m_i$.

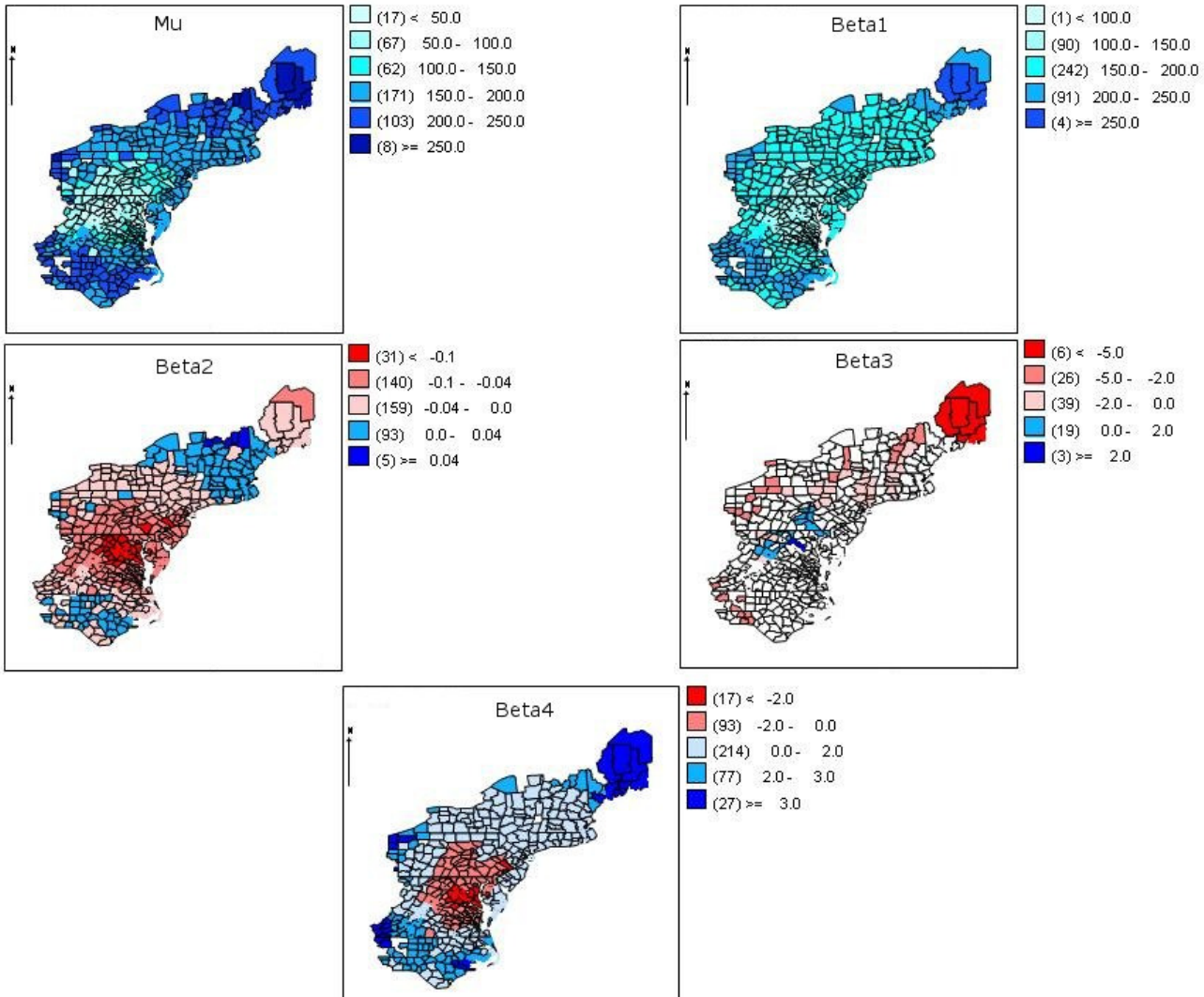
- $\Omega \sim$ Inverse-Wishart$(\nu, 0.02 \cdot I_{p \times p})$.

Including covariates

- Include effects of (mean) elevation, presence of a major river, and the natural log of the (human) population density.
- Best fitting (via DIC) model includes spatial variation in all three (and intercept).

$$E[Y_i] = \beta_{i1} + \beta_{i2}(\text{mean elev}) + \beta_{i3}(\text{river}) + \beta_{4i}(\log(\text{pop dens}))$$

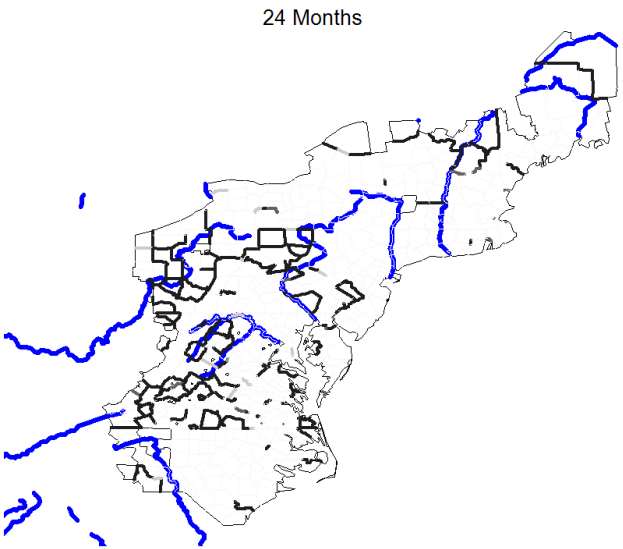
β_1: int, β_2: elev, β_3: river, β_4: log pop

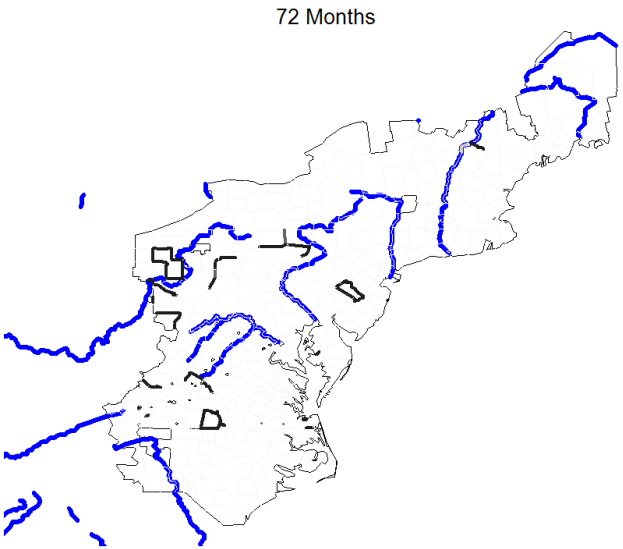


Findings/interpretations

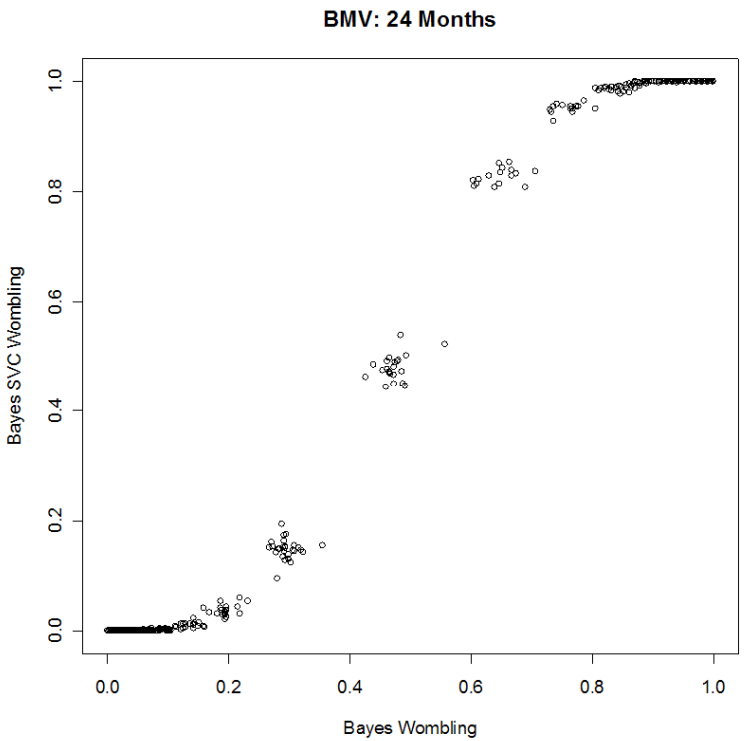
- Map of posterior mean (MU): shows the overall wave or spread.
- Random intercept reveals local adjustments.
- River effect indicates increases in time until first appearance across Potomac and Susquehanna Rivers, decreases time for Hudson River and others.
- Elevation is not difference in elevation so not directly informing on elevation gradients as barriers, simply elevation impact on time until appearance.

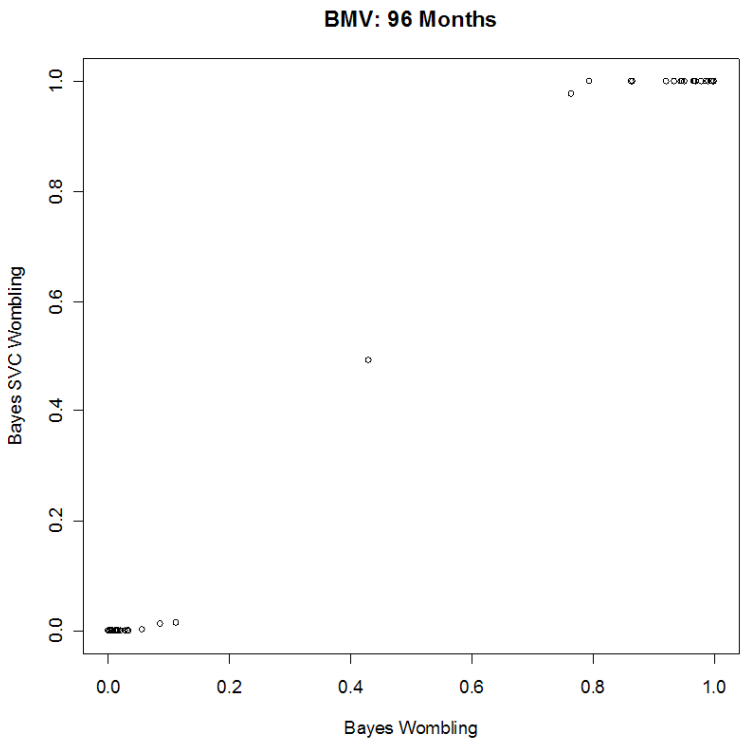
SVC wombled boundaries: $p(\Delta_{ij} > c|\mathbf{y})$





Including covariates $\rightarrow$ better wombling?





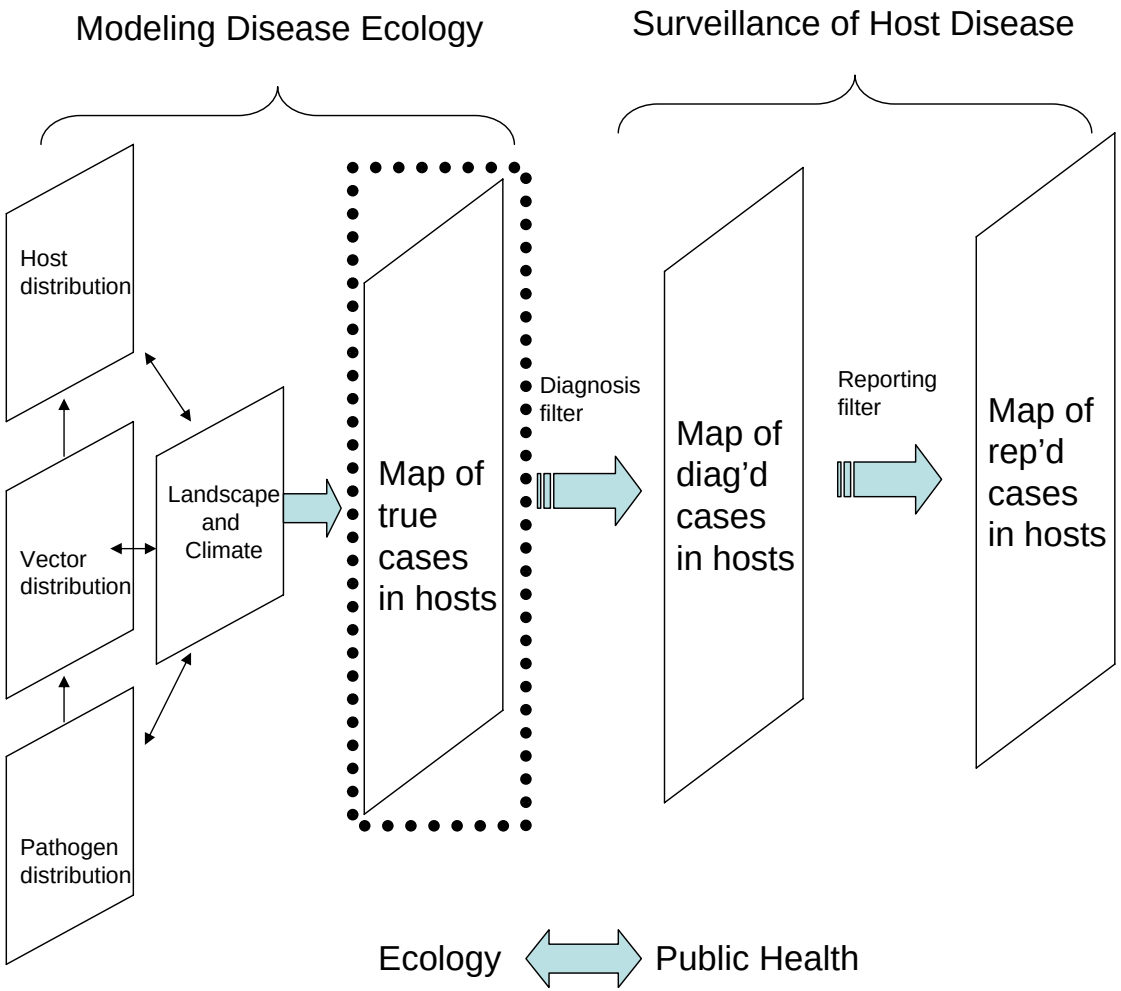
Overall Conclusions

- Much to be done to link mathematical models to statistical ideas.
- Disease ecology offers great setting for exploration.
- Models of transmission, interaction, observation.
- Mathematical models can inform statistics, statistics can inform models.
- Room to move past "ad-hockery".
- Linking landscape features in a more meaningful (inferential) and spatial way.
- Perfect opportunity for future dissertations and post-docs.

Next steps: Surveillance

- WHO: Surveillance is "on ongoing, systematic collection, analysis, and interpretation of health-related data essential to planning, implementation, and evaluation of public health practice".
- How do we detect an outbreak as it is happening?
- What data do we have?
- Can the data tell us where to target increased surveillance efforts as well as what is going on?
- Back to the "big picture".

The "big picture"



What is an epidemic?

- Above a baseline?
- Here: any occurrence of an infectious disease detected in a novel geographic location that poses a public health risk.
- Want to spot new cases in new places to plan prevention and response.
- Challenge: Surveillance of animal reservoirs.

Reality check

- Ill raccoon in my back yard last fall. Dead in morning. Thought: Animal control might want to know and test for rabies.
- Algorithm:
  - Call animal control: "Unless it bit you or your pets, we don't respond to dead animals."
  - Call sanitation: "We won't come into your yard but we can schedule curbside pick-up."
  - Call poison control (as suggested on CDC website): "Sounds rabid, don't touch it. Call animal control, they will want to know."
  - Repeat.
- Result: No testing, no data.

Gerardo-Giorda et al. 2013, *J R Soc Interface*

- Raccoon rabies in New York State.
- SIR (actually SEI) model + model of surveillance (function of reported cases).
- Goal: How to use reporting data (positive and negative occurrences) to identify geographic areas where surveillance levels are potentially insufficient to detect outbreaks.
- Two approaches: constant reporting rate and time-varying reporting rate.

Dynamics

- S (healthy) E (latent) I (infectious) model.

$$\begin{aligned} S' &= aA - bNS - \beta IS, \\ E' &= \beta IS - bNS - \sigma E, \\ I' &= \sigma E - \alpha I. \\ A &= S + E, \text{and} \\ N &= S + E + I \end{aligned}$$

- No reproduction by I, density dependent mortality, $\beta =$ contact rate.
- $\sigma E =$ rate of new infections (unknown source of Is).

Aggregate (reduce) to model of N and I

- Little information on E state (not observed).
- Aggregate to mode of N and I (maintains essential dynamics, assessed via simulation).

$$\begin{aligned} N' &= aN - (a+\alpha)I - bN(N-I) \\ I' &= -\alpha I + \sigma E \end{aligned}$$

- Replace σE by Φ source of new infections.
- Estimate Φ by $F(R_+, R_-)$, function of reported positive and negative cases.

What do we know about raccoon rabies dynamics?

- Birth rate, contact rate, latency, infectious period, death rate.
- $R_0 \in (1.2, 1.4)$
- R_0 (function of model parameters) suggests initial population drop of 16% to 28% (compatibility constraint).
- Next steps:
  - Propose $F(R_+, R_-)$, apply to reports from initial outbreak in New York.
  - Simulate outcomes for initial outbreak in New York.
  - Calibrate parameters in $F(R_+, R_-)$ to yield population drop within compatibility constraint.

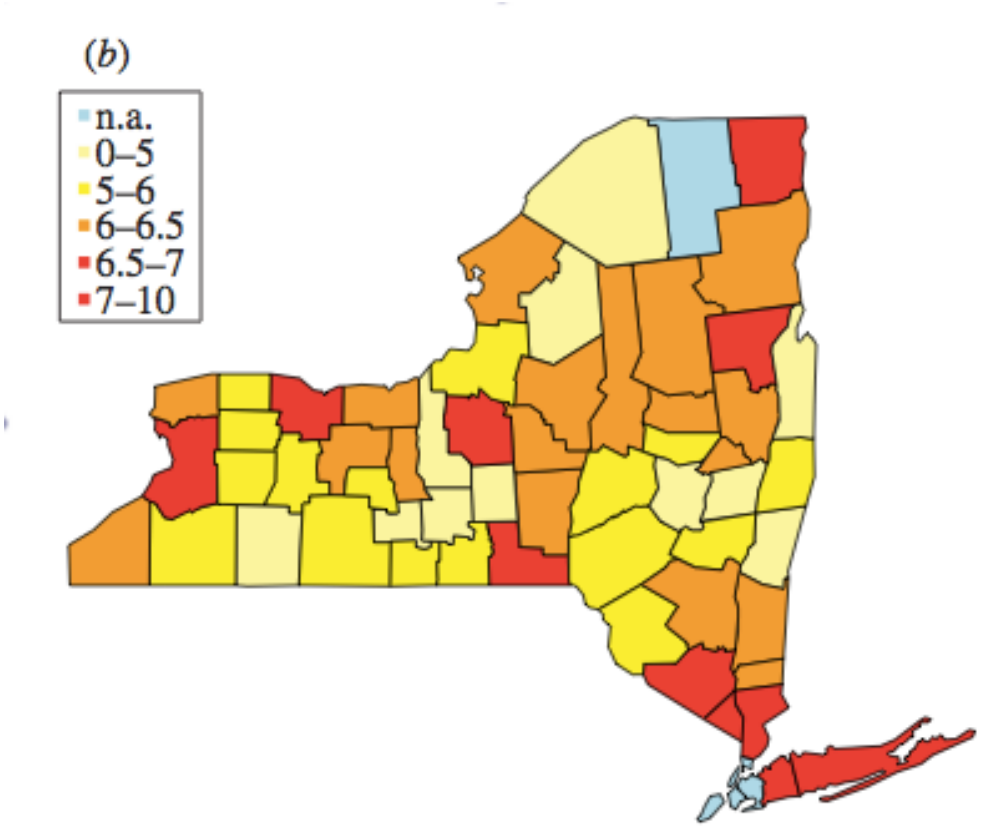
Modeling surveillance

- Gerardo-Giorda et al. (2013) consider two $F(R_+, R_-)$ surveillance functions.
- Constant surveillance (function of R_+ alone):

$$\begin{aligned} F_{\text{const}}(R_+) &= (1/\gamma)R_+ \\ \gamma &= (1 + K/h)^{-1} \end{aligned}$$

  $h =$ population density, $K \uparrow$ reporting rate per density $\downarrow$.
- Map local K values for each county.

K map (red = good surveillance)



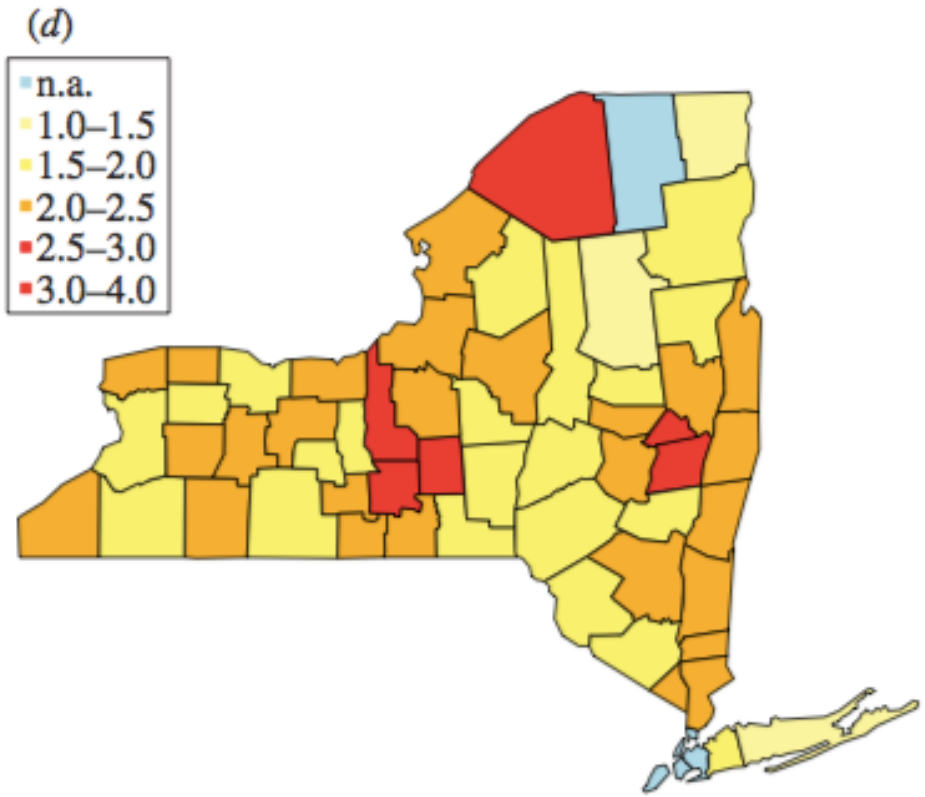
Modeling surveillance (dynamic)

- Dynamic surveillance:

$$F_{\text{dyn}}(R_+, R_-) = \left(\frac{N}{R_+ + R_-}\right)^{1/\theta} R_+$$

- Small θ: high level of surveillance in the area.
- Large θ: risk that an outbreak could go undetected in this area.
- Find θ consistent with local R_+ and R_- and meeting compatibility constraint.
- Map local θ values for each county.

θ map (red = good surveillance)



Overall Conclusions

- Much to be done to link mathematical models to statistical ideas.
- Disease ecology offers great setting for exploration.
- Models of transmission, interaction, observation.
- Mathematical models can inform statistics, statistics can inform models.
- Room to move past "ad-hockery".
- Linking landscape features in a more meaningful (inferential) and spatial way.

References

- Smith et al. (2002) Predicting the spatial dynamics of rabies epidemics on heterogeneous landscapes. *PNAS* **99**, 3668-3672.
- Waller et al. (2003) Monte Carlo assessments of fit for ecological simulation models. *Eco Mod* **164**, 49-63.
- Waller (2010) Bridging gaps between statistical and mathematical modeling in ecology. *Ecology* **91**, 3500-3502.
- Gerado-Giorda et al. (2013) Structuring targeted surveillance for monitoring disease emergence by mapping observational data onto ecological process. *J R Soc Interface* 10: 20130418.

References

- Wheeler and Waller (2008) Mountains, valleys, and rivers: The transmission of raccoon rabies over a heterogeneous landscape. *JABES* **13**, 388-406.