# Spatial statistical analysis of viruses and hosts in geographic and genetic space

Lance A. Waller, Leslie A. Real, Serena Reeder, Roman Biek
(Emory University)
with David Smith (Fogarty Institute, NIH)

Smögen 2006

Disease Ecology: What do we want to do?

Raccoon Rabies: What have we done so far?

- Example 1: Raccoon rabies in CT
- Cellular automata model
- Monte Carlo assessments of fit

Genetic structure: What we are doing now?

- Landscape genetics
- Example 2: FIV in cougars

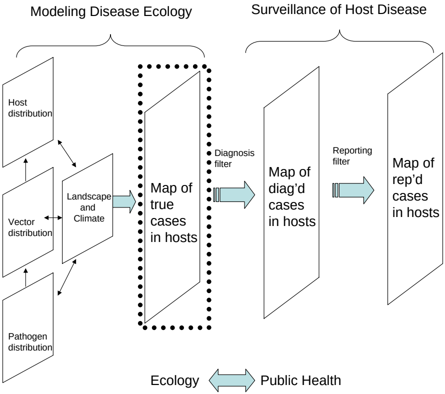
Conclusions

- Spatial landscape genetics

# Disease Ecology

- Interactions between virus, host, landscape.
- Landscape ecology (Manel et al. 2003), landscape genetics (host and virus) (Biek et al. 2006)
- People, animals, ecology, environment!
- Epidemiology, epizoology, environment interactions.
- Spatio-temporal data, mathematical models, genetic sequences, missing data, GIS!
- Fun, fun, fun!

# The “big picture”



# Raccoon rabies

# What is rabies?

- Virus in family of Lyssa virus.
- Reportable disease.
- Various strains associated with primary host (bat, dog, coyote, fox, skunk, and raccoon).
- Host cross-over, typically transmitted via bite/scratch.

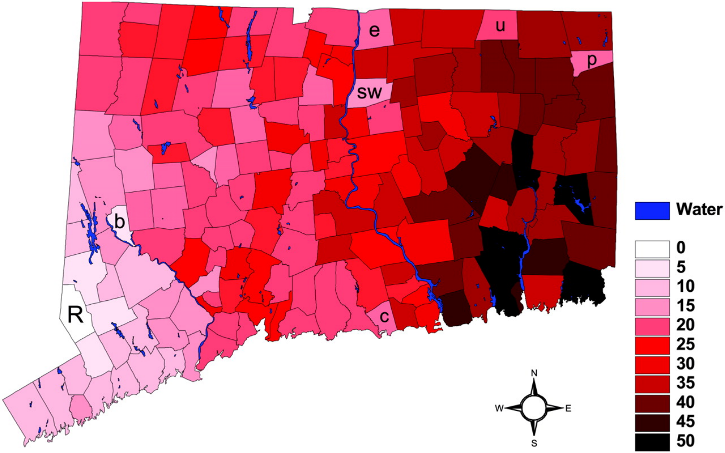
# Raccoon rabies

- 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.

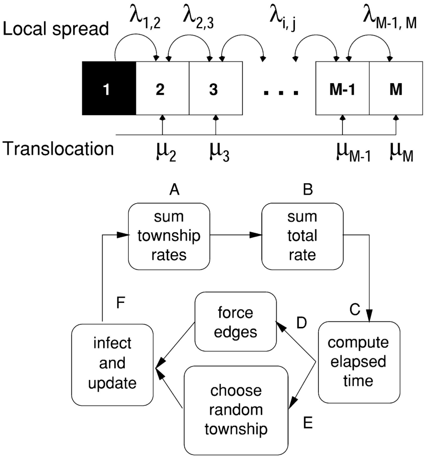
# Raccoon rabies in CT

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

# Data: Months to first appearance



# Cellular automata stochastic model



# 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 $\lambda_{ij}$) + translocation.

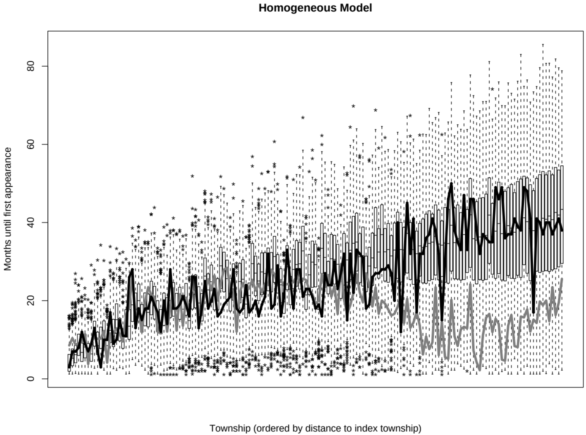
# What do we have?

- 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?

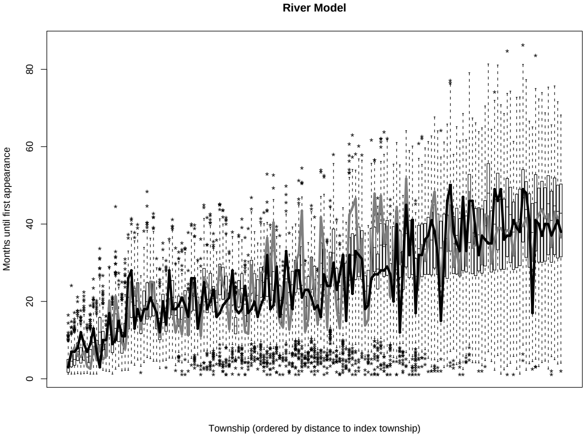
# Monte Carlo testing

- 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



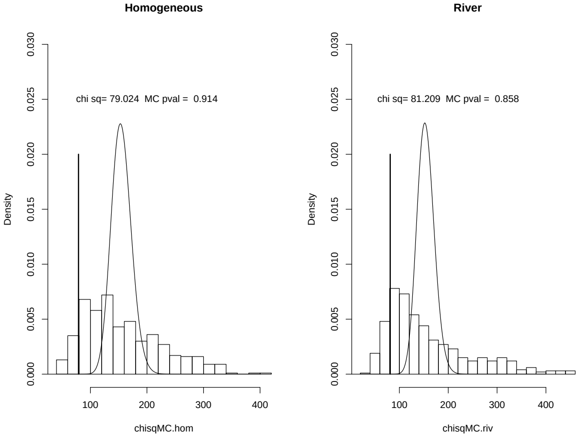
# Model realizations: River model



# 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 $V_i$.

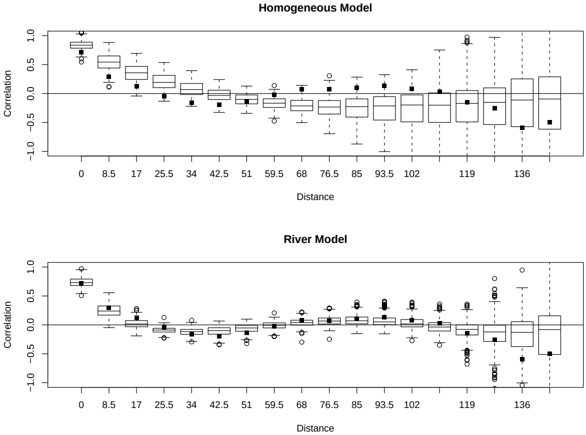
# Adjusted Pearson results



# 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.

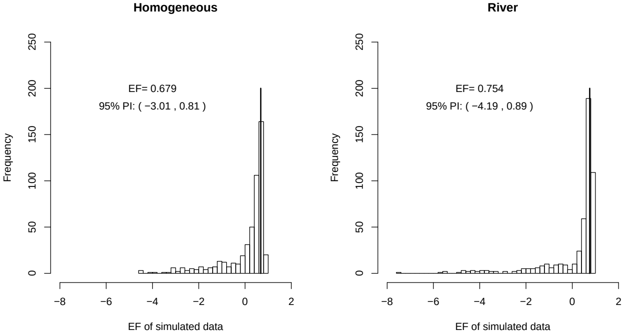
# Modelling efficiency

- EF(Homogeneous) = 67.9%, EF(River) = 75.9%
- Variability under $H_0$, cross-validate again!
- For $r$th 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 $r$th simulation, $-r$ excluding $r$th 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.
- Raccoon rabies moved into Ohio in last year.
- Why is it moving faster in Northeast than it did in Southeast?
- Susceptible hosts? Molecular evolution of virus?
- Tissue samples of hosts and viruses at CDC.
- Sequencing genes from hosts and viruses.

# Landscape genetics

- Two key steps:
  - Detection and location of genetic discontinuities.
  - Correlation (association) of discontinuities with landscape features
- Landscape ecology: Manel et al. (2003, *Trends Ecol Evol*)
- Spatial epidemiology: Ostfeld et al. (2005, *Trends Ecol Evol*)
- Conservation medicine: Aguirre et al. (2002, Oxford Univ. Press)

# Spatial landscape genetics

- Guillot et al. (2005, *Genetics*)
- Hierarchical Bayes spatial model to determine:
  - How many population subgroups (phylogenies).
  - Where subgroups are.
  - Posterior probability of belonging to subgroups.
- Endgame: Link to environmental features.

# Guillot's model

- Data:
    - Locations: $\mathbf{t} = (t_1, \ldots, t_n)$
    - Genotypes: $\mathbf{z} = (\mathbf{z}_1, \ldots, \mathbf{z}_n)$ where $\mathbf{z}_i$ = vector of $L$ allele pairs for each of $L$ loci.
- Assume $K$ subpopulations in subdomains $\Delta_1, \ldots, \Delta_K$ partitioning overall study area.
- Throw down a bunch of points (nuclei) across study area, define Voronoi tessellation.
- Classify nuclei in groups $1, \ldots, K$, *with spatial correlation*.
- Aggregate Voronoi cells to identify subpopulation areas.

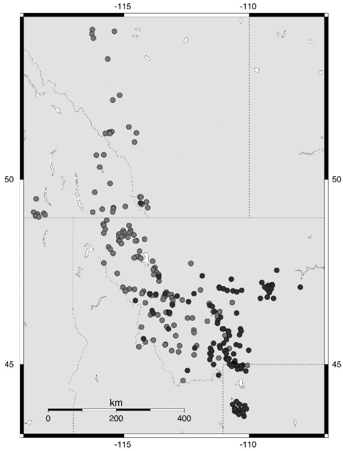
# Guillot's model (continued)

- Number of subpopulations.
- Number, location, and "color" of nuclei. (Marked Poisson Process).
- Spatial prior on "color".
- Ancestral allele frequencies.
- Present allele frequencies given ancestral frequencies.
- Likelihood from $\mathbf{z}|\mathbf{t}$.
- R library Geneland.

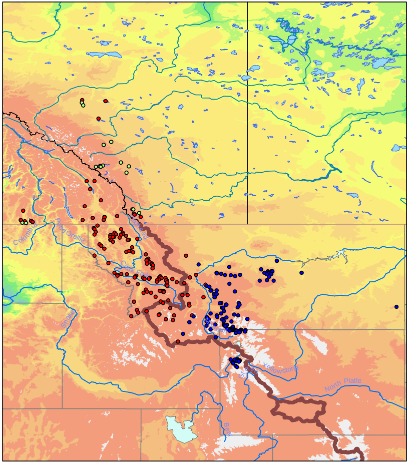
# FIV in cougars

- Sequencing ongoing for raccoons and virus, especially in Ohio samples.
- To illustrate methods, consider FIV data in cougars.
- Poss et al. (2002, *Conservation Medicine*), Biek et al. (2006, *Science*).
- Cougar samples from hunters in western U.S. and Canada.
- Biek et al. (2006) use `Structure` to categorize host samples into two subgroups (7 groups for virus).
- We apply Guillot's `R` library `Geneland` to same data.

# Non-spatial assignment (Structure)



# Population assignment, 3 populations



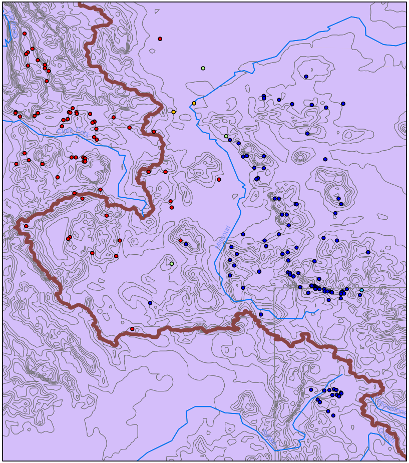
**Legend**

**3 population posterior**

- ● 1.00000
- ● 1.00001 - 2.00000
- ● 2.00001 - 3.00000

μ

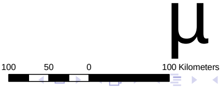
# Closer look with elevation



**Legend**

**3 population posterior values**

- ● 0.00000 - 0.20000
- ● 0.20001 - 0.40000
- ● 0.40001 - 0.60000
- ● 0.60001 - 0.80000
- ● 0.80001 - 1.00000



# Overall Conclusions

- Much to be done to link mathematical models to statistical ideas.
- Disease ecology offers a myriad of interesting statistical problems.
- Models of transmission, models of interaction, models of data collection.
- Mathematical models can inform statistics, statistics can inform models.
- Room to move past “ad-hockery”.

# Moving on...

- Sequencing virus and host for raccoon rabies in eastern US.
- Nagging questions: How to incorporate model selection into model fit.
- Guillot's spatial prior too strong?
- Incorporating geographic and genetic space in models?
- Linking landscape features in a more meaningful (inferential) way.
- Perfect opportunity for future dissertations and post-docs.

# References

- Lucey et al. (2002) Spatiotemporal analysis of epizootic raccoon rabies propagation in Connecticut, 1991-1995. *Vector Borne and Zoonotic Diseases* **2**, 77-86.
- Smith et al. (2002) Predicting the spatial dynamics of rabies epidemics on heterogeneous landscapes. *PNAS* **99**, 3668-3672.
- Russell et al. (2003) *A priori* predictino of disease invasion dynamics in a novel environment. *Proc. R. Soc. Lond. B* **271**, 21-25.
- Waller et al. (2003) Monte Carlo assessments of fit for ecological simulation models. *Ecological Modelling* **164**, 49-63.
- Real et al. (2005) Unifying the spatial population dynamics and molecular evolution of epidemic rabies virus. *PNAS* **102**,

# Guillot's model (continued)

- Likelihood: $[\mathbf{t}, \mathbf{z}|\boldsymbol{\theta}] = [\mathbf{z}|\mathbf{t}, \boldsymbol{\theta}] = \prod_{i=1}^{n} \prod_{\ell=1}^{L} [z_{i,\ell}|\boldsymbol{\theta}]$
- Parameters: $\boldsymbol{\theta} = (K, m, \mathbf{u}, \mathbf{c}, d, \mathbf{f}, \mathbf{f}_A, s)$
- $[z_{i,\ell} = (\alpha, \beta)|\boldsymbol{\theta}] = 2 f_{k\ell\alpha} f_{k\ell\beta}, (\alpha \neq \beta)$ or $f_{k\ell\alpha}^2 (\alpha = \beta)$.
- $K =$ number of subpopulations.
- $(m, \mathbf{u}) =$ number, location of nuclei. (Poisson Process).
- $\mathbf{c} =$ "color" (marks).
- $\mathbf{f} =$ present allele frequencies given ancestral frequencies.
- $\mathbf{f}_A =$ ancestral allele frequencies (Falush et al. (2003)).
- $d =$ genetic drift parameter (linearly related to $F_{ST}$).
- $t_i = s_i + \epsilon_i$ (location noise).

# Guillot's model (priors)

- Likelihood: $[\mathbf{t},\mathbf{z}|\boldsymbol{\theta}] = [\mathbf{z}|\mathbf{t},\boldsymbol{\theta}] = \prod_{i=1}^{n}\prod_{\ell=1}^{L}[z_{i,\ell}|\boldsymbol{\theta}]$
- Parameters: $\boldsymbol{\theta} = (K, m, \mathbf{u}, \mathbf{c}, d, \mathbf{f}, \mathbf{f}_A, s)$
- $K \sim \text{Unif}(K_{min}, K_{max})$
- $(m, \mathbf{u}) = \text{Poisson Process}(\lambda), \lambda \sim \text{Unif}(0, \lambda_{max})$
- $\mathbf{c} : Pr[c_{u_1} = c_{u_2}] \downarrow$ as $d_{1,2} \uparrow$
- $\mathbf{f} \sim \text{Dirichlet}\left(f_{A\ell 1}\left(\frac{1-d_k}{d_k}\right), \ldots, f_{A\ell J_\ell}\left(\frac{1-d_k}{d_k}\right)\right)$
- $\mathbf{f}_A \sim \text{Dirichlet}(1, \ldots, 1)$
- $d \sim \text{Beta}(2, 20)$
- $t_i = s_i + \epsilon_i$