Predicting the Spread of Invasive Species

Mevin B. Hooten

Department of Statistics
University of Missouri

March 15-16, 2006

Title Page [notes]

- The basic idea here involves a method for incorporating a scientifically meaningful deterministic model in a more general probabilistic framework for estimation and prediction while accounting for uncertainty at multiple levels.

Acknowledgements

- Christopher Wikle
- Robert Dorazio
- J. Andrew Royle

Acknowledgements [notes]

Various components of this work can be found in:

- Wikle, C.K., (2003). Hierarchical Bayesian models for predicting the spread of ecological processes. Ecology , 84, 1382-1394
- Royle, J.A., and C.K. Wikle, (2005). Efficient Statistical Mapping of Avian Count Data. Ecological and Environmental Statistics , 12, 225-243.
- Royle, J.A. and R. Dorazio, (2006). Hierarchical models of animal abundance and occurrence. In Review.
- Hooten, M.B., C.K. Wikle, R.M. Dorazio, and J.A. Royle (2006). Hierarchical matrix models for characterizing invasions. In Review.

Characteristics of Invasive Species

- Invasive: quickly spreads and becomes abundant.
- Can be naturally introduced or imported.
- Successful Invasions:

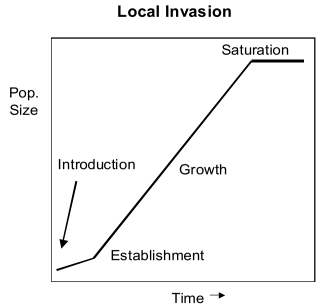
Invasive Species: phases of a successful invasion

1.) Introduction

2.) Establishment

3.) Dispersal / Growth

4.) Saturation

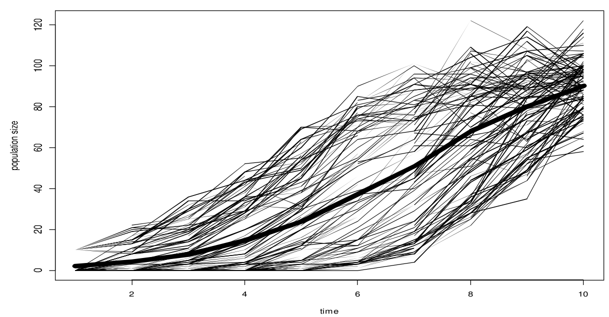


Characteristics of Invasive Species [notes]

- The idea with this figure is that it represents growth in the population size for an organism over time.
- The growth curve shown is very generalized, of course there are all manner of more complex forms of population growth. The basic idea is that after introduction, population size grows rapidly until resources become limiting.
- As the population size approaches the carrying capacity (i.e., saturation) other forms of dynamical behavior could ensue (e.g., stability, periodicity, chaos).

Characteristics of Invasive Species (cont'd)

- Multiple growth curves for various locations:



Characteristics of Invasive Species (cont'd) [notes]

- Studying total population size is useful, but we want to make inference about the population size at numerous locations over time.
- These plots with multiple growth curves representing the growth in population size at each location of interest are informative, but it's difficult to see the interaction between locations (that is, the movement of organisms between locations).
- A sequence of maps is helpful here, such as those in the results section of this presentation.

Characteristics of Invasive Species (cont'd)

Impacts of exotic species:

- Pests can attack humans and livestock (e.g., Killer Bee).
- Cause or transmit disease (e.g., West Nile Virus and Avian Flu).
- Disrupt native food webs (e.g., Peacock Bass and the exotic zooplankton, *Daphnia lumholtzi*.

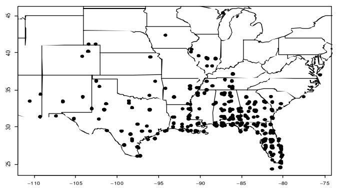
Characteristics of Invasive Species (cont'd) [notes]

- Obviously these are some of the more prominent examples in the media.
- The point of this slide is to provide some justification for wanting to study these processes in more detail in order to better understand them and thus make better management decisions.

History

Eurasian Collared-Dove (ECD):

- Invaded Europe in 1930's.
- Introduced to Florida mid-1980's.
- Count data collected through N. Amer. Breeding Bird Survey, documenting invasion.
- Imperfect detection.



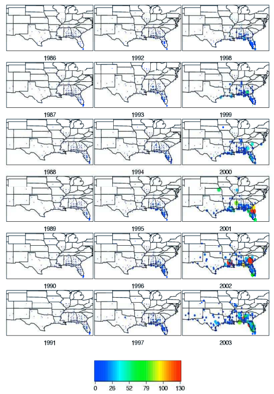
History [notes]

There are several good references for this species:

- Hengeveld, R. (1993) What to do about the North American invasion by the Collared-Dove? Journal of Field Ornithology 64:477-489.
- Romagosa, C., and R. Labisky. (2000) Establishment and dispersal of the Eurasian Collared-Dove in Florida. Journal of Field Ornithology 71:159-166.

Characteristics

- No replicate data through BBS.
- Separate dataset used to estimate detection probability.
- Data for years: 1986-2003



Characteristics [notes]

- We desire to estimate the “true” population size, that is, the real number of birds in a given location at a given time.
- Our data represents only the “observed” number of birds. In this type of data collection we could miss a few birds even though they were there.
- Treating the probability of missing a bird as a parameter in our model, we would need more than one observation to estimate that parameter as well as the “true” population size.
- In the case where we only have the one space-time observation (as with the BBS data) we must estimate the probability of detection separately.

Impacts

ECD biological threats (Romagosa and Labisky 2000):

- Competition for resources with native avifauna.
- Transmission of disease.

“ECD will probably colonize all of North America within a few decades.”

- Just how probable is it?

Impacts [notes]

- One of the most important game animals in this country is the mourning dove.
- It would not be good if the ECD causes problems for the mourning dove.
- The goal here is to associate (determine) some level of probability with the ongoing invasion at various locations and times.

Overview

Environmental/Ecological Sciences

- Common classes of scientifically meaningful behavior.
- Often with non-linear and spatially varying dynamics.
- A hierarchical modeling framework can be employed to accommodate such behavior.

Overview [notes]

- The question is: How do we make use of all of this scientific knowledge while characterizing complex dynamics in a rigorous statistical model?
- A hierarchical framework allows us to characterize very complex systems by breaking the problem down into simpler and more intuitive components.
- It also allows us to incorporate scientific knowledge (e.g., functional model forms and parameter spaces) into the model.

Spatio-temporal Processes

Examples:

- Diffusion: Spreading process; similar to “dispersal” in ecology.
- Growth: Process increasing in intensity; a simple form of population growth in ecology.
- Density Dependent Growth: Process increasing in intensity non-linearly; a more realistic form of population growth.

Spatio-temporal Processes [notes]

- These processes, when formulated mathematically, can be written as Partial Differential Equations (in continous space and time) or difference equations (in discrete space and time).
- Difference equations can be derived as approximations to partial differential equations.
- There are many other deterministic models capable of exhibiting dynamical behavior (as discussed in the methods).

Examples

- Diffusion: spreading

diffusion movie...

Examples [notes]

- The movies that are shown on the current and following slides are only example simulations of these types of processes to give you some idea of what they might look like.
- Alone they appear quite simple, but in combination they can represent more realistic invasive behavior.
- You will see from the movie in the results section, that the behavior is a combination of both diffusion and non-linear growth.

Examples

- Growth: increasing in intensity

growth movie...

Examples [notes]

- This is a linear form of growth where the population size increases at a constant rate in a given area.
- The differences are subtle between this movie and the next but very important for exhibiting realistic behavior.
- In these two growth movies, the process is not spreading out (diffusing), but rather growing independently at each location.

Examples

- Density Dependent Growth: non-linear increase in intensity

non-linear growth movie...

Examples [notes]

- Again, the differences in types of growth are subtle here.
- In this movie, the growth rate slows down as a function of intensity in the process (population size) and after reaching a carrying capacity it ceases to grow further.

Definitions and Notation

- "Population": loosely, the true number (N) of organisms at a place and time (also let $\lambda =$ mean population $= E(N)$).
- "Count": the observed number (n) of organisms at a place and time where $n \leq N$.
- Bolded variables denote vectors and matrices (e.g., $\mathbf{x} = [x_1, \ldots, x_m]'$).
- "$|$" = given; as in conditional probability.
- Square bracket notation refers to a probability distribution (e.g., $[x|\beta] = \text{Prob}(x|\beta) = f_x(\beta)$).
- "$\sim$" = is distributed as ... (e.g., $x|\beta \sim [x|\beta]$).
- "$\propto$" = is proportional to ... (e.g., $[\beta|x] \propto [x|\beta][\beta]$).

Definitions and Notation [notes]

- Another way of saying: $x|\beta \sim f_x(\beta)$ is that x is a sample from the probability distribution f given the parameter β.
- These kind of expressions: $[\beta|x] \propto [x|\beta][\beta]$ will be used later to illustrate the hierarchical nature of the models.
- In this general case we may be interested in estimating the parameter β given the data (x). To do so, we need only know the distribution of the data given the parameter $[x|\beta]$ (often called the likelihood) and any prior knowledge about the distribution of the parameter $[\beta]$.

Hierarchical Specification

- We want to characterize real environmental processes in the presence of data.
- We have *a priori* scientific knowledge about the process evolution.
- If we assume the data are a realization from such a process, which is latent and evolves dynamically, then a hierarchical probability model is useful:

$$[\text{data}|\text{process}][\text{process}]$$

- Our knowledge of the process contains uncertainty, so we must learn about the process parameters as well:

$$[\text{data}|\text{process}][\text{process}|\text{parameters}][\text{parameters}]$$

Hierarchical Specification [notes]

- This is just a very general representation of a hierarchical model.
- In the specific application of modeling invasive species, each of these components will have specific probability distributions associated with them.

Hierarchical Components

- [data|process]: Specified in the usual statistical sense. Accounts for possible observational uncertainty and/or measurement error.
- [process|parameters]: Specified with discretized scientific model (for computation).
- [parameters]: Specified according to *a priori* scientific knowledge or lack thereof.
- [process, parameters|data]: We want to learn about the true process given the data (via Bayes).

Hierarchical Components [notes]

- These models are very data and process specific.
- Each different scientific problem (and dataset) will require a different model specification. That is, different probability distributions and process models.
- The specification given in the following slides is relevant to the spatio-temporal ECD model only, though the general framework holds for many similar problems.

Data Model

$$n_{i,t}|N_{i,t},\boldsymbol{\theta} \sim \text{Beta-Binomial}(N_{i,t},\boldsymbol{\theta}),$$

where, $i = 1,\ldots,m$, $t = 1,\ldots,T$, and

- $n_{i,t}$: sample count at location i and time t.
- $N_{i,t}$: Population size at location i and time t.
- $\boldsymbol{\theta}$: probability of detection parameters (assumed to be known).

Data Model [notes]

- Here $n_{i,t}$ for all locations i and times t, are the data.
- $N_{i,t}$ is the "true" population size and the thing we want to estimate.
- $\boldsymbol{\theta} = \{\alpha,\beta\}$ are the parameters corresponding to the probability of detection. In this case we have estimated them using a separate model (see Royle and Dorazio 2006).
- The Beta-Binomial model is an "over-dispersed" binomial model where the $n_{i,t}$ is a random integer from zero to $N_{i,t}$. This data model allows us to account for the uncertainty in the probability of detection through the parameters α and β.

Process Model

$$N_{i,t}|\lambda_{i,t} \sim \text{Poisson}(\lambda_{i,t}) \quad \text{,for} \quad t = 1,\ldots,T,$$

where, $\boldsymbol{\lambda}_t$ is the mean and variance of the population size at time t and is modeled via a latent dynamic process:

$$\begin{aligned}\boldsymbol{\lambda}_t &= \mathbf{H}\boldsymbol{\lambda}_{t-1} \quad ,\\ &= \mathbf{MG}\boldsymbol{\lambda}_{t-1} \quad \text{,for} \quad t = 1,\ldots,T,\end{aligned}$$

- $\mathbf{G} = \mathbf{G}(a,b,\boldsymbol{\lambda})$ is the growth matrix.
- $\mathbf{M} = \mathbf{M}(\delta)$ is the movement (dispersal) matrix.

Process Model [notes]

- The Poisson model is a common model for "relative abundance" and allows for a substantial amount of variability in the "true" population size.
- The dynamical process model is a version of a "matrix model", as it is known in ecology.
- Conventionally, matrix models are used to study demographics in population growth. Here we modify it to study dispersal in population growth.
- Matrix Models are thoroughly discussed in:

Caswell, H. (2001) Matrix Population Models. Sinauer Associates, Inc., Sunderland, MA.

Process Model (growth and dispersal)

Growth Model:

$$\mathbf{G}(a, b, \lambda_{i,t}) = \exp\{b\left(1 - \frac{\lambda_{i,t}}{a}\right)\},$$

Dispersal Model:

$$\mathbf{M}(\boldsymbol{\delta}) = [(M_{i,j})]_{m \times m},$$

$$M_{i,j} \propto \exp\{-\frac{d_{i,j}^2}{\delta_j}\}.$$

Process Model (growth and dispersal) [notes]

- This slide illustrates how these matrices are parameterized.
- The growth model is a Ricker growth equation; one form of a non-linear or density dependent growth equation. For further information see:

Kot, M. (2003) Elements of Mathematical Ecology. Cambridge University Press, Cambridge, UK.

- The dispersal model is a Gaussian (i.e., Normal) dispersal kernel. That is, it is a function that weights the dispersal of organisms from one location to another based on the distance between them ($d_{i,j}$). This is the component of the model that allows for **spatial** effects.
- The parameters controlling the rate of dispersal (δ_j) vary by location, allowing for organisms to move more easily in different areas. This also allows for a heterogeneous environment.

Parameter Model

Probability distributions for the parameters can be specified based on prior scientific knowledge (or lack thereof):

$$a \sim Gamma(\alpha_a, \beta_a)$$
$$b \sim Normal(\mu_b, \sigma_b^2)$$
$$log(\boldsymbol{\delta}) \sim Normal(\boldsymbol{\mu}_\delta, \boldsymbol{\Sigma}_\delta)$$
$$log(\boldsymbol{\lambda}_1) \sim Normal(\boldsymbol{\mu}_\lambda, \boldsymbol{\Sigma}_\lambda)$$

Parameter Model [notes]

- These prior probability distributions represent *a priori* scientific knowledge about the parameters.
- They would be different for different problems of interest and should be specified vaguely (i.e., with large varibility) if little is known about the parameters.

Implementation

- We want to estimate the model parameters as well as characterize and predict the population size over time given the data.
- Bayes theorem and various sampling methods can be utilized to do this:

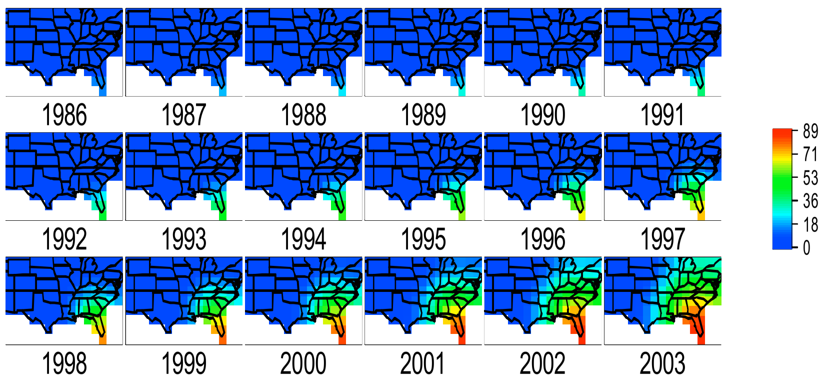
$$[\mathbf{N}, \boldsymbol{\lambda}, a, b, \delta | \mathbf{n}] \propto [\mathbf{n} | \mathbf{N}, \theta][\mathbf{N} | \boldsymbol{\lambda}, a, b, \delta][\boldsymbol{\lambda}, a, b, \delta]$$

Implementation [notes]

- So we're interested in learning about $\mathbf{N}$, $\boldsymbol{\lambda}$, a, b, and δ given the data ($\mathbf{n}$). That is, the we want to estimate the stuff on the left hand side of the equation (what's known as the posterior distribution).
- What allows us to do so is by writing the complicated joint model as the series of simpler probability models on the right hand side of the equation.
- We can't find these exactly, but we can get pretty close by taking numerous samples from these distributions and calculating our statistics based on the samples. One such method for sampling is called Markov Chain Monte Carlo (MCMC).

Estimation (maps)

Posterior means for **N**

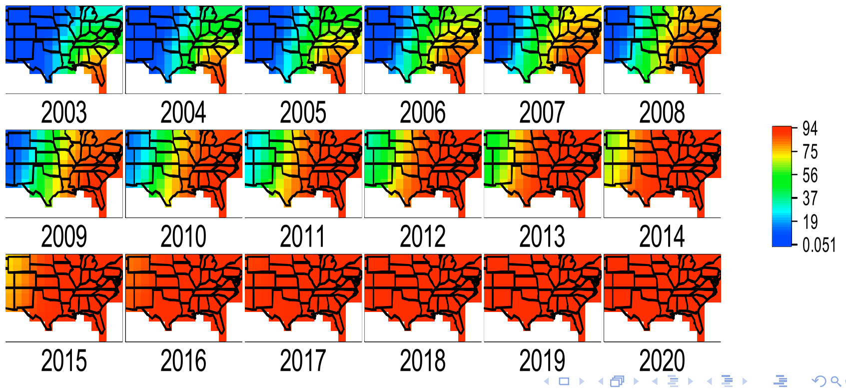


Estimation (maps) [notes]

- These maps show the mean of the posterior distribution for the "true" population size.
- Notice how the invasion starts in South Florida and spreads out as well as grows in intensity over time.
- Also notice how the population in South Florida ceases to grow after about 2001. It has reached its carrying capacity there.

Prediction (maps)

Posterior prediction means for **N**



Prediction (maps) [notes]

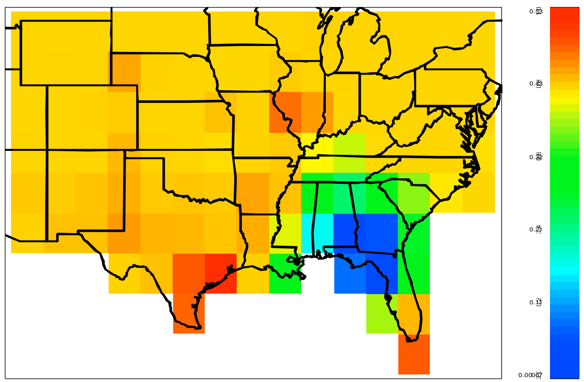
- Recall that speculation in the year 2000 suggested that the ECD would invade most of North America within a few decades.
- These maps provide a forecast by displaying the mean posterior predictions for future years.
- This model provides some statistical justification for such speculations.

Estimation and Prediction (movie)

Estimation and Prediction (movie) [notes]

- This movie just combines the previous two slides into a sequence of images.

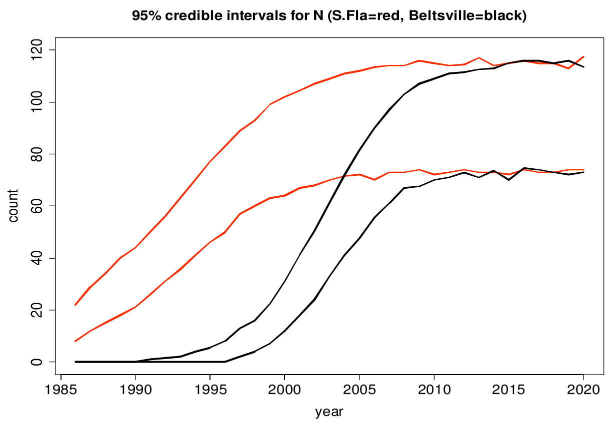
Dispersal (δ)



Dispersal (δ) [notes]

- This image represents the mean of the posterior distribution for the dispersal parameters (δ).
- Notice how there is a pocket of low dispersal in Northern Florida.
- This area of low dispersal is only marginally significant (based on the variability of the posterior distribution; not shown), but the effect of which can be seen in the previous movie.
- Essentially, ECDs are dispersing slower in that area than in some others.

Estimation and Prediction (growth curves)

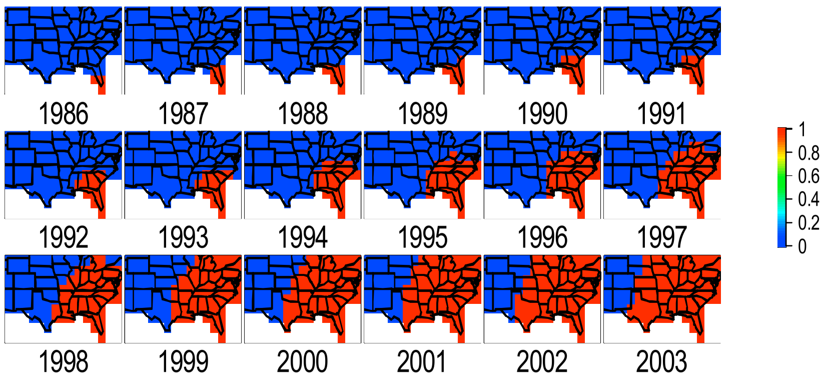


Estimation and Prediction (growth curves) [notes]

- These curve envelopes allow us to compare the population growth for two locations simultaneously.
- For each location, the lower line represents the 2.5^{th} percentile of the posterior distribution for population size and the upper represents the 97.5^{th} percentile.

Presence/Absence

$P(\mathbf{N} > 0) \geq 0.95$



Presence/Absence [notes]

- An advantage of obtaining the output from Bayesian models is that we can easily calculate probabilities.
- These maps show in red all areas where the probability of presence is at least 0.95.
- These can be viewed as probabilistic range maps.
- They are not to be confused with political election maps.

Summary

In this setting, the matrix model specification accommodates:

- *a priori* scientific knowledge.
- Flexible dynamical behavior.
- Multiple sources of uncertainty.
- Long-range predictions (assuming no population collapse).

In addition to:

- More intuitive parameterization than other models (e.g., partial differential equation based models).
- More accessible to ecologists and managers.

Summary [notes]

- In addition, there are many extensions to this model that can be (and were) implemented.
- These include things like: letting other parameters vary spatially and the comparison of models with different specifications.