sim.basic: A first step in testing the suitability of trawl data for detecting spatial dyanmics of diversity

Ryan Batt

2015-06-02

1 Introduction

As water temperatures change, species may shift the size and location of their geographical ranges, bearing consequences for the food webs and economies linked to those species. However, species don't always respond similarly to shifting temperatures (different thermal tolerances, e.g.), which means that changing temperature may remix the composition and diversity of ecological communities.

The biological, spatial, and temporal scale of community diversity shifting in response to climate is massive. A functional definition of a community may consist of 100's or 1000's of species, each of which may be shifting its range at a scale of decades and 100's kilometers. As a result, we need statistical methods for estimating biodiversity that don't rely on replicate samples and that make efficient use of available data. Enter the superstars: on the data side the trawl data set has amazing spatiotemporal and taxonomic extent and resolution; on the statistical side multispecies occurrence models (MSOM) are hierarchical state space models that are designed to estimate species richness and don't require consistent or extensive "replication". Although they're superstars, even these data and models have their limitations and pitfalls.

Can we estimate the dyanmics of species richness from trawl data using an MSOM? It's a hard question to answer, because we can never know the "truth" for sure —- and it's not useful to say that the answer we get from the model+data isn't "truth", because it never will be (always a little bit of error). The trawl data set is generated by two distinct processes: Nature's data generating process (NDGP), and the process by which humans observe the result of NDGP. So we ask: to what extent is the accuracy of estimates from an MSOM dependent on characteristics of NDGP, and in particular, the way in which we observe the result of NDGP? The strategy for answering this question is to simulate fake data where we approximate Nature but gain knowledge of "truth", "observe" the results of the true process, then try to recover the true species richness from these simulated data.

The first step in this procedure is to simulate a basic version of NDGP. It is this simulation that I will be discussing in this document.

2 Approach

To understand the simulation, I need to provide context concerning the data (bottom trawl survey), what information we want from those data (richness), and how how we intend to derive that information from the data (MSOM). After introducing those basics, I'll describe my approach to simulating a basic process that involves a 7x5 spatial grid monitored over 12 years, how I filled that grid with temperatures that changed over time and space, and how I put 200 species on that grid and let them move around.

2.1 The Basic Context: Model, Data, Richness

Bottom Trawl Survey Data

Malin and Jim, you guys are familiar with the data, so I wont say much here. What might be important to point out is that I've "snapped" all of the hauls to a grid by redefining a sampling stratum as a 1° grid. Other than that, you know the other most important/ prominent features of the data.

Basics of the MSOM

There are four very important things to know about the MSOM:

1. It is designed to estimate an "index" of species richness from binary data (presence/ absence data)
2. The model has two levels: a *process* level (NDGP), and an *observation* level
3. The model assumes a hierarchical structure for parameters
4. The model can make use of covariates in either the process or observation levels

Item #**1** is important because it means we shouldn't dwell on how the model output compares to the exact number of species we might think are out there. Item #**2** distinguishes between what we think drives species presence vs. what we think affects our ability to detect it; it also emphasizes why the simulation needs a NDGP and an observation process. Item #**3** implies that species and strata are expected to be different but not independent; e.g., for a given species, the probability (in logit space) it'll present is drawn from the same normal distribution for each stratum. Item #**4** means that the model can harness temperature (e.g.) to predict the probability a species is present, which is helpful for improving estimates of richness (if temperature is a good predictor) and for combatting the constraint of #**3**.

Species Richness

Richness = the number of species. Particularly, the number of species in a place, time, and/or ecological community. The MSOM estimates richness as the asymptote of the species accumulation curve (you see more species as you sample more, but there are diminishing returns). Thus, richness seems to be a simple term, but it can require careful interpretation. I consider the demersal community to consist of the species that could feasibly be caught by any of the gear ever used in a region's bottom trawl survey, and richness to be the enumeration of these species in a 1° grid in a given year.

2.2 What to simulate: Species, temperature, space, and time

My simulation includes many species moving on a 2D spatial grid over a couple years, with the presence/absence and the movment of the species governed by temperature. I considered these elements of species, temperature, space, and time to be essential to testing the MSOM. The nature of the overarching question is such that we need to simulate many **species** (to have non-trivial richness), and that these species must exist on some sort of spatial grid. The **spatial** requirement is because of the form of the MSOM – we need different strata to satisfy the hierarchical nature of the model. I previously mentioned that the MSOM can incorporate statistical covariates (measured predictor variables) to improve estimates of species presence/ absence, and thus richness. Because environmental covariates like **temperature** vary over space (and time), they can also help to make meaningful distinctions between the probability of a species being present in different strata. Finally, I wanted **time** to be a component of the simulation because we are interested in temporal dynamics of richness; right now the MSOM is analyzing each year separately, but I think I could better leverage the data by linking the years.

It is worth mentioning that by incorporating these elements of species, temperature, space, and time into the model, I have set up a structure that can be flexible and extensible —- e.g., future simulations could incorporate additional environmental variables. If I had excluded an element like time from the simulation it would have require substantial effort to add it in later on. One component that the model *does not* have is species interactions; I thought this to be too complicated for the first cut. Similarly, the model does not include any physiological mechanisms, density dependence, resource constrains, etc. This is **very** basic.

2.3 How to simulate: Random walks, dispersal, and graph theory

In this section I will describe the statistics/ framework I used in a semi-generic form. It is in this section where I explain the framework of the analysis and the statistics and how they relate to achieving the goal of the basic simulation. The next section (See Section 3) will reiterate many of these points in terms of what I actually simulated (the code, choices for model parameters, and simulated output), but I will not explain the statistics as much.

The goal of the simulation model is to create 200 species that each have their own thermal preferences, and to allow these species to move between grid cells (7x5 grid) between years. I wanted the ability for a species to occupy by a grid cell to be restricted by some sort of thermal tolerance, and I wanted to be able to describe this tolerance using observations (more on that later).

To explain how I achieve such a simulation, I will start with a simple time series model of a single variable (i.e., 1 species), expand that time series model to temporal dynamics in 2-dimensional space, add in environmental constraint, and generalize the model to many species.

AR(*p*) & random walks: A simple time series model

$$X_t = \beta_1 X_{t-1} + \epsilon_t \tag{1}$$

where X is our "state" variable of interest (e.g., biomass), t a subscript denoting the time step ($t = 1, 2, ...T$ where T is the length of the time series), β_1 is known as the autoregressive coefficient ("auto" because it describes how the variable is related to itself), and ϵ_t are normally distributed errors ($\epsilon \sim \mathcal{N}(0, \sigma^2)$). In R, we might program an autoregressive process like so:

```r
beta <- 0.8 # AR(1) coefficient
Tn <- 100 # length of time series
X <- rep(NA, Tn) # create empty vector to store results
X[1] <- 0 # set starting value
sigma2 <- 1 # variance of epsilon
for(i in 2:Tn){
        X[i] <- beta*X[i-1] + rnorm(1, 0, sqrt(sigma2))
}
par(mar=c(2,2,0.1,0.1), mgp=c(1.15, 0.15, 0), tcl=-0.15, ps=8, cex=1)
plot(X, xlab="time", ylab="X", type="o")
abline(h=0, lty="dashed")
```

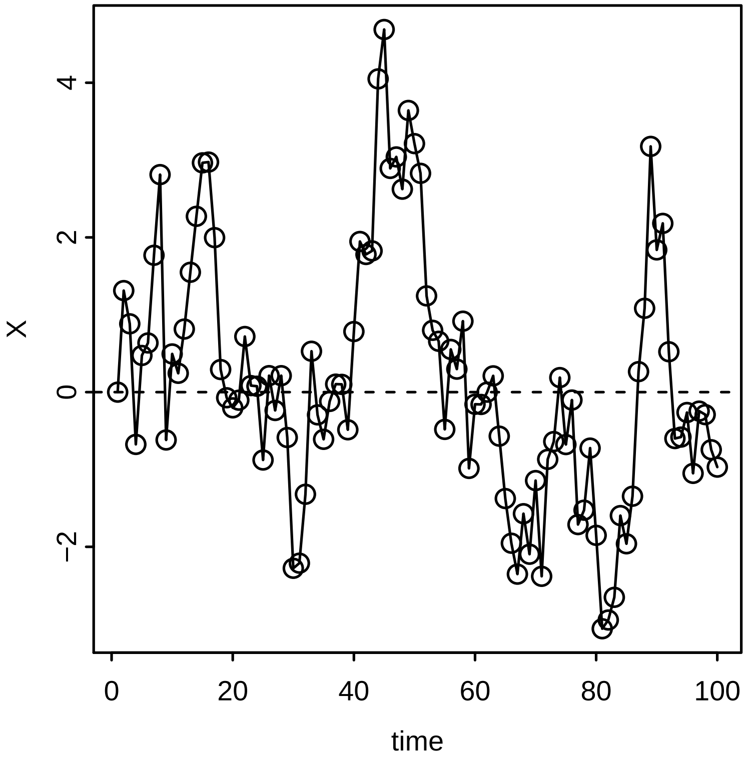

Figure 1: Example AR(1) process with autocorrelation of 0.8

Autoregressive processes are generic time series models that can approximate a wide range of complex, and even nonlinear, processes. When the model is AR(1) and the coefficient is 1, then the model is simply a so-called random walk. Conveniently, Eq. 1 is very similar in form to the Gompertz growth equation when the Gompertz is on a log scale and in discrete time ($X_t = a + bX_{t-1}$), and the intrinsic rate of increase (a) is set to 1 and there is no noise term (ϵ). The key here is that autoregressive models, which will make an appearence in the final simulation methods, can be useful approximations to nonlinear models of ecological populations.

The previous example of an AR(1) time series is known as a univariate time series model because X is only of one variable. However, the AR model can be generalized to multivariate autoregressive (MAR) models when the response variable includes multiple, potentially interacting, state variables. For example S species may be interacting over time:

$$\boldsymbol{X}_t = \boldsymbol{X}_{t-1}\boldsymbol{B} + \boldsymbol{E}_t \tag{2}$$

where $\boldsymbol{X}_t$ is a 1xS vector of S populations at time t, $\boldsymbol{B}$ is an SxS square matrix of parameters, and $\boldsymbol{E}$ is the variance-covariance matrix. Although the basic simulation does not (yet) include interacting species, it includes interacting grid cells, and in general this notation can be used to describe any set of interacting vertices of a graph (whether those vertices be species, places, or both). Of particular importance are the elements of $\boldsymbol{B}$, whose diagonal elements describe density dependence (i.e., the coefficients relating $\boldsymbol{X}_{t,s}$ to $\boldsymbol{X}_{t-1,s}$), and the other elements describe the interactions between pairs of species. E.g., if $S = 2$, then $\boldsymbol{B} = \begin{pmatrix} b_{1,1} & b_{1,2} \\ b_{2,1} & b_{2,2} \end{pmatrix}$. Thus, $b_{1,1}$ tells us how the past value of species 1 ($X_{T-k,1}$) is related to itself over time, and $b_{1,2}$ tells us how a past value of species 2 gives rise to the current value of species 1. Similarly, the second row of $\boldsymbol{B}$ tells us what gives rise to the current value of species 2. These statements could be rephrased to say that the elements of $\boldsymbol{B}$ tell us how biomass is transferred between and species and/ or time steps.

The transition matrix: using graph theory to model dispersal

It turns out to be a fairly smally leap to go from a MAR model to a model that is discrete and explicit in space and time. As was mentioned earlier, the simulation will operate on a 2D grid. However, it is not wholly constructive to think of a 2D grid as a matrix upon which we can perform algebraic operations. Rather, it is better to think of the grid as instead being a “graph” —- graphs consist of vertices (nodes) and edges (links connecting nodes). The vertices would be things like species, who might have a certain biomass —- this is not unlike the values of the species in $\boldsymbol{X}$ from the MAR example. The edges connecting those vertices describe the transfer of mass between the species, like $\boldsymbol{B}$ from the MAR model.

For example, consider Figure 2, which as 4 vertices (A through D), and a series of edges. All vertices are connected to at least 1 other vertex, and those connections are directed (i.e., not all relationships are symmetrical).

Imposing thermal constrains on occupancy

The many species generalization

3 Implementation and Results

Content.

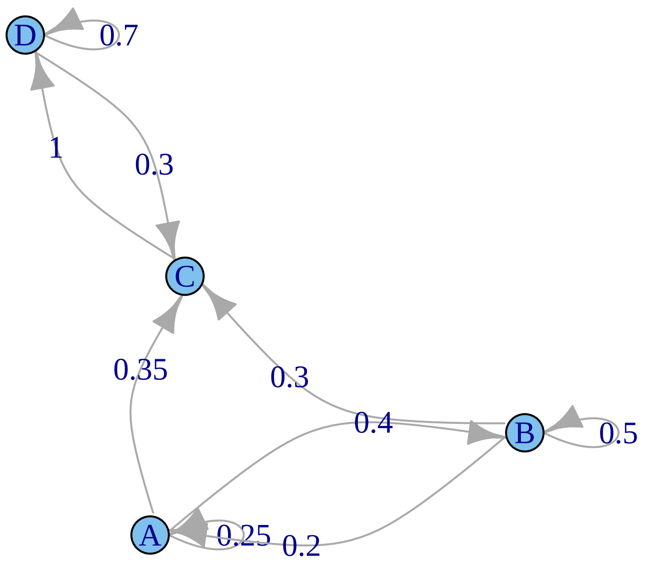

Figure 2: A Markov chain as an example graph.

4 Conclusion

5 Appendix