# Predicting biodiversity dynamics in response to environmental change

## Can we do it? A report from assess.sim.basic.R

Ryan Batt

2015-08-23

**Abstract**

I use simulated data to evaluate the ability of a multispecies occupancy model (MSOM) to estimate species richness. An emphasis is placed on the potential to use MSOMs to estimate richness over time and space, although the current version of the simulation and of the MSOM is somewhat simplistic in this regard. The principal finding is that the MSOM can be sensitive to changes in detectability. Not emphasized explicitly in this document, but discovered through experimenting in different versions, is that these sensitivies can depend quite heavily on sample size. Therefore, it may be prudent to analyze all years of trawl data at once, rather than separately. Finally, some of the methods for evaluating the performance of the MSOM on simulated data may provide a useful guide for how to gauge the reliability of the results when the MSOM is applied to empirical data. This document represents a first step towards what I think we want to do in a paper, and also serves to get everyone on the same page. I look forward to further progress and conversation.

# Contents

# Introduction

## Overview

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 heavy replication 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 occupancy 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 dynamics 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, but we can get an idea of how reliable our analysis is by simulating fake data, for which we know true values because we created them. 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 Simulation

The goal of this simulation was to use a very basic process to generate presences and absences of species in space and time. In this version of the simulation, there is no explicit connection between years (they are independent). There is a modest spatial connection, because in the simulation an environmental variable determines habitat suitability. I think of this environmental variable as temperature, and I filled a grid with temperatures that ranged from the coldest at the top of the grid (north) and the warmest at the bottom (south) and added random variation among columns in the same row (among longitudes at the same latitude).

One level of the simulation mimics NDGP. In this level, NDGP is best characterized by $\psi$, which is the product of a temperature and species' response curves. I.e., temperatures were used to determine the suitability of each grid cell to each simulated species. This suitability is known as $\psi$ throughout this document.

A second level of the simulation mimics human observation of NDGP — what we do when we collect data. This process was simulated by assigning each species has a unique probability of being observed or "detected" (this variable is $p$). The observation process gets several attempts at observing a given species in a given grid cell; think of this as subdividing each site into subsites, and when you visit each subsite you have probability $p$ of observing a particular species (each species has its own $p$). Depending on the settings used in the analysis that this document summarizes, the maximum number of subsites can vary, as can the number of subsites per site (OK, fine; the maximum number of subsites in this version is 4, the number of subsites per site varied between 1 and 4, and overall 50% of total possible subsites were sampled).

As previously mentioned, the simulation included "time". In this basic version, not much changes between the "years" for the true process (temperature doesn't change, nor do the response curves), but the mean of $p$ does change. In a given year, the entire community has an overall mean probability of being detected, and each species randomly deviates from that mean.

The simulation also has replicates. To understand the replicates, it needs to be clear that even when a parameter in the simulation does not change, the outcome can change. The replicates hold the realization of the simulated NDGP constant, and draw new realizations of the observation process. I.e., both $\psi$ and $p$ are constant among replicates, and the binary *outcome* of $\psi$ is also held constant, but the outcome for $p$ can change. Furthermore, although each replicate has same values of $p$ (both the mean $p$ and each species' individualized random draw from that distribution), each replicate switches which year is associated with which $p$'s. In this way we can observe each outcome of Nature's data generating process under a series of settings for the human observation process.

## Multispecies Occupancy Models (MSOMs)

Multispecies occupancy models are Bayesian statespace hierarchical models. They distinguish between truth and observation of the truth, and many parameters share a common "parent" distribution. They are very flexible models, and can be adapted to include new types of processes. The MSOM being used here is a relatively simple version of these models. It predicts the probability of each species existing in a grid cell from a logistic regression equation that uses a second-order polynomial of the environmental variable as a covariate. The parameters in this level of the model are hierarchical, with species having their own paramter values, but these individual parameters are not wholly independent in the sense that they share a common parent distribution, which sort of acts to both limit how different they can be and to inform one another. The model also has an observation level, which only has a hierarchical intercept (just a mean) as a predictor variable.

The MSOM makes guesses of the true state of the system (whether a species is actually present or not). It then makes guesses at how the observation of that true state might turn out, which is effectively a prediction of what our data will be. The Bayesian model fitting process then uses this comparison of the observed data to the estimate of the observation to tweak the parameters in the MSOM. This process is repeated until the choice of paramters boils down to what is essentially the posterior distribution of the estimated parameters.

Right now the MSOM model is fit separately to each year and each replicate. So the model never gets to see multiple years or multiple replicates at the same time. Furthermore, when referring to a parameter value fitted in the MSOM, it is implied that it can be subscripted with time or replicate (because all years and replicates are fit independently).

The parameters in the logistic regression that predicts the value of $\psi$ vary among species, although *psi* itself varies among species and space, because the regression parameters (subscripted by species) are multiplied by the environmental variable (subscripted by space). More or less, it can be said that, for a given species, $\psi$ varies among space because of the environmental variable, and in a given location it varies among species because of the regression parameters.

---

# Conventions and Settings

In this section I outline the subscripting and notation used in the MSOM analysis and for the simulation. I also outline various settings (number of species simulated, replicates, etc.). Most of the numbers you see (and some of the text) is dynamically generated based on the code that produced the statistics and figures. Therefore, you can refer back to these sections to see what settings may have changed since the last version of this document.

***Note:*** *I've often found myself having to get creative with subscripts and superscripts. I've tried to be clear an consistent, but small inconsistencies likely exist, so don't be confused by them. For example, if you see* $max(Z)$ *and* $Z_{max}$ *in two different sections, they are probably referring to the same thing. If you see something confusing, let me know (preferably by creating an issue on GitHub), and I'll fix it.*

## Dimension Conventions

### Summary

1. Site ($j = 1, 2, \ldots, j_{max} = 9 \times 9 = 81$)
    - Sites are unique combinations of latitude and longitude
    - The spatial arrangement of sites is *not* arbitrary, although importance depends on settings (see `dynamic` below)
    - The environmental variable $X$ varies among sites (and years, below)
2. Sub-sites ($k = 1, 2, ...$)
    - Sub-sites are only relevant to the "observation" process
    - Each site has the same number of possible sub-sites, but the number of sub-sites observed can vary
    - In this simulation, $k_{max} = 4$, $k_{min}^{observed} = 1$, and $k_{max}^{observed} = 4$
    - Substrata are are primarily useful for determining $p$, the detection probability
3. Species ($i = 1, 2, \ldots i_{max} = R = 30$)
    - Does not include "augmented" species
    - For this MSOM analysis, the species array was padded with 10 0's
4. Time ($t = 1, 2, \ldots 4$)
    - Time is primarily used to vary the parameters controlling the "true" process
    - When those parameters don't change, time provides independent*realizations of the same "true" process
        - **Note: only when* ***`dynamic=FALSE`*** *in* ***`sim.spp.proc`***
5. Replicates ($r = 4$)
    - Replicates are *simulated* repeated human observations of the same *realization* of the "true" process at $\text{Time}_t$
    - Replicates are used to vary the parameters that control the "observation" process
    - When those parameters don't change, each replicate provides an independent*realization of the same "observation" process

### In Code

The MSOM analyzes each year$_t$–replicate$_r$ combination independently. Parameters subscripted by these dimensions are derived from separate analyses.

In my code, I've tried to be consistent in my use of these indices to describe arrays, matrices, and rasters. Rows are dimension 1, columns dimension 2, etc. The precedent of subscripted dimensions follows the numeric ordering of the list above. E.g., in the matrix $X_{i,t}$ each row will refer to a different species, and each column a different year (note that site$_j$ is skipped, so species$_i$ is "promoted" to dimension 1, the row.). By default, R fills matrices and arrays by column, whereas the `raster` package fills them by row. In most cases where an R object needs to split sites into the lat/ lot components, I make use of the `raster` package. Therefore, the numbering of the sites proceeds row-wise, where each site is numbered according to the order in which it is filled, as in this $2 \times 3$ matrix: $J = \begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{pmatrix}$

Note that even though this matrix is numbered row-wise, it is still indexed as $J_{row,column}$, such that $J_{1,2} = 2$. As mentioned previously, this information is primarily important for understanding the code involved with this project, and in most cases it is not crucial to be explicitly aware of the spatial arrangement of sites.

---

## Settings

### Simulation Settings

I created a class called `"spp"`, which has methods for `print()`. The `Dimensions` are the number of sites, the number of species, then the number of years.
Also `print`ed are some richness summary statistics. `All cells` refers to the collective richness over all $j$ taken together. The meaning of `One cell` differs slightly between the true and observed printouts: in the true printout the richness is of a particular site ($j$), and in the observed printout it is of a particular sub-site ($k$).

```
## Dimensions: 81, 30, 4
## grid.h = 9
## grid.w = 9
## grid.t = 4
##
## Number Species Possible (ns):
## 30
## Total Species Richness:
## 30
## Total Observed Species Richness:
## 30
##
## Annual Species Richness:
##           Min. 1st Qu. Median  Mean 3rd Qu. Max.
## All cells   29      29     29 29.25   29.25   30
## One cell     5      10     14 13.63   17.00   24
##
##
## Observed Annual Species Richness:
##           Min. 1st Qu. Median   Mean 3rd Qu. Max.
## All cells   27   27.75   28.5 28.250      29   29
## One cell     0    0.00    0.0  4.008       7   23
```

In the MSOM, detectability ($p_i$) is determined in the form of a logistic regression, which currently only has an intercept ($v_0$) as a predictor (so just a mean). That intercept varies among species (i.e., $v_{0,i}$), and that variation is generated by drawing each individual species's intercept ($v_{0,i}$) from a parent distribution: $v_{0,i} \sim \mathcal{N}(\mu_{v_0}, \sigma^2_{v_0})$. See section about $p$ for more info.

| year | mu.v0 | sigma.v0 |
|---|---|---|
| 1 | -2 | 2 |
| 2 | 0 | 2 |
| 3 | 2 | 2 |
| 4 | 4 | 2 |

---

**Settings for JAGS & MSOM**

| nChains | nIter | n0s | nSamples |
|---|---|---|---|
| 3 | 50000 | 10 | 500 |

In the table above, `nChains`, `nIter`, and `nSamples` are all variables that are strictly pertinent to the Bayesian analysis carried out in JAGS. The `n0s` value refers to the the degree of "data augmentation". In this process, you add extra species to the data set, and say that they were never observed. For our purposes, this is employed for purely technical reasons, although it can be used to extra further inferences about species richness.

The posterior samples from JAGS consist of 500 samples (above). However, to save memory and hard drive space, I have often only saved measures of central tendency for each of these. In this assessment, I have performed all calculations on the `centralT`=**median** of the posterior samples.

---

# Species Richness

## Definition of species richness

Species richness is the number of different species, or more generically, unique taxa. The point is moot in the simulation study, and in the empirical trawl data it refers to species.

Estimates of richness ($R$) can be made spatially or temporally explicit (or neither, or both). In the following figures, different levels of aggregation are performed – for most figures $R$ is split by year (this is true for all figures but the Boxplot Figure). The Time Series of Richness Figure emphasizes temporal dynamics and keeps replicates separated, but aggregates over space (the $j$ sites). The `Nsite` Scatter Figure doesn't aggregate over space or time, but it does aggregate over "replicate" observations; importantly, while the figure does present any spatial aggregation, it does not retain the spatial relationship (you can't tell which sites are next to others). The final two figures of the section (Heatmap of of Richness Figure) are similar to the previous figure, except that spatial relationship among points is retained via a heatmap representation.

None of these estimates of richness include the 10 species that were part of the "data augmented"/ "adding 0's" process. Richness values can either be true (true simulated NDGP; $R^{true}$), observed (true simulated human observation of NDGP; $R^{obs}$), or MSOM estimates of one of those two ($\hat{R}^{true}$ or $\hat{R}^{obs}$).

---

## Regional Richness

These estimates of species richness only distinguish between replicates and years. They do not contain any site-specific information.

### Richness Boxplots

With the boxplots we're mostly looking to see if the estimates of richness vary with the mean probability of detection, $p$. In the empirical data, we know that taxonomic identification changed over time (it improved; generally, more species were ID'd in later years). We also suspect that gear might change, which affects the probability of observing a species. The "Average Detection Probability" category in the boxplots is the cross-species average of $p$ (which with large sample size approach the hyperparameter $p_\mu$).

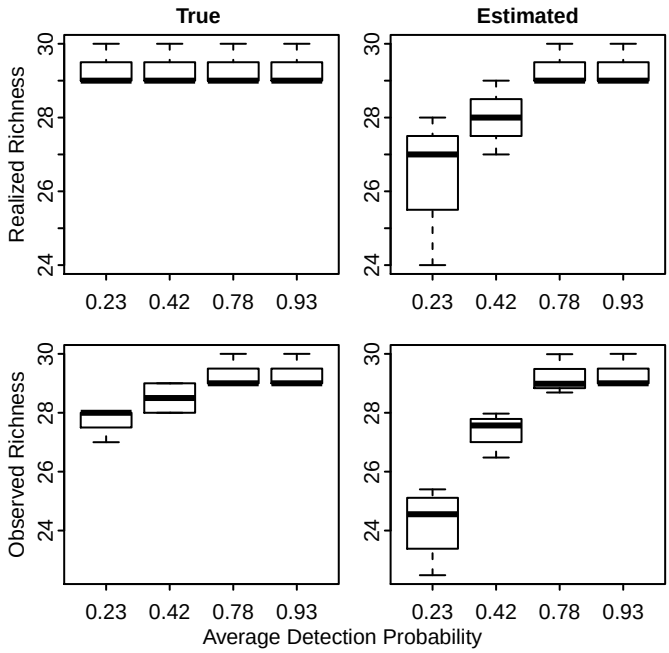


Figure 1: Boxplots of species richness. Numeric groupings indicate the average value of $p$ across species during a given year–replicate combination. The panels in the left column are the true simulated values, and the panels on the right are the corresponding MSOM estimates. The top row indicates the latent realized species richness or MSOM estimates of the richness. The bottom row's panels are the simulated observed values of richness (the response variable in the MSOM) and the MSOM estimates of the observed values.

## Richness Time Series

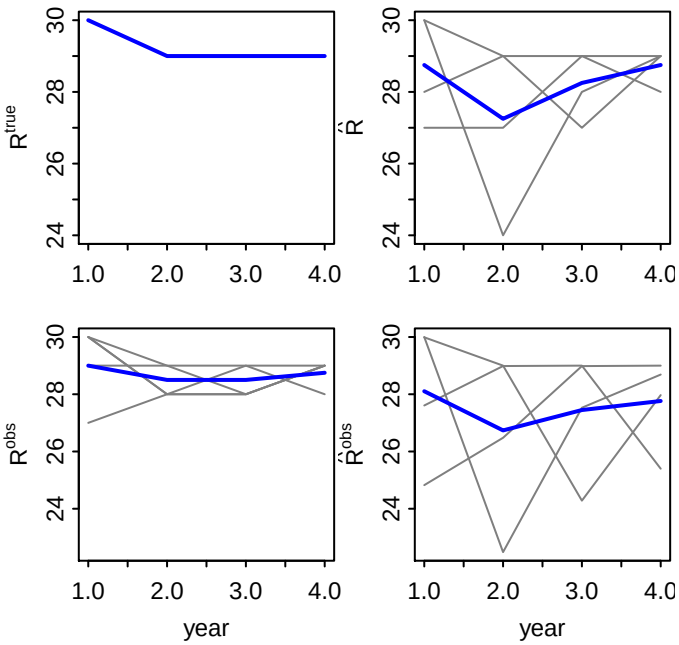


Figure 2: Time series of richness. Gray lines are individual replicates. Blue lines are averages. Note that detection probabilities ($p_{i,t,r}$, see simulation settings above, as well as definition of $p$ below) change over time, and their temporal ordering differs among replicates.

1. $R^{true}$ is just true richness (among all sites) in each year

2. $\hat{R}$ is from $\sum_{i=1}^{R^{true}} max(\hat{Z}_{i,\nabla j \in J})$;

    - i.e., basically same as for $R^{true}$, except from $\hat{Z}$ instead of $Z$
    - only difference is that $\hat{Z}$ is contains extra (fake, always absent) species compared to $Z$ (see next)
    - $\hat{R}$ doesn't include the augmented species introduced to the MSOM occurrence matrix ($Y$)

3. $R^{obs}$ are the values of $R^{true}$ after they pass through the observation process

- Note that "true" and "observed" and "estimated" can be confusing here; the last term can prefix either of the first 2 terms. This notation needs some work still.

4. $\hat{R}^{obs}$ is the MSOM estimate of what was observed

- this the estimate that is compared to the data in the fitting process

- true values are latent (unobserved)

- thus, we need this extra step to connect our estimates of what's actually going on to our data

- although, it should be noted that our data don't arrive in terms of "richness", but in presences/ absences In these "time series" plots (so short!), the replciates (grey lines) look so jagged and incosistent because each replicate shifts which year gets which value of $p_\mu$ (average detectability). It's interesing to note here that the bottom panel tells us that the MSOM expects our data to have lower richness than it really does.

---

## Site Specific Richness (`Nsite`)

### Scatter Plots of `Nsite` Split by Year

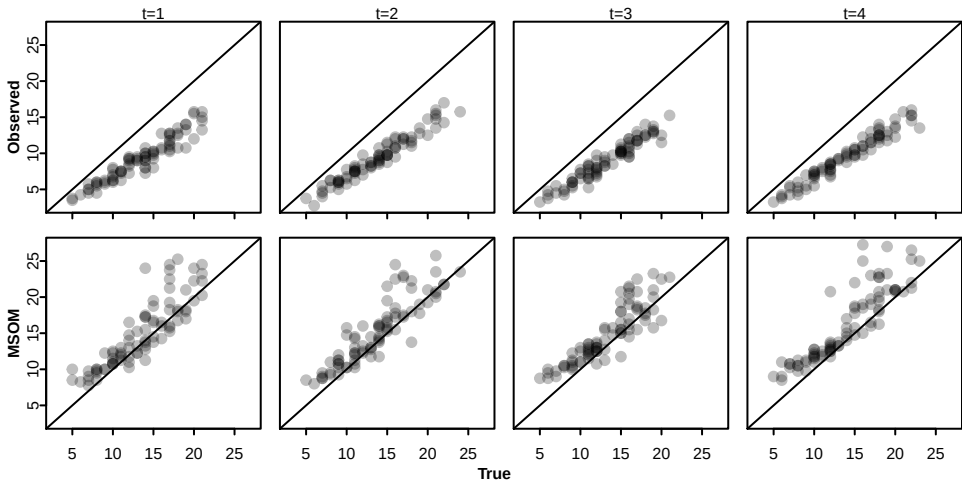


Figure 3: Site-specific richness (`Nsite`, $N_j$) from simulated observations (vertical axis, top row; $N_j^{obs}$) and from MSOM estimates (vertical axis, bottom row, $\hat{N}_j$) vs true site-specific richness (horizontal axis; $N_j^*$). The panel columns delineate the years of the simulation. Each point is site-specific species richness that has been averaged over the simulated replicate observations.

The first thing I notice in the `Nsite` scatter plot figure is that the observations tend to underestimate true richness, and the absolute magnitude of this underestimate increases as true richness increases (i.e., with the observation as the response and the true value as the predictor, the intercept is fine, but the slope is too shallow). The MSOM tends to do a much better job of staying near the 1:1 line, but there's is more variance in the residuals at high richness. I'm not sure I understand why (in later figures, $\psi$ estimates don't get worse or more uncertain at high values of true $\psi$).

---

### Maps of Richness (space and time)

In these maps, the environmental variable $X$ changes linearly across the y-axis, and randomly (and much less) across the x-axis. The different columns represent separate years. The environmental variable changes

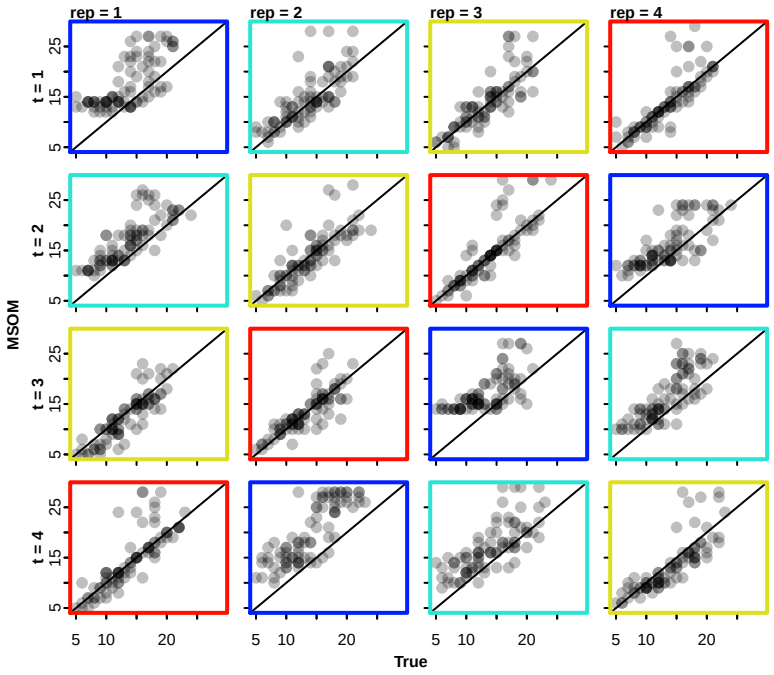


Figure 4: Scatter plots of MSOM-estimated species richness (y-axis) and the true species richness (x-axis), with each panel representing a particular year-replicate combination. The color of the panel border indicates the value of $p_\mu$, the community-wide mean detectability.

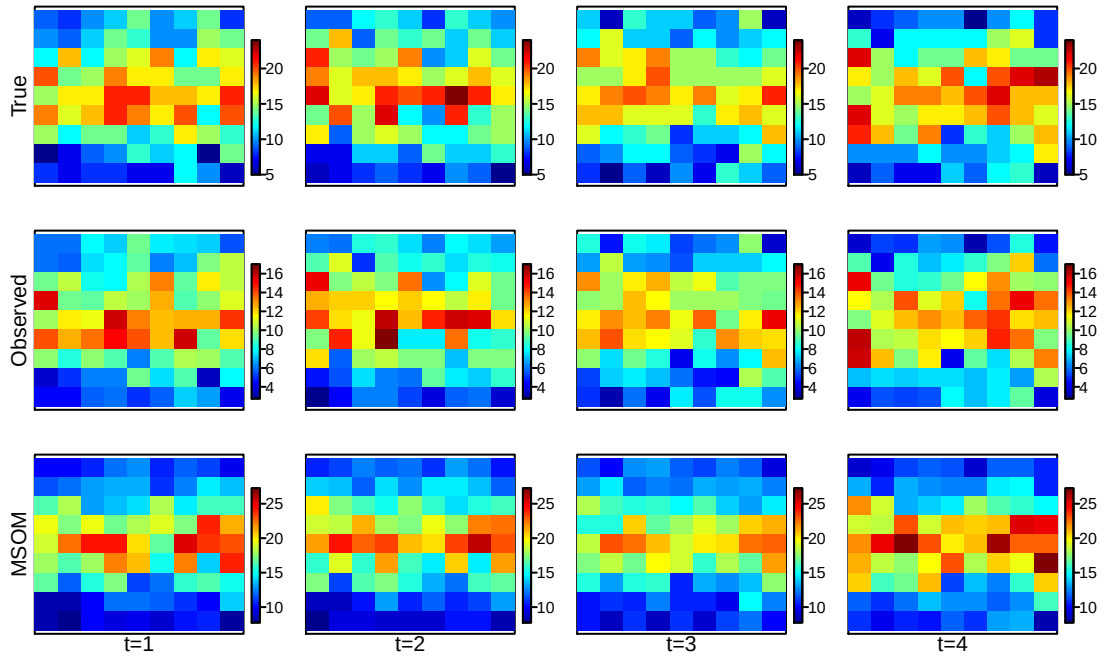


Figure 5: Maps of site- and year-specific species richness (`Nsite`) from the simulation of the True process (top row), simulation of the Observed process (middle row), and the MSOM estimates (bottom row). X-axis and Y-axis indicate position in 2 dimensional space. Colors indicate species richness (warm colors are higher richness than cool colors), averaged over the simulated replicate observations. Horizontal and vertical axes Each row of panels is scaled independently, columns within a row are scaled equally.

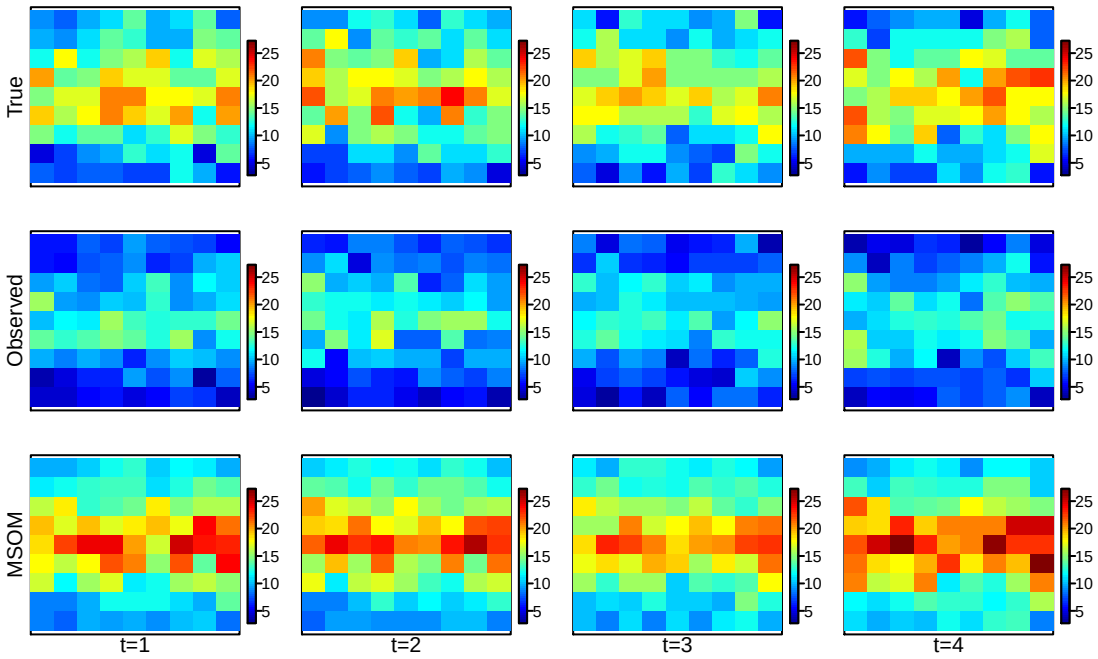


Figure 6: Same as previous figure, but all panels are on the same scale.

linearly among years (the rate of change is the same for all x-y locations), and in this basic simulation that rate of change is actually just 0 units/year.

Because so many things are static in this simulation the maps of `Nsite` are not much more informative than the scatter plots of `Nsite`. However, makign the spatial aspect of richness visually explicit does emphasize that richness is highly dependent on the envionrmental variable —

# Occupancy Probability, $\psi$

## Definition of $\psi$

The probability that a particular location $j$ will be occupied by species $i$ is $\psi$ (omitting subscripts). This probability is a function of the environment and 3 species-specific parameters. To calculate $\psi_i$ under a set of known conditions in an environmental variable $X$ at time $t$ and site $j$, we can express it as the log of the odds ratio (logit link) resulting from the the linear combination of three terms:

$$logit(\psi_{j,i,t}) = \begin{pmatrix} 1 \\ x_{j,t} \\ x_{j,t}^2 \end{pmatrix}^{\intercal} \begin{pmatrix} a_{0,i} \\ a_{3,i} \\ a_{4,i} \end{pmatrix}$$

and

$$\begin{aligned} a_{0,i} &\sim \mathcal{N}(\mu_{a0}, \sigma_{a0}^2) \\ a_{3,i} &\sim \mathcal{N}(\mu_{a3}, \sigma_{a3}^2) \\ a_{4,i} &\sim \mathcal{N}(\mu_{a4}, \sigma_{a4}^2) \end{aligned}$$

Thus, the parameters are hierarchical, and for this reason the response curves vary but are somewhat clustered around a central value.

In my mind, $\psi$ is the Holy Grail of parameters to recover. It tells us the odds that a species will be in a certain place at a certain time. If I could know this perfectly, I would be very pleased (and we'd all be very famous).

When this analysis (using an MSOM on the trawl data) was originally crafted, $\psi$ was more of a means to an end than it was the objective – $\psi$ lets us get at richness, and we have hypotheses about how richness should change with climate that we'd like to test. But if you know what controls $\psi$, you know what controls richness.

---

## Scatter Plot of $\psi$

In a general sense, the MSOM can distinguish between instances (sites/ years) when a species is likely to be present, and when it's not. However, in every simulation I've done (varying many parameters that aren't compared in this document), the scatter plot of $\psi$ always makes it apparent that

1. There is a lot of variability around the 1:1 line

2. The residuals are not normal, and they are not independent

    i. In general, I've found that $\hat{\psi}$ exhibits an upward bias, overestimating $\psi^{true}$

    ii. Smoothly-curving excursions from the 1:1 line are often prominent

These patterns are somewhat concerning. The curve-like sequence of residuals is probably a byproduct of slightly incorrect estimates of the parameters in the logistic regression ($[a_0, a_1, a_2]$), resulting in estimated response curves that deviate non-randomly from the true response curve. For a heuristic of how these smoothe excursions can occur, in R try something as simple as `d <- rnorm(100); plot(dnorm(d), dt(d, 1))` to see the relationship between the density estimate from the correct distribution and that from the wrong distrubtion (the density is analogous to $\psi$); or for really crazy patterns, try `d <- rnorm(100); plot(dnorm(d), do.call(approxfun, density(d)[c("x","y")])(d))`. So the curves are explainable, but I cannot explain the consistent overestimation; I could understand how underestimating detectability ($p$)

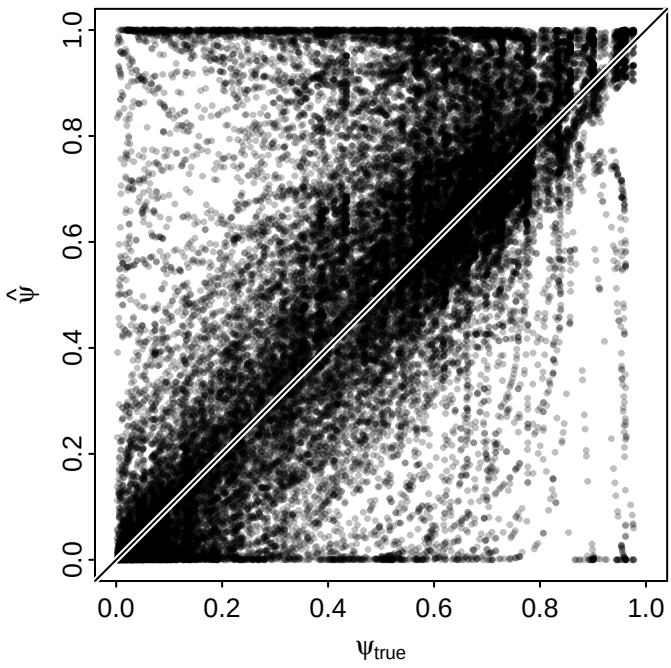


Figure 7: MSOM estimates of $\psi$ ($\hat{\psi}$) vs. true values of $\psi$ ($\psi_{true}$). Each point is a $\psi$ value for a particular site-species-year-replicate. The white and black line is the 1:1 line.

would result in overestimating $\psi$, but the MSOM appears to recover true $p$ values rather well (e.g., see $p$ Scatter Figure), so that's not a satistfying explanation.

In the next section I drill into $\psi$ a bit more to try and understand what causes the largest deviations from true values.

---

## Scatter Plots for Each $\psi_{t,r}$

The estimates and true values of $\psi$ are best correlated when $p$ is high. When the average species has a low chance of being detected (when $p_\mu$ is, say, 20%), the estimates of $\psi$ are a mess.

Does this help explain an apparent positive bias the the aggregated scatter plot of *psi*? Maybe. When detectability ($p_\mu$, the variability indicated by panel colors) is low, that's when we run into trouble. With low detectability, you have fewer observations. You have a poorer sense of what's going on. So that adds uncertainty. Perhaps when detectability is super low, it's entirely too easy to conflate an absence with an undetected presence — you start assuming that 0's are just because you didn't look hard enough, not because it's really absent. I'm not convinced by this logic, though; I'd want to see a better explanation. Alternatively, maybe the chains aren't converging yet; I didn't run diagnostics. So for now this is a mystery to me.

*Note: what I refer to as p here is really just the probability that a species will be detected if an occupied site is sampled, so the number of substrata sampled per site isn't relfected in p. In this simulation, 50% of substrata were sampled, and while this doesn't influence p, it could add noise to its estimates.*

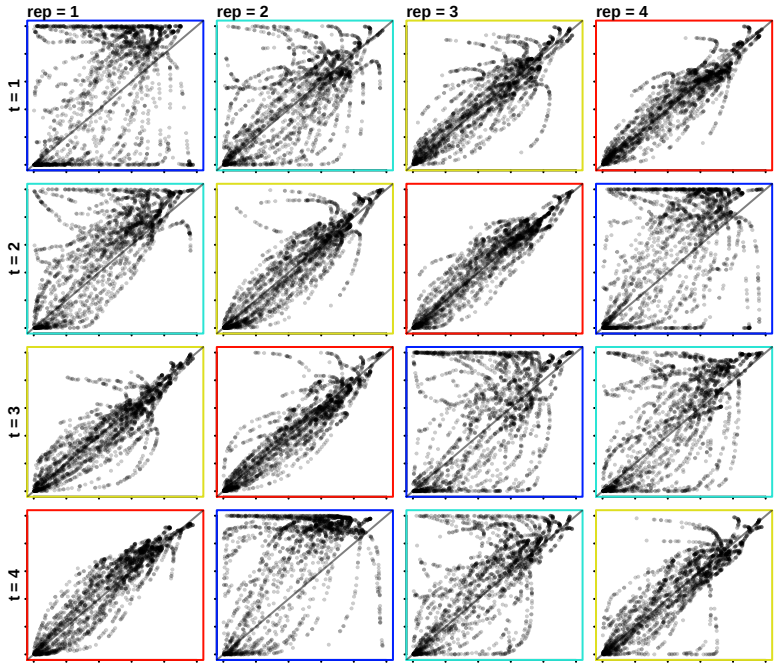


Figure 8: True (horizontal axes) and MSOM estimates (vertical axes) of occupancy probabilities ($\psi_{j,i,t,r}$) of species $i$ occupying a location $j$. Years ($t$) are separated by rows, replicates ($r$) are separated by columns. The border color of each panel indicates the community-level mean probability of detection ($p_\mu$; where $p_i \sim \mathcal{N}(p_\mu, \sigma^2)$), with warm colors indicating high detectability, and cool colors low. The species-specific detectabilities are **not** re-randomized among replicates, but even when the probabilities associated with the observation process do not change, the outcome of the process can change. The year $t$ of the simulated true process changes across the rows of panels, and the simulated replicate observation $r$ changes across columns.

## Occupancy Response Curves

Occupancy response curves are calculated as $logit(\psi_i) = \mathbf{X} \times \mathbf{a_i}$, where

$$\mathbf{X} = \begin{pmatrix} 1 & X_{min} & X_{min}^2 \\ \vdots & \vdots & \vdots \\ 1 & X_{max} & X_{min}^2 \end{pmatrix}; \mathbf{a_i} = \begin{pmatrix} a_{0,i} \\ a_{3,i} \\ a_{4,i} \end{pmatrix}$$

Therefore, these curves are tantamount to values of $\psi$, except that $\psi$ generally pertains to a simulated, observed, or true occupancy probability, whereas the occupancy probability in the response curves is calculated over hypothetical conditions (i.e., over hypothetical values of the environmental gradient $X$). This formulation yields value of $\psi$ for each species at all temperatures over the interval $(X_{min}, X_{max})$ (this implies that the number of rows of $X$ is infinite, which it isn't, but it is quite large).

### True Occupancy Response Curves

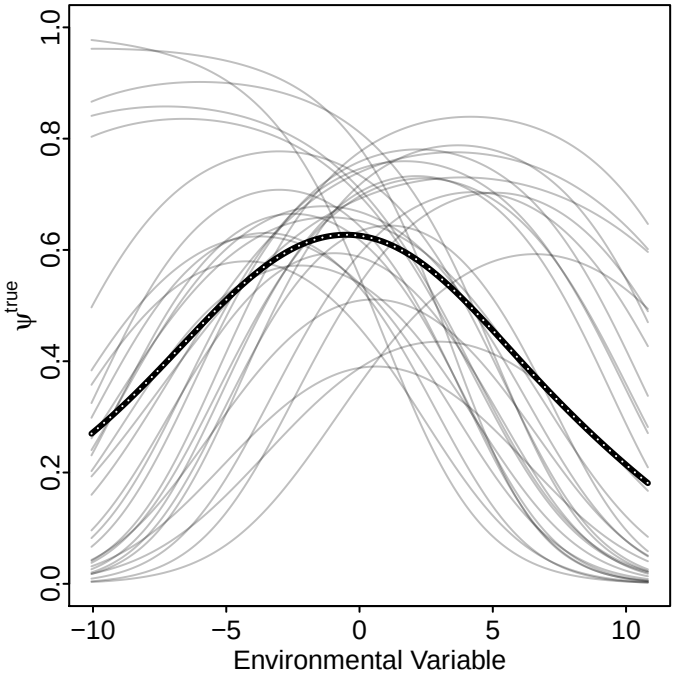


Figure 9: True simulated response curves. Vertical axis is the value of $\psi^{true}$, horizontal axis is the value of the environmental variable that, along with species-specific regression parameters, determines $\psi^{true}$. The thick line is the among-species mean value of $\psi^{true}$ at a given value of the environmental variable.

In the response curve, the values of the environmental variable are an arbitrary gradient, and do not necessarily correspond to what was observed in the simulated environment (although they are intended to cover the same range). The formulation $logit(\psi_i) = \mathbf{X} \times \mathbf{a_i}$ is useful for producing the response curve, but it is inherently discrete. This is actually how I simulated the true $\psi$ in the model, but it's not how the MSOM analyzes it (although the difference is negligible because I use so many rows in $\mathbf{X}$).

***

## Estimated Occupancy Response Curves

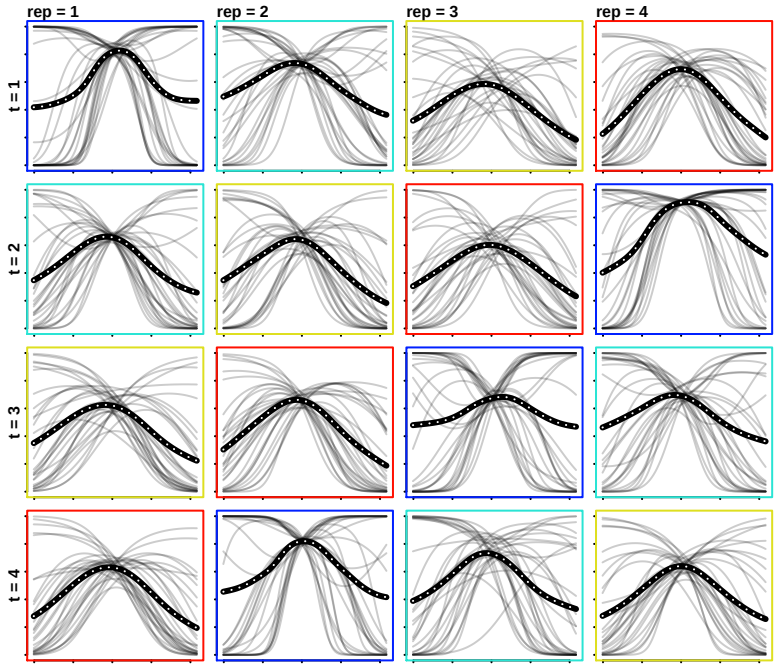


Figure 10: Response curves of species' probability of occupancy ($\psi_i$, vertical axis) across the full range of temperatures in the simulation. The color of the boxes around each panel refer to the among-species average of the probability of detection; warm colors indicate that the mean detection probability is high, whereas cool colors indicate that $p$ was low. The year $t$ of the simulated true process changes across the rows of panels, and the simulated replicate observation $r$ changes across columns.

It is rather evident in these figures that the estimated response curves look less like the true response curves for low detection probabilities (low $p_\mu$). It is important to remember this effect when evaluate the esimates of `Nsite`, and especially when looking at the heat maps of `Nsite` in Figures 4 and 5 because in those figures each grid cell represents the richness averaged over replicates, and the value of $p_\mu$ varies among replicates. Furthermore, the inaccurate response curves at low $p_\mu$ are consistent with the relationships between $\hat{\psi}$ and $\psi^{true}$ in Figure 7.

It is my suspicion that, at a large enough sample size of both preseneces and abseneces across a full gradient of temperatures, these deficiencies will diminish. However, this analysis, in its basic form, was not designed to test the influence of sample size (I've run separate versions of this model, not presented here, and can informally confirm my suspicion – sample size matters a lot). Furthermore, most regions in the trawl data set have a few dozen sites (site being defined on a 1° grid), so the $9 \times 9$ grid simulated here is approximately the spatial sample size we'd have to work with in the empirical analysis.

Two ways to change the sample size in the empirical analysis would be to

1. reduce grid size to increase the number of sites
    - could do a half-degree grid
    - make substrata 1/4 degree
    - downside is that each site would be represented in a fewer proportion of years
2. Use multiple years of data

- drastically increases sample size
- would require a new model
- the model would need to conider factors that change among years

---

# Probability of Detection, $p$

## Definition of $p$

The probability of detection ($p$), is a species specific parameter in the MSOM model. The MSOM analyzes all years ($t$) and replicates ($r$) separately, so I am going to leave those subscripts out of this description. In the simulation, the probability of observing a species is a function of two independent factors:

1. The probability that site $j$ is occupied by species $i$; this is $\psi_{j,i}$
    - $\psi_{j,i}$ is a function of species-specific niche and an environmental variable that changes over space and time
    - $Z_{j,i}$ is the species- and site-specific richness, which is a function of $\psi$ (given that we're only talking about species that are in the pool of possible species, determined by $w_i$)

2. A species-specific ($i$) chance of being identified (`taxChance`), given that it is present in a location that was sampled (i.e., detectability does not reflect the probability of sampling a place); this detectability parameter is $p_i$
    - Detectability changed between years.
    - In a given year, $logit(p_i) \sim \mathcal{N}(p_\mu, \sigma^2)$. $p_\mu$ changed between years (taking on values of -2, 0, 2, and 4), $\sigma^2 = 2$ in all years.
    - The value of $p$ only changes between species (and years), but the observation process occurs at the substratum ($k$) level. Thus, the parameter is really $p_{j,k,i}$, but for a given $i$, all $p_{j,k}$ are constant. I represent this probability as $p_i$ with the understanding that this value is repeated over space.
    - $Y_{j,i}$ is the observed version of $Z_{j,i}$.
    - $Y_{j,i} \sim Bern(p_i \times Z_{j,i})$.
        - *Note: Because p is actually subscripted to k, the Y are also actually subscripted to k. Maybe leaving these subscripts out is making things more confusing. I've only excluded them to emphasize how parameters are estimated.*
    - Our data about species presence/ absence correspond to $Y_{j,i}$. So it might be useful to think of the MSOM as estimating $\hat{Y}_{j,i}$, which is compared to the observed data $Y_{j,i}^{obs}$.

---

## Demo: Effect of MSOM Hierarchy on $p$

The above plot is interesting because it shows that exceptionally high or exceptionally low chances to be observed only occur for the species that were observed at least once; i.e., this says that if you didn't observe it, it just takes on the mean. That's reasonable, I guess; but I also would think that the things that were never observed could also be things that had a low chance of observability; but they could also have just a low chance of actually being present. So I suppose in the end it just doesn't get informed, and reverts to the mean?

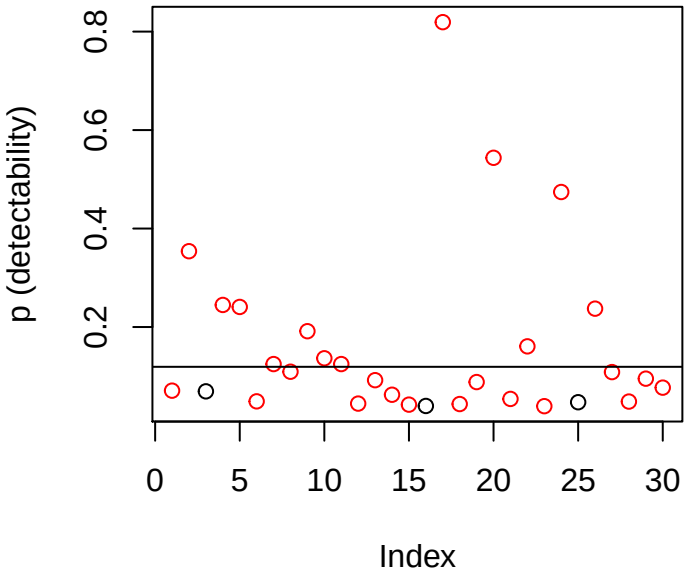


Figure 11: Probability of being detected, $p$. Horizontal line is mean probability. Figure only shows results for the first year of the simulation/ observation, and only 1 replicate. Different points are different species. Probability of being detected is a species-specific parameter (does not vary among sites, e.g.). Red points are species that were observed, black points are species that were never observed.

## Scatter Plot of $p$

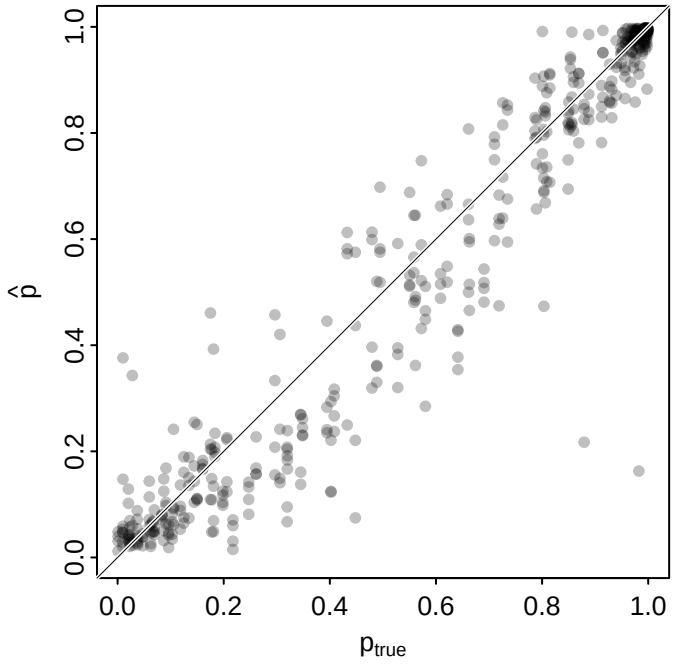


Figure 12: MSOM estimates (vertical axis) and true values of $p_i$, the species-specific ($i$) detection probability. Each point is subscripted by species $i$, year $t$, and observation replicate $r$.

The MSOM does a pretty good job of recovering $p$. This is true in other conditions I've simulated — conditions where it failed miserably to recover $\psi$ (mainly because of small sample sizes). It's worth noting here that most of the points are either very close to 1 or very close to 0. In the next figures, the reason for this will become more apparent.

## Scatter Plot of $\hat{p}$ vs $p_{true}$, split by year and replicate

The grouping in this figure is a bit odd because the panels with the highest detectability also don't show very much range (because both axes are sitting at the high detectability!). However, in general, it looks like the estimates of $p$ are pretty good overall. There are some weird points, but overall, both the multi-panel and the combined scatter plot indicate that we are recovering $p$ fairly reliably.

In the previous section I pointed out that points group close to 0 or 1. That's just due to the different values of $p_\mu$. I also wonder what a real value for $p$ would look like. I have no idea which of these values are reasonable.

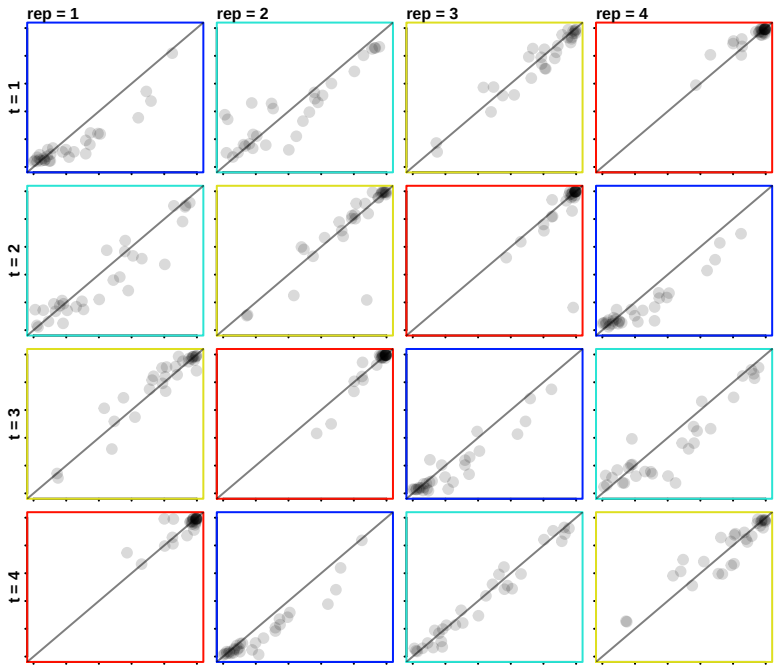


Figure 13: Probability of detection ($p$), separated by year and replicate. True values ($p^{true}$) are on the horizontal axis, MSOM estimates ($\hat{p}$) are on the vertical axis. The color of the panel borders corresponds to the value of $p_\mu$, with warm colors being high and cool colors being low values of $p_\mu$. In a given panel, each point is a different species. The rows are different years, the columns are the simulated replicate observations of those years. The same set of true values for $p$ is used for each replicate, just in a different order. If this simulation shows the same color panel box more than once per column, those panels share the same $p_\mu$, but they have independent realizations of $p^{true}$. See the section entitled The Simulation for further clarification.

451

# Assessment with Mixed Effects Models

## Describe Motivation for Mixed Effects Models

**Motivation**: MSOM skill might differ across dimensions, trying to figure out what patterns I should expect to pick out (spatial patterns in richness, temporal?) E.g., Is the correlation between MSOM and True the same comparing across sites as comparing across years? Species, reps, also.

**Motivation**: What factors influence MSOM skill in a given dimension? E.g., Skill in finding differences in $\psi$ across species may depend on $p$, the chance of being identified. If $p$ changes among years, might also explain

Read more about specifying mixed effects models using `lmer` in R here

This example is looking at $\psi$, probability of an individual species being present

## Example LMER Analysis for $\psi$

```
# Just exploration/ starting point
library(car)
library(lme4)

# psi
# psi true
dat.psi.true <- reshape2:::melt.array(
    psi.true,
    varnames=c("site","spp","time","rep"),
    value.name="psi.true",
    as.is=T
)
# psi hat
dat.psi.hat <- reshape2:::melt.array(
    psi.hat,
    varnames=c("site","spp","time","rep"),
    value.name="psi.hat",
    as.is=T
)

# p
# p true
dat.p.true <- reshape2:::melt.array(
    aperm(array(p.true, dim=c(dim(p.true), dim(psi.true)[1])),c(4,1,2,3)),
    varnames=c("site","spp","time","rep"),
    value.name="p.true",
    as.is=T
)
# p hat
dat.p.hat <- reshape2:::melt.array(
    aperm(array(p.hat, dim=c(dim(p.hat), dim(psi.hat)[1])),c(4,1,2,3)),
    varnames=c("site","spp","time","rep"),
    value.name="p.hat",
    as.is=T
)
```

```r
# n.hauls
n.hauls <- sapply(big.out.obs, function(x)attributes(x)$n.haul)
n.hauls.dim <- c(grid.w*grid.h, n.obs.reps, ns, grid.t)
dat.n.hauls <- reshape2:::melt.array(
    aperm(array(n.hauls, dim=n.hauls.dim), c(1,3,4,2)),
    varnames=c("site","spp","time","rep"),
    value.name="n.hauls",
    as.is=T
)

# grid.X
# same structure (dims) as n.hauls
temp <- values(grid.X)
temp.dim <- c(grid.w*grid.h, n.obs.reps, ns, grid.t)
dat.temp <- reshape2:::melt.array(
    aperm(array(temp, dim=temp.dim), c(1,3,4,2)),
    varnames=c("site","spp","time","rep"),
    value.name="temp",
    as.is=T
)

# tax chance
tax.chance <- simplify2array(
    lapply(big.out.obs, function(x)(attributes(x)$obs.params)$tax.chance)
)
tax.chance.dim <- c(grid.t, ns, n.obs.reps, grid.w*grid.h)
dat.tax.chance <- reshape2:::melt.array(
    aperm(array(tax.chance, dim=tax.chance.dim), c(4,2,1,3)),
    varnames=c("site","spp","time","rep"),
    value.name="tax.chance",
    as.is=T
)


mod.dat <- cbind(
    dat.psi.true,
    psi.hat=dat.psi.hat[,"psi.hat"],
    p.true=dat.p.true[,"p.true"],
    p.hat=dat.p.hat[,"p.hat"],
    n.hauls=dat.n.hauls[,"n.hauls"],
    tax.chance=dat.tax.chance[,"tax.chance"],
    temp=dat.temp[,"temp"]
)
mod.dat[,"psi.error"] <- mod.dat[,"psi.hat"]-mod.dat[,"psi.true"]
mod.dat[,"p.error"] <- mod.dat[,"p.hat"] - mod.dat[,"p.true"]

mod.dat$site <- as.factor(mod.dat$site)
mod.dat$spp <- as.factor(mod.dat$spp)
mod.dat$time <- as.factor(mod.dat$time)
mod.dat$rep <- as.factor(mod.dat$rep)


# ====================
```

```
# = Do LMER Analysis =
# ====================
mod1 <- lmer(psi.error~temp+(1|spp)+(1|site), data=mod.dat)
mod2 <- lmer(psi.error~n.hauls+(1|spp)+(1|site), data=mod.dat)
mod3 <- lmer(psi.error~p.error+(1|spp)+(1|site), data=mod.dat)
# mod4 <- lmer(psi.error~n.hauls-1+(1|spp)+(n.hauls-1|spp)+(1|site), data=mod.dat)
mod4 <- lmer(psi.error~temp+(1|spp)+(temp-1|spp)+(1|site), data=mod.dat)
```

```
## Warning in checkConv(attr(opt, "derivs"), opt$par, ctrl =
## control$checkConv, : Model failed to converge with max|grad| = 0.0116548
## (tol = 0.002)
```

```
mod5 <- lmer(psi.error~p.error+(1|spp)+(p.error-1|spp)+(1|site), data=mod.dat)
```

```
## Warning in checkConv(attr(opt, "derivs"), opt$par, ctrl =
## control$checkConv, : Model failed to converge with max|grad| = 0.00275493
## (tol = 0.002)
```

```
mod6 <- lmer(psi.error~p.error+(p.error|spp)+(1|site), data=mod.dat)
```

```
## Warning in checkConv(attr(opt, "derivs"), opt$par, ctrl =
## control$checkConv, : Model failed to converge with max|grad| = 0.00268069
## (tol = 0.002)
```

```
# Calculate covariance matrix (for spp)
mod6.varcor.spp <- attr(summary(mod6)$varcor$spp, "correlation")
mod6.varcor.spp <- format(mod6.varcor.spp, digits=2)
mod6.varcor.spp[!lower.tri(mod6.varcor.spp)] <- ""

mod.cap <- c(
    "Mixed effect models assessing sensitivity of $\\psi_{\\epsilon}$ to simulation conditions"
)
```

The goal with the mixed effects models was to understand what causes errors in $\psi$. I focused on $\psi$ because it has all the information needed to understand variability in richness, but it has more information than the actual richness (richness is a community level statistic, $\psi$ is species-specific). In these models, the response variable is $\hat{\psi} - \psi^{true}$, which we'll call $\psi_\epsilon$. If we understand the source of variability in $\psi_\epsilon$, then we can understand what leads to inaccuracies in our model.

When analyzing the trawl data, we will not know $\psi_\epsilon$ – we can obtain model residuals, but these are distinct from $\psi_\epsilon$, because calculating $\psi_\epsilon$ requires knowing $\psi^{true}$ which is a latent, unobserved variable. An empirical analysis would also lack some of the explantory variables made available to us in the simulation. However, if we can explain variability in $\psi_\epsilon$ using simulated information that will also be available to the trawl analysis, then we can build intuition about the sources of error in our estimate of species richness even when we don't know the true value. And that is the fundamental goal of this document.

I'll highlight some of the things I learned from this analysis:

1. Neither the environmental variable (`temp`) nor the number of subsites sampled per site (`n.hauls`) were strongly related to $\psi_\epsilon$
    i. But `temp` might be a better predictor if transformed into an absolute value

Table 3: Mixed effect models assessing sensitivity of $\psi_\epsilon$ to simulation conditions

| | Model 1 | Model 2 | Model 3 | Model 4 | Model 5 | Model 6 |
|---|---|---|---|---|---|---|
| (Intercept) | 0.062*** | 0.060*** | 0.046*** | 0.062*** | 0.040*** | 0.040*** |
| | (0.010) | (0.011) | (0.008) | (0.010) | (0.009) | (0.009) |
| temp | 0.001*** | | | 0.001 | | |
| | (0.000) | | | (0.002) | | |
| n.hauls | | 0.001 | | | | |
| | | (0.001) | | | | |
| p.error | | | −0.618*** | | −0.867*** | −0.868*** |
| | | | (0.012) | | (0.136) | (0.136) |
| AIC | -6331.371 | -6314.550 | -8840.187 | -8685.580 | -11637.684 | -11636.548 |
| BIC | -6288.530 | -6271.709 | -8797.346 | -8634.170 | -11586.274 | -11576.570 |
| Log Likelihood | 3170.686 | 3162.275 | 4425.094 | 4348.790 | 5824.842 | 5825.274 |
| Num. obs. | 38880 | 38880 | 38880 | 38880 | 38880 | 38880 |
| Num. groups: site | 81 | 81 | 81 | 81 | 81 | 81 |
| Num. groups: spp | 30 | 30 | 30 | 30 | 30 | 30 |
| Variance: site.(Intercept) | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| Variance: spp.(Intercept) | 0.003 | 0.003 | 0.002 | | | 0.002 |
| Variance: Residual | 0.049 | 0.049 | 0.046 | 0.046 | 0.043 | 0.043 |
| Variance: spp.temp | | | | 0.000 | | |
| Variance: spp.1.(Intercept) | | | | 0.003 | 0.002 | |
| Variance: spp.p.error | | | | | 0.543 | 0.543 |

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

ii. When the model does not contain a site-specific random intercept (not shown here), `n.hauls` acounts for more variability

2. The `spp` random intercept explains much more variability than `site` equivalent

3. `p.error` explains a ton of variability

   i. inversely related to $\psi_\epsilon$; intuition: if you have perfect detectability but you think it's terrible, you'll overestimate the true value
   ii. model 6 is a worse fit than model 5, meaning that adding covariance structure between `spp`-specific intercept and `spp`-specific `p.error` does not explain much variability

As stated above, a lot of variance is explained by the model term `(p.error|spp)`, which allows the parameter associated with `p.error` and the intercept (both fixed effects) to vary randomly among species. The interpretation of these model terms is that

- Each species gets to draw it's own intercept ("Variance: spp.1.(Intercept)" in the table) from a parent distribution of intercepts
- A unit of error in the estimate of $p$ has an influence on $\psi_\epsilon$ that, similar to the intercept, varies among species ("Variance: spp.p.error" in the table).

Therefore, the effect that a bad estimate of $p$ has on $\psi_\epsilon$ is not the same among species. It is not clear what causes some species to be more sensitive to a poorly estimated $p$ than others; one possibility is that $p$ is poorly estimated for species that have not been observed much, and it is this lack of observation that is also responsible for generating uncertainty in $\hat{\psi}$. Regardless, a bad (good) estimate of $p$ is a good predictor of a

bad (good) estimate of $\hat{\psi}$.

# Conclusion

## Discussion of Results

Overall, the MSOM performed well. It is definitely data-hungry in the sense that it needs to observe 1's and 0's for each species in many places under different conditions. This may seem counter to the goal of the MSOM – to make efficient use of hard-won data. Would other richness methods be better? But remember, what we're getting is site-specific richness, and even species-specific presences and absences. Also, we don't have the "replicates" in the trawl data set needed for the other richness methods.
This seems obvious in retrospect, but Figure 1 (Boxplots) shows us that the sensitivity of our richness estimates to detectability is highly nonlinear. It'll be important to gauge where we think we are along that spectrum. Regardless, we were never too far off the real richness — e.g., when 29 species exist, we estimate 24. On the upside, we're not *too* much worse off if we want to know $\psi$ instead of just R. Of greater concern is the tendency to overestimate R or $\psi$ when $p$ is low However, these low values for $p$ might be *really* low, so perhaps that "problem" is not in a relevant region of parameter space.

## Next Steps

I think the next important step is to decide if we want to analyze all of the years together. It'll require a more complicated model, but it'll give a lot more statistical power (and better estimates). I'm leaning towards doing this.

We also need to decide what regions from the trawl data set we'll want to use for this analysis. Right now the Alaskan regions are the front runners (just Eastern Berring Sea, or Aleutians and Gulf of Alaska as well?) simply because they have pdf file ranking each species on a scale 1-3 for each year according how likely it was to be identified. This isn't $p$ per se (because other factors affect detectability as well), but this could give insight into a big contributor to temporal shifts in $p$.

We may also want to upgrade the realism of the simulation. I'm already set up to link years together if we decide to do a multi-year MSOM (MY MSOM?).

We should also think about what exactly our story/ selling points will be. I tried to word the introduction to this document as something that might sound like a paper Introduction. But this is tough when you don't actually know what you're introducing. I think the key to this paper may to emphasize an ecological finding the Alaskan data sets, but back up that finding with a careful evaluation of the method we're using. My goal would be to highlight how it's important but difficult to understand biodiversity dynamics, and then discover something about biodiversity dynamics while showing that the MSOM is a valid tool for doing so with trawl (and similar) data.

## Concluding Remarks

I think we have a good group of people to discover some cool stuff in the trawl data set. What we have in this document is just the prelude – I think the real exciting stuf will come once we have estimates of richness for the trawl data that we trust. I'm looking foward to hearing your thoughts and working with all of you on this project!

---

# Report Generation Notes

## R Session Information

```
## R version 3.1.2 (2014-10-31)
## Platform: x86_64-apple-darwin13.4.0 (64-bit)
##
## locale:
## [1] en_US.UTF-8/en_US.UTF-8/en_US.UTF-8/C/en_US.UTF-8/en_US.UTF-8
##
## attached base packages:
## [1] parallel  grid      stats     graphics  grDevices utils     datasets
## [8] methods   base
##
## other attached packages:
##  [1] memisc_0.97      MASS_7.3-35      lme4_1.1-6       Rcpp_0.11.6
##  [5] Matrix_1.1-4     car_2.0-24       rbLib_0.0.2      texreg_1.35
##  [9] stargazer_5.2    kfigr_1.2        xtable_1.7-4     rmarkdown_0.7
## [13] knitr_1.10.5     doParallel_1.0.8 iterators_1.0.7  foreach_1.4.2
## [17] R2jags_0.5-6     rjags_3-15       coda_0.16-1      lattice_0.20-29
## [21] igraph_0.7.1     fields_6.9.1     maps_2.3-6       spam_0.41-0
## [25] data.table_1.9.4 raster_2.3-24    sp_1.0-17
##
## loaded via a namespace (and not attached):
##  [1] abind_1.4-0         boot_1.3-17         chron_2.3-45
##  [4] codetools_0.2-9     digest_0.6.8        evaluate_0.7
##  [7] formatR_1.2         highr_0.5           htmltools_0.2.6
## [10] mgcv_1.8-3          minqa_1.2.3         nlme_3.1-118
## [13] nnet_7.3-8          numbers_0.5-6       pbkrtest_0.4-2
## [16] plyr_1.8.1          quantreg_5.11       R2WinBUGS_2.1-19
## [19] RcppEigen_0.3.2.1.1 reshape2_1.4.1      SparseM_1.6
## [22] splines_3.1.2       ssh.utils_1.0       stringr_0.6.2
## [25] tools_3.1.2         yaml_2.1.13
```

## Date Document Last Compiled

```
## Last compiled on: 2015-08-29
```

---