assess.sim.basic.R

Ryan Batt

2015-08-22

Abstract

Lorem ipsum dolor sit amet, consectetur adipisicing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enimad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.

Contents

Conventions and Settings

Dimension Conventions

Summary

1. Site ($j = 1, 2, \ldots, j_{max} = 20 \times 20 = 400$)
    - 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, $k_{max} = 4$, $k_{min}^{observed} = 1sdf$, and $k_{max}^{observed} = 4$
    - Substrata are are primarily useful for determining p, the detection probability
3. Species ($i = 1, 2, \ldots i_{max} = R = 40$)
    - 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 ($r = 8$)
    - Replicates are *simulated* repeated human observations of the same *realization* of the "true" process at 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 2×3 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.

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: 400, 40, 4
## grid.h = 20
## grid.w = 20
## grid.t = 4
##
## Number Species Possible (ns):
## 40
## Total Species Richness:
## 40
## Total Observed Species Richness:
## 40
##
## Annual Species Richness:
##           Min. 1st Qu. Median   Mean 3rd Qu. Max.
## All cells   39      39     39 39.250   39.25   40
## One cell     0       2      5  5.642    9.00   18
##
##
## Observed Annual Species Richness:
```

```
##           Min. 1st Qu. Median   Mean 3rd Qu. Max.
## All cells   38   38.75     39 38.750      39   39
## One cell     0    0.00      1  2.549       4   16
```

In the MSOM, detectability (p_i) is determined in the form of a logistic regression, which currently only has an intercept (v_0) as 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

JAGS Settings for MSOM

nChains	nIter	n0s	nSamples
3	50000	10	500

Assessment Settings

Central Tendency

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 emprical trawl data it refers to species.

Estimates of richness can be made spatially or temporally explicit (or neither, or both), but obviously a

---

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 "Fraction Capable of Being ID'd" category in the boxplots is essentially the cross-species average of p.

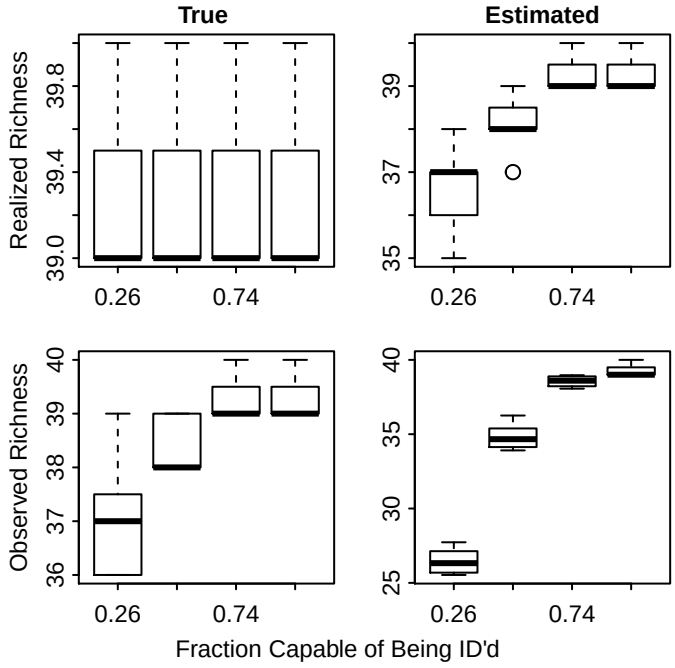


Figure. Boxplots of species richness. Numeric groupings indicate the average value of p across species during a given year–replicate combination. The panels in the left column are the true simulated values, and the panels on the right are the corresponding MSOM estimates. The top row indicates the latent 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

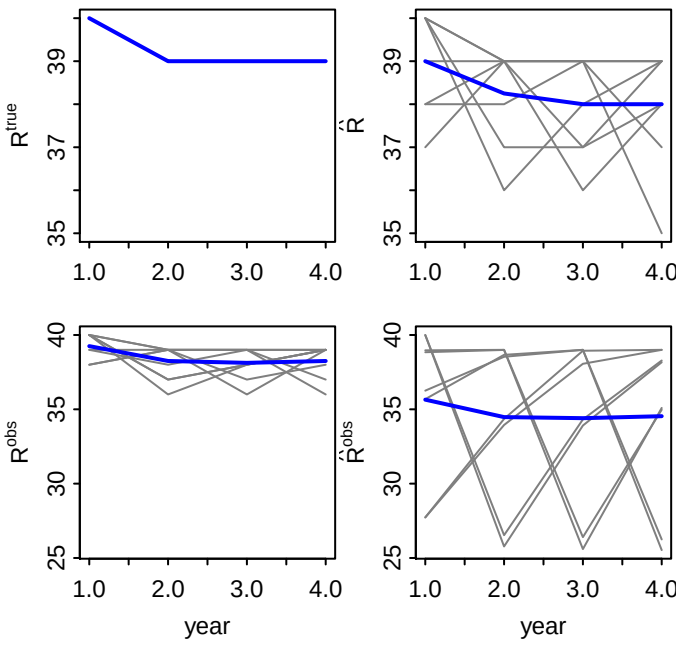


Figure. Time series of richness. Gray lines are individual replicates. Blue lines are averages. Note that detection probabilities ($p_{i,t,r}$, see simulation settings above, as well as definition of p below) change over time, and their temporal ordering differs among replicates. Text explanation goes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

Need explanations for how each panel was calculated.

1. R^{true} is straightforward

2. $\hat{R}$ is from $\sum_{i=1}^{R^{true}} max(\hat{Z}_{i,\nabla j \in J})$; and to be clear, $\hat{R}$ does not include the "unobserved" species introduced to the MSOM occurrence matrix (Y)

Site Specific Richness (`Nsite`)

Scatter Plots of `Nsite` Split by Year

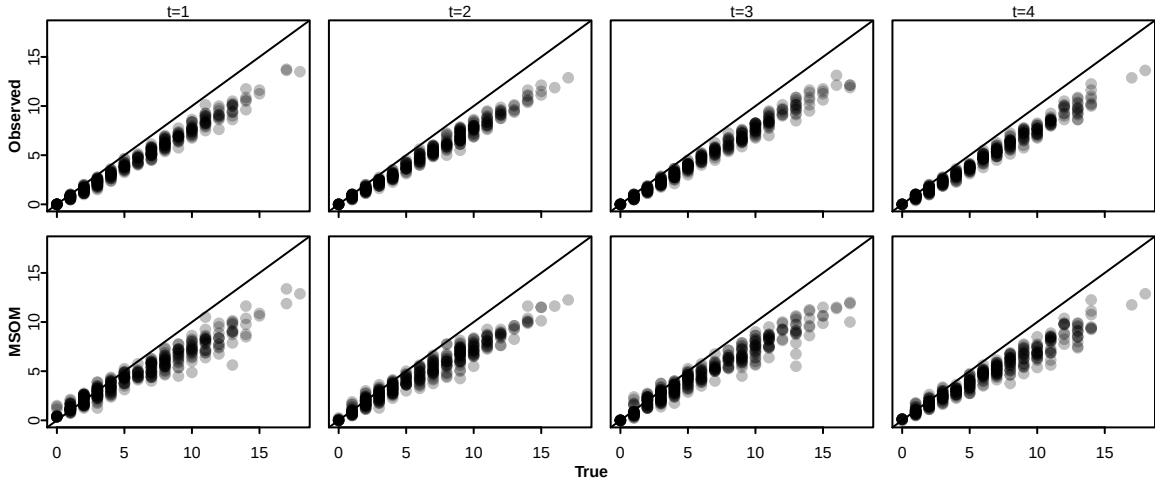


Figure. Site-specific richness (`Nsite`, N_j) from simulated observations (vertical axis, top row; N_j^{obs}) and from MSOM estimates (vertical axis, bottom row, $\hat{N}_j$) vs true site-specific richness (horizontal axis; N_j^*). The panel columns delineate the years of the simulation. Each point is site-specific ($j = 20 \times 20 = 400$) species richness that has been averaged over the simulated replicate observations ($r = 8$).

Maps of Richness (space and time)

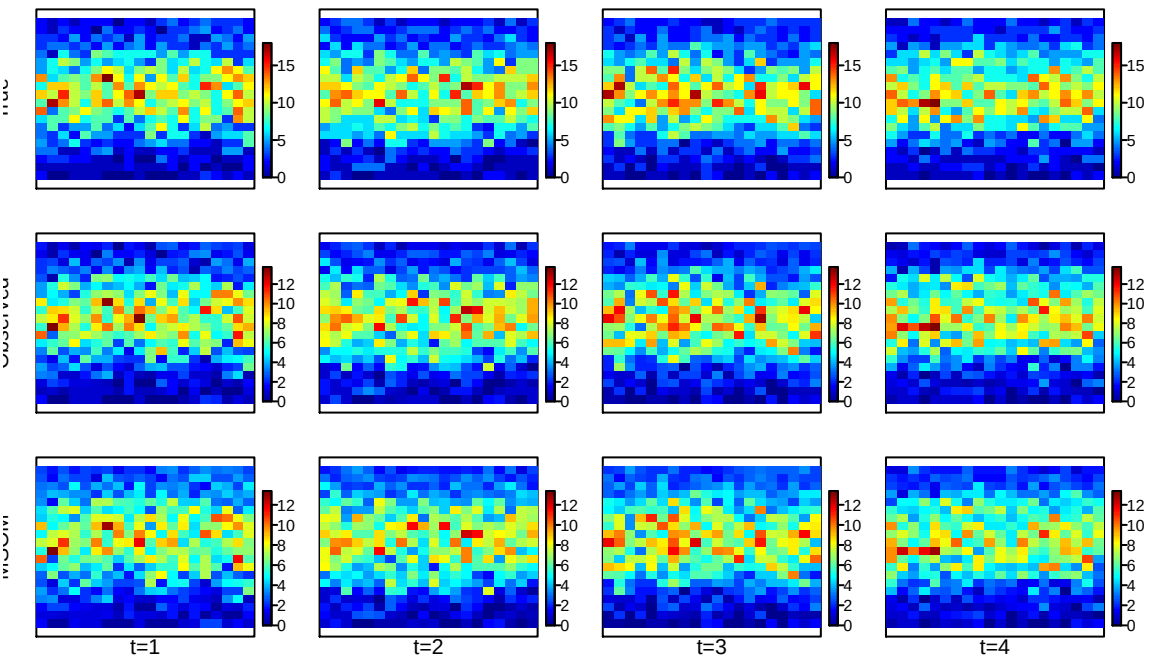


Figure. Maps of site- and year-specific species richness (`Nsite`) from the simulation of the True process (top row), simulation of the Observed process (middle row), and the MSOM estimates (bottom row). X-axis and Y-axis indicate position in 2 dimensional space; it is important to note that the environmental variable 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 linearly among years (the rate of change is the

same for all x-y locations). Colors indicate species richness (warm colors are higher richness than cool colors), averaged over the simulated replicate observations ($r = 8$). Horizontal and vertical axes Each row of panels is scaled independently, columns within a row are scaled equally.

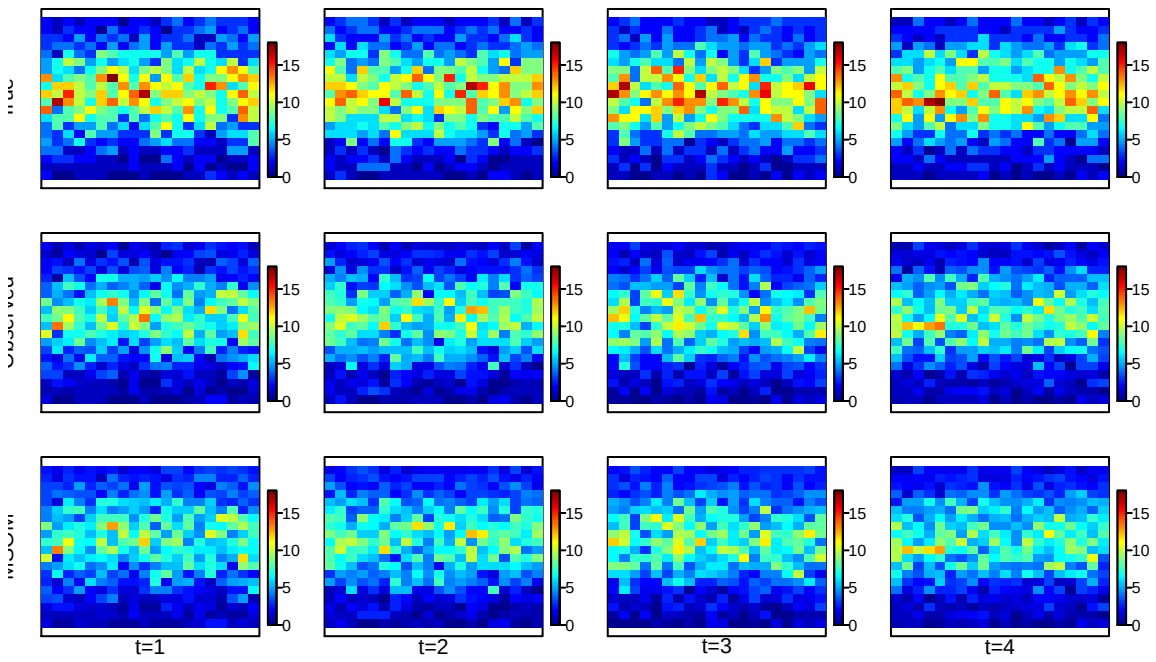


Figure. Same as previous figure, but all panels are on the same scale.

Text explanation goes here

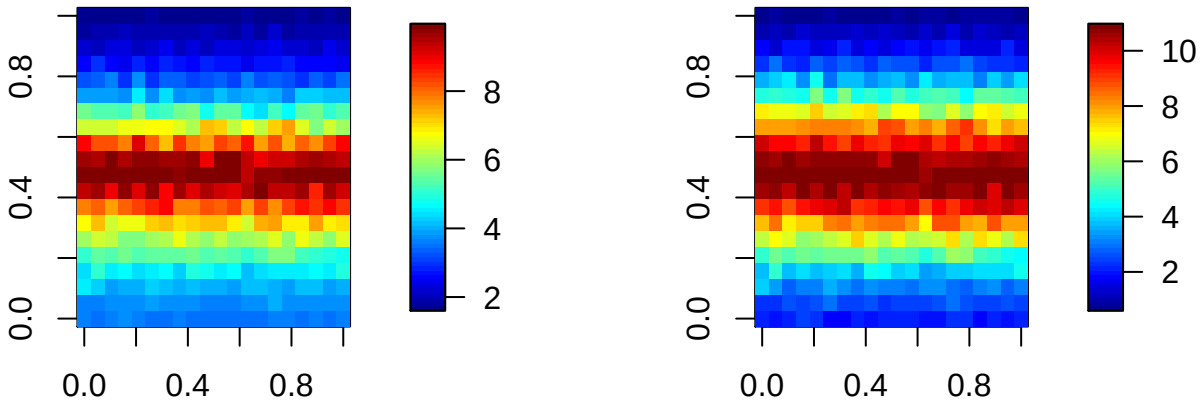
Occupancy Probability, ψ

Definition of ψ

Definition description goes here
Probably need to describe how it's generated in the simulation
As well as how it's estimated in the MSOM
In particular, important to point out that they may or may not match

---

Scatter Plot of Aggregated ψ



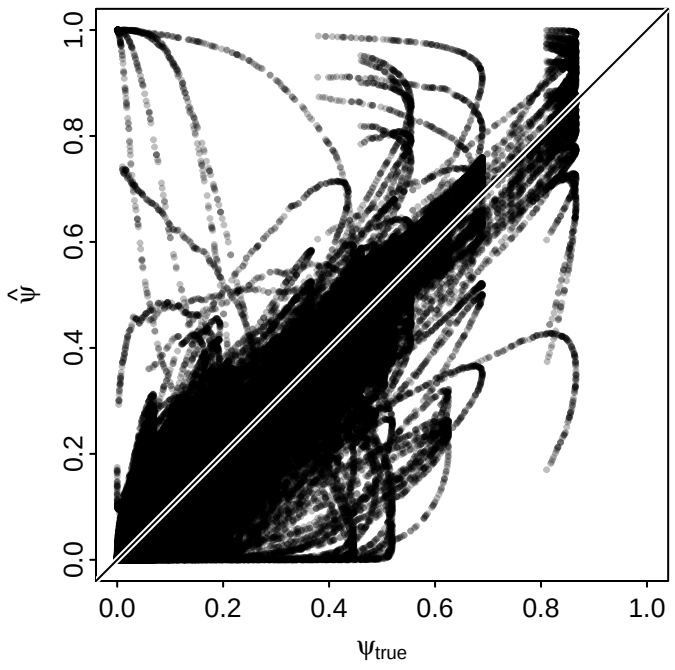


Figure. MSOM estimates of ψ ($\hat{\psi}$) vs. true values of ψ (ψ_{true}). Each point is a ψ value for a particular site-species-year~~, averaged across $r = 8$ simulated replicate observations (i.e., the "true" value is the same, but each simulated replicate has a different outcome of how the same true process was observed)~~. The white and black line is the 1:1 line.

Scatter Plot of $\hat{\psi}$ vs ψ_{true}, split by year and replicate

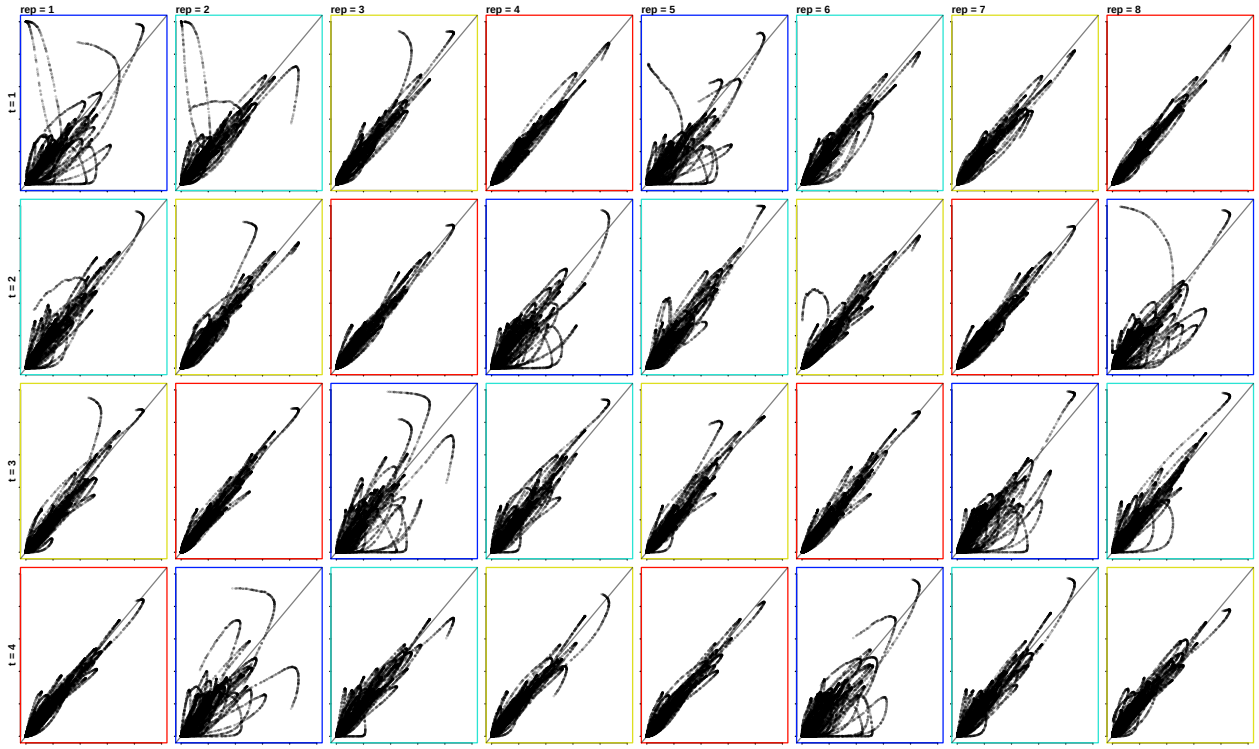


Figure. True (horizontal axes) and MSOM estimates (vertical axes) of occupancy probabilities ($\psi_{j,i,t,r}$) of species i occupying a location j in year t. In our simulation, ψ is a function of individual species characteristics (niche) and the environment, the latter of which changes among years. The simulated (true) outcome of each year was subject to r replicate observations of the true process. Each simulated observation (r) was an independent realization, but the r replicates also differed in the probability that a species would be detected (p): 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 (red; $p_{max} = 0.91$), whereas cool colors indicate that p was low (blue; $p_{min} = 0.26$). The year t of the simulated true process changes across the rows of panels, and the simulated replicate observation r changes across columns. *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. In this simulation, 75% of substrata were sampled, which doesn't influence p, but can add noise to its estimates.*

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 ψ, except that ψ 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).

True Occupancy Response Curves

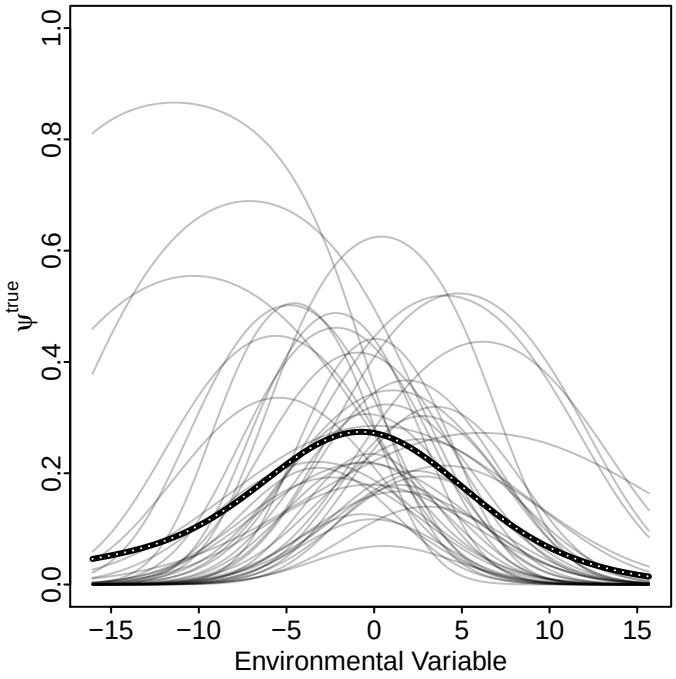


Figure. True simulated response curves. Vertical axis is the value of ψ^{true}, horizontal axis is the value of the environmental variable that, along with species-specific regression parameters, determines ψ^{true}. The thick line is the among-species mean value of ψ^{true} at a given value of the environmental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

Estimated Occupancy Response Curves

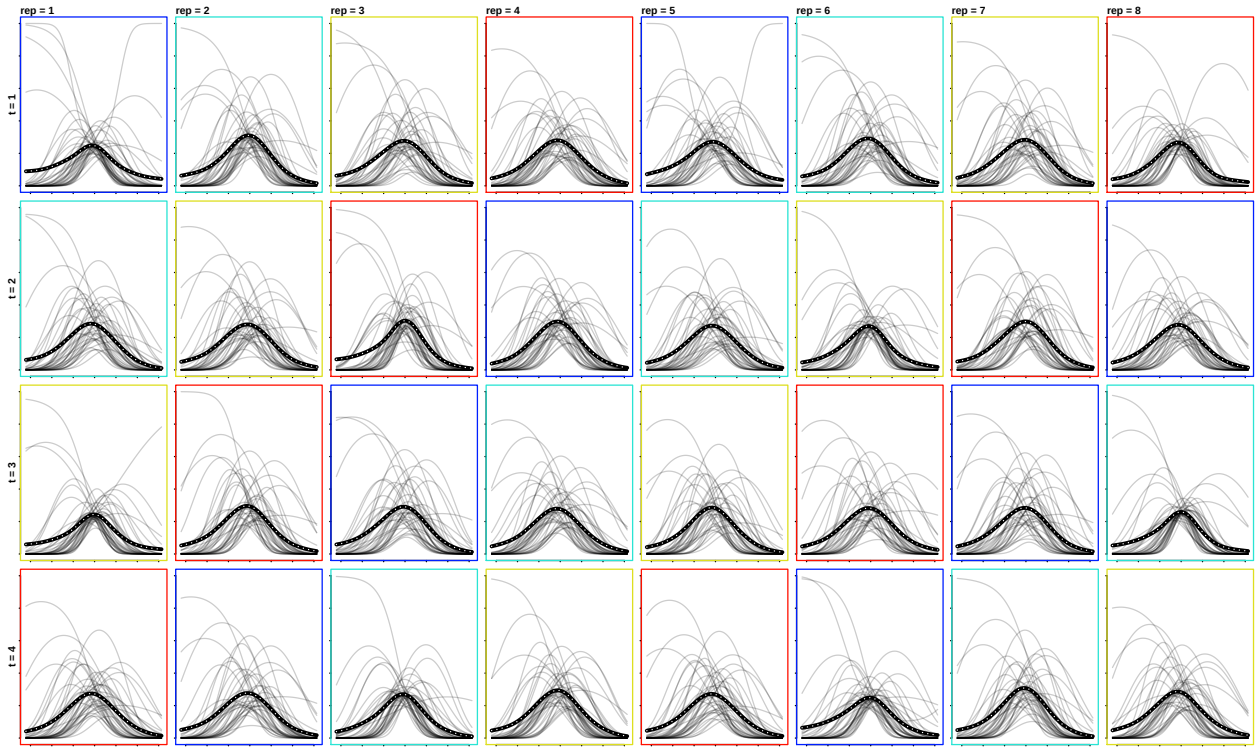


Figure. Response curves of species' probability of occupancy (ψ_i, vertical axis) across the full range of temperatures in the simulation ($min(X) = -16.1$, and $max(X) = 15.7$). The color of the boxes around each panel refer to the among-species average of the probability of detection; warm colors indicate that the mean detection probability is high (red; $p_{max} = 0.91$), whereas cool colors indicate that p was low (blue; $p_{min} = 0.26$). The year t of the simulated true process changes across the rows of panels, and the simulated replicate observation r changes across columns.

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 ψ (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, $logit(p_i) \sim \mathcal{N}(\sqrt{}^{\mu}, \sigma^{\in})$. p_μ 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

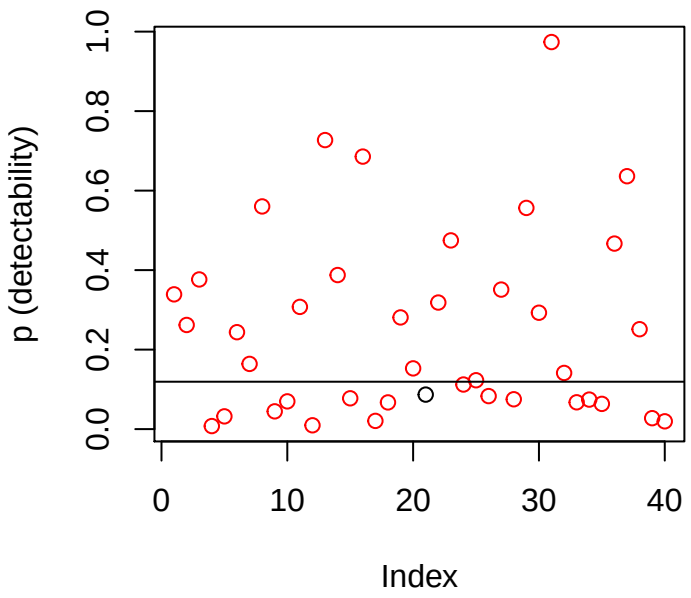


Figure. Probability of being detected, p. Horizontal line is mean probability. Figure only shows results for the first year of the simulation/ observation, and only 1 replicate. Different points are different species. Probability of being detected is a species-specific parameter (does not vary among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

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?

Scatter Plot of $\hat{p}$ vs p_{true}

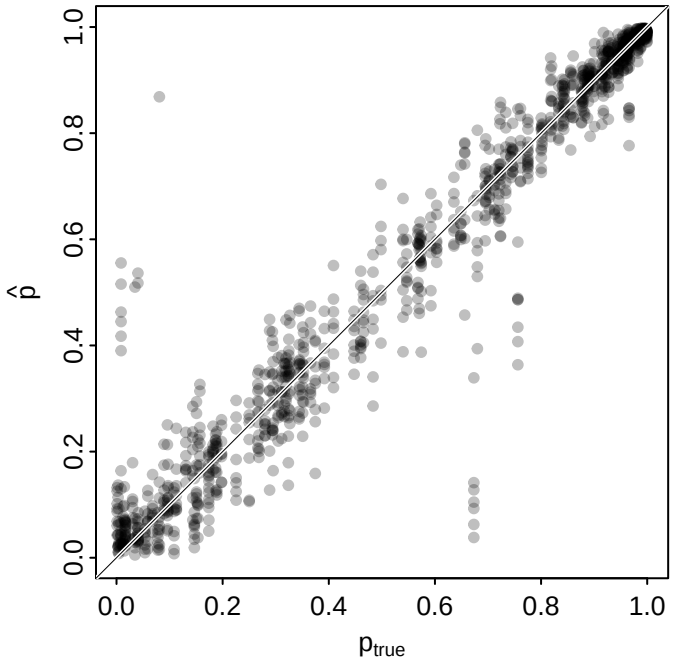


Figure. MSOM estimates (vertical axis) and true values of p_i, the species-specific (i) detection probability. Each point is subscripted by species i, year t, and observation replicate r.

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

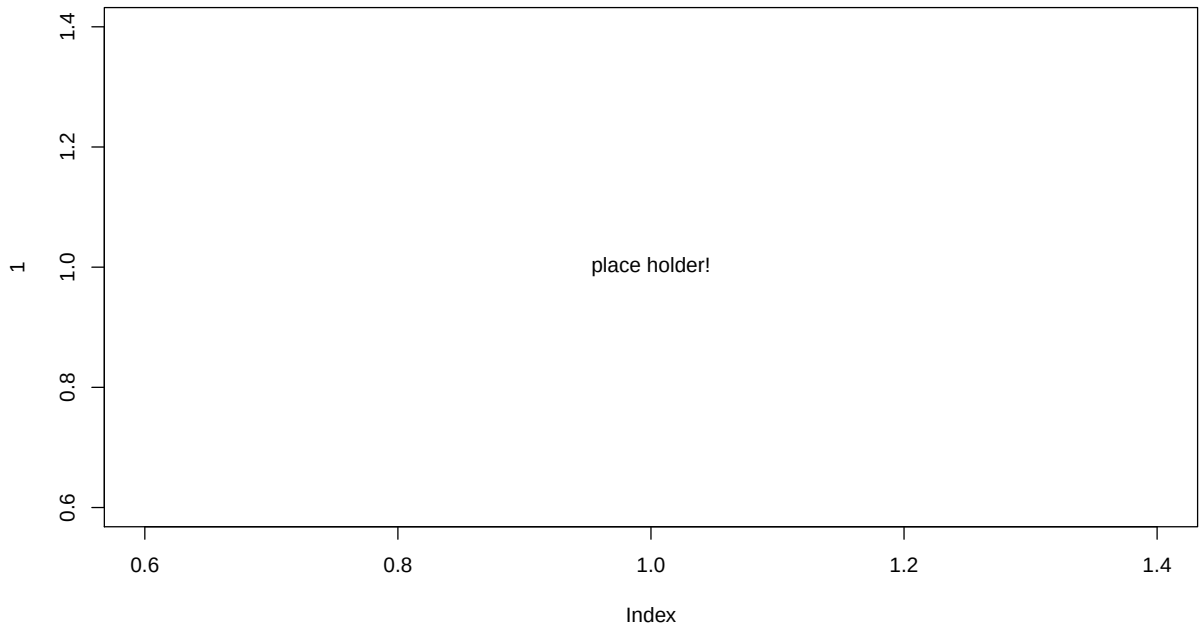


Figure. Caption goes here.

Text explanation goes here

Assessment with Mixed Effects Models

E.g. LME for ψ Evaluation

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 ψ 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 ψ, probability of an individual species being present

```
# ====================
# = LME Model on Psi =
# ====================
# Just exploration/ starting point
library(car)
library(lme4)

blah <- reshape2:::melt.array(psi.true, varnames=c("site","spp","time","rep"), value.name="true", as.is=
blah.hat <- reshape2:::melt.array(psi.hat, varnames=c("site","spp","time","rep"), value.name="hat", as.i
blah <- cbind(blah, hat=blah.hat[,"hat"])

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

(blah.mod <- lmer(hat~true+(1|spp)+(1|time), data=blah))
```

```
## Warning in checkConv(attr(opt, "derivs"), opt$par, ctrl =
## control$checkConv, : Model failed to converge with max|grad| = 0.00539003
## (tol = 0.002)

## Linear mixed model fit by REML ['lmerMod']
## Formula: hat ~ true + (1 | spp) + (1 | time)
##    Data: blah
## REML criterion at convergence: -1373439
## Random effects:
##  Groups   Name        Std.Dev.
##  spp      (Intercept) 0.016385
##  time     (Intercept) 0.002759
```

```
##  Residual             0.063261
## Number of obs: 512000, groups: spp, 40; time, 4
## Fixed Effects:
## (Intercept)         true
##     0.00385      0.95525
```

```
Anova(blah.mod)
```

```
## Analysis of Deviance Table (Type II Wald chisquare tests)
##
## Response: hat
##        Chisq Df Pr(>Chisq)
## true 2378819  1  < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```
