

1

# assess.sim.basic.R

2

Ryan Batt

3

2015-08-22

4

## **Abstract**

5

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

<sup>10</sup> **Contents**

<sup>11</sup>	<b>Conventions and Settings</b>	<b>3</b>
<sup>12</sup>	Dimension Conventions . . . . .	3
<sup>13</sup>	Simulation Settings . . . . .	4
<sup>14</sup>	JAGS Settings for MSOM . . . . .	5
<sup>15</sup>	Assessment Settings . . . . .	5
<sup>16</sup>	<b>Species Richness</b>	<b>6</b>
<sup>17</sup>	Definition of species richness . . . . .	6
<sup>18</sup>	Regional Richness . . . . .	6
<sup>19</sup>	Site Specific Richness ( <code>Nsite</code> ) . . . . .	7
<sup>20</sup>	<b>Occupancy Probability, <math>\psi</math></b>	<b>10</b>
<sup>21</sup>	Definition of $\psi$ . . . . .	10
<sup>22</sup>	Scatter Plot of Aggregated $\psi$ . . . . .	10
<sup>23</sup>	Scatter Plot of $\hat{\psi}$ vs $\psi_{true}$ , split by year and replicate . . . . .	12
<sup>24</sup>	Occupancy Response Curves . . . . .	12
<sup>25</sup>	<b>Probability of Detection, <math>p</math></b>	<b>15</b>
<sup>26</sup>	Definition of $p$ . . . . .	15
<sup>27</sup>	Demo: Effect of MSOM Hierarchy on $p$ . . . . .	16
<sup>28</sup>	Scatter Plot of $\hat{p}$ vs $p_{true}$ . . . . .	17
<sup>29</sup>	Scatter Plot of $\hat{p}$ vs $p_{true}$ , split by year and replicate . . . . .	18
<sup>30</sup>	<b>Assessment with Mixed Effects Models</b>	<b>19</b>
<sup>31</sup>	E.g. LME for $\psi$ Evaluation . . . . .	19

33 **Conventions and Settings**

34 **Dimension Conventions**

35 **Summary**

36 1. Site ( $j = 1, 2, \dots, j_{max} = 20 \times 20 = 400$ )

- 37 • Sites are unique combinations of latitude and longitude

- 39 • The spatial arrangement of sites is *not* arbitrary, although importance depends on settings (see  
40 `dynamic` below)

- 42 • The environmental variable  $X$  varies among sites (and years, below)

43 2. Sub-sites ( $k = 1, 2, \dots$ )

- 44 • Sub-sites are only relevant to the “observation” process
- 45 • Each site has the same number of possible sub-sites, but the number of sub-sites observed can vary
- 46 • In this simulation,  $k_{max} = 4$ ,  $k_{min}^{observed} = 4$ , and  $k_{max}^{observed} = 4$

- 48 • Substrata are primarily useful for determining  $p$ , the **detection probability**

50 3. Species ( $i = 1, 2, \dots i_{max} = R = 20$ )

- 51 • Does not include “augmented” species
- 52 • For this MSOM analysis, the species array was padded with 10 0’s

53 4. Time ( $t = 1, 2, \dots 2$ )

- 54 • Time is primarily used to vary the parameters controlling the “true” process
- 55 • When those parameters don’t change, time provides independent\*realizations of the same “true”  
56 process

57 — \*Note: only when `dynamic=FALSE` in `sim.spp.proc`

58 5. Replicates ( $r = 4$ )

- 59 • Replicates are *simulated* repeated human observations of the same *realization* of the “true” process  
60 at Time $_t$
- 61 • Replicates are used to vary the parameters that control the “observation” process
- 62 • When those parameters don’t change, each replicate provides an independent\*realization of the  
63 same “observation” process

64 **In Code**

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

67 In my code, I've tried to be consistent in my use of these indices to describe arrays, matrices, and rasters.  
 68 Rows are dimension 1, columns dimension 2, etc. The precedent of subscripted dimensions follows the numeric  
 69 ordering of the list above. E.g., in the matrix  $X_{i,t}$  each row will refer to a different species, and each column  
 70 a different year (note that site $_j$  is skipped, so species $_i$  is "promoted" to dimension 1, the row.). By default, R  
 71 fills matrices and arrays by column, whereas the **raster** package fills them by row. In most cases where an R  
 72 object needs to split sites into the lat/ lot components, I make use of the **raster** package. Therefore, the  
 73 numbering of the sites proceeds row-wise, where each site is numbered according to the order in which it is  
 74 filled, as in this  $2 \times 3$  matrix:  $J = \begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{pmatrix}$   
 75 Note that even though this matrix is numbered row-wise, it is still indexed as  $J_{row,column}$ , such that  $J_{1,2} = 2$ .  
 76 As mentioned previously, this information is primarily important for understanding the code involved with  
 77 this project, and in most cases it is not crucial to be explicitly aware of the spatial arrangement of sites.

78

---

## 79 Simulation Settings

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

```

86 ## Dimensions: 400, 20, 2
87 ## grid.h = 20
88 ## grid.w = 20
89 ## grid.t = 2
90 ##
91 ## Number Species Possible (ns):
92 ## 20
93 ## Total Species Richness:
94 ## 20
95 ## Total Observed Species Richness:
96 ## 20
97 ##
98 ## Annual Species Richness:
99 ##          Min. 1st Qu. Median Mean 3rd Qu. Max.
100 ## All cells   19    19.25   19.5 19.50   19.75   20
101 ## One cell    0     2.00    3.0  3.29    5.00    11
102 ##
103 ##
104 ## Observed Annual Species Richness:
```

```

105 ##          Min. 1st Qu. Median   Mean 3rd Qu. Max.
106 ## All cells    19    19.25    19.5 19.500    19.75    20
107 ## One cell     0     1.00     2.0  2.436    3.00    11

```

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

year	mu.v0	sigma.v0
1	0	2
2	4	2

---

112

### 113 JAGS Settings for MSOM

nChains	nIter	n0s	nSamples
3	50000	10	500

---

114

### 115 Assessment Settings

#### 116 Central Tendency

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

---

120

121 **Species Richness**

122 **Definition of species richness**

123 Species richness is the number of different species, or more generically, unique taxa. The point is moot in the  
124 simulation study, and in the empirical trawl data it refers to species.  
125 Estimates of richness can be made spatially or temporally explicit (or neither, or both), but obviously a

126

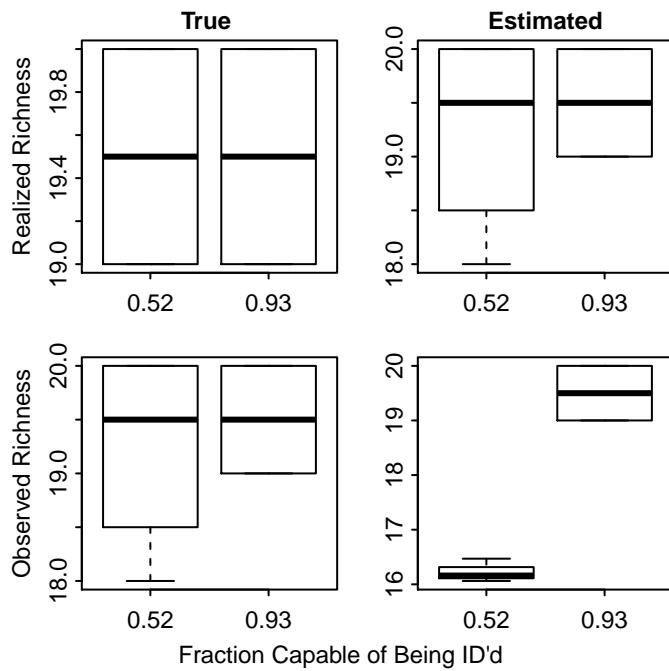
---

127 **Regional Richness**

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

130 **Richness Boxplots**

131 With the boxplots we're mostly looking to see if the estimates of richness vary with the mean probability of  
132 detection,  $p$ . In the empirical data, we know that taxonomic identification changed over time (it improved;  
133 generally, more species were ID'd in later years). We also suspect that gear might change, which affects  
134 the probability of observing a species. The "Fraction Capable of Being ID'd" category in the boxplots is  
135 essentially the cross-species average of  $p$ .



136

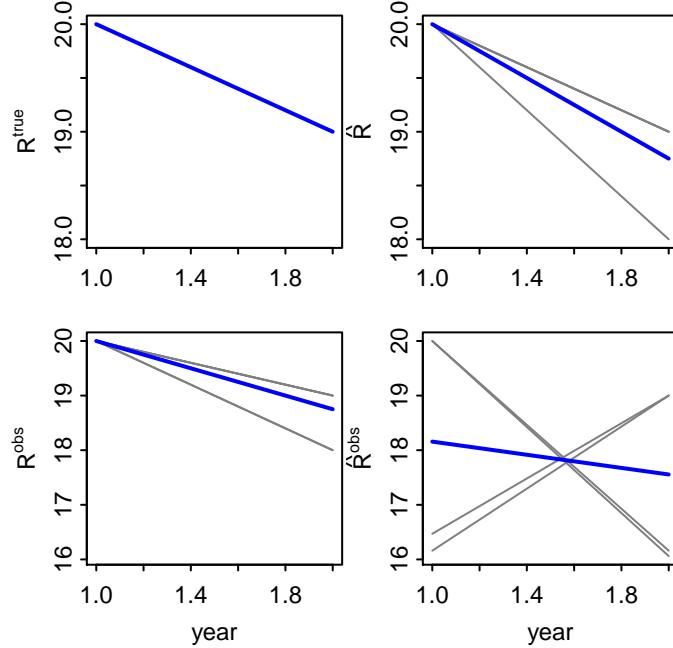
137 **Figure.** Boxplots of species richness. Numeric groupings indicate the average value of  $p$  across species during  
138 a given year-replicate combination. The panels in the left column are the true simulated values, and the  
139 panels on the right are the corresponding MSOM estimates. The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

<sup>140</sup> richness or MSOM estimates of the richness. The bottom row's panels are the simulated observed values of  
<sup>141</sup> richness (the response variable in the MSOM) and the MSOM estimates of the observed values.

<sup>142</sup>

---

<sup>143</sup> **Richness Time Series**



<sup>144</sup>

<sup>145</sup> **Figure.** Time series of richness. Gray lines are individual replicates. Blue lines are averages. Note that  
<sup>146</sup> detection probabilities ( $p_{i,t,r}$ , see [simulation settings above](#), as well as [definition of  \$p\$  below](#)) change over time,  
<sup>147</sup> and their temporal ordering differs among replicates. Text explanation goes here

<sup>148</sup> Need explanations for how each panel was calculated.

<sup>149</sup> 1.  $R^{true}$  is straightforward

<sup>150</sup>

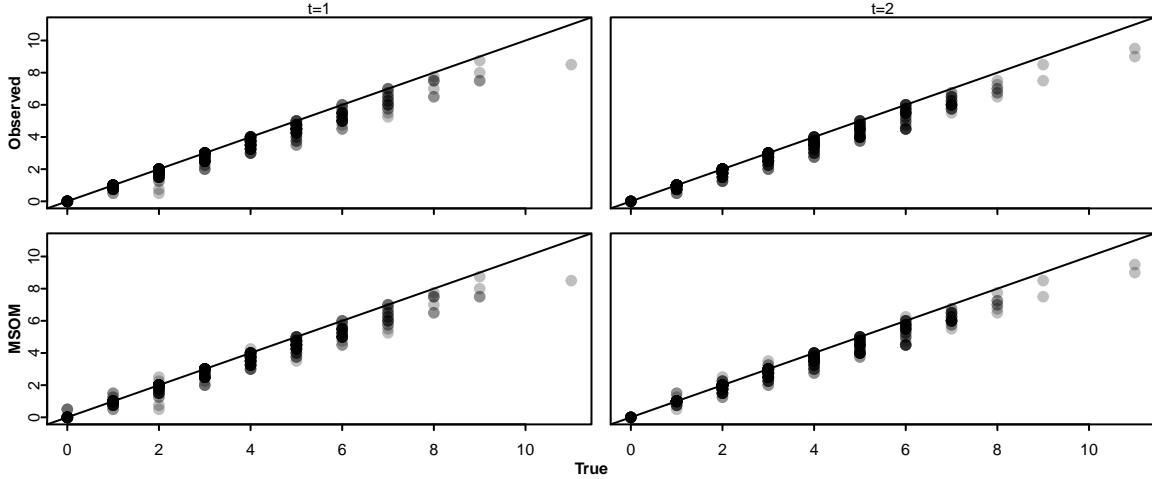
<sup>151</sup> 2.  $\hat{R}$  is from  $\sum_{i=1}^{R^{true}} \max(\hat{Z}_{i,\nabla j \in J})$ ; and to be clear,  $\hat{R}$  does not include the “unobserved” species introduced  
<sup>152</sup> to the MSOM occurrence matrix ( $Y$ )

<sup>153</sup>

---

<sup>154</sup> **Site Specific Richness (Nsite)**

<sup>155</sup> **Scatter Plots of Nsite Split by Year**

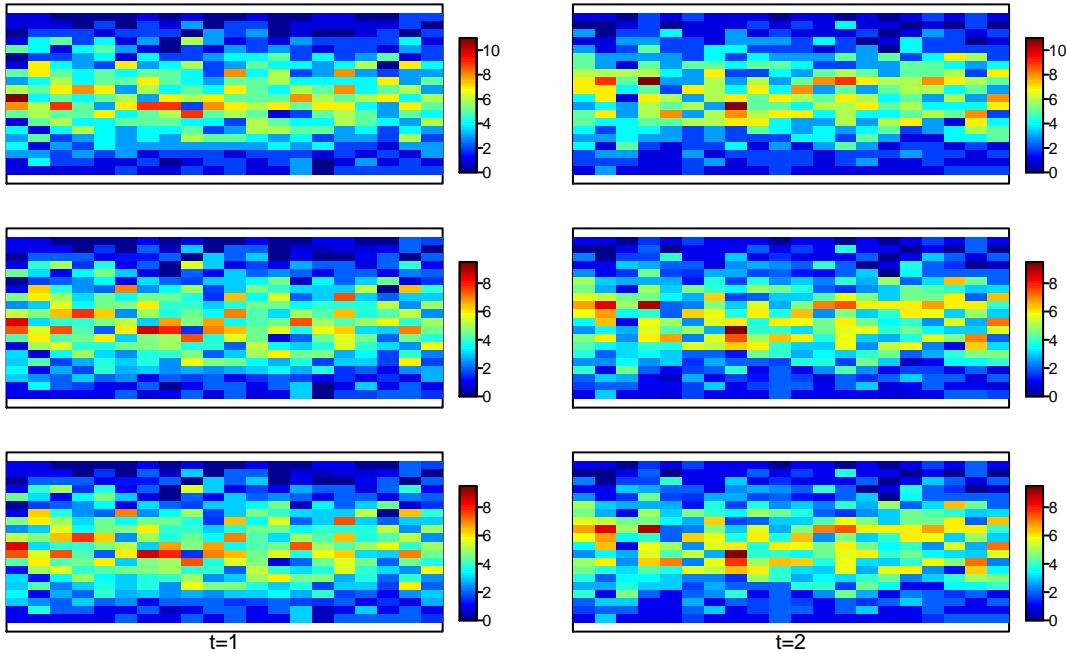


156    **Figure.** Site-specific richness ( $N_{site}$ ,  $N_j$ ) from simulated observations (vertical axis, top row;  $N_j^{obs}$ ) and  
 157    from MSOM estimates (vertical axis, bottom row,  $\hat{N}_j$ ) vs true site-specific richness (horizontal axis;  $N_j^*$ ).  
 158    The panel columns delineate the years of the simulation. Each point is site-specific ( $j = 20 \times 20 = 400$ )  
 159    species richness that has been averaged over the simulated replicate observations ( $r = 4$ ).  
 160

161

---

162    **Maps of Richness (space and time)**

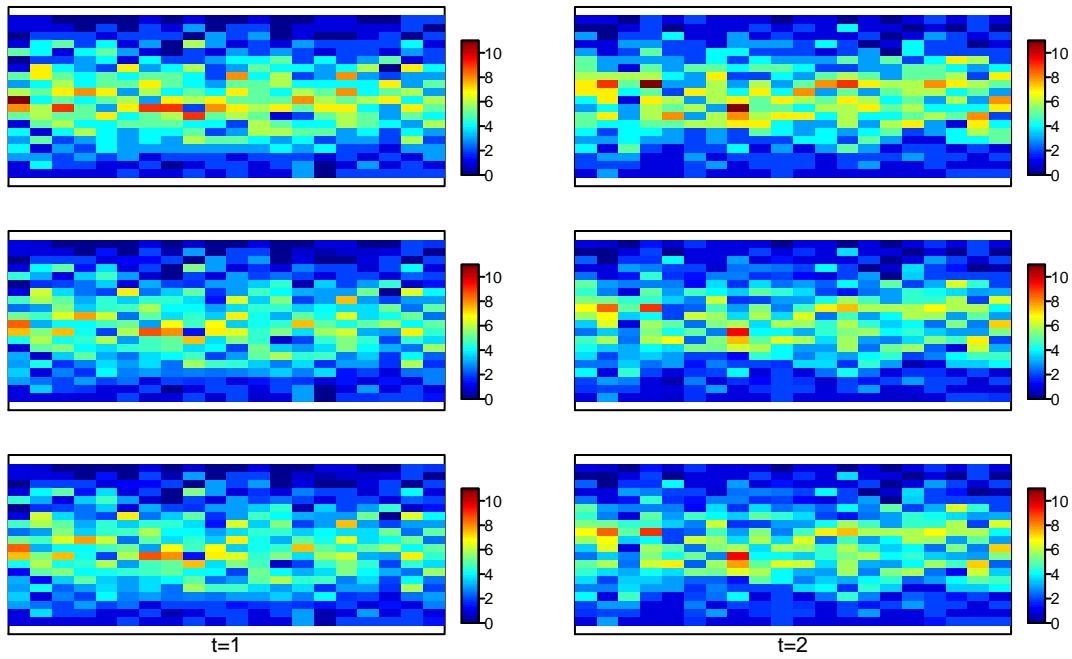


163    **Figure.** Maps of site- and year-specific species richness ( $N_{site}$ ) from the simulation of the True process  
 164    (top row), simulation of the Observed process (middle row), and the MSOM estimates (bottom row). X-axis  
 165    and Y-axis indicate position in 2 dimensional space; it is important to note that the environmental variable  
 166    changes linearly across the y-axis, and randomly (and much less) across the x-axis. The different columns  
 167    represent separate years. The environmental variable changes linearly among years (the rate of change is the  
 168

169 same for all x-y locations). Colors indicate species richness (warm colors are higher richness than cool colors),  
170 averaged over the simulated replicate observations ( $r = 4$ ). Horizontal and vertical axes Each row of panels  
171 is scaled independently, columns within a row are scaled equally.

172

---



173

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

175 Text explanation goes here

176

---

<sub>177</sub> **Occupancy Probability,  $\psi$**

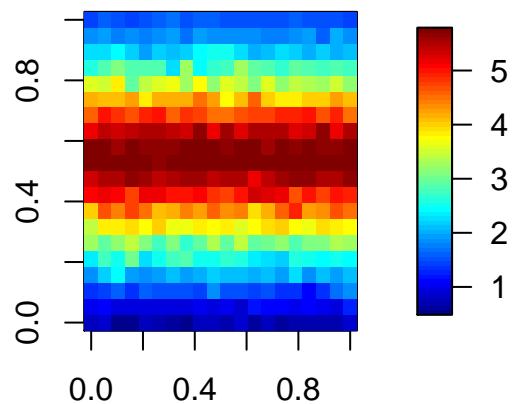
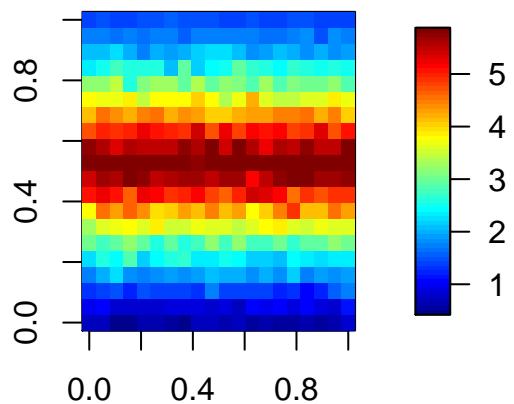
<sub>178</sub> **Definition of  $\psi$**

- <sub>179</sub> Definition description goes here
- <sub>180</sub> Probably need to describe how it's generated in the simulation
- <sub>181</sub> As well as how it's estimated in the MSOM
- <sub>182</sub> In particular, important to point out that they may or may not match

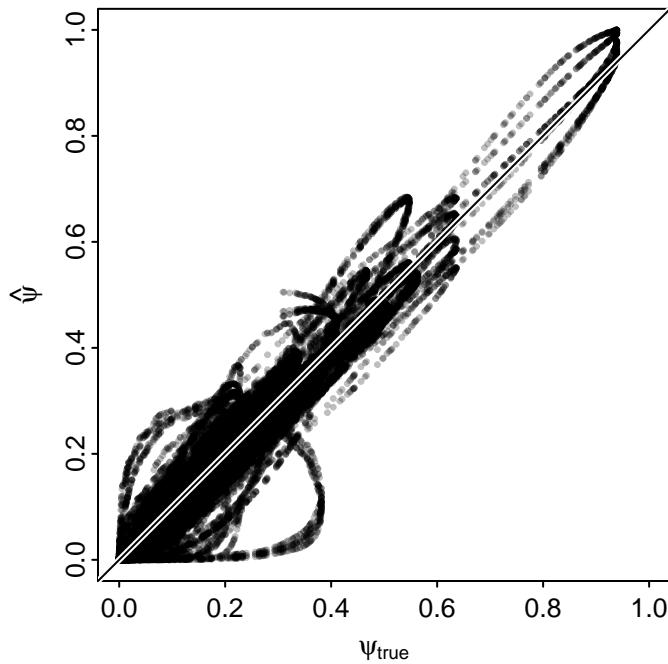
<sub>183</sub>

---

<sub>184</sub> **Scatter Plot of Aggregated  $\psi$**



<sub>185</sub>



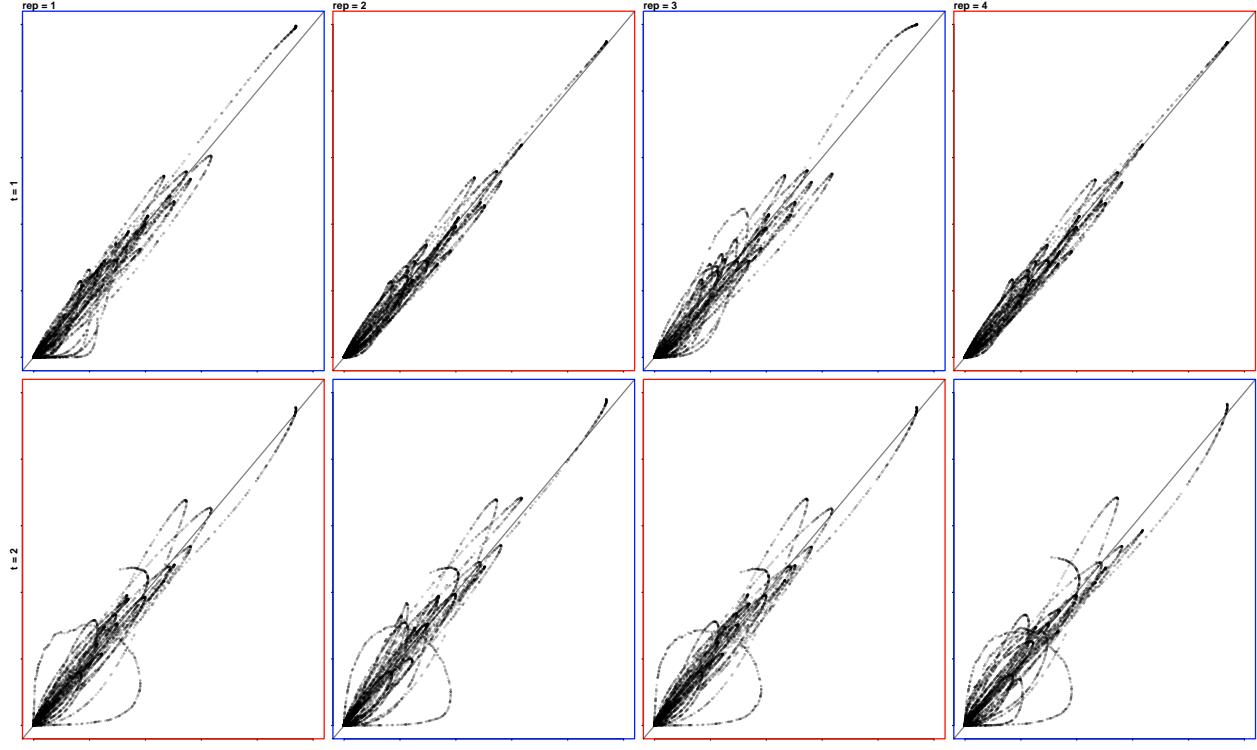
186

187 **Figure.** MSOM estimates of  $\psi$  ( $\hat{\psi}$ ) vs. true values of  $\psi$  ( $\psi_{true}$ ). Each point is a  $\psi$  value for a particular  
 188 site-species-year, averaged across  $r = 4$  simulated replicate observations (i.e., the “true” value is the same;  
 189 but each simulated replicate has a different outcome of how the same true process was observed). The white  
 190 and black line is the 1:1 line.

191

---

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



193 **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,  $\psi$  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;  $p_{max} = 0.93$ ), whereas cool colors indicate that  $p$  was low (blue;  $p_{min} = 0.52$ ). 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, 100% of substrata were sampled, which doesn't influence  $p$ , but can add noise to its estimates.

205

---

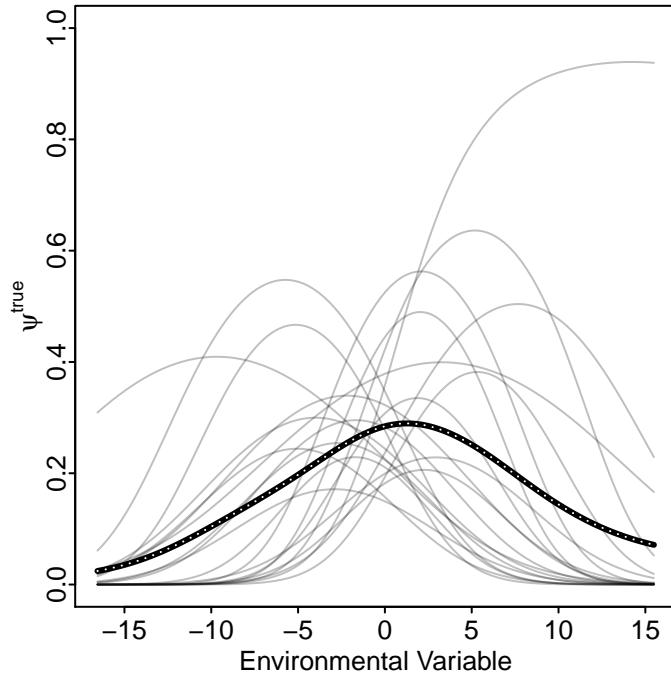
206 Occupancy Response Curves

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

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

207 Therefore, these curves are tantamount to values of  $\psi$ , except that  $\psi$  generally pertains to a simulated,  
208 observed, or true occupancy probability, whereas the occupancy probability in the response curves is calculated  
209 over hypothetical conditions (i.e., over hypothetical values of the environmental gradient  $X$ ).

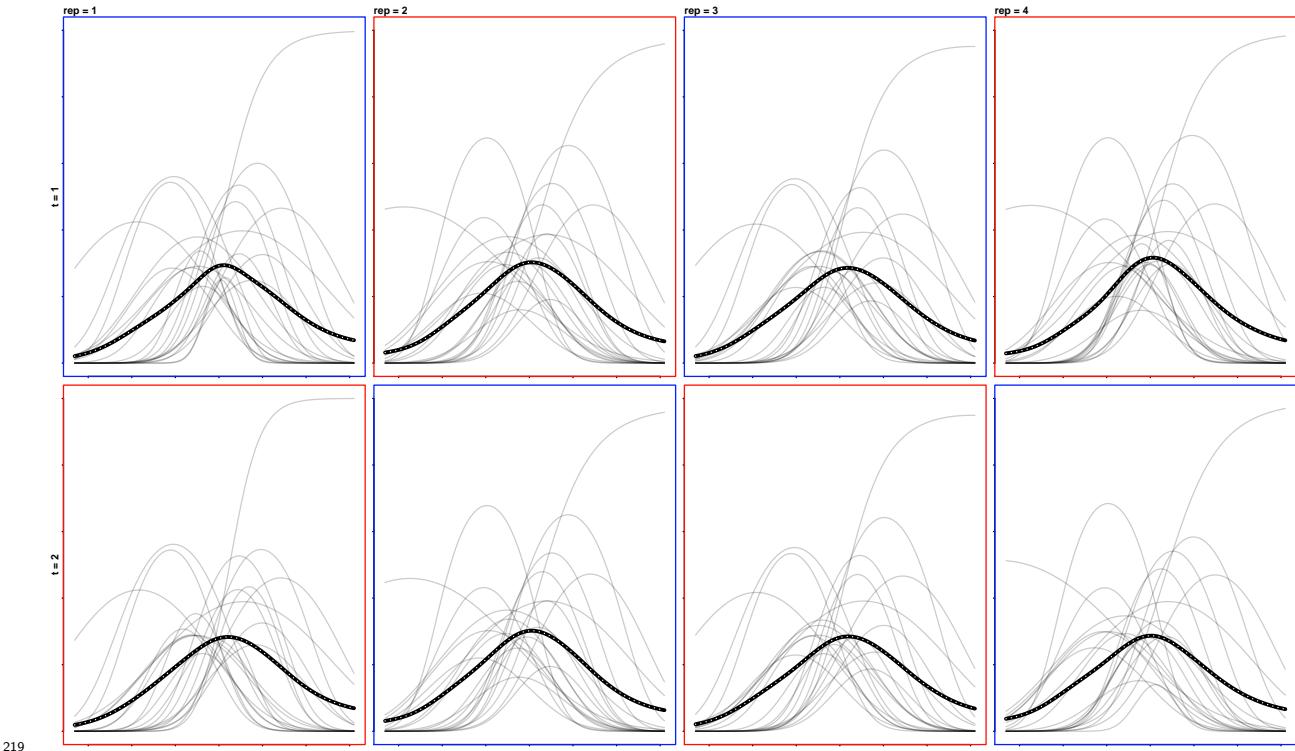
210 **True Occupancy Response Curves**



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

214 In the response curve, the values of the environmental variable are an arbitrary gradient, and do not necessarily  
215 correspond to what was observed in the simulated environment (although they are intended to cover the  
216 same range).

217 **Estimated Occupancy Response Curves**



219  
220 **Figure.** Response curves of species' probability of occupancy ( $\psi_i$ , vertical axis) across the full range of  
221 temperatures in the simulation ( $\min(X) = -16.6$ , and  $\max(X) = 15.5$ ). The color of the boxes around each  
222 panel refer to the among-species average of the probability of detection; warm colors indicate that the mean  
223 detection probability is high (red;  $p_{\max} = 0.93$ ), whereas cool colors indicate that  $p$  was low (blue;  $p_{\min} =$   
224 0.52). The year  $t$  of the simulated true process changes across the rows of panels, and the simulated replicate  
225 observation  $r$  changes across columns.

226

---

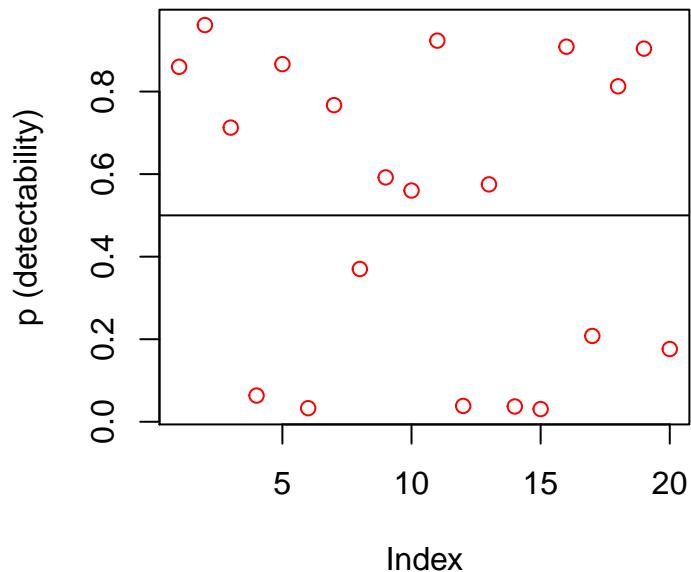
227 **Probability of Detection,  $p$**

228 **Definition of  $p$**

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

- 232 1. The probability that site  $j$  is occupied by species  $i$ ; this is  $\psi_{j,i}$ 
  - 233 •  $\psi_{j,i}$  is a function of species-specific niche and an environmental variable that changes over space  
234 and time
  - 235 •  $Z_{j,i}$  is the species- and site-specific richness, which is a function of  $\psi$  (given that we're only talking  
236 about species that are in the pool of possible species, determined by  $w_i$ )
- 238 2. A species-specific ( $i$ ) chance of being identified (`taxChance`), given that it is present in a location that  
239 was sampled (i.e., detectability does not reflect the probability of sampling a place); this detectability  
240 parameter is  $p_i$ 
  - 241 • Detectability changed between years.
  - 242 • In a given year,  $\text{logit}(p_i) \sim \mathcal{N}(\mu, \sigma^2)$ .  $\mu$  changed between years (taking on values of 0, and 4),  
243  $\sigma^2 = 2$  in all years.
  - 244 • The value of  $p$  only changes between species (and years), but the observation process occurs at the  
245 substratum ( $k$ ) level. Thus, the parameter is really  $p_{j,k,i}$ , but for a given  $i$ , all  $p_{j,k}$  are constant. I  
246 represent this probability as  $p_i$  with the understanding that this value is repeated over space.
  - 247 •  $Y_{j,i}$  is the observed version of  $Z_{j,i}$ .
  - 248 •  $Y_{j,i} \sim \text{Bern}(p_i \times Z_{j,i})$ .
    - 249 – Note: Because  $p$  is actually subscripted to  $k$ , the  $Y$  are also actually subscripted to  $k$ . Maybe  
250 leaving these subscripts out is making things more confusing. I've only excluded them to  
251 emphasize how parameters are estimated.
  - 252 • Our data about species presence/ absence correspond to  $Y_{j,i}$ . So it might be useful to think of the  
253 MSOM as estimating  $\hat{Y}_{j,i}$ , which is compared to the observed data  $Y_{j,i}^{obs}$ .

255 Demo: Effect of MSOM Hierarchy on  $p$

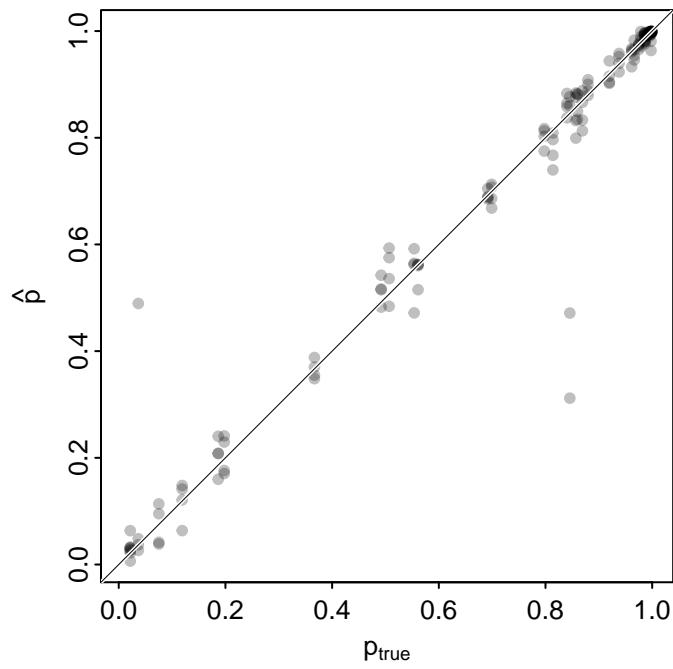


256

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

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

<sup>267</sup> Scatter Plot of  $\hat{p}$  vs  $p_{true}$



<sup>268</sup>

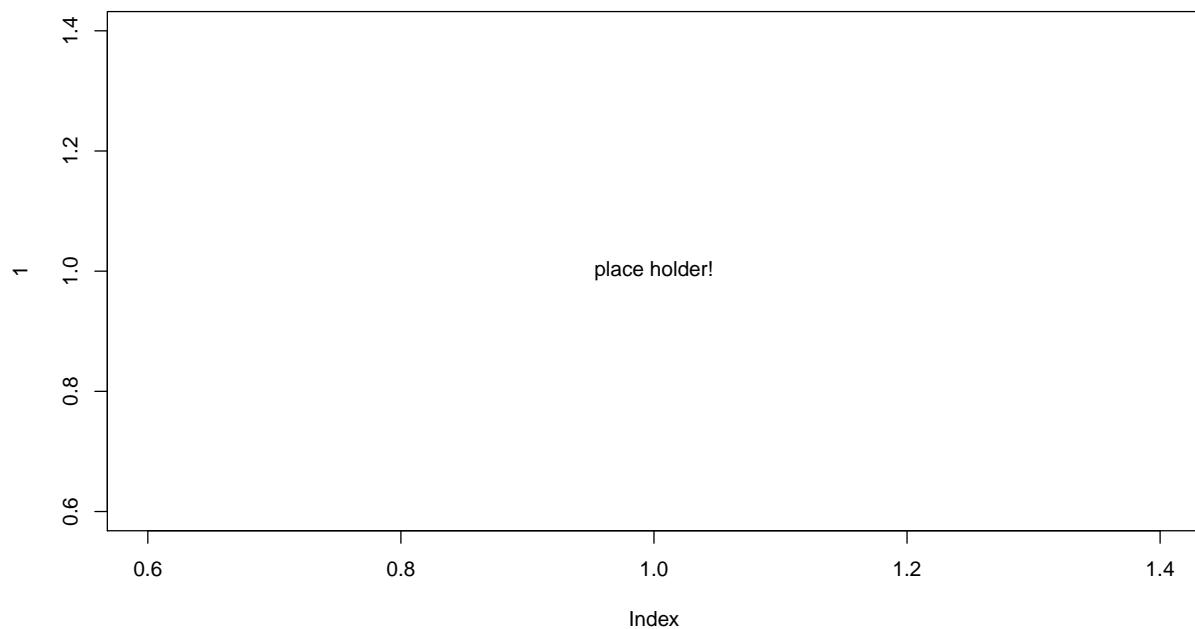
<sup>269</sup> **Figure.** MSOM estimates (vertical axis) and true values of  $p_i$ , the species-specific ( $i$ ) detection probability.

<sup>270</sup> Each point is subscripted by species  $i$ , year  $t$ , and observation replicate  $r$ .

<sup>271</sup>

---

<sup>272</sup> Scatter Plot of  $\hat{p}$  vs  $p_{true}$ , split by year and replicate



<sup>273</sup>

<sup>274</sup> **Figure.** Caption goes here.

<sup>275</sup> Text explanation goes here

<sup>276</sup>

## 277 Assessment with Mixed Effects Models

### 278 E.g. LME for $\psi$ Evaluation

- 279 **Motivation:** MSOM skill might differ across dimensions, trying to figure out what patterns I should expect  
280 to pick out (spatial patterns in richness, temporal?) E.g., Is the correlation between MSOM and True the  
281 same comparing across sites as comparing across years? Species, reps, also.
- 282 **Motivation:** What factors influence MSOM skill in a given dimension? E.g., Skill in finding differences in  $\psi$   
283 across species may depend on  $p$ , the chance of being identified. If  $p$  changes among years, might also explain  
284 Read more about [specifying mixed effects models using lmer in R here](#)
- 285 This example is looking at  $\psi$ , 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=TRUE)
blah.hat <- reshape2:::melt.array(psi.hat, varnames=c("site","spp","time","rep"), value.name="hat", as.is=TRUE)
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))

286 ## Linear mixed model fit by REML ['lmerMod']
287 ## Formula: hat ~ true + (1 | spp) + (1 | time)
288 ##   Data: blah
289 ## REML criterion at convergence: -227716.4
290 ## Random effects:
291 ##   Groups    Name        Std.Dev.
292 ##     spp      (Intercept) 0.010779
293 ##     time     (Intercept) 0.004635
294 ##   Residual           0.040802
295 ## Number of obs: 64000, groups: spp, 20; time, 2
296 ## Fixed Effects:
297 ##   (Intercept)      true
```

```
298 ## -0.001704 0.988997
Anova(blah.mod)

299 ## Analysis of Deviance Table (Type II Wald chisquare tests)
300 ##
301 ## Response: hat
302 ##      Chisq Df Pr(>Chisq)
303 ## true 1043475 1 < 2.2e-16 ***
304 ## ---
305 ## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

306

---