

1 Predicting biodiversity dynamics in response to environmental
2 change

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

4 Ryan Batt

5 2015-08-23

6 **Abstract**

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

¹² Contents

¹³	Introduction	4
¹⁴	Overview	4
¹⁵	The Simulation	4
¹⁶	Multispecies Occupancy Models (MSOMs)	5
¹⁷	Conventions and Settings	7
¹⁸	Dimension Conventions	7
¹⁹	Settings	8
²⁰	Species Richness	11
²¹	Definition of species richness	11
²²	Regional Richness	11
²³	Site Specific Richness (N_{site})	13
²⁴	Occupancy Probability, ψ	16
²⁵	Definition of ψ	16
²⁶	Scatter Plot of Aggregated ψ	16
²⁷	Scatter Plot of $\hat{\psi}$ vs ψ_{true} , split by year and replicate	17
²⁸	Occupancy Response Curves	17
²⁹	Probability of Detection, p	21
³⁰	Definition of p	21
³¹	Demo: Effect of MSOM Hierarchy on p	21
³²	Scatter Plot of \hat{p} vs p_{true}	23
³³	Scatter Plot of \hat{p} vs p_{true} , split by year and replicate	23
³⁴	Assessment with Mixed Effects Models	25
³⁵	Describe Motivation for Mixed Effects Models	25
³⁶	Example LMER Analysis for ψ	25

37	Report Generation Notes	27
38	Check that all Figures were Cited	27
39	R Session Information	27
40	Date Document Last Compiled	28

41

42 Introduction

43 Overview

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

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

56 Can we estimate the dynamics of species richness from trawl data using an MSOM? It's a hard question to
57 answer because we can never know the "truth" for sure, but we can get an idea of how reliable our analysis
58 is by simulating fake data, for which we know true values because we created them. The trawl data set is
59 generated by two distinct processes: Nature's data generating process (NDGP), and the process by which
60 humans observe the result of NDGP. So we ask: to what extent is the accuracy of estimates from an MSOM
61 dependent on characteristics of NDGP, and in particular, the way in which we observe the result of NDGP?
62 The strategy for answering this question is to simulate fake data where we approximate Nature but gain
63 knowledge of "truth", "observe" the results of the true process, then try to recover the true species richness
64 from these simulated data.

65 The Simulation

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

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

75 A second level of the simulation mimics human observation of NDGP — what we do when we collect data.
76 This process was simulated by assigning each species has a unique probability of being observed or "detected"
77 (this variable is p). The observation process gets several attempts at observing a given species in a given grid
78 cell; think of this as subdividing each site into subsites, and when you visit each subsite you have probability

79 p of observing a particular species (each species has its own p). Depending on the settings used in the analysis
80 that this document summarizes, the maximum number of subsites can vary, as can the number of subsites per
81 site (OK, fine; the maximum number of subsites in this version is 4, the number of subsites per site varied
82 between 1 and 4, and overall 50% of total possible subsites were sampled).

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

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

95 Multispecies Occupancy Models (MSOMs)

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

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

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

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

₁₁₇ varies among space because of the environmental variable, and in a given location it varies among species
₁₁₈ because of the regression parameters.

₁₁₉

120 Conventions and Settings

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

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

131 Dimension Conventions

132 Summary

133 1. Site ($j = 1, 2, \dots, j_{\max} = 8 \times 8 = 64$)

- 134 • Sites are unique combinations of latitude and longitude
- 135
- 136 • The spatial arrangement of sites is *not* arbitrary, although importance depends on settings (see
- 137 *dynamic* below)
- 138
- 139 • The environmental variable X varies among sites (and years, below)

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

- 141 • Sub-sites are only relevant to the “observation” process
- 142 • Each site has the same number of possible sub-sites, but the number of sub-sites observed can vary
- 143 • In this simulation, $k_{\max} = 4$, $k_{\min}^{\text{observed}} = 1$, and $k_{\max}^{\text{observed}} = 4$
- 144
- 145 • Substrata are primarily useful for determining p , the **detection probability**

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

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

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

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

154 — *Note: only when *dynamic=FALSE* in *sim.spp.proc*

155 5. Replicates ($r = 4$)

- 156 • Replicates are *simulated* repeated human observations of the same *realization* of the “true” process
157 at Time $_t$
- 158 • Replicates are used to vary the parameters that control the “observation” process
- 159 • When those parameters don’t change, each replicate provides an independent* realization of the
160 same “observation” process

161 **In Code**

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

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

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

175

176 **Settings**

177 **Simulation Settings**

178 I created a class called "spp", which has methods for **print()**. The **Dimensions** are the number of sites, the
179 number of species, then the number of years.

180 Also **printed** are some richness summary statistics. **All cells** refers to the collective richness over all j
181 taken together. The meaning of **One cell** differs slightly between the true and observed printouts: in the
182 true printout the richness is of a particular site (j), and in the observed printout it is of a particular sub-site
183 (k).

184 **## Dimensions:** 64, 50, 4
185 **## grid.h** = 8
186 **## grid.w** = 8
187 **## grid.t** = 4
188 **##**
189 **## Number Species Possible (ns):**

```

190 ## 50
191 ## Total Species Richness:
192 ## 50
193 ## Total Observed Species Richness:
194 ## 49
195 ##
196 ## Annual Species Richness:
197 ##      Min. 1st Qu. Median Mean 3rd Qu. Max.
198 ## All cells   49     49    49 49.25 49.25   50
199 ## One cell    8     17    22 22.10 28.00   37
200 ##
201 ##
202 ## Observed Annual Species Richness:
203 ##      Min. 1st Qu. Median Mean 3rd Qu. Max.
204 ## All cells   44    47.75 49 47.750 49    49
205 ## One cell    0     0.00   0  6.979 14    33

```

206 In the MSOM, detectability (p_i) is determined in the form of a logistic regression, which currently only
207 has an intercept (v_0) as predictor (so just a mean). That intercept varies among species (i.e., $v_{0,i}$), and
208 that variation is generated by drawing each individual species's intercept ($v_{0,i}$) from a parent distribution:
209 $v_{0,i} \sim \mathcal{N}(\mu_{v_0}, \sigma_{v_0}^2)$. See [section about \$p\$](#) for more info.

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

210

211 Settings for JAGS & MSOM

nChains	nIter	n0s	nSamples
3	50000	10	500

212 In the table above, `nChains`, `nIter`, and `nSamples` are all variables that are strictly pertinent to the Bayesian
213 analysis carried out in JAGS. The `n0s` value refers to the the degree of “data augmentation”. In this process,
214 you add extra species to the data set, and say that they were never observed. For our purposes, this is
215 employed for purely technical reasons, although it can be used to extra further inferences about species
216 richness.
217 The posterior samples from JAGS consist of 500 samples (above). However, to save memory and hard drive

218 space, I have often only saved measures of central tendency for each of these. In this assessment, I have
219 performed all calculations on the **centralT=median** of the posterior samples.

220

221 **Species Richness**

222 **Definition of species richness**

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

225 Estimates of richness (R) can be made spatially or temporally explicit (or neither, or both). In the following
226 figures, different levels of aggregation are performed – for most figures R is split by year (this is true for all
227 figures but [Figure 1](#)). [Figure 2](#) emphasizes temporal dynamics and keeps replicates separated, but aggregates
228 over space (the j sites). [Figure 3](#) doesn't aggregate over space or time, but it does aggregate over “replicate”
229 observations; importantly, while the figure does present any spatial aggregation, it does not retain the spatial
230 relationship (you can't tell which sites are next to others). The final two figures of the section ([Figure 4](#) and
231 [Figure 5](#)) are similar to the previous figure, except that spatial relationship among points is retained via a
232 heatmap representation.

233 None of these estimates of richness include the 10 species that were part of the “data augmented”/ “adding
234 0's” process. Richness values can either be true (true simulated NDGP; R^{true}), observed (true simulated
235 human observation of NDGP; R^{obs}), or MSOM estimates of one of those two (\hat{R}^{true} or \hat{R}^{obs}).

236

237 **Regional Richness**

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

240 **Richness Boxplots**

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

246

247 **Richness Time Series**

248 Text explanation goes here

249 Need explanations for how each panel was calculated.

250 1. R^{true} is straightforward

251

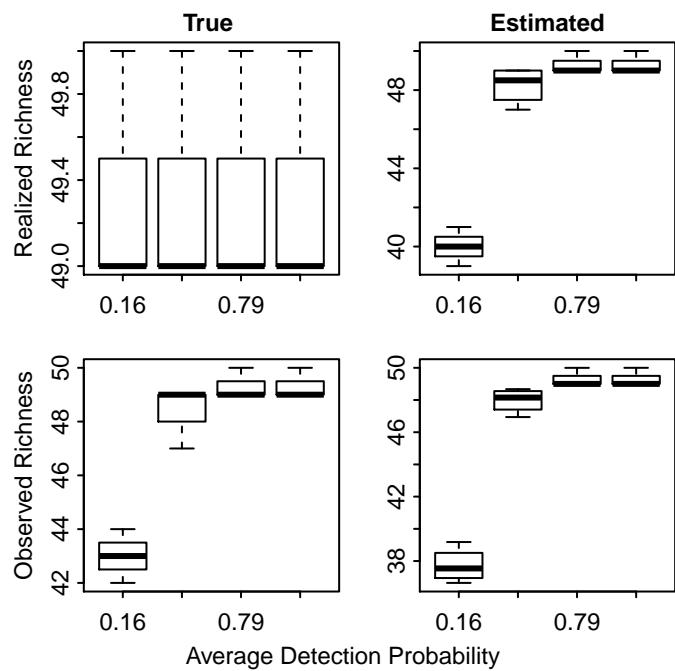


Figure 1: **Figure 1.** Boxplots of species richness. Numeric groupings indicate the average value of p across species during a given year-replicate combination. The panels in the left column are the true simulated values, and the panels on the right are the corresponding MSOM estimates. The top row 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.

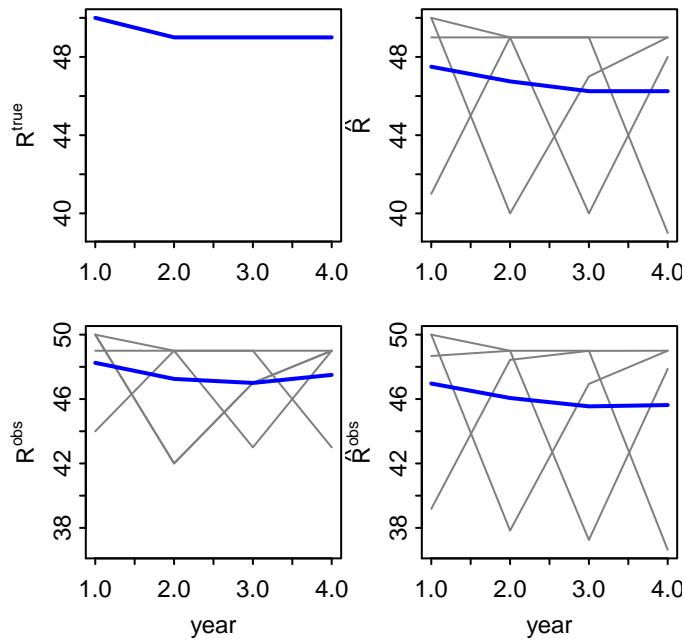


Figure 2: **Figure 2.** Time series of richness. Gray lines are individual replicates. Blue lines are averages. Note that detection probabilities ($p_{i,t,r}$, see [simulation settings above](#), as well as [definition of \$p\$ below](#)) change over time, and their temporal ordering differs among replicates.

252 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
 253 to the MSOM occurrence matrix (Y)

254

255 **Site Specific Richness (Nsite)**

256 **Scatter Plots of Nsite Split by Year**

257

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

259 Text explanation goes here

260

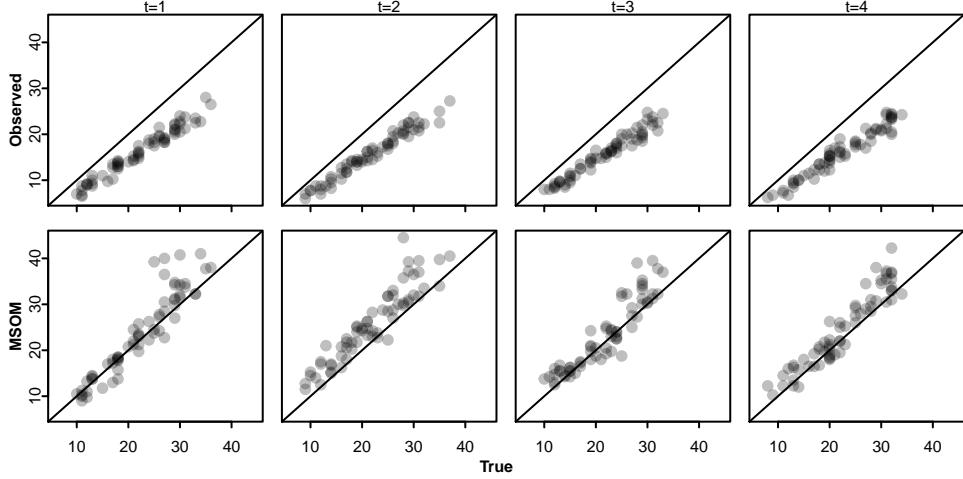


Figure 3: **Figure 3.** Site-specific richness (N_{site} , N_j) from simulated observations (vertical axis, top row; N_j^{obs}) and from MSOM estimates (vertical axis, bottom row, \hat{N}_j) vs true site-specific richness (horizontal axis; N_j^*). The panel columns delineate the years of the simulation.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

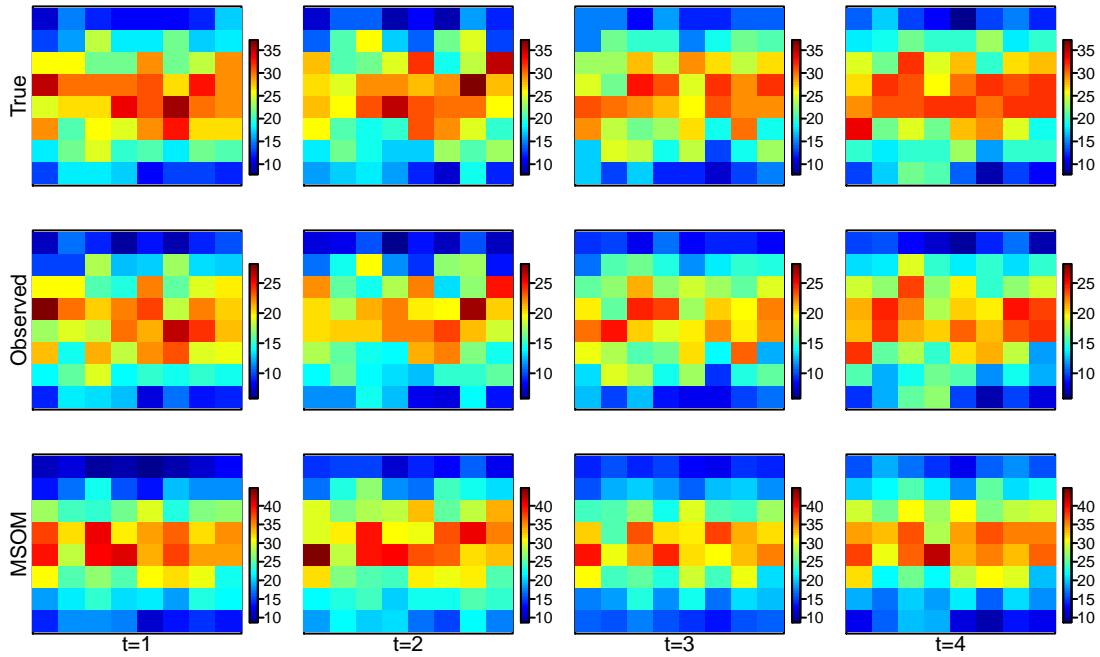


Figure 4: **Figure 4.** Maps of site- and year-specific species richness (N_{site}) from the simulation of the True process (top row), simulation of the Observed process (middle row), and the MSOM estimates (bottom row). X-axis and Y-axis indicate position in 2 dimensional space; it is important to note 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. Horizontal and vertical axes Each row of panels is scaled independently, columns within a row are scaled equally.

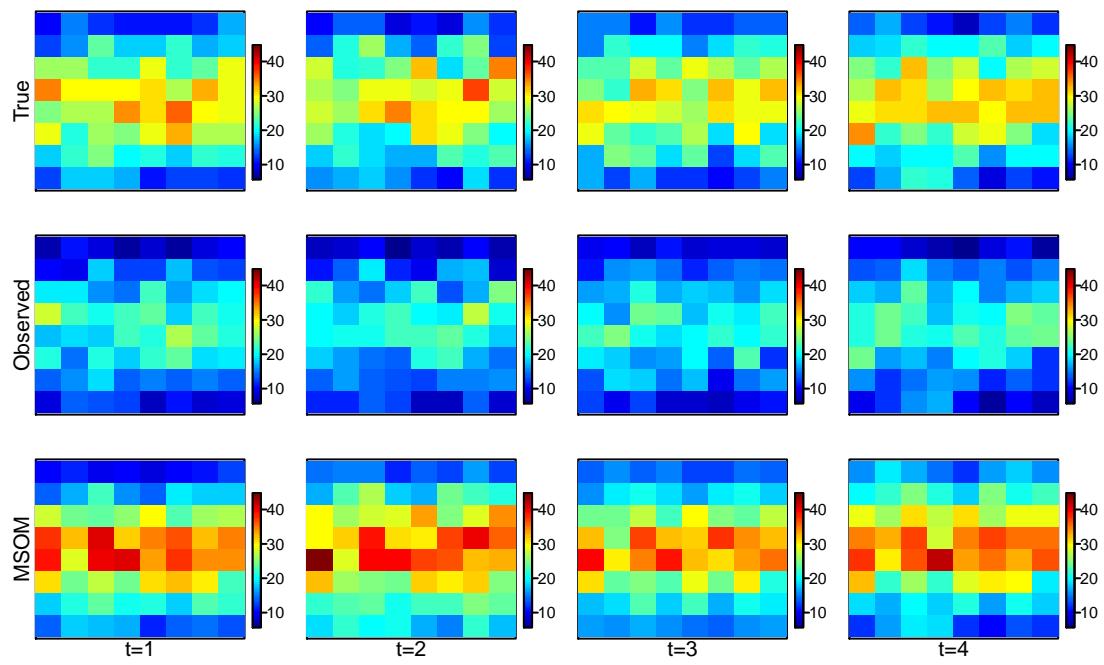


Figure 5: **Figure 5.** Same as previous figure, but all panels are on the same scale.

261 **Occupancy Probability, ψ**

262 **Definition of ψ**

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

267

268 **Scatter Plot of Aggregated ψ**

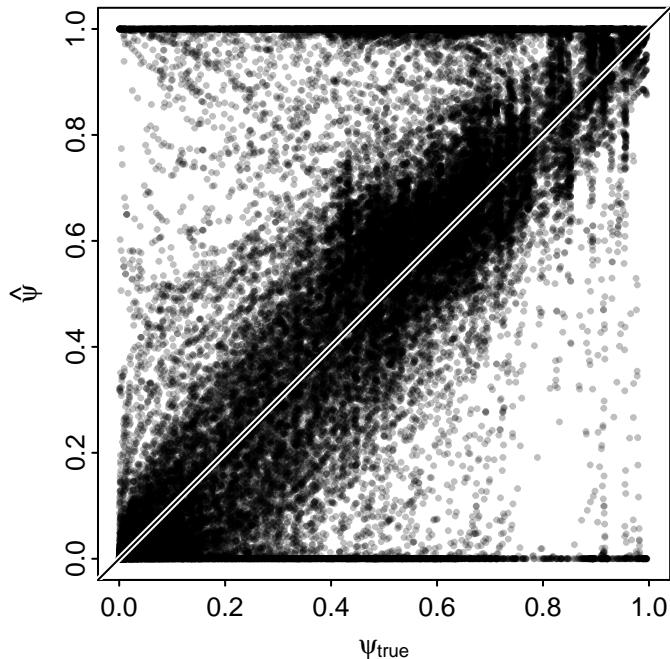


Figure 6: **Figure 6.** MSOM estimates of ψ ($\hat{\psi}$) vs. true values of ψ (ψ_{true}). Each point is a ψ value for a particular site-species-year-replicate. The white and black line is the 1:1 line.

- 269 In a general sense, the MSOM can distinguish between instances (sites/ years) when a species is likely to be
270 present, and when it's not (Figure 6). However, in every simulation I've done (varying many parameters that
271 aren't compared in this document), Figure 6 always makes it apparent that

- 272 1. There is a lot of variability around the 1:1 line
273
274 2. The residuals are not normal, and they are not independent
275 i. In general, I've found that $\hat{\psi}$ exhibits an upward bias, overestimating ψ^{true}

276

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

278 These patterns are somewhat concerning. The curve-like sequence of residuals is probably a byproduct of
279 slightly incorrect estimates of the parameters in the logistic regression ($[a_0, a_1, a_2]$), resulting in estimated
280 **response curves** that deviate non-randomly from the true response curve. For a heuristic of how these
281 smooth excursions can occur, in R try something as simply as `d <- rnorm(100); plot(dnorm(d), dt(d,`
282 `1))` to see the relationship between the density estimate from the correct distribution and that from
283 the wrong distribution (the density is analogous to ψ); or for really crazy patterns, try `d <- rnorm(100);`
284 `plot(dnorm(d), do.call(approxfun, density(d)[c("x", "y")])(d))`. So the curves are explainable, but
285 I cannot explain the consistent overestimation; I could understand how underestimating detectability (p)
286 would result in overestimating ψ , but the MSOM appears to recover true p values rather well (e.g., see Figure
287 7), so that's not a satisfying explanation.

288 In the next section I drill into ψ a bit more to try and understand what causes the largest deviations from
289 true values.

290

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

292 The estimates and true values of ψ are best correlated when p is high ()

293 Note: what I refer to as p here is really just the probability that a species will be detected if an occupied site is
294 sampled. In this simulation, 50% of substrata were sampled, which doesn't influence p , but can add noise to
295 its estimates.

296

297 Occupancy Response Curves

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

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

298 Therefore, these curves are tantamount to values of ψ , except that ψ generally pertains to a simulated,
299 observed, or true occupancy probability, whereas the occupancy probability in the response curves is calculated
300 over hypothetical conditions (i.e., over hypothetical values of the environmental gradient X).

301 True Occupancy Response Curves

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

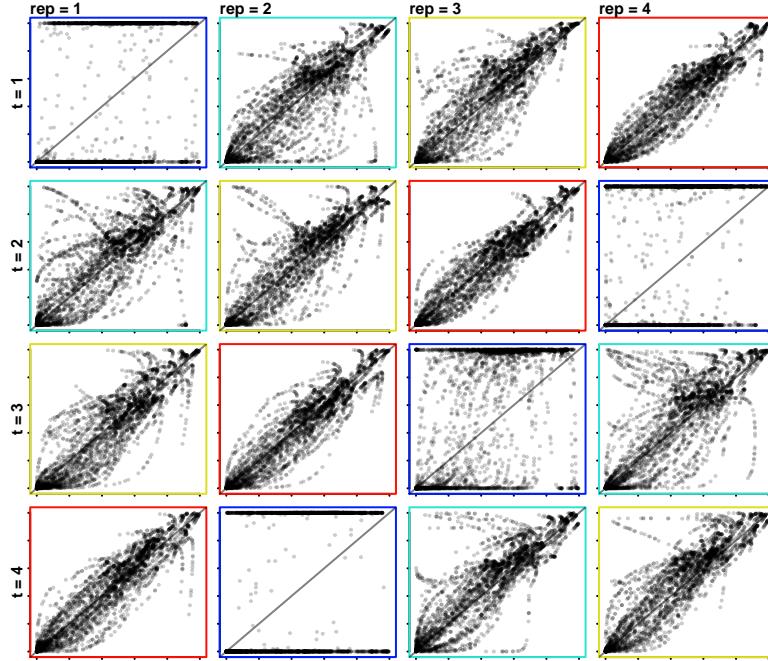


Figure 7: **Figure 8.** True (horizontal axes) and MSOM estimates (vertical axes) of occupancy probabilities ($\psi_{j,i,t,r}$) of species i occupying a location j . Years (t) are separated by rows, replicates (r) are separated by columns. The border color of each panel indicates the community-level mean probability of detection (p_μ ; 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.

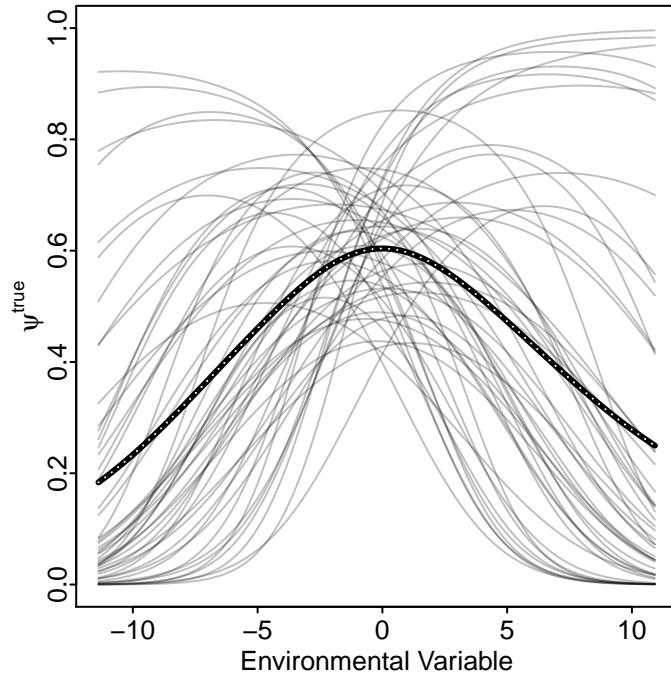


Figure 8: **Figure 9.** True simulated response curves. Vertical axis is the value of ψ^{true} , horizontal axis is the value of the environmental variable that, along with species-specific regression parameters, determines ψ^{true} . The thick line is the among-species mean value of ψ^{true} at a given value of the environmental variable.

305 **Estimated Occupancy Response Curves**

306 ($\min(X) = -11.4$, and $\max(X) = 10.9$)

307 (red; $p_{max} = 0.96$)

308 (blue; $p_{min} = 0.16$)

309

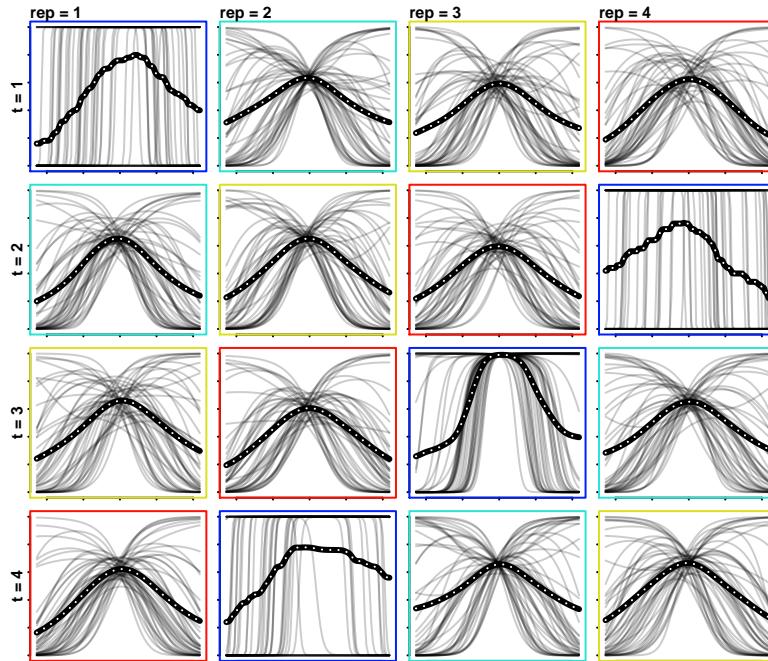


Figure 9: **Figure 10.** Response curves of species' probability of occupancy (ψ_i , vertical axis) across the full range of temperatures in the simulation. The color of the boxes around each panel refer to the among-species average of the probability of detection; warm colors indicate that the mean 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.

310 **Probability of Detection, p**

311 **Definition of p**

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

- 315 1. The probability that site j is occupied by species i ; this is $\psi_{j,i}$
 - 316 • $\psi_{j,i}$ is a function of species-specific niche and an environmental variable that changes over space
317 and time
 - 319 • $Z_{j,i}$ is the species- and site-specific richness, which is a function of ψ (given that we're only talking
320 about species that are in the pool of possible species, determined by w_i)
- 321 2. A species-specific (i) chance of being identified (`taxChance`), given that it is present in a location that
322 was sampled (i.e., detectability does not reflect the probability of sampling a place); this detectability
323 parameter is p_i
 - 324 • Detectability changed between years.
 - 325 • In a given year, $\text{logit}(p_i) \sim \mathcal{N}(\sqrt{\mu}, \sigma^\epsilon)$. μ changed between years (taking on values of -2, 0, 2,
326 and 4), $\sigma^2 = 2$ in all years.
 - 327 • The value of p only changes between species (and years), but the observation process occurs at the
328 substratum (k) level. Thus, the parameter is really $p_{j,k,i}$, but for a given i , all $p_{j,k}$ are constant. I
329 represent this probability as p_i with the understanding that this value is repeated over space.
 - 330 • $Y_{j,i}$ is the observed version of $Z_{j,i}$.
 - 331 • $Y_{j,i} \sim \text{Bern}(p_i \times Z_{j,i})$.
 - 332 – Note: Because p is actually subscripted to k , the Y are also actually subscripted to k . Maybe
333 leaving these subscripts out is making things more confusing. I've only excluded them to
334 emphasize how parameters are estimated.
 - 335 • Our data about species presence/ absence correspond to $Y_{j,i}$. So it might be useful to think of the
336 MSOM as estimating $\hat{Y}_{j,i}$, which is compared to the observed data $Y_{j,i}^{obs}$.

337

338 **Demo: Effect of MSOM Hierarchy on p**

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

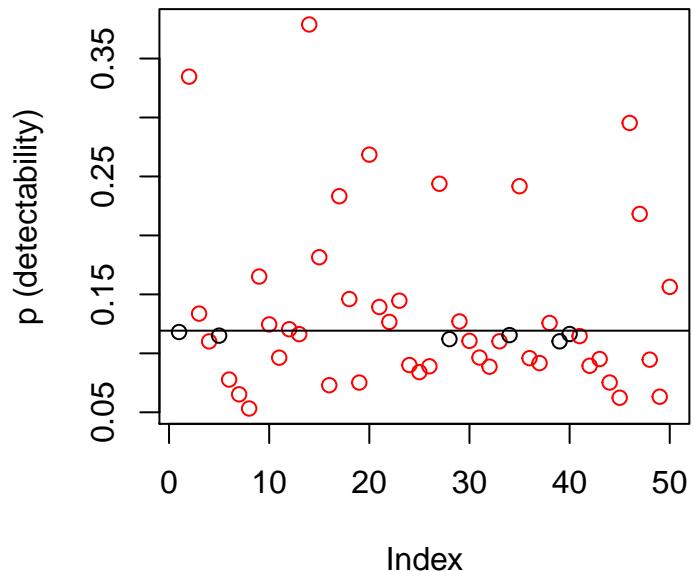


Figure 10: **Figure 11.** Probability of being detected, p . Horizontal line is mean probability. Figure only shows results for the first year of the simulation/ observation, and only 1 replicate. Different points are different species. Probability of being detected is a species-specific parameter (does not vary among sites, e.g.). Red points are species that were observed, black points are species that were never observed.

³⁴⁵ **Scatter Plot of \hat{p} vs p_{true}**

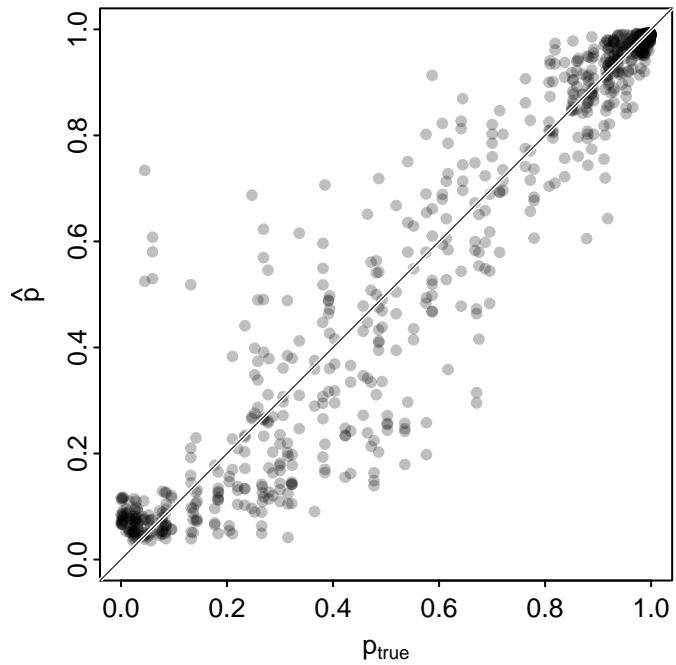


Figure 11: **Figure 7.** MSOM estimates (vertical axis) and true values of p_i , the species-specific (i) detection probability. Each point is subscripted by species i , year t , and observation replicate r .

³⁴⁶

³⁴⁷ **Scatter Plot of \hat{p} vs p_{true} , split by year and replicate**

³⁴⁸ Text explanation goes here

³⁴⁹

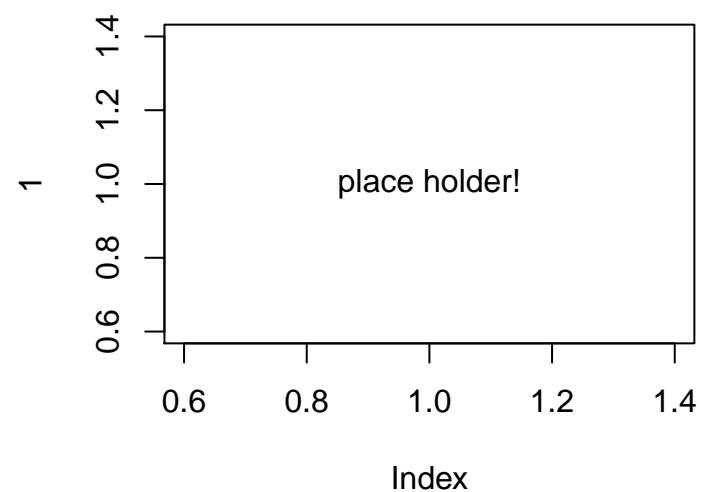


Figure 12: **Figure 12.** Caption goes here.

350 Assessment with Mixed Effects Models

351 Describe Motivation for Mixed Effects Models

- 352 **Motivation:** MSOM skill might differ across dimensions, trying to figure out what patterns I should expect
353 to pick out (spatial patterns in richness, temporal?) E.g., Is the correlation between MSOM and True the
354 same comparing across sites as comparing across years? Species, reps, also.
- 355 **Motivation:** What factors influence MSOM skill in a given dimension? E.g., Skill in finding differences in ψ
356 across species may depend on p , the chance of being identified. If p changes among years, might also explain
357 Read more about [specifying mixed effects models using lmer in R here](#)
- 358 This example is looking at ψ , probability of an individual species being present

359 Example LMER Analysis for ψ

```
# 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=T
)

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

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)
```

```

# =====
# = Do LMER Analysis =
# =====
(blah.mod <- lmer(hat~true+(1|spp)+(1|time), data=blah))

360 ## Linear mixed model fit by REML ['lmerMod']
361 ## Formula: hat ~ true + (1 | spp) + (1 | time)
362 ## Data: blah
363 ## REML criterion at convergence: 1135.817
364 ## Random effects:
365 ## Groups   Name        Std.Dev.
366 ##   spp      (Intercept) 0.07691
367 ##   time     (Intercept) 0.01563
368 ##   Residual           0.24405
369 ## Number of obs: 51200, groups: spp, 50; time, 4
370 ## Fixed Effects:
371 ## (Intercept)      true
372 ##             0.07055    0.95294

```

Anova(blah.mod)

```

373 ## Analysis of Deviance Table (Type II Wald chisquare tests)
374 ##
375 ## Response: hat
376 ##          Chisq Df Pr(>Chisq)
377 ## true 48463  1 < 2.2e-16 ***
378 ## ---
379 ## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

```

# =====
# = Explain LMER =
# =====

```

381 Report Generation Notes

382 Check that all Figures were Cited

```
383 ## Warning: Figure(s) 89101112 with label(s)
384 ## 'psiPlot.splitScatterresponseCurve.trueresponseCurve.msomegHierarchpPlot.splitScatter'
385 ## are present in the document but are never referred to in the text.
```

386 R Session Information

```
387 ## R version 3.2.0 (2015-04-16)
388 ## Platform: x86_64-apple-darwin13.4.0 (64-bit)
389 ## Running under: OS X 10.10.5 (Yosemite)
390 ##
391 ## locale:
392 ## [1] en_US.UTF-8/en_US.UTF-8/en_US.UTF-8/C/en_US.UTF-8/en_US.UTF-8
393 ##
394 ## attached base packages:
395 ## [1] parallel      grid       stats      graphics    grDevices   utils      datasets
396 ## [8] methods      base
397 ##
398 ## other attached packages:
399 ## [1] lme4_1.1-9      Rcpp_0.11.6      Matrix_1.2-0      car_2.0-26
400 ## [5] lattice_0.20-31 kfigr_1.2        xtable_1.7-4      rmarkdown_0.7
401 ## [9] knitr_1.11      doParallel_1.0.8 iterators_1.0.7    foreach_1.4.2
402 ## [13] R2jags_0.5-6    rjags_3-15      coda_0.17-1      igraph_0.7.1
403 ## [17] fields_8.2-1    maps_2.3-9      spam_1.0-1       data.table_1.9.4
404 ## [21] raster_2.3-40   sp_1.1-0       rbLib_0.0.2      taxize_0.5.2
405 ##
406 ## loaded via a namespace (and not attached):
407 ## [1] reshape2_1.4.1    splines_3.2.0    htmltools_0.2.6   yaml_2.1.13
408 ## [5] mgcv_1.8-6       chron_2.3-45    XML_3.98-1.3    nloptr_1.0.4
409 ## [9] plyr_1.8.2       stringr_1.0.0    codetools_0.2-11 evaluate_0.7.2
410 ## [13] Taxonstand_1.7   SparseM_1.7     permute_0.8-4    quantreg_5.11
411 ## [17] pbkrtest_0.4-2   numbers_0.6-1    highr_0.5       formatR_1.2
412 ## [21] vegan_2.3-0     jsonlite_0.9.16  abind_1.4-3     digest_0.6.8
413 ## [25] stringi_0.5-5   ssh.utils_1.0    tools_3.2.0     bitops_1.0-6
414 ## [29] magrittr_1.5     RCurl_1.95-4.6   bold_0.2.6     cluster_2.0.1
415 ## [33] ape_3.3         MASS_7.3-40     minqa_1.2.4    assertthat_0.1
416 ## [37] reshape_0.8.5    httr_0.6.1      boot_1.3-16    R2WinBUGS_2.1-21
417 ## [41] nnet_7.3-9      nlme_3.1-120
```

418 Date Document Last Compiled

419 ## Last compiled on: 2015-08-25

420
