

The R package `mobsim`

1 [Title](#)

2 `mobsim`: An R package for the simulation and measurement of biodiversity

3 across spatial scales

4 [Article type:](#)

5 Application

6 [Authors](#)

7 Felix May^{1,2}, Katharina Gerstner¹, Dan McGlenn³, Xiao Xiao⁴, Jonathan M. Chase^{1,2}

8 [Contact details](#)

9 Felix May

10 E-mail: felix.may@idiv.de

11 Phone: +49-(0)341-9733192

12 [Affiliations](#)

13 ¹German Centre for Integrative Biodiversity Research (iDiv) Halle-Jena-Leipzig, Deutscher Platz 5e, D-
14 04103 Leipzig, Germany

15 ²Institute of Computer Science, Martin-Luther University Halle-Wittenberg, D-06099 Halle (Saale),
16 Germany

17 ³Biology Department, College of Charleston, Charleston, SC, USA

18 ⁴School of Biology and Ecology, and Senator George J. Mitchell Center for Sustainability Solutions,
19 University of Maine, Orono, ME, USA 04469

20

21

22 Abstract

23 1. Estimating biodiversity and its changes in space and time poses serious methodological
24 challenges. First, there has been a long debate on how to quantify biodiversity, and second,
25 measurements of biodiversity change are scale-dependent. Therefore comparisons of biodiversity
26 metrics between communities are ideally carried out across scales. Simulation can be used to study
27 the utility of biodiversity metrics across scales, but most approaches are system specific and plagued
28 by large parameter spaces and therefore cumbersome to use and interpret. However, realistic spatial
29 biodiversity patterns can be generated without reference to ecological processes, which suggests a
30 simple simulation framework could provide an important tool for ecologists.

31 2. Here, we present the R package `mobsim` that allows users to simulate the abundances and
32 the spatial distribution of individuals of different species. Users can define key properties of
33 communities, including the total numbers of individuals and species, the relative abundance
34 distribution, and the degree of spatial aggregation. Furthermore, the package provides functions that
35 derive biodiversity patterns from simulated communities, or from observed data, as well as functions
36 that simulate different sampling designs.

37 3. We show several example applications of the package. First, we illustrate how species
38 rarefaction and accumulation curves can be used to disentangle changes in the fundamental
39 biodiversity components: (i) total abundance, (ii) relative abundance distribution, (iii) and species
40 aggregation. Second, we demonstrate how `mobsim` can be used to assess the performance of species-
41 richness estimators. The latter indicates how spatial aggregation challenges classical non-spatial
42 species-richness estimators.

43 4. `mobsim` allows the simulation and analysis of a large range of biodiversity scenarios and
44 sampling designs in an efficient and comprehensive way. The simplicity and control provided by the
45 package can also make it a useful didactic tool. The combination of controlled simulations and their
46 analysis will facilitate a more rigorous interpretation of real world data that exhibit sampling effects
47 and scale-dependence.

48 Key-words

- 49 1. diversity indices
- 50 2. sampling
- 51 3. scale-dependence
- 52 4. simulation
- 53 5. species-abundance distribution
- 54 6. species-accumulation curve
- 55 7. species-area relationship
- 56 8. rarefaction curve
- 57

58 Introduction

59 Understanding how biodiversity varies in space and time poses one of the greatest challenges in
60 ecology. One reason this challenge is difficult to overcome is that observed biodiversity changes
61 depend on spatial scale (Rosenzweig 1995; Rahbek 2005) and on the specific biodiversity measure used
62 (reviewed in Magurran & McGill 2011). Reasons for the complexity in measuring biodiversity across
63 scales are twofold. First, any measure of biodiversity (e.g., species richness, Shannon, or Simpson
64 diversity) transforms the numbers of individuals and species in a given sample into a single univariate
65 metric that necessarily only captures a portion of information about the underlying abundances and
66 spatial distribution. Second, biodiversity measures vary non-linearly with spatial scale and thus any
67 comparisons among two or more sites will typically be highly scale-dependent (i.e., their difference or
68 ratio will depend on the scale in which it is measured) (Chase & Knight 2013). Despite continued
69 discussion approaches for estimating and comparing diversity measures (e. g. Jost 2006; Colwell *et al.*
70 2012), no single measure can capture all of the relevant information and multiple measures and
71 approaches will provide more complete information about biodiversity and its change. All biodiversity
72 patterns, including local and regional measures of diversity (α , γ -diversities) and their scaling
73 relationships (measures of β -diversity) depend on three biodiversity components, namely (i) the total
74 abundance of individuals, (ii) relative species abundance distribution, and (iii) the spatial distributions
75 of individuals and species (He & Legendre 2002; McGill 2011). Although we focus on taxonomic
76 diversity measures here, the same issues apply for measurements of functional and (phylo-)genetic
77 diversity (Chao, Chiu & Jost 2014).

78 Here, we introduce the software package `mobsim` to facilitate understanding and
79 interpretation of biodiversity changes across scales. `mobsim` includes spatially explicit simulation
80 tools, which allow user-defined manipulations of the three biodiversity components. Of course, in
81 nature the components emerge from species traits and dynamic ecological processes such as
82 competition, dispersal limitation, or habitat filtering. However, we suggest that direct simulations and

The R package `mobsim`

83 manipulations of the emerging biodiversity components provide generality that embraces different
84 taxa and ecosystems.

85 Often, when discussing processes of sampling and resulting biodiversity measurements,
86 analogies such as pulling jellybeans from a jar are used (e.g., Gotelli & Colwell 2001). These analogies
87 also provide useful pedagogical tools for teaching biodiversity concepts (e.g., Heard 2016). `mobsim`
88 simulates individuals of different species (i.e., the proverbial “jelly beans”) in a spatially explicit
89 landscape and thus allows studying the influence of sampling and scale, as well as the interrelatedness
90 between different biodiversity descriptors and patterns in a comprehensive and efficient way. The
91 package provides functions for three purposes (1) the simulation of communities in space, (2) the
92 analysis of biodiversity patterns, and (3) the simulation and assessment of different sampling designs
93 (Fig. 1, Table 1).

94 Specifically, in spatially-explicit simulations users define the total number of individuals, the
95 species richness, the shape and evenness of the relative species-abundance distribution, and the
96 intraspecific aggregation of species. Functions for the analysis of biodiversity patterns, such as
97 rarefaction-curves (Gotelli & Colwell 2001) and species-area relationships (Rosenzweig 1995) allow
98 users to assess how different biodiversity indices vary with spatial scale and/or sampling effort. Finally,
99 the package provides functions to simulate sampling processes and to convert the spatially-explicit
100 simulated data into classical community matrices (i.e. sites-by-species abundances matrices). These
101 matrices can then be analysed using standard analytical tools (Legendre & Legendre 2012) to assess
102 how the simulated changes are expressed in measures of biodiversity and influenced by the sampling
103 design. The package is currently available on GitHub (<https://github.com/MoBiodiv/mobsim>) and also
104 as interactive shiny application with graphical user interface
105 (https://github.com/KatharinaGerstner/mobsim_app). We also plan to publish the package on CRAN
106 (<https://cran.r-project.org>).

107 [Package description](#)

108 [Simulation of community data](#)

109 An ecological community is characterised by its species-abundance distribution (SAD) and by the
110 spatial distribution of individuals. In `mobsim`, users can use a predefined SAD and add simulated
111 positions of individuals, or simulate both the SAD and the positions (Table 1). For the simulation of
112 SADs, a wrapper around the function `rsad` from the R package `sads` is provided, which offers many
113 options for the underlying statistical distribution (Prado, Miranda & Chalom 2016). In contrast to
114 `sads::rsad`, the function `mobsim::sim_sad` allows the simultaneous specification of the
115 simulated number of individuals and the number of species in the pool. Due to random sampling, there
116 can be fewer species in the simulated community than in the user-defined species pool, but we also
117 provide an argument in `sim_sad` to allow the number of simulated species equal that of the pool
118 size. In this case there can be deviations from the underlying statistical distribution, because rare
119 species are added until the required species richness is reached, leading to a longer tail than expected
120 from the underlying statistical SAD model.

121 The spatial coordinates of individuals are simulated as 2-dimensional point processes in
122 `mobsim` following Wiegand & Moloney (2014) either using a Poisson process, where individuals are
123 placed randomly, or a Thomas process, where individuals of the same species are clustered. For the
124 Thomas process, users define the numbers and sizes of the clusters, and the number of individuals per
125 cluster, either independently or jointly for all species. The Thomas-process only considers intraspecific
126 aggregation. Individuals of different species are distributed randomly with respect to each other
127 (McGill 2010).

128 [Analysis of community data](#)

129 `mobsim` offers several functions to derive spatial and non-spatial patterns from simulated or empirical
130 spatial data. Conversion of empirical data to the format required by `mobsim` is facilitated by auxiliary
131 functions. The function `spec_sample_curve` derives the expected number of species given a

The R package `mobsim`

132 certain number of sampled individuals. Individuals are sampled either randomly, giving the well-known
133 species rarefaction-curve (Gotelli & Colwell 2001), or sampling proceeds always from a focal individual
134 to the nearest neighbour, which results in the spatial species-accumulation curve (spatial SAC)
135 (Chiarucci *et al.* 2009). Note that this is different from the sample-based accumulation curve described
136 in Gotelli & Colwell (2001), which considers the distribution of individuals among plots, but not the
137 spatial location of plots.

138 The function `divar` (diversity-area relationships) estimates several diversity indices for
139 sampling plots of user-defined areas. Accordingly, this function can be used to derive the species-area
140 and endemics-area relationships (Rosenzweig 1995; Harte & Kinzig 1997). In addition, `divar`
141 estimates the Shannon- and Simpson-diversity indices, and their corresponding effective numbers of
142 species (ENS) (Jost 2006).

143 The function `dist_decay` estimates the distance decay of community similarity (Nekola &
144 White 1999; Morlon *et al.* 2008). The calculation is based on pairwise similarity indices of species
145 composition in square sampling plots using the function `vegdist` from the R package `vegan`
146 (Oksanen *et al.* 2017). `mobsim` includes standard plotting functions for community data and for all the
147 biodiversity patterns introduced. See the online documentation for detailed information on all main
148 and auxiliary functions.

149 [Sampling of community data](#)

150 `mobsim` also provides functionality to simulate sampling processes by distributing square plots in a
151 community. The data type provided by the sampling is a sites-by-species matrix, which is a classical
152 data type in community ecology (Legendre & Legendre 2012). Users can choose the size and number
153 of sampling quadrats, as well as the spatial design.

154

The R package `mobsim`

155

156

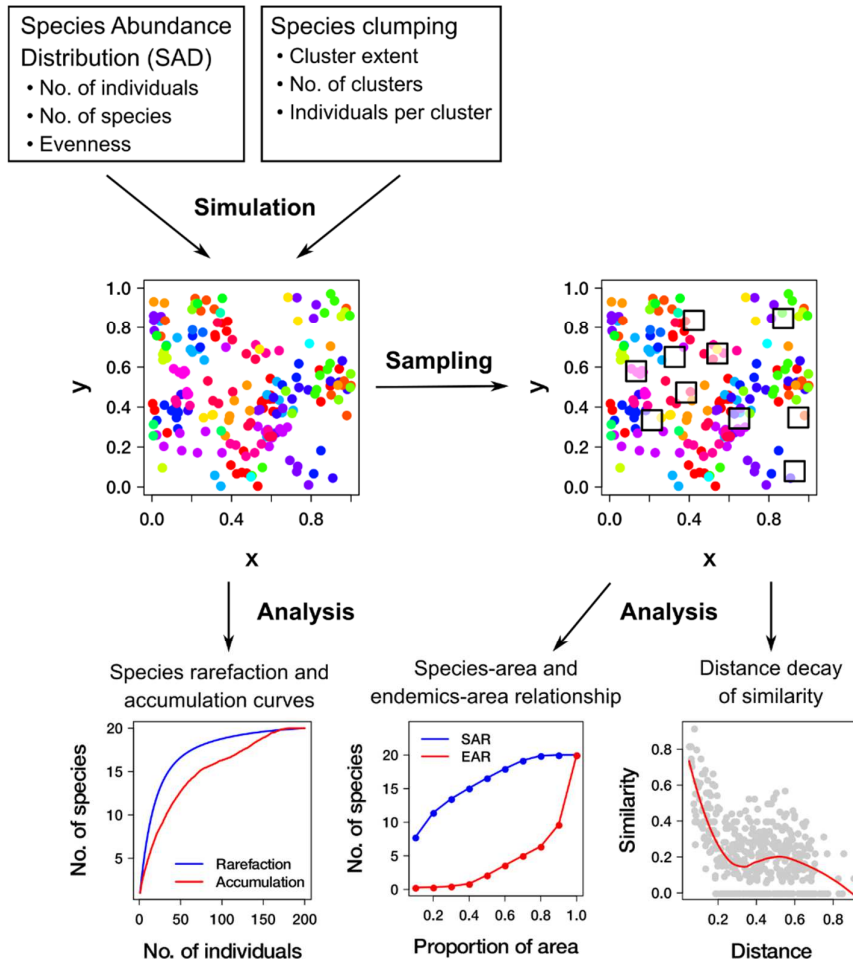
Table 1: List of main functions in `mobsim`

Function name	Category	Description
<code>sim_sad</code>	simulation	Simulate species-abundance distributions (SADs)
<code>sim_poisson_coords</code>	simulation	Add spatially random coordinates to a SAD
<code>sim_thomas_coords</code>	simulation	Add coordinates with intraspecific aggregation to an SAD
<code>sim_poisson_community</code>	simulation	Simulate a community with certain SAD and spatially random coordinates
<code>sim_thomas_community</code>	simulation	Simulate a community with certain SAD and intraspecific aggregation
<code>spec_sample_curve</code>	analysis	Derive rarefaction and individual-based species-accumulation curves
<code>divar</code>	analysis	Derive diversity-area relationships from square sampling plots of different sizes. Diversity indices include species richness, no. of endemics, Shannon index, Simpson index and the effective numbers of species based on Shannon and Simpson indices
<code>dist_decay</code>	analysis	Derive the distance-decay function from pairwise community similarity measures of square sampling plots
<code>sample_quadrats</code>	sampling	Virtual sampling of different communities using square plots of user-defined sizes. Users can choose different sampling designs, including random sampling, transect and lattice designs.

157

The R package `mobsim`

158



159

160 **Figure 1:** General overview on the purposes of `mobsim`. The package provides functions to simulate
161 species abundances and spatial distributions based on user-defined parameters, including the
162 numbers of species and individuals, the species-abundance distributions, and the aggregation of
163 conspecific individuals. Simulated or observed distributions can be analysed and visualised using
164 `mobsim` functions. Alternatively, sampling processes can be simulated using `mobsim` and the analysis
165 can be done with additional software for classical site-by-species abundance matrices.

166

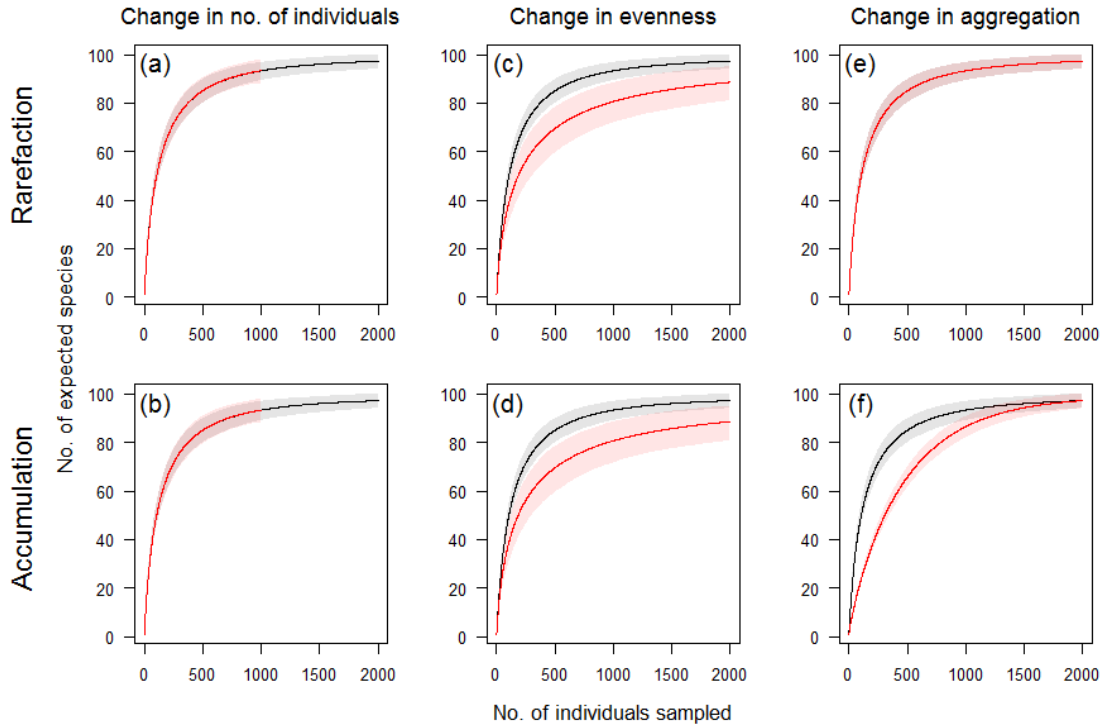
167 Example applications

168 Here we present two example applications of `mobsim`: (i) on changes of biodiversity components
169 and (ii) an assessment of species richness estimators. A third example, on extinctions due to habitat
170 loss is provided in the online supporting information.

171 Changes of single biodiversity components

172 The biodiversity in a sampled area depends on three components that can vary independently: (1) the
173 total number of individuals, (2) the SAD of the species pool, and (3) the spatial distribution of
174 individuals and species (McGill 2011; Chase & Knight 2013). Using simulations, we show how the
175 combination of rarefaction and accumulation curves can be used to disentangle changes in these three
176 biodiversity components (Fig. 2). The key point is that the shape of the rarefaction curve only depends
177 on the underlying SAD, but not on the spatial distribution (Gotelli & Colwell 2001), while the shape of
178 the accumulation curve depends on both the SAD and the spatial distribution. First, we randomly
179 removed 50% of all individuals. This does not affect the underlying SAD, as indicated by overlapping
180 rarefaction and accumulation curves that end at different numbers of individuals (Fig. 2a, b). Second,
181 we simulated communities with lower evenness by increasing the variation of species abundances.
182 This resulted in changes in both the rarefaction and the accumulation curves (Fig. 2c, d). Here, the
183 difference (or ratios) between the curves changes with sampling effort, which indicates scale and/or
184 sampling effort dependent effect sizes (Chase & Knight 2013). Despite having the same number of
185 species in the pool, the simulated communities differ in species richness even for the maximum
186 number of individuals, because the rarest species are not sampled into the local community. See Fig.
187 S1 for the same figure that used a fixed species richness of the simulated community, where the curves
188 converge at the largest number of individuals. Third, we add intraspecific aggregation by using a
189 Thomas-process instead of a Poisson-process, which does not affect the rarefaction curve, but leads
190 to lower expected species richness in the spatial species-accumulation curve (Fig. 2e,f).

The R package `mobsim`



191

192 **Figure 2:** Simulated species rarefaction- and accumulation-curves for changes of three different

193 biodiversity components. The black lines and intervals show the reference community with 2,000

194 individuals and 100 species in the species pool, a log-normal species-abundance distribution (SAD) with

195 $\text{meanlog} = 3$ and $\text{sdlog} = 1$, and spatially random positions (Poisson distribution). The red lines

196 show changed communities for half the number of individuals (first column), a decrease in evenness

197 (second column), simulated as higher variation in abundances with $\text{sdlog} = 1.5$, and a higher

198 intraspecific aggregation (third column), simulated with a Thomas-process with cluster extent

199 $\text{sigma} = 0.02$. Ribbons indicate 95% confidence intervals derived from 1,000 replicate simulations.

200

201 [Testing species-richness estimators](#)

202 We are often interested in inferring biodiversity and/or its change in a larger region based on a limited

203 amount of samples. Species-richness estimators offer approaches for estimating the biodiversity of a

204 large community based on local samples (Colwell & Coddington 1994; Chiu *et al.* 2014). However, it

205 remains an open question of how well these estimators perform for different communities and for

The R package `mobsim`

206 different sampling strategies (Colwell & Coddington 1994; ter Steege *et al.* 2017). The simulation tools
207 of `mobsim` are well-suited to address this issue.

208 We used `mobsim` to assess the performance of a bias-corrected version of the well-known
209 Chao1-estimator (Chiu *et al.* 2014), in the face of spatial aggregation and different sampling designs.
210 We simulated a community with 1,000 species and 1,000,000 individuals. Then we used the function
211 `sample_quadrats` to sample from the community. We varied the proportion of total area sampled
212 between 0.01% - 1% as well as the number of sampling quadrats (1 – 100) that jointly represent the
213 total sampling effort. These combinations of sampling strategies were applied to communities with
214 the same SAD, but with different intraspecific aggregations. We examined four scenarios for
215 aggregation: (1) a random distribution; (2) several large clusters per species; (3) several small clusters
216 per species; (4) one large cluster per species. We used the function `vegan::estimateR` to calculate
217 the species-richness estimator of Chiu *et al.* (2014).

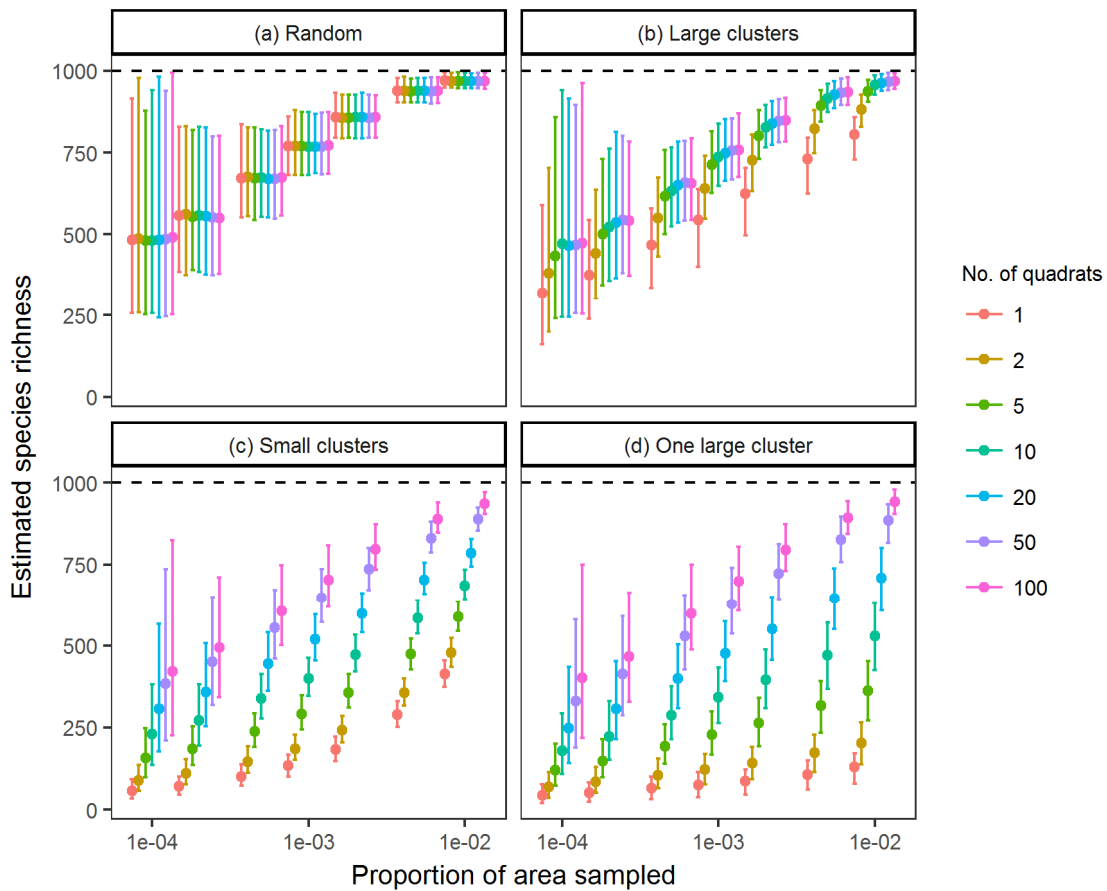
218 For the community with a random distribution we found no influence of whether a single large
219 or several small quadrats were sampled (Fig. 3a). However, the estimated richness and its uncertainty
220 strongly varied with total sampling effort. The bias of the estimator decreased with increasing sampling
221 effort, but the estimated and true values only converged at the highest effort. At the same time, the
222 uncertainty decreased drastically with sampling effort (Fig. 3a).

223 For aggregated distributions, the spatial configuration of sampling mattered and a sampling
224 strategy with several small quadrats was less biased than few large quadrats (Fig. 3b-d). For high
225 aggregation, species richness and its uncertainty were strongly underestimated (Fig. 3c, d). This is an
226 important finding, because aggregated species distributions tend to be the rule in nature (McGill 2010,
227 2011).

228 Our simulation results underline the recommendation by developers of species-richness
229 estimators, that the estimated values should be only interpreted as lower bounds (Chao 1987; Chiu *et al.*
230 *et al.* 2014). Furthermore, our findings indicate that for aggregated species distributions, both sampling
231 design and sampling effort have a large influence.

The R package `mobsim`

232



233

234 **Figure 3:** Performance of an asymptotic species-richness estimator for communities with different
235 intraspecific aggregation and for different sampling strategies. The panels show the estimated species
236 richness (`chao1` from function `vegan::estimateR`) vs. the proportion of total area sampled. The
237 colour indicate different numbers of randomly distributed sampling quadrats that together form the
238 total amount of sampled area. The different panels show results for communities with the same SAD
239 but different intraspecific aggregation. The points and error bars represent means and 95% confidence
240 intervals from 1,000 replicate simulations. The horizontal lines indicates the true species richness. The
241 following parameter values were used for the simulations: species pool richness: `s_pool = 1000`;
242 number of individuals: `n_sim = 1,000,000`; a log-normal species-abundance distribution (SAD)
243 with `meanlog = 3` and `sdlog = 1`; large cluster size: `sigma = 0.05`, small cluster size:
244 `sigma = 0.01`; one cluster per species: `mother_points = 1`.

245 Conclusions

246 The total number of individuals, the distributions of species relative abundances, and intraspecific
247 aggregation are key components of community structure and any changes in biodiversity are mediated
248 through changes in these components. Furthermore, biodiversity changes are scale-dependent. The
249 combination of tools for simulation and analysis of biodiversity patterns provided in `mobsim` is well-
250 suited to foster understanding on the emergence and consequences of scale-dependent biodiversity
251 changes. The package integrates key tools of community ecology so that ecologists can derive valid
252 and robust interpretations of biodiversity patterns and changes observed in real-world data. We also
253 believe that the use of controlled simulation experiments is highly beneficial for education in
254 biodiversity science.

255 Acknowledgements

256 We gratefully acknowledge the support of the German Centre for Integrative Biodiversity
257 Research (iDiv) Halle-Jena-Leipzig funded by the German Research Foundation (FZT 118).

258 Authorship statement

259 FM and JMC conceived the package concept and structure. FM implemented the first package version.
260 KG, XX, and DM contributed code and supported the package revision. KG implemented the shiny
261 online application. FM wrote the first manuscript draft and all authors critically revised the text and
262 gave final approval for publication.

263 References

- 264 Chao, A. (1987). Estimating the population size for capture-recapture data with unequal catchability.
265 *Biometrics*, **4**, 783–791.
- 266 Chao, A., Chiu, C.-H. & Jost, L. (2014). Unifying Species Diversity, Phylogenetic Diversity, Functional
267 Diversity, and Related Similarity and Differentiation Measures Through Hill Numbers. *Annual*
268 *Review of Ecology, Evolution, and Systematics*, **45**, 297–324.
- 269 Chase, J.M. & Knight, T.M. (2013). Scale-dependent effect sizes of ecological drivers on biodiversity:
270 why standardised sampling is not enough. *Ecology Letters*, **16**, 17–26.
- 271 Chiarucci, A., Bacaro, G., Rocchini, D., Ricotta, C., Palmer, M. & Scheiner, S. (2009). Spatially constrained
272 rarefaction: incorporating the autocorrelated structure of biological communities into sample-
273 based rarefaction. *Community Ecology*, **10**, 209–214.
- 274 Chiu, C.-H., Wang, Y.-T., Walther, B.A. & Chao, A. (2014). An improved nonparametric lower bound of
275 species richness via a modified good–turing frequency formula. *Biometrics*, **70**, 671–682.
- 276 Colwell, R.K., Chao, A., Gotelli, N.J., Lin, S.-Y., Mao, C.X., Chazdon, R.L. & Longino, J.T. (2012). Models
277 and estimators linking individual-based and sample-based rarefaction, extrapolation and
278 comparison of assemblages. *Journal of Plant Ecology*, **5**, 3–21.
- 279 Colwell, R.K. & Coddington, J.A. (1994). Estimating terrestrial biodiversity through extrapolation.
280 *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, **345**, 101–118.
- 281 Gotelli, N.J. & Colwell, R.K. (2001). Quantifying biodiversity: procedures and pitfalls in the
282 measurement and comparison of species richness. *Ecology letters*, **4**, 379–391.
- 283 Harte, J. & Kinzig, A.P. (1997). On the implications of species-area relationships for endemism, spatial
284 turnover, and food web patterns. *Oikos*, **80**, 417–427.
- 285 He, F. & Legendre, P. (2002). Species diversity patterns derived from species-area models. *Ecology*, **83**,
286 1185–1198.
- 287 Heard, M.J. (2016). Using a Problem-Based Learning Approach to Teach Students about Biodiversity,
288 Species Distributions & the Impact of Habitat Loss. *The American Biology Teacher*, **78**,
289 733–738.
- 290 Jost, L. (2006). Entropy and diversity. *Oikos*, **113**, 363–375.
- 291 Legendre, P. & Legendre, L.F.J. (2012). *Numerical Ecology, Volume 24, Third Edition*, 3 editionn.
292 Elsevier, Amsterdam; Boston.
- 293 Magurran, A.E. & McGill, B.J. (2011). *Biological Diversity: Frontiers in Measurement and Assessment*, 1
294 editionn. Oxford University Press, Oxford.
- 295 McGill, B.J. (2011). Linking biodiversity patterns by autocorrelated random sampling. *American Journal*
296 *of Botany*, **98**, 481–502.
- 297 McGill, B.J. (2010). Towards a unification of unified theories of biodiversity. *Ecology Letters*, **13**, 627–
298 642.

The R package `mobsim`

- 299 Morlon, H., Chuyong, G., Condit, R., Hubbell, S., Kenfack, D., Thomas, D., Valencia, R. & Green, J.L.
300 (2008). A general framework for the distance–decay of similarity in ecological communities.
301 *Ecology letters*, **11**, 904–917.
- 302 Nekola, J.C. & White, P.S. (1999). The distance decay of similarity in biogeography and ecology. *Journal*
303 *of Biogeography*, **26**, 867–878.
- 304 Oksanen, J., Blanchet, F.G., Friendly, M., Kindt, R., Legendre, P., McGlenn, D., Minchin, P.R., O’Hara,
305 R.B., Simpson, G.L., Solymos, P., Stevens, M.H.H., Szoecs, E. & Wagner, H. (2017). *vegan*:
306 *Community Ecology Package*.
- 307 Prado, P.I., Miranda, M.D. & Chalom, A. (2016). *sads: Maximum Likelihood Models for Species*
308 *Abundance Distributions*.
- 309 Rahbek, C. (2005). The role of spatial scale and the perception of large-scale species-richness patterns.
310 *Ecology Letters*, **8**, 224–239.
- 311 Rosenzweig, M. (1995). *Species diversity in space and time*. Cambridge University Press.
- 312 ter Steege, H., Sabatier, D., Mota de Oliveira, S., Magnusson, W.E., Molino, J.-F., Gomes, V.F., Pos, E.T.
313 & Salomão, R.P. (2017). Estimating species richness in hyper-diverse large tree communities.
314 *Ecology*, n/a-n/a.
- 315 Wiegand, T. & Moloney, K.A. (2014). *Handbook of Spatial Point-Pattern Analysis in Ecology*, 1 editionn.
316 Chapman and Hall/CRC, s.l.
- 317