



HAL
open science

Rarity and sparseness in plant communities: impact of minor species removal on beta diversity and canonical ordination

François Gillet, Adeline Rouzet, Daniel Borcard, Pierre Legendre

► **To cite this version:**

François Gillet, Adeline Rouzet, Daniel Borcard, Pierre Legendre. Rarity and sparseness in plant communities: impact of minor species removal on beta diversity and canonical ordination. *Journal of Vegetation Science*, 2025, 36 (1), pp.e70013. 10.1111/jvs.70013 . hal-04924382

HAL Id: hal-04924382

<https://hal.science/hal-04924382v1>

Submitted on 31 Jan 2025

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons Attribution - NonCommercial - NoDerivatives 4.0 International License

1 **Rarity and sparseness in plant communities: impact of minor** 2 **species removal on beta diversity and canonical ordination**

3 François Gillet¹, Adeline Rouzet^{1,2}, Daniel Borcard³, Pierre Legendre³

4 ¹ UMR 6249 Chrono-environnement, Université Marie et Louis Pasteur – CNRS, Besançon,
5 France

6 ² Service de Parasitologie-Mycologie, CHRU Jean-Minjoz, Besançon, France

7 ³ Département de Sciences biologiques, Université de Montréal, Montréal QC, Canada

8 Correspondence: François Gillet, francois.gillet@univ-fcomte.fr

9

10 **Abstract**

11 **Question** – Among the ‘minor’ species present in communities, we distinguish between true
12 ‘rare’ species, with infrequent occurrence (low occupancy) in a given regional data set, and
13 ‘sparse’ species, which may be present over most of the study area, but with low local
14 abundance. Do rare and sparse species play a different role in the evaluation of beta diversity
15 and in the constrained ordination of plant community data sets?

16 **Methods** – Based on their positions in the abundance-occupancy scatterplots of six contrasted
17 vegetation data sets, we distinguished core, rural, urban and satellite species. To disentangle
18 the role of rarity and sparseness, we applied to each data set a progressive removal of either
19 the least frequent or the least locally abundant species. We assessed impacts on beta diversity
20 ($q = 0, 1$ and 2), and on model performance of RDA, without or after pre-transformation of
21 absolute cover values.

22 **Results** – Multiplicative beta diversity decreased with the number of removed rare species,
23 with slightly higher values for $q = 2$, whereas it increased when removing sparse species, with
24 much higher values for $q = 0$. With raw data or after binary or by-site transformation, the
25 fraction of variation explained by RDA increased only slightly when removing rare species,
26 with a more sensible increase of the relative contribution of the first canonical axis. By
27 contrast, progressive elimination of sparse species, which mimics a lower sampling effort
28 within each community, negatively affected model performance. Generally, the removal of
29 rare species clearly improved the performance of RDA after double transformation (chi-
30 square transformation), contrary to the removal of sparse species.

31 **Conclusions** –The frequently observed positive correlation between occupancy and
32 abundance hides profound differences with critical impacts on vegetation analysis. Providing
33 that meaningful transformations are applied, there is no need to remove rare species prior to
34 RDA. Focusing only on abundant species during sampling is likely to limit the performance
35 of ecological empirical models.

36

37 **Key-words:** abundance-occupancy relationship; beta-diversity; core-satellite species
38 hypothesis; minor species; rare species; redundancy analysis

39

40 **Introduction**

41 Vegetation data sets generally include many species with a low frequency of occurrence, their
42 number depending on plot size and on vegetation heterogeneity within and among sites. They
43 are responsible for a high proportion of zeros in the community data matrix and thus for a
44 high inertia in the multivariate response to be explained by constrained ordination. It has been
45 argued that such infrequent species should be removed because they add noise to the analysis
46 and reduce model performance (Cao et al., 2001; Poos & Jackson, 2012; Jing et al., 2015;
47 Brasil et al., 2020). However, Legendre & Gallagher (2001) argued that removal of the most
48 infrequent species may be relevant before correspondence analysis (CA) and canonical
49 correspondence analysis (CCA), but not before principal component analysis (PCA), after
50 showing that these species only play minor roles in PCA, without or with data transformation.

51 These infrequent species often show low abundances in local communities, whereas
52 frequent species are often locally dominant, so that there is generally a positive correlation
53 between the frequency of occurrence of a species (also called ‘occupancy’ or ‘distribution’)
54 throughout a data set and its average abundance in occupied plots (‘abundance-when-
55 present’), i.e. a positive interspecific abundance-occupancy relationship (Collins et al., 1993;
56 Gaston, 1998; Gaston et al., 2000; Blackburn et al., 2006; Borregaard & Rahbek, 2010;
57 Eriksson, 2013; Guedo & Lamb, 2013).

58 In its broad sense, ‘rarity’ is often defined using the framework of Rabinowitz (1981),
59 which considers the three axes of range size, local abundance and habitat specialism. In the
60 British flora, Rabinowitz et al. (1986) found these three axes to be independent, whereas other
61 studies often find high proportions of species that are rare in all three dimensions (Harrison et
62 al., 2008). Habitat specialism, i.e. the specificity of a species to a few habitats, is more

63 difficult to measure than the two other dimensions and is sometimes interpreted by the
64 availability of suitable sites for a species in the landscape, measured by the rarity of its
65 preferred habitat (Broennimann et al., 2005). Habitat specificity can be better viewed as an
66 underlying cause of rarity than a rarity dimension (Crisfield et al., 2024). Here, we will not
67 consider this controversial dimension of ‘rarity’ and will focus on the difference between the
68 two other dimensions.

69 They correspond to the conceptual distinction between regionally *rare* species with low
70 occupancy (frequency of occurrence) in a given regional data set, and locally *sparse* species
71 with low relative abundance (measured by density, cover, biomass or local frequency) in each
72 local community. These sparse species are sometimes neglected during sampling, e.g. when
73 recording dominant species only or when working on small plots, where the probability of
74 observing an abundant species is higher. In contrast, rare species may be omitted when
75 restricting the range of habitats in the study area, as they are expected to have introgressed
76 from surrounding habitats. We argue that the expression *rare* species should be restricted to
77 species with a low frequency of occurrence in a given regional community data set, although
78 this term is often also used to refer to locally *sparse* species. A generic term for both rare and
79 sparse species may be ‘minor’ species.

80 Adding to the confusion, abundance measured as local frequency – e.g. the fraction of
81 small quadrats located within each site that is occupied by a given species, or the proportion
82 of contacts with this species in a point-intercept method – can be seen as similar to regional
83 occupancy, i.e. the fraction of large plots where a species is found, but at a finer scale
84 (Eriksson, 2013). Thus, from this strict methodological point of view, the distinction between
85 rare and sparse species would be dependent on the scale (grain and extent) considered in the
86 study from which the data set was built. However, the distinction between rare and sparse
87 species is also supported by an ecological argument: as far as each relevé corresponds to a
88 local plant community, with strong biotic interactions and homogeneous habitat conditions
89 (Lortie et al., 2004), species abundance results from fine-scale environmental and biotic
90 filtering, while species occupancy depends on dispersal, biogeographical filtering, and on the
91 range of habitats considered in the regional data set. Hence, rarity and sparseness, as well as
92 commonness and abundance, deserve to be distinguished in order to correctly assess the role
93 of ‘minor’ species in the analysis of a vegetation data set.

94 To put the distribution-abundance relationship in the dynamic context of metapopulation
95 models, Hanski (1982) developed his core-satellite species hypothesis by distinguishing two
96 main kinds of species with respect to their average local abundance N and the fraction of

97 occupied patches, i.e. occupancy p : ‘core’ species with large N and p (abundant and common,
98 relatively well spaced-out in niche space), and ‘satellite’ species with small N and p (sparse
99 and rare, with a narrow niche). Following Söderström (1989), Hanski (1991) further
100 considered two other distribution types of species which he named by analogy to human
101 populations: ‘rural’ species with small N but large p (sparse but common), and ‘urban’ species
102 with small p but large N (rare but abundant). Urban species are supposed to have high local
103 growth rate and low dispersal rate, leading to population concentration, contrary to rural
104 species. Thus, for a given data set, species may be classified into core, satellite, urban and
105 rural species according to their position in the abundance-occupancy scatterplot, revealing
106 distinct niche patterns (Collins et al., 1993). Indeed, many data sets include urban and rural
107 species in addition to core and satellite species that shape the popular positive distribution-
108 abundance relationship.

109 Hanski’s core and satellite species are somewhat equivalent to Grime’s ‘dominant’ and
110 ‘transient’ species, respectively (Grime, 1998; Gibson et al., 1999). ‘Subordinate’ (i.e.
111 frequent but never dominant) species of Grime’s classification can be viewed as equivalent to
112 rural species. Real-world species removal experiments showed that subordinate plant species
113 play an overlooked but important role in grassland biodiversity and ecosystem functioning
114 (Mariotte, Buttler, et al., 2013; Mariotte, Vandenberghe, et al., 2013). The ‘subordinate
115 insurance hypothesis’ suggests that subordinate species may assist dominant species or
116 compensate for their loss on ecosystem functions via complex plant-soil feedbacks (Mariotte,
117 2014). Another typology of concepts based on abundance-occupancy patterns has been
118 proposed recently (Avolio et al., 2019), somewhat equivalent to Hanski’s classification, core
119 species being called ‘common’, urban species ‘restricted’, satellite species ‘rare’, and rural
120 species ‘sparse’. In this conceptual framework, locally abundant species (either ‘common’ or
121 ‘restricted’) may be ‘dominant’ or ‘subordinate’ depending on their impact on their
122 surrounding environment, community and ecosystem functioning. In the absence of a
123 consensus on vocabulary and for methodological reasons, we prefer to restrict the meaning of
124 ‘rare’ and ‘sparse’ species to each dimension of occupancy and abundance.

125 Equivalences among these different concepts and theories should be considered with
126 caution, however. Urban species are not considered in Grime’s dominant-subordinate-
127 transient classification. Grime’s classification is based on frequency-abundance patterns
128 inside a plant community, whereas Hanski’s classification basically considers a
129 metacommunity, i.e. a network of patches of the same habitat occupied by local communities
130 at a broader spatial scale. Moreover, discussions about abundance-distribution relationships

131 often refer to macroecology and biogeography, involving a variety of habitats in a wide
132 region. Thus, as often in ecology, semantic shifts are likely to obscure the comparison of these
133 concepts.

134 The questions addressed in this paper are the following: given the potential ecological
135 roles of rare and sparse species, what are the consequences of discarding them when
136 analyzing abundance-occupancy patterns, diversity patterns and community-environment
137 relationships in vegetation science? Do rare and sparse species play the same role in these
138 analyses? Therefore, considering a variety of vegetation data sets differing in their spatial
139 extent and ecological diversity, the specific question examined in this paper is: how does a
140 progressive removal of *either* the regionally least frequent species *or* the locally least
141 abundant species affect (1) the proportion of core, rural, urban and satellite species, (2) the
142 evaluation of taxonomic β -diversity and (3) the performance of constrained ordination? In
143 addition, does the impact of species removal on RDA model performance depend on the
144 choice of the prior transformation of abundance-cover data?

145 **Methods**

146 *Data sets*

147 We selected six contrasted vegetation data sets for which species absolute cover and
148 explanatory variables were available (Table 1), i.e. three small and homogeneous data sets
149 (hereafter called ‘vare’, ‘catgrass’ and ‘dune’) and three bigger and more heterogeneous data
150 sets (‘trufe’, ‘vltava’, ‘bryce’). Each species data set is associated with an environmental data
151 set containing explanatory variables to be used in constrained ordination.

152 We describe each original species data set by the species frequency distribution (Fig. 1)
153 and by the abundance-occupancy scatterplot (Fig. 2). According to the core-satellite species
154 hypothesis, frequency distribution is supposed to be bimodal (U-shaped), but it depends on
155 plot size and on the heterogeneity of the studied vegetation (Collins & Glenn, 1997). An
156 abundance-occupancy scatterplot represents the average relative cover of each species in all
157 occupied sites, without taking into account absences (its mean abundance-when-present, on a
158 log scale) vs. the proportion of sites occupied by the species (its frequency of occurrence). It
159 includes a regression line from the linear model explaining the logarithm of average
160 abundance-when-present by the frequency of occurrence, as well as the delimitation of core,
161 satellite, urban and rural species.

162 Fifty percent frequency is the threshold considered by Mariotte (2014) to first distinguish
163 transient (i.e. satellite) from both dominant and subordinate species within a plant community.
164 Above 50% frequency, this author suggested that species with a cumulative relative cover
165 below 2% should be considered as transients, comprised between 2% and 12% as
166 subordinates, and above 12% as dominants. Collins & Glenn (1997) considered as core
167 species those present in more than 90% of the sites and as satellite species those present in
168 less than 10% of the sites, all other species being classified as ‘intermediate’ without
169 distinguishing rural and urban. To our knowledge, no clear boundaries have been specified so
170 far to delimit these four categories. Since we consider here both mono- and multi-site data
171 sets at various spatial scales, we retained Hanski’s classification, but we adapted Mariotte’s
172 typology to delimit species groups. Therefore, we counted the number of core, satellite, urban
173 and rural species by considering common arbitrary boundaries to delimit these four groups:
174 50% frequency and 5% relative cover (Fig. 2 and 3).

175 The Spearman rank correlation was used to measure and to test the abundance-occupancy
176 relationship (Table 1). The heterogeneity of the species matrix was measured by (i) the
177 multiplicative β -diversity of species richness, related to the differentiation among habitats
178 (Whittaker, 1972; Tuomisto, 2010), (ii) the relative MacArthur’s homogeneity measure,
179 which compares communities based on the Hill-Shannon beta diversity (Jost, 2007, equation
180 22), (iii) the proportion of zeros in the vegetation table (species absences), and (iv) the
181 number of species observed at only one site or plot (Table 1).

182 The ‘vare’ data set describes the understorey vegetation of boreal pine forests (Väre et al.,
183 1995). ‘varespec’ (species) and ‘varechem’ (environmental variables) data frames are
184 available in the R package *vegan* (Oksanen et al., 2024). Abundance was recorded as the
185 average absolute cover of 44 lichen, bryophyte and vascular plant species in 15 or 10 quadrats
186 of 1 m² within 24 sites. This relatively homogeneous data set shows a low β -diversity, with a
187 high proportion of core and rural species, but no urban species. The environmental matrix is
188 made of 14 quantitative variables: soil concentration of 11 chemical elements, bare soil cover,
189 humus thickness and pH.

190 The ‘catgrass’ data set (Kohler, 2004) consists of 90 1-m² quadrats equally distributed
191 over three experimental sites of montane pastures in the Swiss Jura Mountains, in which
192 absolute cover of 94 vascular plant species was measured by a point-intercept method after
193 two years of treatment (simulation of various combinations of cattle activities). This rather
194 homogeneous data set contains few core and urban species. The environmental matrix is made

195 of 7 binary and semi-quantitative variables representing each factor of the treatments: site,
196 shading, mowing, manuring, trampling, abandonment and grazing.

197 The ‘dune’ data set is a classic example of dune meadows in the Netherlands (Jongman et
198 al., 1995). ‘dune’ (species) and ‘dune.env’ (environmental variables) data frames are available
199 in the R package *vegan*. Original 9-class cover values of 30 vascular and bryophyte species
200 recorded in 20 quadrats of $2 \times 2 \text{ m}^2$ were transformed into mid-class percentages of absolute
201 cover (van der Maarel, 1979; Dengler & Dembicz, 2023). β -diversity is low, with many urban
202 and no rural species. The environmental data frame is composed of five variables: thickness
203 of soil A1 horizon (quantitative), soil moisture and manure (semi-quantitative, treated as
204 numeric variables instead of ordered factors to avoid overfitting in ordination analysis),
205 management and use (qualitative).

206 The ‘trufe’ data set was obtained from a systematic sampling of the herb layer by 110
207 quadrats of 1 m^2 within a 1-ha area of heterogeneous wood-pasture at La Sagne in the Swiss
208 Jura Mountains (Béguin, 2007). Original 6-class Braun-Blanquet dominance values of 111
209 vascular plant species were transformed into mid-class percentages of absolute cover. The
210 community data set encompasses a high proportion of satellite and rural species. The
211 environmental data frame is made of 6 quantitative variables: slope, eastern aspect, shrub
212 cover, mean soil depth, soil depth variance and yearly potential sunshine hours.

213 The ‘vltava’ data set is made of 97 forest plots of $10 \times 15 \text{ m}^2$ arranged in transects across
214 the Vltava river valley in the Czech Republic (Zelený & Chytrý, 2007). The transects were
215 perpendicular to the map contour lines and the river course but the plots along the transects
216 were rectangles parallel to the contour lines. Within each plot, the absolute cover of vascular
217 plant species was recorded separately for each of the three layers of the forest vegetation
218 (herbs, shrubs and trees). Cover was estimated using the 9-class ordinal Braun-Blanquet scale
219 and these codes were consequently transformed into mid-point percentages of cover classes.
220 The data are available at <https://www.davidzeleny.net/anadat-r/doku.php/en:data:vltava>. The
221 β -diversity is high with very few core and rural species. The environmental data frame is
222 made of 12 quantitative or binary variables: elevation, slope, aspect, heat load, landform
223 shape (2 variables), soil type (4 dummy variables), soil depth and pH.

224 Finally, the ‘bryce’ data set is another large-extent, heterogeneous vegetation data set from
225 the Bryce Canyon National Park in Utah (USA). The ‘bryceveg’ (species) and ‘brycesite’
226 (environmental variables) data frames are included in the R package *labdsv* (Roberts, 2023).
227 Absolute cover of the herbaceous species was recorded in various vegetation types (badlands,
228 savannas and forests) using an 8-class dominance scale in 145 circular plots of 375 m^2 and

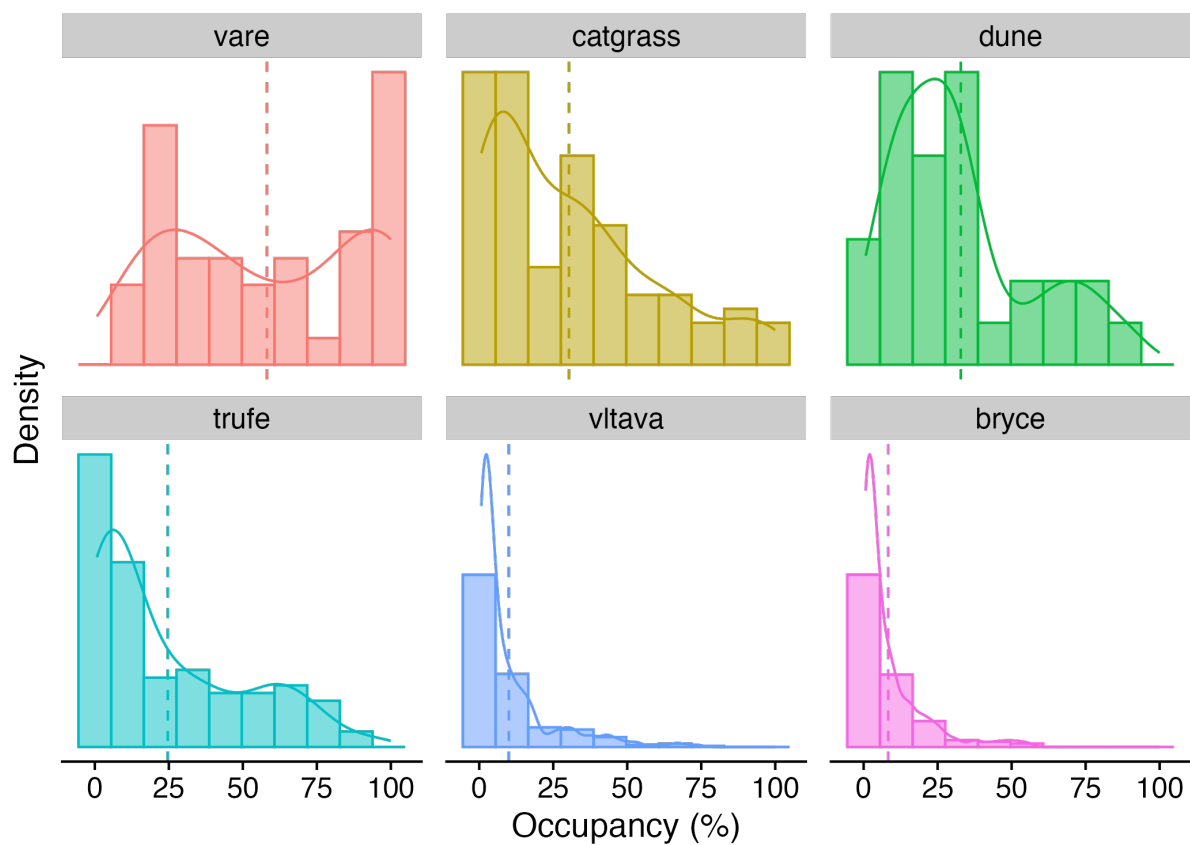
229 transformed into mid-point percentages. The β -diversity is very high with very few species
 230 occurring in more than half of the sites. The environmental data frame is a selection of 7
 231 quantitative or qualitative variables: elevation, soil depth, topographic position, annual and
 232 growing-season solar radiation, slope and aspect. 15 sites with missing values for soil depth
 233 were removed from the original data set.

234 Data sets ‘vare’, ‘dune’ and ‘trufe’ show a bimodal frequency distribution, due to their
 235 relative homogeneity, whereas ‘vltava’, ‘bryce’ and ‘catgrass’ show a unimodal distribution
 236 (Fig. 1). However, the modes of the distribution do not always correspond to a clear
 237 opposition between satellite and core species.

238
 239 Table 1. Characteristics of the six vegetation data sets. Unique species are those observed in only one plot.
 240 Abundance/occupancy correlation is estimated by the Spearman rank correlation coefficient between species
 241 mean absolute cover and frequency of occurrence. Alpha and gamma diversities were calculated from raw
 242 absolute cover data as Hill numbers for $q = 0, 1$ and 2 . Relative McArthur homogeneity measure was deduced
 243 from Hill numbers for $q = 1$. More explanations in text.

Data set	vare	catgrass	dune	trufe	vltava	bryce
Plot size (m ²)	10 to 15	1	4	1	150	375
Number of plots	24	90	20	110	97	145
Number of core species	7	4	7	4	3	1
Number of rural species	17	16	0	19	3	1
Number of urban species	0	5	10	12	22	36
Number of satellite species	20	69	13	76	246	128
Proportion of zeros	41.9%	69.7%	67.2%	75.4%	90.0%	91.7%
Number of unique species	0	9	3	9	55	29
Abundance/occupancy correlation	0.495	0.577	0.497	0.351	0.448	0.233
αN_0 (mean number of species)	25.6	28.5	9.9	27.4	27.4	13.8
αN_1 (alpha Hill-Shannon diversity)	5.8	12.3	4.2	11.0	13.2	4.2
αN_2 (alpha Hill-Simpson diversity)	3.8	7.2	2.9	7.0	7.0	2.4
γN_0 (total number of species)	44	94	30	111	274	166
γN_1 (gamma Hill-Shannon diversity)	11.7	25.9	16.3	36.6	70.0	31.9
γN_2 (gamma Hill-Simpson diversity)	8.4	10.5	13.1	18.7	33.0	12.3
βN_0 (beta species richness diversity)	1.72	3.30	3.05	4.06	10.00	12.02
βN_1 (beta Hill-Shannon diversity)	2.02	2.11	3.86	3.34	5.30	7.53
βN_2 (beta Hill-Simpson diversity)	2.25	1.46	4.53	2.68	4.74	5.14
Relative MacArthur homogeneity	0.474	0.469	0.220	0.293	0.180	0.127

244

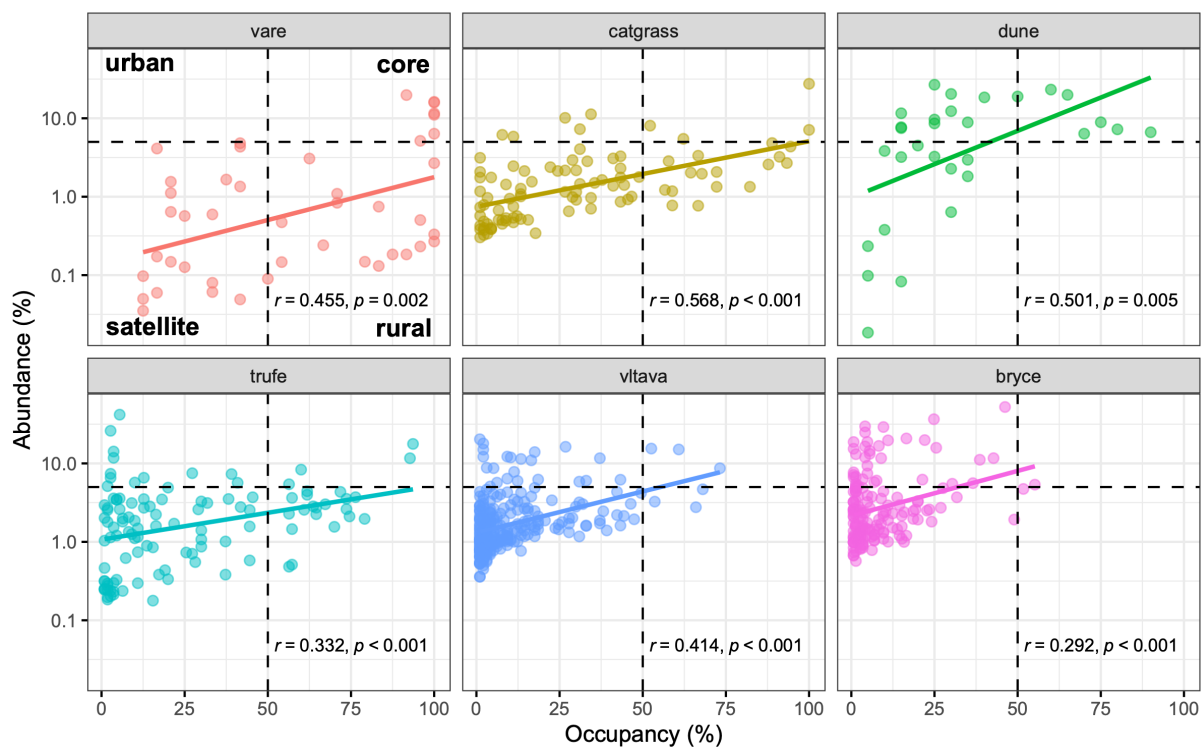


245

246 Figure 1. Species frequency distribution (occupancy percentage) in the six data sets. Histograms and smooth

247 curves represent density. A vertical dashed line shows the mean species frequency of occurrence.

248



249

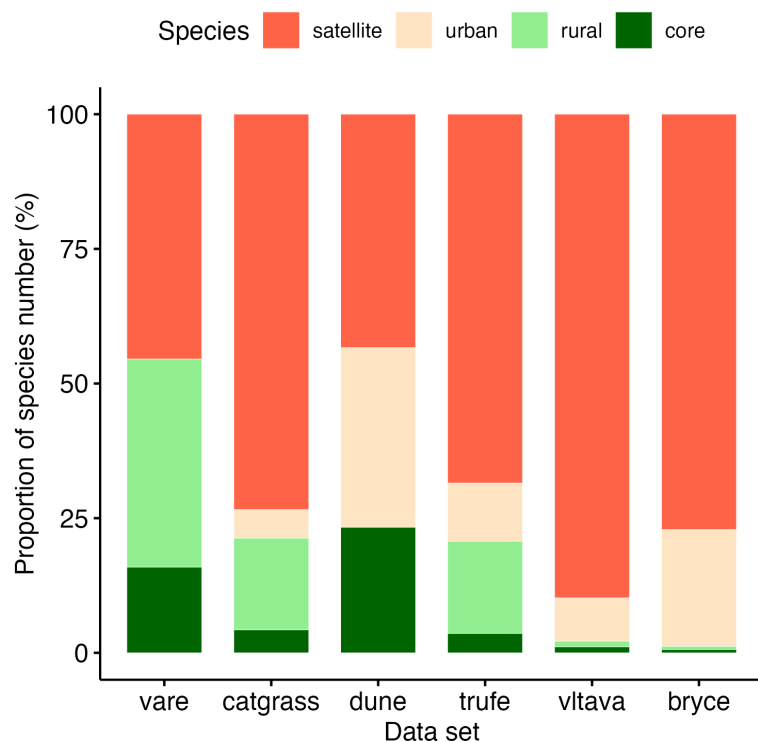
250 Figure 2. Abundance-occupancy scatterplots for the six data sets. Each point represents a species. The abscissa is
 251 occupancy, measured by the species frequency in the data set, i.e. the percentage of sites occupied by the
 252 species. The ordinate (on a log-scale) is the abundance-when-present, measured by the average relative cover of
 253 the species in occupied sites. Dashed lines represent boundaries between core (top-right quadrant), satellite
 254 (bottom-left quadrant), urban (top-left quadrant) and rural (bottom-right quadrant) species. The horizontal
 255 dashed line is positioned at 5% of relative cover and the vertical dashed line at 50% of frequency of occurrence.
 256 The solid straight line is the linear regression model of the logarithm of abundance-when-present by occupancy
 257 (with Pearson linear correlation coefficients r and p -values).

258

259 These six examples support the common report of significant positive correlations
 260 between the frequencies of the species and their average relative covers in occupied plots
 261 (Table 1, Fig. 2). However, the variation in local abundance may be very high among
 262 regionally rare species.

263 Most of the species belong to the satellite category in each data set. Core species never
 264 reach 25% of the species pool. Several data sets, such as ‘dune’ and ‘bryce’, contain a large
 265 proportion of urban species (Fig. 3).

266



267

268 Figure 3. Proportion of core, rural, urban and satellite species in each data set.

269 *Progressive removal of species records*

270 In each of the selected data sets, species records were progressively removed according to
 271 either their increasing frequency of occurrence in the whole community data set (*rare species*
 272 *removal*) or their increasing relative cover in each site (*sparse species removal*). As for sparse
 273 species removal, we fixed a constant interval between two successive minimum abundance
 274 values, that is 2% relative cover, and at each step all species records with an abundance lower
 275 than the minimum value were assigned the value zero. The phasing out was stopped before a
 276 site became empty and continued until the remaining dataset contained only five species.

277 We assessed impacts of species removal on abundance-occupancy patterns, on β -diversity
 278 estimates and on the model performance of redundancy analysis, without or after pre-
 279 transformation of absolute cover data (Fig. 4).

280 For each data set, the number of core, urban, rural and satellite species was plotted in a
 281 stacked area plot against the minimum species occupancy or abundance applied during the
 282 progressive removal of rare or sparse species, respectively.

283 α , β and γ taxonomic diversity indices were computed using the $d()$ function available in
 284 the R package *vegetarian* (Charney & Record, 2012), with equal community weights. Such
 285 diversity indices are referred to as species numbers equivalents or Hill diversity numbers
 286 (Hill, 1973; Jost, 2007; De Bello et al., 2010). Three orders were considered ($q = 0, 1$ and 2),
 287 corresponding to an increasing sensitivity to differences in species abundances (Jost, 2006).
 288 More specifically, we compared three multiplicative β taxonomic diversities: βN_0 , or
 289 Whittaker β -diversity, is the inter-site diversity of all species ignoring their abundance; βN_1 or
 290 Hill-Shannon β -diversity is the inter-site diversity of abundant species; βN_2 or Hill-Simpson
 291 β -diversity is the inter-site diversity of dominant species. Multiplicative β -diversity was
 292 calculated as the ratio between γ -diversity and mean α -diversity for $q = 0, q = 1$ and $q = 2$,
 293 respectively. Finally, additive β -diversity was calculated as the difference between γ -diversity
 294 and mean α -diversity for $q = 0, 1$ and 2 .

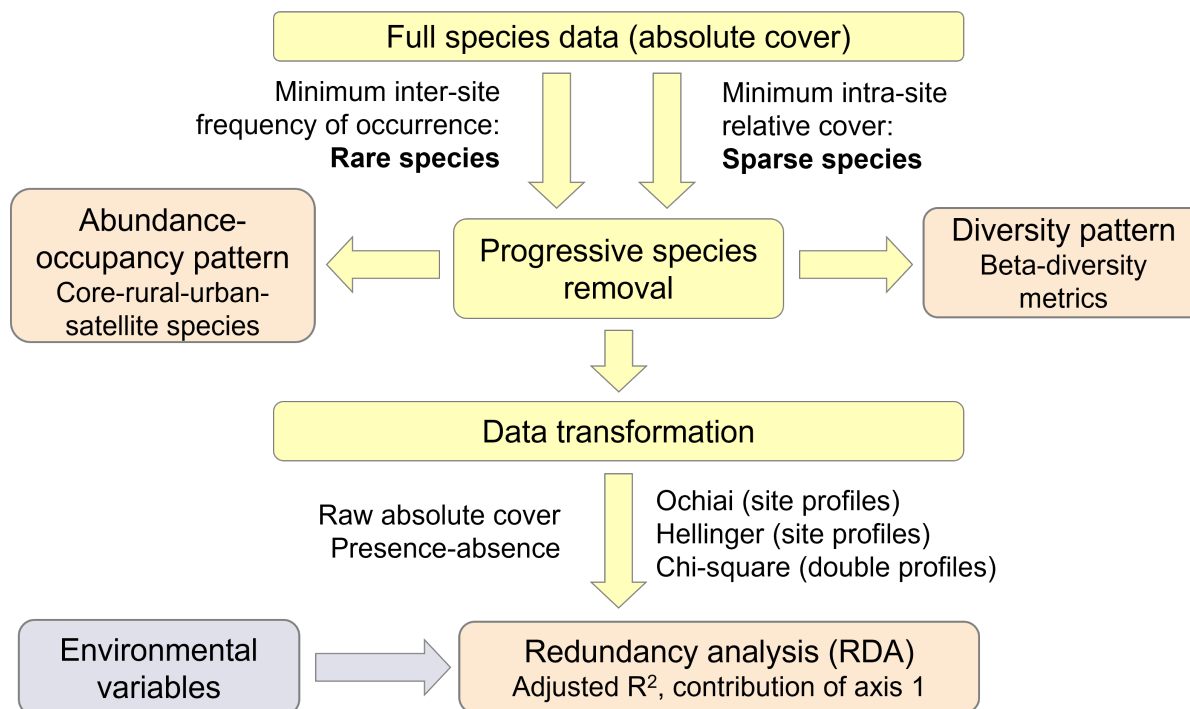
295 At each step of the progressive species removal, we compared the performance of
 296 redundancy analysis (RDA) after five optional pre-transformations of the species abundance
 297 data: raw absolute percentage cover (no data transformation), presence-absence
 298 transformation (binary data), chord transformation of the presence-absence data
 299 (normalization of binary site vectors), Hellinger transformation (square root of relative cover
 300 by site), and χ^2 (chi-square) double transformation (by site and by species). Site profiles
 301 (chord or Hellinger transformation) or double profiles (χ^2 transformation) make double

302 absences invisible in the computation of distances between sites; this transformation is
303 recommended before the RDA of species data sets with many zeros (Legendre & Gallagher,
304 2001). The Euclidean distance computed after a chord transformation of the presence-absence
305 data is proportional to the Ochiai dissimilarity, which allows double zeros in the presence-
306 absence data to be ignored (Borcard et al., 2018, p. 42); therefore and for simplicity, this
307 transformation will be named Ochiai transformation hereafter. Double profiles based on χ^2
308 distance make RDA similar (but not equivalent because eigenvalues and eigenvectors are
309 computed in different ways) to canonical correspondence analysis (CCA), another popular
310 constrained ordination method in community ecology (Borcard et al., 2018). Since we wanted
311 to compare transformations and not constrained ordination methods, we have chosen to only
312 apply RDA on our test data and not CCA.

313 Every vegetation data set was associated with an environmental data set used to constrain
314 the ordinations with the same explanatory variables at each step of the rare or sparse species
315 removal. Performance of RDA was assessed by (i) the unbiased percentage of variation
316 explained by the environmental variables (adjusted R^2 , Peres-Neto et al. (2006)) and (ii) the
317 contribution of the first canonical axis to the explained variation (expressing the efficiency of
318 the ordination space reduction).

319 All analyses were performed in the R environment (RCoreTeam, 2024), using packages
320 `vegan`, `labdsv`, `vegetarian`, `tidyverse` (Wickham et al., 2019) and `ggpubr` (Kassambara, 2023).
321 The data and the R code used to perform these analyses are provided as electronic
322 Supplementary Material (Appendix S2).

323



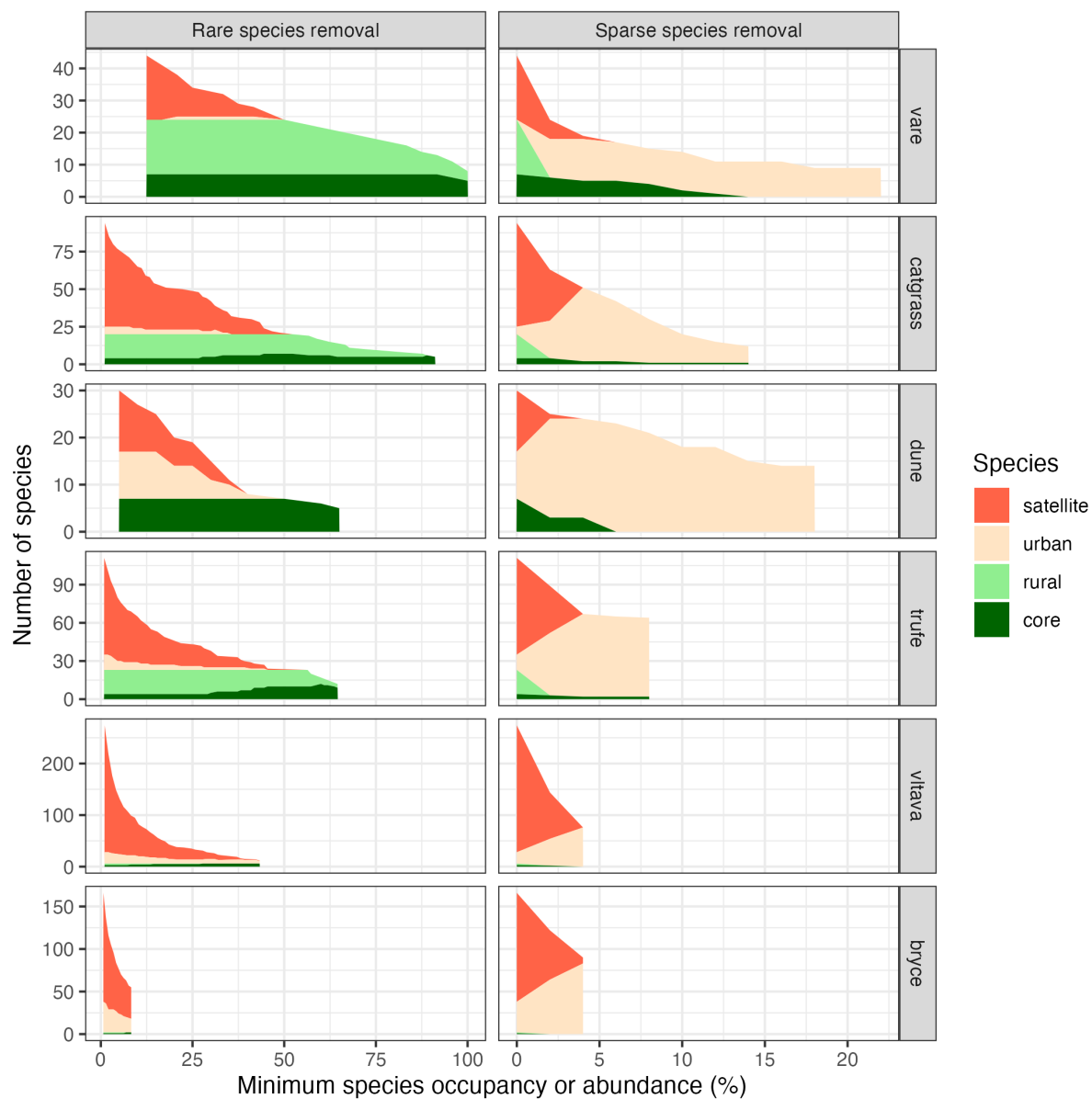
324
325 Figure 4. Summary of the methodology applied to assess the impact of progressive species removal on
326 abundance-occupancy pattern, diversity pattern and constrained ordination.

327 Results

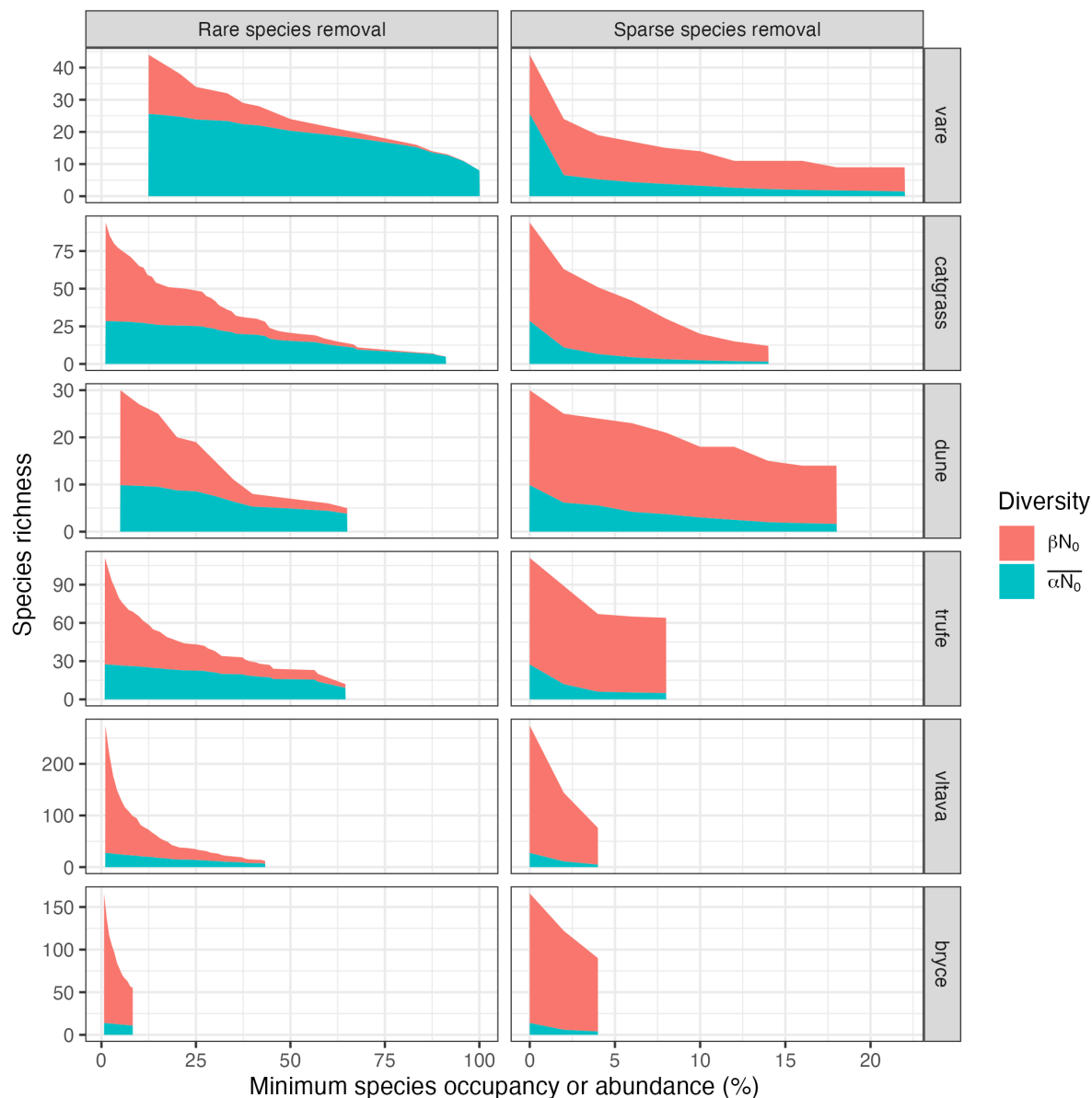
328 *Impact of progressive species removal on abundance-occupancy patterns*

329 The impact of excluding the least frequent (rare) species or the least abundant (sparse) species
330 from the original data set is consistent among the data sets (Fig. 5). When removing rare
331 species, satellite and urban species disappear first, and the proportion of core species tends to
332 increase. Rural species, when present, disappear more slowly. By contrast, the elimination of
333 locally sparse species leads to a dramatic increase of the proportion of urban species. Rural
334 species first disappear, followed by satellite and core species. Note that the exclusion of rare
335 or sparse species in large and heterogeneous data sets ('vltava' and 'bryce') quickly leads to
336 empty sites, and therefore short series.

337



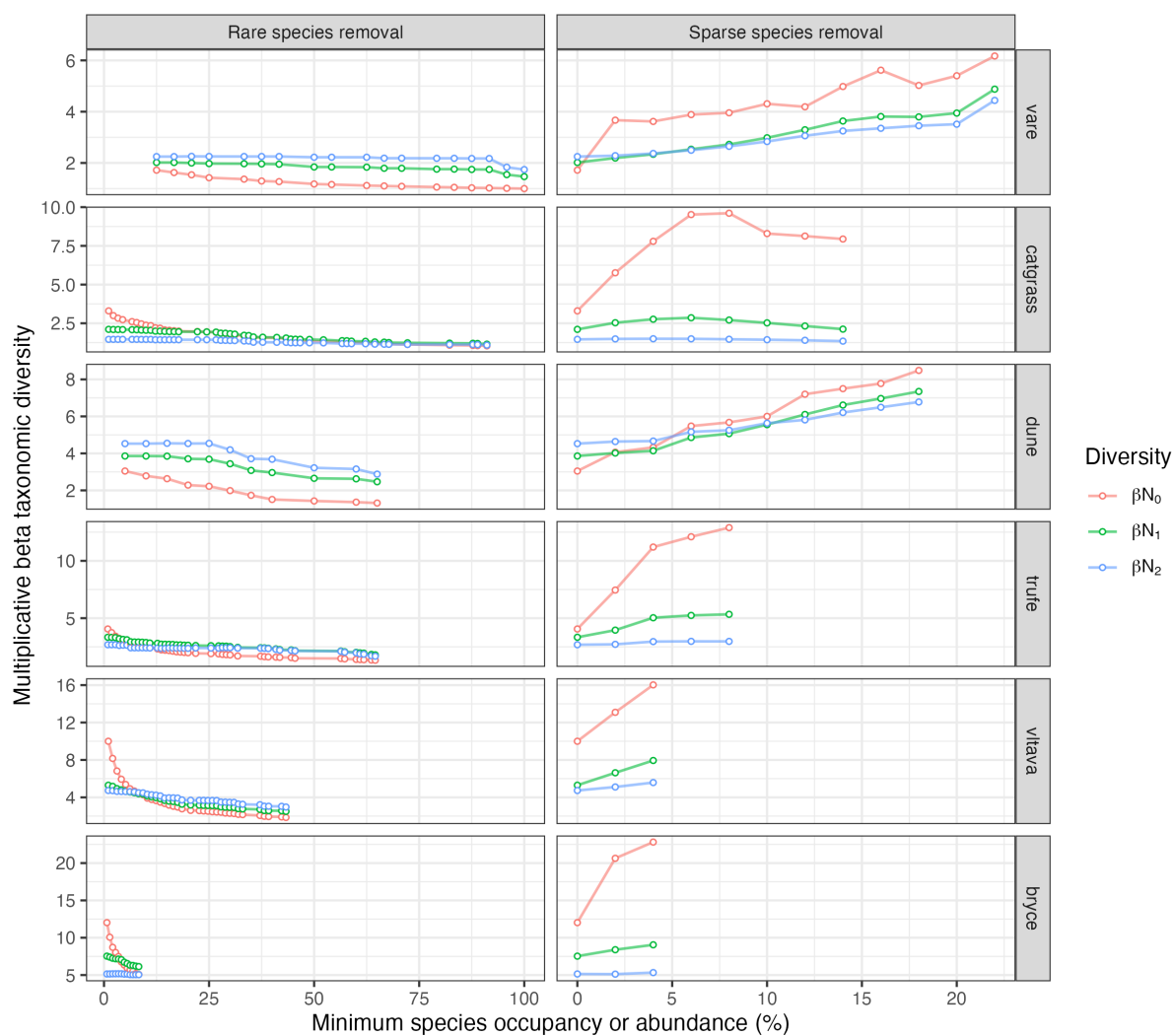
338
 339 Figure 5. Impact of rare (left) or sparse (right) species exclusion from each data set on the number of core, rural,
 340 urban and satellite species. For rare species removal, percentages on the x axis represent minimum frequency of
 341 occurrence (occupancy) and for sparse species removal they represent minimum relative cover (abundance-
 342 when-present).

343 *Impact on beta diversity assessment*

344
 345 Figure 6. Impact of rare (left) and sparse (right) species exclusion from each data set on the additive partitioning
 346 of gamma species richness ($q = 0$). Blue area: mean alpha species richness (average number of species present at
 347 each site of the data set); red area: additive beta species richness (average number of species absent from each
 348 site but present in other sites of the data set); total area: gamma species richness (total number of species in the
 349 data set).

350
 351 When excluding rare species based on minimum frequency, additive beta species richness (q
 352 $= 0$) decreases more than mean alpha species richness, contrary to exclusion of sparse species
 353 based on minimum relative cover, for which alpha diversity declines more rapidly (Fig. 6).
 354 Similar patterns are obtained for Hill-Shannon ($q = 1$) and Hill-Simpson ($q = 2$) diversity
 355 (Supplementary Material, Appendix S1).

356



357

358 Figure 7. Impact of rare (left) or sparse (right) species exclusion on the multiplicative beta taxonomic diversity.

359 βN_0 : multiplicative beta diversity of all species (Whittaker); βN_1 : multiplicative beta diversity of abundant360 species (Hill-Shannon); βN_2 : multiplicative beta diversity of dominant species (Hill-Simpson).

361

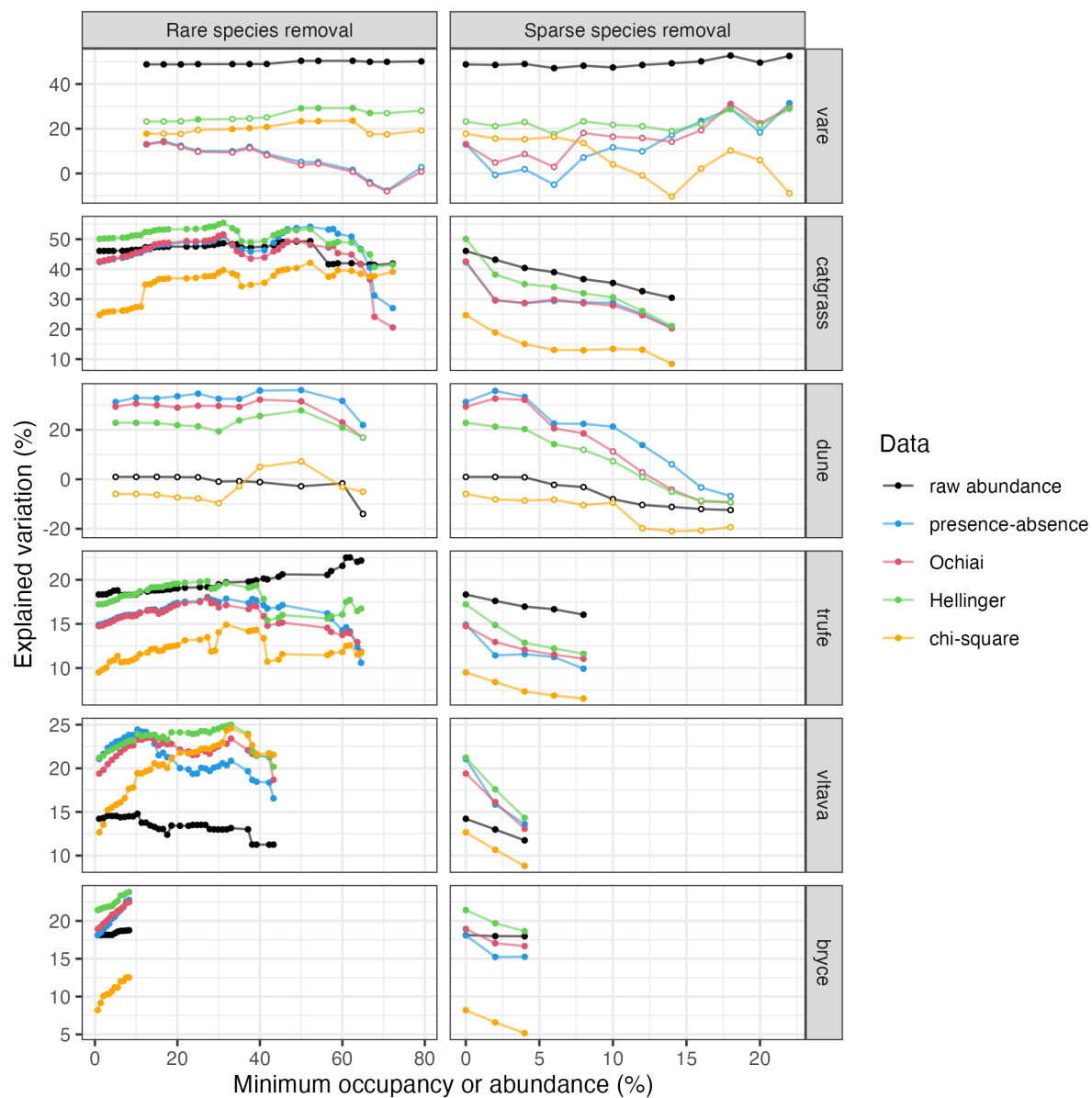
362 The difference is even more evident when considering multiplicative beta diversity of
 363 orders 0, 1 and 2 (Fig. 7). All beta diversity indices decline when removing rare species,
 364 whereas they tend to increase when removing sparse species. Apart from these general trends,
 365 we observe strong differences in the relative impact of species removal on these indices
 366 among data sets, depending on their initial structure. However, the multiplicative beta
 367 diversity of dominant or abundant species (βN_2 and βN_1 , respectively) is generally less
 368 affected by species removal than beta diversity computed on presence-absence data (βN_0).
 369 Irrespective of the initial values of beta diversity indices, beta species richness βN_0 becomes
 370 always the lowest at the end of rare species exclusion and the highest at the end of sparse

371 species exclusion. This confirms that the exclusion of rare species leads to an increasingly
372 apparently homogeneous community composition, whereas the exclusion of sparse species
373 leads to greater apparent heterogeneity among the sites.

374 *Impact on RDA and community-environment relationship assessment*

375 The impact of species removal on RDA results was assessed by the variation of the
376 percentage of explained variation (model performance expressed by adjusted R^2 , Fig. 8) and
377 of the contribution of the first axis to this explained variation (Fig. 9) following a progressive
378 exclusion of rare species (left) or of sparse species (right) records.

379 In the case of 'catgrass', the data set for which the percentage of explained variation was
380 initially the highest, both model performance and contribution of the first axis are slightly
381 improved by rare species removal until the suppression of most satellite and urban species (at
382 about 30% frequency), whatever the pre-transformation of relative cover data. By contrast,
383 the elimination of sparse species tends to decrease model performance in all simulations.

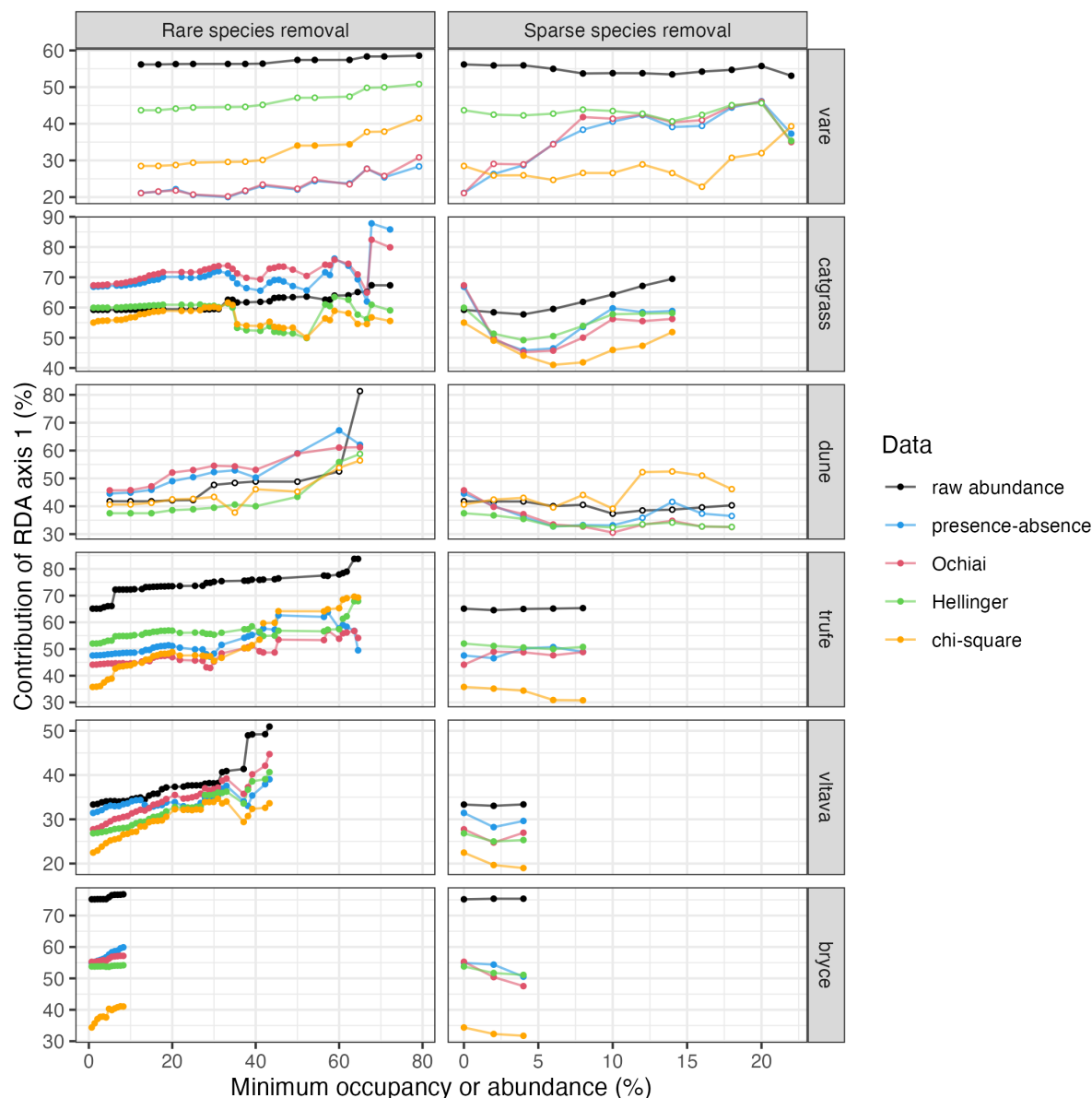


384

385 Figure 8. Impact of rare (left) and sparse (right) species exclusion on the percentage of explained variation of

386 RDA (adjusted R^2) without or after several pre-transformations of each species matrix. Solid circles indicate387 significant results of permutation tests ($p < 0.05$, 999 permutations).

388



389
 390 Figure 9. Impact of rare (left) and sparse (right) species exclusion on the contribution of axis 1 of RDA without
 391 or after several pre-transformations of each species matrix. Solid circles indicate significant results of
 392 permutation tests ($p < 0.05$, 999 permutations).
 393

394 For 'vare', a homogeneous data set in which β -diversity is mainly due to dominant
 395 species, the best results are obtained with raw absolute cover data, which are not affected by
 396 rare or sparse species exclusion. Most prior transformations of the data lead to non-significant
 397 results. Nevertheless, the performance of the model can be slightly improved by eliminating
 398 rare species after Hellinger or chi-square transformation.

399 With 'dune', the best results are obtained after presence-absence or Ochiai transformation.
 400 They are improved by the elimination of rare species, but worsened by the exclusion of sparse
 401 species.

402 For 'trufe', conclusions are similar to those obtained from the 'catgrass' example, but with
403 better results for raw abundance, even at the end of rare species removal.

404 For 'vltava', Ochiai and Hellinger site profiles, as well as presence-absence data produce
405 the highest explained variation slightly improved by the removal of species below 10%
406 occupancy, but strongly impaired by the loss of sparse species.

407 Finally, in the most heterogeneous data set ('bryce'), the adjusted R^2 of the model and the
408 contribution of axis 1 are positively affected by the removal of rare species whatever the
409 transformation applied, whereas the removal of sparse species affects model performance
410 negatively, except for raw abundance data.

411 Despite strong differences among data sets, some general conclusions may be drawn after
412 comparing the results of these simulations:

- 413 (1) With raw data or site profiles of abundance data (i.e. Hellinger transformation), model
414 performance is not or only slightly enhanced when removing the least frequent species.
- 415 (2) By contrast, model performance is generally negatively affected by the progressive
416 exclusion of sparse species, which mimics a decreasing sampling effort in each
417 community.
- 418 (3) The ranking of prior transformation of absolute cover data according to model
419 performance varies strongly among the data sets, yet without affecting the general effect
420 of rare or sparse species removal. Interestingly, the Ochiai transformation of presence-
421 absence data does not improve model performance, as compared to the simple binary
422 transformation, in any data set. However, canonical redundancy analysis on raw absolute
423 cover data or on presence-absence data considers double zeros in the calculation of
424 distances between sites and for this reason should only be applied to matrices with very
425 few zeros (Borcard et al., 2018).
- 426 (4) In large and heterogeneous datasets, the removal of rare species clearly improves the
427 performance of RDA after double transformation (chi-square), contrary to the removal of
428 sparse species.

429 **Discussion**

430 *Minor species, β -diversity and ecological assessment*

431 To assess the impact of reduced sampling effort and incomplete representation of local
432 plant diversity on mean ecological indicator values of vegetation relevés, Ewald (2003)

433 simulated a reduction of an original compositional matrix by randomly deleting 1, 10, 20, 40
434 and 80% of species records with low abundance. The random omission of low-abundance
435 species affected the correlation between log-abundance weighted average ecological indicator
436 values and measured environmental variables only weakly. Incomplete recording of
437 taxonomic information increased multiplicative β -diversity, due to the non-proportional
438 relationship between α -diversity, which declined linearly with sparse species removal, and γ -
439 diversity, which was little affected until ca. 40% of the records had been deleted (Ewald,
440 2003). These last findings are consistent with the results of our progressive exclusion of
441 sparse plant species. In addition, the author applied Mantel correlation to test the overall
442 multivariate association between distance matrices of species composition and environment;
443 however, this method is now considered inappropriate for this kind of study and should be
444 replaced by the computation of RV coefficients (Escoufier, 1973; Legendre & Legendre,
445 2012; Omelka & Hudecová, 2013; Legendre et al., 2015).

446 As for animal communities, Sgarbi et al. (2020) compared the effect of sequential removal
447 of genera from the most to the least locally abundant, sequential removal from the least to the
448 most abundant or random removal on the ordination patterns of stream benthic invertebrate
449 communities. By comparing Procrustes correlations based on abundance data, they found that
450 taxon-based reduction in sampling effort, consisting in the omission of up to 50% of the least
451 abundant species (confusingly called “rarest” by these authors), affected only weakly the
452 multivariate pattern observed in the complete assemblage data, contrary to the removal of the
453 most abundant species (called “commonest”).

454 However, focusing on the effect on ecological assessment of excluding stream dwelling
455 macroinvertebrate taxa with low abundances or with small distribution ranges, Nijboer &
456 Schmidt-Klüber (2004) concluded that neither sparse nor rare taxa should be excluded,
457 because of their essential ecological indicative power. In addition, our study shows that the
458 omission of rare species had an opposite (i.e. negative) effect on multiplicative β -diversity.
459 This finding supports the determinant role of these infrequent species in the differentiation of
460 habitats. The removal of rare species may have similar or greater influence in biological
461 assessment as other choices inherent in its computation, such as the choice of ordination
462 method or measures of multivariate resemblance (Poos & Jackson, 2012).

463 Indeed, rare species may be indicators of rare environmental conditions or specific
464 habitats, to which these species could be adapted better than common species. According to
465 Rabinowitz et al. (1986), habitat specificity (ecological specialization) is the most important

466 dimension of ‘rarity’ (in a broad sense). It could explain why many rare species often co-
467 occur in one or few sites, where they may be dominant. For example, the ‘vltava’ data set
468 contains a plot with nine unique species including one dominant. Such sites and species
469 appear as outliers in ordinations and can hide the main gradients, especially when applying
470 chi-square-based methods such as CA and CCA.

471 *Minor species in canonical ordination*

472 By removing rare species from fish and odonatan communities Brasil et al. (2020) found that
473 rare species are of little importance for understanding the relationships with spatial and
474 environmental gradients using variation partitioning based on partial RDA (Borcard et al.,
475 1992; Peres-Neto et al., 2006). As shown in our study, the only improvement one can expect
476 from rare species removal in canonical ordination is with double profiles (after chi-square
477 transformation).

478 Of course, RDA model performance cannot be reduced to an adjusted R^2 and a
479 contribution of axis 1 to the analysis. The purpose here was not to use these indicators to
480 grade the different pre-transformation options, since their choice depends upon the questions
481 and hypotheses of the modeling framework. Rather, the purpose was to figure out whether the
482 effect of species removal on the RDA results was dependent on this choice. Indeed, whatever
483 the various ranking of the four options for abundance transformation (presence-absence,
484 Ochiai, Hellinger or chi-square) in terms of variation explained, a common general trend
485 emerged from the six examples, i.e. a slight improvement after a moderate exclusion of rare
486 species, and a clear degradation after any exclusion of sparse species. The only exception is
487 the ‘vare’ data set after presence-absence or Ochiai transformation, for which sparse species
488 removal has a positive effect on both variation explained and contribution of axis 1. In this
489 homogeneous, species-poor data set, species-environment relationships are mainly expressed
490 by variations in absolute cover among frequent species.

491 **Conclusion**

492 The positive correlation between inter-site occupancy (frequency of occurrence) and average
493 intra-site dominance (relative abundance-when-present) should not hide the profound
494 differences between rare (uncommon) and sparse (undominant) species. They must not be
495 confounded because they play a completely different role in vegetation analysis and
496 ecological studies of other groups of organisms. As for biodiversity assessment, removing

497 rare species decreases the proportion of satellite species, hence decreasing multiplicative beta
498 diversity, whereas removing sparse species increases the proportion of urban species, hence
499 dramatically increasing multiplicative beta diversity. As for ecological assessment, focusing
500 only on dominant species during sampling or multivariate analyses by ignoring sparse species
501 is likely to limit the performance of ecological empirical models. These conclusions are also
502 applicable to animal and microbial communities, to which we applied the same methodology
503 in a preliminary study (results not shown here).

504 One of the main findings of our simulation study was that the multiplicative beta diversity
505 decreased when removing rare species, but increased when removing the sparse ones. A
506 possible explanation is that when you ignore the most infrequent species, the whole dataset
507 becomes more homogenous because rare species, which are often numerous, are the most
508 responsible for differences in community composition, hence *decreasing* beta diversity. By
509 contrast, when you ignore sparse species (based on their average relative cover, not based on
510 their absolute cover in each separate site), you focus more on dominant and abundant species,
511 irrespective to their occupancy, while increasing the contrast between communities dominated
512 by different species, hence *increasing* beta diversity.

513 Provided that meaningful transformations are applied, there is no need to remove any
514 species prior to RDA. For homogeneous data sets (“short gradients”) in which total vegetation
515 cover and differences in species cover among sites are of interest, absolute cover data should
516 not be transformed. For heterogeneous data sets (“long gradients”) and a focus on local
517 dominance relationships, site profiles (e.g. Hellinger or other Box-Cox family transformation)
518 should be used. Removing rare species is only useful for the ordination of double profiles,
519 achieved by RDA after chi-square transformation or, more commonly, by canonical
520 correspondence analysis (CCA). If CCA is preferred, down-weighting of rare species has
521 proved to be more efficient than eliminating them (Jing et al., 2015).

522 If canonical ordination is used to extract the main ecological gradients while taking into
523 account differences in species abundances among sites, then RDA should be applied without
524 pre-transformation of abundance data, but only in case of “short gradients” (species-poor
525 communities with a small proportion of absences in the data set). To preserve the differences
526 in total abundance among sites, reflecting site productivity, in a data set containing a high
527 proportion of zeros, distance-based redundancy analysis (db-RDA) should be preferred, based
528 on a percentage difference or Ružička dissimilarity matrix (Legendre & Anderson, 1999). If
529 the community data set is very heterogeneous (“long gradient” or multiple gradients) and
530 researchers are most interested in relative proportions of species within a community rather

531 than absolute comparisons of species abundances across multiple sites, then site
532 transformation (e.g. Hellinger) should be applied prior to RDA.

533 By encouraging researchers to recognize the contrasting roles of rare and sparse species in
534 community assemblages, the results of the present study will motivate them to formulate more
535 precise working hypotheses in their own studies.

536 Our study focused on plant communities, but the question is more general and may be
537 investigated for any animal or microbial communities at a given trophic level. Recently, new
538 tools have been developed to standardize Hill diversity indices based on sample completeness
539 or coverage (Roswell et al., 2021, 2023; Chao et al., 2023). In ecology, the concept of sample
540 coverage has been only described and applied to species assemblages for which we can count
541 individuals (typically animal or tree communities), which is not the case for most plant
542 communities, where species “abundance” is generally measured by the percentage of
543 aboveground cover (or classes of absolute or relative cover). As a matter of fact, all current
544 formulae and implementations of the coverage-based or size-based standardisation of Hill
545 diversity indices (iNEXT.beta3D; Chao et al., 2023) apply to individual-based or incidence-
546 based data (point-based or grid-based frequency) and cannot be applied to continuous
547 abundance measures, such as cover or biomass. From a perspective of statistical inference,
548 rare and sparse species carry the most essential information about sample coverage. However,
549 we cannot presently apply such standardisation to our cover-based vegetation data.

550 Future work is needed to assess the role of rare and sparse species on other beta diversity
551 indices, such as dissimilarities or variance-based methods (Legendre & De Cáceres, 2013)
552 and, as soon as it is available for any type of abundance data, coverage-based standardisation
553 of Hill diversity indices, including phylogenetic and functional facets of community diversity.

554 **References**

- 555 Avolio, M.L., Forrestel, E.J., Chang, C.C., La Pierre, K.J., Burghardt, K.T., & Smith, M.D.
556 (2019) Demystifying dominant species. *New Phytologist*, 223(3), 1106–1126.
557 <https://doi.org/10.1111/nph.15789>
- 558 Béguin, D. (2007) Tree regeneration and growth in wood pastures: patterns and processes.
559 PhD thesis, University of Neuchâtel, Neuchâtel, Switzerland.
- 560 Blackburn, T.M., Cassey, P., & Gaston, K.J. (2006) Variations on a theme: Sources of
561 heterogeneity in the form of the interspecific relationship between abundance and
562 distribution. *Journal of Animal Ecology*, 75(6), 1426–1439.
563 <https://doi.org/10.1111/j.1365-2656.2006.01167.x>

- 564 Borcard, D., Gillet, F., & Legendre, P. (2018) *Numerical Ecology with R*. Springer
565 International Publishing: Cham. <https://doi.org/10.1007/978-3-319-71404-2>
- 566 Borcard, D., Legendre, P., & Drapeau, P. (1992) Partialling out the spatial component of
567 ecological variation. *Ecology*, 73(3), 1045–1055. <https://doi.org/10.2307/1940179>
- 568 Borregaard, M.K., & Rahbek, C. (2010) Causality of the relationship between geographic
569 distribution and species abundance. *Quarterly Review of Biology*, 85(1), 3–25.
570 <https://doi.org/10.1086/650265>
- 571 Brasil, L.S., Vieira, T.B., Andrade, A.F.A., Bastos, R.C., Montag, L.F. de A., & Juen, L.
572 (2020) The importance of common and the irrelevance of rare species for partition the
573 variation of community matrix: implications for sampling and conservation. *Scientific*
574 *Reports*, 10(1), 1–8. <https://doi.org/10.1038/s41598-020-76833-5>
- 575 Broennimann, O., Vittoz, P., Moser, D., & Guisan, A. (2005) Rarity types among plant
576 species with high conservation priority in Switzerland. *Botanica Helvetica*, 115(2),
577 95–108. <https://doi.org/10.1007/s00035-005-0713-z>
- 578 Cao, Y., Larsen, D., & Thorne, R. (2001) Rare species in multivariate analysis for
579 bioassessment: some considerations. *Journal of the North American Benthological*
580 *Society*, 20(1), 144–153.
- 581 Chao, A., Thorn, S., Chiu, C., Moyes, F., Hu, K., Chazdon, R.L., et al. (2023) Rarefaction and
582 extrapolation with beta diversity under a framework of Hill numbers: The iNEXT .
583 BETA3D standardization. *Ecological Monographs*, 93(4), e1588.
584 <https://doi.org/10.1002/ecm.1588>
- 585 Charney, N., & Record, S. (2012) vegetarian: Jost Diversity Measures for Community Data.
586 R package version 1.2 <https://CRAN.R-project.org/package=vegetarian>.
- 587 Collins, S.L., & Glenn, S.M. (1997) Effects of organismal and distance scaling on analysis of
588 species distribution and abundance. *Ecological Applications*, 7(2), 543–551.
589 <https://doi.org/10.1890/1051-0761>
- 590 Collins, S.L., Glenn, S.M., & Roberts, D.W. (1993) The hierarchical continuum concept.
591 *Journal of Vegetation Science*, 4(2), 149–156. <https://doi.org/10.2307/3236099>
- 592 Crisfield, V.E., Guillaume Blanchet, F., Raudsepp-Hearne, C., & Gravel, D. (2024) How and
593 why species are rare: towards an understanding of the ecological causes of rarity.
594 *Ecography*, e07037. <https://doi.org/10.1111/ecog.07037>
- 595 De Bello, F., Lavergne, S., Meynard, C.N., Lepš, J., & Thuiller, W. (2010) The partitioning of
596 diversity: Showing Theseus a way out of the labyrinth. *Journal of Vegetation Science*,
597 21(5), 992–1000. <https://doi.org/10.1111/j.1654-1103.2010.01195.x>
- 598 Dengler, J., & Dembiczy, I. (2023) Should we estimate plant cover in percent or on ordinal
599 scales? *Vegetation Classification and Survey*, 4, 131–138.
600 <https://doi.org/10.3897/VCS.98379>
- 601 Eriksson, O. (2013) A closer look at the species behind abundance-occupancy relationships.
602 *Journal of Vegetation Science*, 24(4), 589–590. <https://doi.org/10.1111/jvs.12063>

- 603 Escoufier, Y. (1973) Le traitement des variables vectorielles. *Biometrics*, 29(4), 751.
604 <https://doi.org/10.2307/2529140>
- 605 Ewald, J. (2003) The sensitivity of Ellenberg indicator values to the completeness of
606 vegetation relevés. *Basic and Applied Ecology*, 4(6), 507–513.
607 <https://doi.org/10.1078/1439-1791-00155>
- 608 Gaston, K.J. (1998) Rarity as double jeopardy. *Nature*, 394(6690), 229–230.
609 <https://doi.org/10.1038/28288>
- 610 Gaston, K.J., Blackburn, T.I.M.M., Greenwood, J.D., Gregory, R.D., Quinn, M., & Lawton,
611 J.H. (2000) Abundance-occupancy relationships. *Journal of Applied Ecology*, 37, 39–
612 59.
- 613 Gibson, D.J., Ely, J.S., & Collins, S.L. (1999) The core-satellite species hypothesis provides a
614 theoretical basis for Grime's classification of dominant, subordinate, and transient
615 species. *Journal of Ecology*, 87(6), 1064–1067. [https://doi.org/10.1046/j.1365-
616 2745.1999.00424.x](https://doi.org/10.1046/j.1365-2745.1999.00424.x)
- 617 Grime, J.P. (1998) Benefits of plant diversity to ecosystems: immediate, filter and founder
618 effects. *Journal of Ecology*, 86, 902–910.
- 619 Guedo, D.D., & Lamb, E.G. (2013) Temporal changes in abundance-occupancy relationships
620 within and between communities after disturbance. *Journal of Vegetation Science*,
621 24(4), 607–615. <https://doi.org/10.1111/jvs.12006>
- 622 Hanski, I. (1982) Dynamics of Regional Distribution: The Core and Satellite Species
623 Hypothesis. *Oikos*, 38(2), 210. <https://doi.org/10.2307/3544021>
- 624 Hanski, I. (1991) Single-species metapopulation dynamics: concepts, models and
625 observations *Biological Journal of the Linnean Society* 42: 17–38.
- 626 Harrison, S., Viers, J.H., Thorne, J.H., & Grace, J.B. (2008) Favorable environments and the
627 persistence of naturally rare species. *Conservation Letters*, 1(2), 65–74.
628 <https://doi.org/10.1111/j.1755-263X.2008.00010.x>
- 629 Hill, M.O. (1973) Diversity and evenness: A unifying notation and its consequences. *Ecology*,
630 54(2), 427–432. <https://doi.org/10.2307/1934352>
- 631 Jing, C., Yan-Ming, M., Fei, F., Qiang, X., Qin-Di, Z., & Run-Cheng, B. (2015) Comparison
632 of different treatments of rare species in canonical correspondence analysis. *Chinese
633 Journal of Plant Ecology*, 39(2), 167–175. <https://doi.org/10.17521/cjpe.2015.0016>
- 634 Jongman, R.H.G., Ter Braak, C.J.F., & Van Tongeren, O.F.R. (1995) *Data analysis in
635 community and landscape ecology*. Cambridge University Press: Cambridge.
636 <https://doi.org/10.2307/2531665>
- 637 Jost, L. (2006) Entropy and diversity. *Oikos*, 113(2), 363–375.
638 <https://doi.org/10.1111/j.2006.0030-1299.14714.x>
- 639 Jost, L. (2007) Partitioning diversity into independent alpha and beta components. *Ecology*,
640 88(10), 2427–2439. <https://doi.org/10.1890/06-1736.1>

- 641 Kassambara, A. (2023) ggpubr: 'ggplot2' based publication ready plots. R package version
642 0.6.0.
- 643 Kohler, F. (2004) Influence of grazing, dunging and trampling on short-term dynamics of
644 grasslands in mountain wooded pastures. PhD thesis, University of Neuchâtel,
645 Neuchâtel, Switzerland.
- 646 Legendre, P., & Anderson, M.J. (1999) Distance-based redundancy analysis: testing
647 multispecies responses in multifactorial ecological experiments. *Ecological*
648 *Monographs*, 69(1), 1–24. <https://doi.org/10.2307/2657228>
- 649 Legendre, P., & De Cáceres, M. (2013) Beta diversity as the variance of community data:
650 dissimilarity coefficients and partitioning. *Ecology Letters*, 16(8), 951–963.
651 <https://doi.org/10.1111/ele.12141>
- 652 Legendre, P., Fortin, M.-J., & Borcard, D. (2015) Should the Mantel test be used in spatial
653 analysis? *Methods in Ecology and Evolution*, 6(11), 1239–1247.
654 <https://doi.org/10.1111/2041-210X.12425>
- 655 Legendre, P., & Gallagher, E.D. (2001) Ecologically meaningful transformations for
656 ordination of species data. *Oecologia*, 129(2), 271–280.
657 <https://doi.org/10.1007/s004420100716>
- 658 Legendre, P., & Legendre, L. (2012) *Numerical Ecology, Third English Edition*. Elsevier:
659 Amsterdam.
- 660 Lortie, C.J., Brooker, R.W., Choler, P., Kikvidze, Z., Michalet, R., & Pugnaire, F.I. (2004)
661 Rethinking plant community theory. *Oikos*, 107(2), 433–438.
662 <https://doi.org/10.1111/j.0030-1299.2004.13250.x>
- 663 van der Maarel, E. (1979) Transformation of cover-abundance values in phytosociology and
664 its effects on community similarity. *Vegetatio*, 39(2), 97–114.
- 665 Mariotte, P. (2014) Do subordinate species punch above their weight? Evidence from above-
666 and below-ground. *New Phytologist*, 203(1), 16–21. <https://doi.org/10.1111/nph.12789>
- 667 Mariotte, P., Buttler, A., Kohler, F., Gilgen, A.K., & Spiegelberger, T. (2013) How do
668 subordinate and dominant species in semi-natural mountain grasslands relate to
669 productivity and land-use change? *Basic and Applied Ecology*, 14(3), 217–224.
670 <https://doi.org/10.1016/j.baae.2013.02.003>
- 671 Mariotte, P., Vandenberghe, C., Kardol, P., Hagedorn, F., & Buttler, A. (2013) Subordinate
672 plant species enhance community resistance against drought in semi-natural
673 grasslands. *Journal of Ecology*, 101(3), 763–773. <https://doi.org/10.1111/1365-2745.12064>
- 675 Nijboer, R.C., & Schmidt-Kloiber, A. (2004) The effect of excluding taxa with low
676 abundances or taxa with small distribution ranges on ecological assessment.
677 *Hydrobiologia*, 516, 347–363.
- 678 Oksanen, J., Simpson, G.L., Blanchet, F.G., Kindt, R., Legendre, P., Minchin, P.R., et al.
679 (2024) vegan: Community Ecology Package. R package version 2.6-6.1.

- 680 Omelka, M., & Hudecová, Š. (2013) A comparison of the Mantel test with a generalised
681 distance covariance test. *Environmetrics*, 24(7), 449–460.
682 <https://doi.org/10.1002/env.2238>
- 683 Peres-Neto, P.R., Legendre, P., Dray, S., & Borcard, D. (2006) Variation partitioning of
684 species data matrices: Estimation and comparison of fractions. *Ecology*, 87(10), 2614–
685 2625. <https://doi.org/10.1890/0012-9658>
- 686 Poos, M.S., & Jackson, D.A. (2012) Addressing the removal of rare species in multivariate
687 bioassessments: The impact of methodological choices. *Ecological Indicators*, 18, 82–
688 90. <https://doi.org/10.1016/j.ecolind.2011.10.008>
- 689 Rabinowitz, D. (1981) Seven forms of rarity. In: Synge, H. (Ed.), *The biological aspects of*
690 *rare plant conservation*. John Wiley and Sons: Chichester, UK, pp. 205–217.
- 691 Rabinowitz, D., Cairns, S., & Dillon, T. (1986) Seven forms of rarity and their frequency in
692 the flora of the British Isles. In: Soule, M.E. (Ed.), *Conservation biology, the science*
693 *of scarcity and diversity*. Sinauer Associates: Sunderland, Massachusetts, pp. 182–204.
- 694 RCoreTeam. (2024) R: A language and environment for statistical computing. Version 4.4.1.
695 <https://www.R-project.org/>.
- 696 Roberts, D.W. (2023) labdsv: Ordination and Multivariate Analysis for Ecology. R package
697 version 2.1-0.
- 698 Roswell, M., Dushoff, J., & Winfree, R. (2021) A conceptual guide to measuring species
699 diversity. *Oikos*, 130(3), 321–338. <https://doi.org/10.1111/oik.07202>
- 700 Roswell, M., Harrison, T., & Genung, M.A. (2023) Biodiversity–ecosystem function
701 relationships change in sign and magnitude across the Hill diversity spectrum.
702 *Philosophical Transactions of the Royal Society B: Biological Sciences*, 378(1881),
703 20220186. <https://doi.org/10.1098/rstb.2022.0186>
- 704 Sgarbi, L.F., Bini, L.M., Heino, J., Jyrkänkallio-Mikkola, J., Landeiro, V.L., Santos, E.P., et
705 al. (2020) Sampling effort and information quality provided by rare and common
706 species in estimating assemblage structure. *Ecological Indicators*, 110(April 2019),
707 105937. <https://doi.org/10.1016/j.ecolind.2019.105937>
- 708 Söderström, L. (1989) Regional distribution patterns of bryophyte species on spruce logs in
709 Northern Sweden. *The Bryologist*, 92(3), 349–355. <https://doi.org/10.2307/3243403>
- 710 Tuomisto, H. (2010) A diversity of beta diversities: Straightening up a concept gone awry.
711 Part 1. Defining beta diversity as a function of alpha and gamma diversity. *Ecography*,
712 33(1), 2–22. <https://doi.org/10.1111/j.1600-0587.2009.05880.x>
- 713 Väre, H., Ohtonen, R., & Oksanen, J. (1995) Effects of reindeer grazing on understory
714 vegetation in dry *Pinus sylvestris* forests. *Journal of Vegetation Science*, 6(4), 523–
715 530. <https://doi.org/10.2307/3236351>
- 716 Whittaker, R.H. (1972) Evolution and measurement of species diversity. *Taxon*, 21(2), 213–
717 251.

718 Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L.D., François, R., et al. (2019)
719 Welcome to the tidyverse. *Journal of Open Source Software*, 4(43), 1686.
720 <https://doi.org/10.21105/joss.01686>

721 Zelený, D., & Chytrý, M. (2007) Environmental control of the vegetation pattern in deep river
722 valleys of the Bohemian Massif. *Preslia*, 79(3), 205–222.

723

724 **Acknowledgements**

725 We thank Florian Kohler, Daniel Béguin, David Zelený, Jari Oksanen and Dave Roberts for
726 making available the data used in this study. We are grateful to the participants of the 56th
727 IAVS Symposium in Tartu for their positive comments after an oral presentation of the first
728 results of this study by F.G. We thank Dave Roberts and an anonymous reviewer for helpful
729 comments on an earlier version of the manuscript.

730

731 **Author contributions**

732 F.G. conceived the research idea; F.G. and A.R. performed statistical analyses; F.G., with
733 contributions from D.B. and P.L., wrote the paper; all authors discussed the results and
734 commented on the manuscript.

735

736 **Data availability statement**

737 The data supporting the article's results and the R scripts used to generate the analyses and the
738 figures are provided in the electronic Supplementary Information Appendix S2.

739

740 **Supporting information**

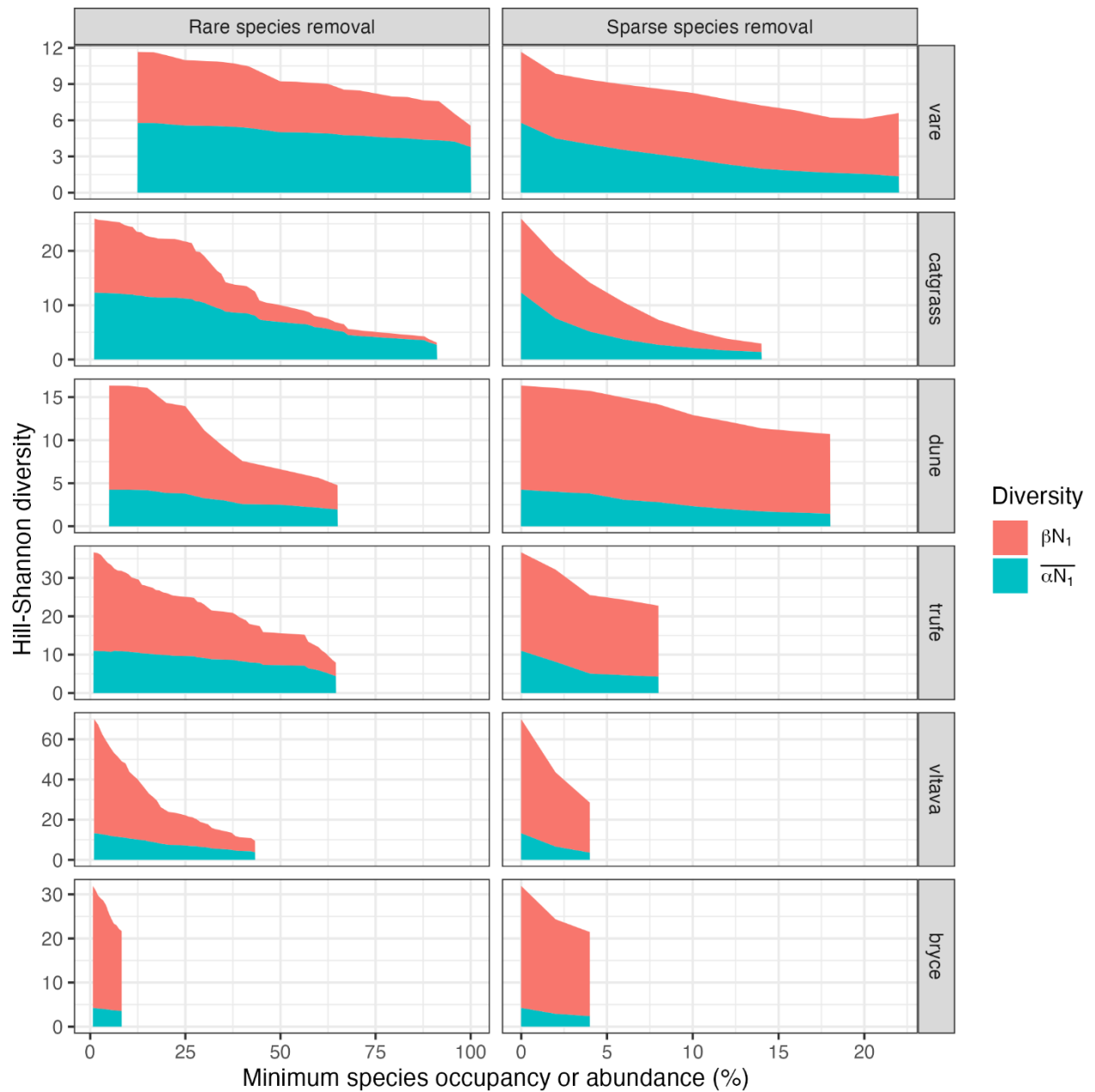
741 Appendix S1. Additional figures.

742 Appendix S2. Data and R code.

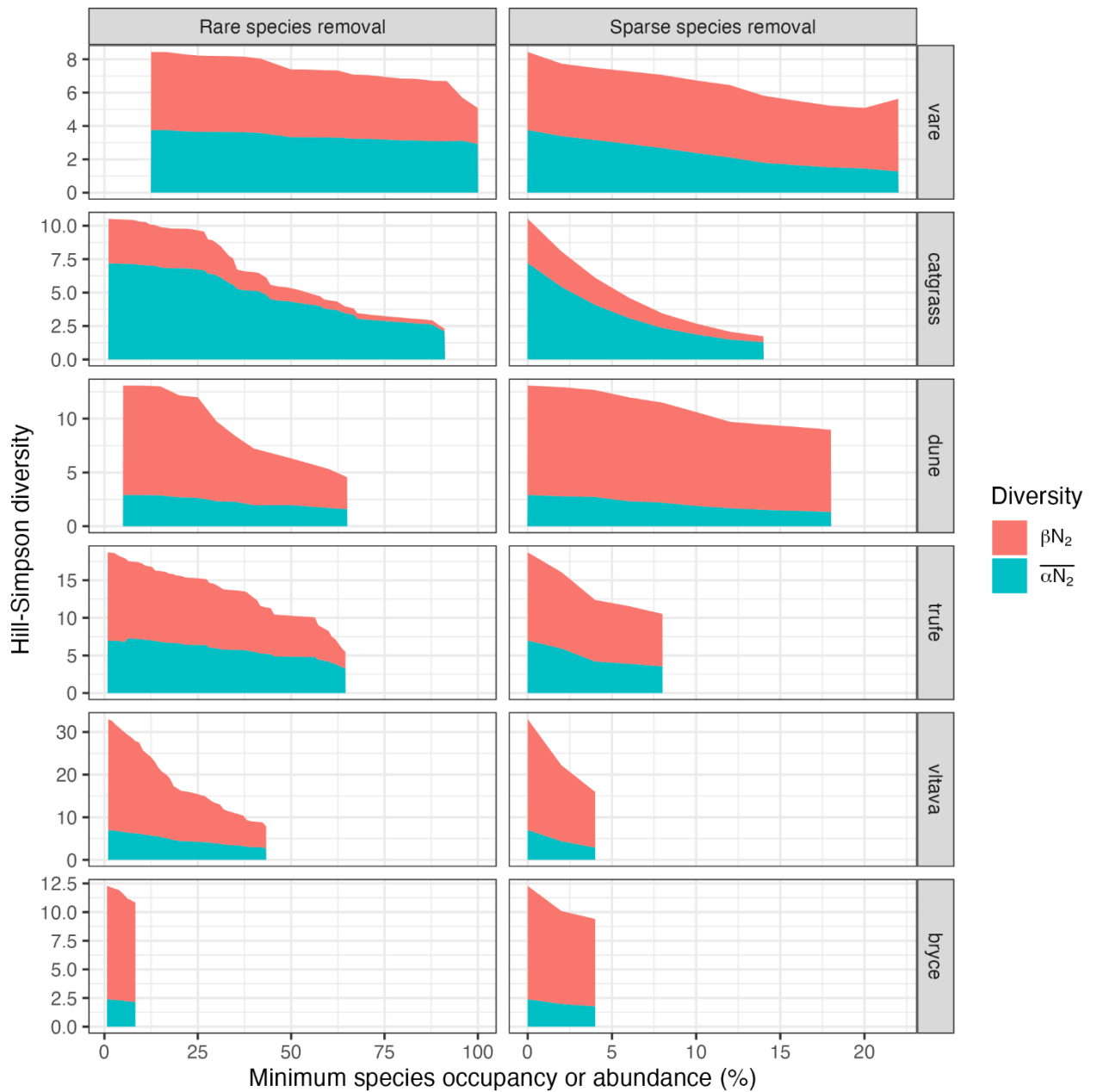
743

744

745 **Appendix S1.** Additional figures.
746



747 **Figure S1.** Impact of rare (left) and sparse (right) species exclusion from each data set on the additive
748 partitioning of gamma Hill-Shannon diversity. Blue area: mean alpha Hill-Shannon diversity (average
749 number of abundant species present at each site of the data set); red area: additive beta Hill-Shannon
750 diversity (average number of abundant species absent from each site but present in other sites of the
751 data set); total area: gamma Hill-Shannon diversity (total number of abundant species in the data set).
752
753



754

755

756

757

758

759

760

Figure S2. Impact of rare (left) and sparse (right) species exclusion from each data set on the additive partitioning of gamma Hill-Simpson diversity. Blue area: mean alpha Hill-Simpson diversity (average number of dominant species present at each site of the data set); red area: additive beta Hill-Simpson diversity (average number of dominant species absent from each site but present in other sites of the data set); total area: gamma Hill-Simpson diversity (total number of dominant species in the data set).