zetadiv

zetadiv: an R package for computing compositional change across multiple sites, assemblages or cases

Guillaume Latombe[1,3,5], Melodie A. McGeoch[1], David A. Nipperess[2], Cang Hui[3,4]

[1]School of Biological Sciences, Monash University, Melbourne 3800, Australia

[2]Department of Biological Sciences, Macquarie University, North Ryde, NSW 2109, Australia

[3]Centre for Invasion Biology, Department of Mathematical Sciences, Stellenbosch University, Matieland 7602, South Africa

[4]African Institute for Mathematical Sciences, Cape Town 7945, South Africa

[5] Email: latombe.guillaume@gmail.com

Highlights

- An R package to analyse compositional change using zeta diversity is presented.
- Zeta diversity is the mean number of species shared by any number of assemblages
- Zeta diversity captures all diversity components produced by assemblage partitioning
- Analyses relate zeta diversity to space, environment and spatial scale
- Analyses differentiate the contribution of rare and common species to biodiversity

Abstract

Spatial variation in compositional diversity, or species turnover, is necessary for capturing the components of heterogeneity that constitute biodiversity. However, no incidence-based metric of pairwise species turnover can calculate all components of diversity partitioning. Zeta (ζ) diversity, the mean number of species shared by any given number of sites or assemblages, captures all diversity components produced by assemblage partitioning. *zetadiv* is an R package for analysing and measuring compositional change for occurrence data using zeta diversity. Four types of analyses are performed on bird composition data in Australia: (i) decline in zeta diversity; (ii) distance decay; (iii) multi-site generalised dissimilarity modelling; and (iv) hierarchical scaling. Some analyses, such as the zeta decline, are specific to zeta diversity, whereas others, such as distance decay, are commonly applied to beta diversity, and have been adapted using zeta diversity to differentiate the contribution of common and rare species to compositional change.

Keywords: species turnover, alpha diversity, beta diversity, zeta diversity, occurrence data.

1. Introduction

1.1. Species turnover in practice

Spatial variation in compositional diversity, or species turnover, is one of the key properties for quantifying the components of heterogeneity that constitute biodiversity, along with total richness and measures of uniqueness, such as endemism and phylogenetic distinctiveness (Magurran and McGill, 2011). Species turnover can show a wide range of responses to environmental changes, and good conservation practice requires the understanding derived from its effective measurement and description for both species that are common and rare (McGeoch and Latombe, 2016; Socolar et al., 2016).

Despite the role of compositional dissimilarity (or similarity) in understanding biodiversity, no single measure previously connected the range of assemblage patterns constructed from species presence-absence data (Hui and McGeoch, 2014). Species turnover is traditionally measured by beta diversity, which quantifies compositional change between pairs of individual assemblages (Chao et al., 2012; Jost, 2007). To compare three or more assemblages, the mean of the pairwise similarities is often used (Jost et al., 2011). However, such incidence-based metrics of pairwise compositional change emphasize the differences in rare species composition between assemblages, and do not capture the characteristics of community structures caused by common species shared by many assemblages. Although multiple-site metrics have also been developed to quantify the heterogeneity in assemblage composition (Baselga, 2013; Diserud and Ødegaard, 2007; Ricotta and Pavoine, 2015), these

74 measures rely on averaging non-independent pairwise values and are difficult to
75 interpret.
76
77 1.2. Necessity of zeta diversity
78 Zeta (ζ) diversity, the mean number of species shared by any given number of sites or
79 assemblages, was proposed as a metric to capture all diversity components produced
80 by assemblage partitioning (Hui and McGeoch, 2014). Computing zeta diversity for
81 combinations of sites from 2 to *n* sites (the orders of zeta), where *n* is the total number
82 of sites, along with ζ_1, the average number of species per site (i.e. alpha diversity), is
83 necessary to obtain a mathematically comprehensive description of species
84 assemblages, and cannot be achieved by only considering alpha and beta diversity.
85 Let us consider a simple example with three sites containing 22 species each (i.e. ζ_1 or
86 α). Let us assume that each site shares exactly 10 species with any of the other two
87 sites (i.e. ζ_2 or β) (Figure 1). There are then multiple ways to partition species
88 diversity between the three sites. At one extreme, there may be no species shared by
89 the three sites simultaneously ($\zeta_3 = 0$). At the other extreme, the 10 species shared by
90 any two sites may actually be extremely common and be shared by all three sites ($\zeta_3 =$
91 10) (Figure 1). These different partitions of species diversity therefore correspond to
92 very different species assemblages, whereas they have the same alpha and beta
93 diversity values. Many other examples are possible, where alternative diversity
94 partitions exist even with the same alpha, beta and gamma diversity values.
95
96 As a consequence of the comprehensive description provided by zeta diversity, as
97 outlined by Hui and McGeoch (2014), zeta diversity enables the computation of a
98 broad range of existing diversity metrics, and the quantification of continuous change

in biodiversity over landscapes. For example, species accumulation curves, endemic-effort relationships and occupancy-frequency distributions can all be derived from zeta diversity. Importantly, from examining an extensive dataset of 291 communities, Hui and McGeoch (2014) identified two most common parametric forms of zeta diversity decline with the increase in the number of given sites – negative exponential and power law, which together account for 80% of examined communities and may differentiate stochastic from deterministic assembly processes, respectively (see for example Roura-Pascual et al., 2016). By providing a common currency for measuring biodiversity from occurrence data, zeta diversity provides an avenue for understanding the mechanistic basis of spatial patterns in diversity. This includes examining if environmental change affects rare and common species differently, or testing hypotheses about the relative importance of deterministic versus stochastic assembly processes in generating patterns of biodiversity.

Because it links different community patterns together, zeta diversity can be used for identifying community assembly processes. The identification of processes generating community assemblages usually relies on community patterns (e.g. Dornelas et al., 2006; Latombe et al., 2015). Since multiple assembly processes can generate the same community pattern, multiple patterns are needed to provide a more comprehensive description of the community and discriminate between processes (Grimm et al., 2005; Grimm and Railsback, 2012). Using multiple, different patterns can nonetheless generate bias due to possible redundancy between their information content (Latombe et al., 2011). Since multiple incidence-based patterns can be derived from zeta diversity, zeta diversity offers a powerful basis for discriminating between community assembly processes while avoiding issues of pattern redundancy. Following this logic,

zeta diversity has been used to compare and provide insights on the nature of compositional change over space and time using 10 datasets encompassing a whole range of levels of biological organisation at various spatial and temporal scales, including birds, insects, plants, microbes, and crop pests, but also intracellular processes in humans, showing the potential of zeta diversity for describing and unveiling the functioning of systems beyond classical site-by-species structure (McGeoch et al., 2017).

Importantly, zeta diversity enables the contribution of rare and common species to compositional change to be disentangled. On average, common (widespread) species are more likely to be present in any site and to be shared by any two sites than rare species. The variation in the number of species shared by different pairs of sites are therefore mostly driven by rare species, and so are analyses based only on alpha and beta diversity. By contrast, since rare species cannot, by definition, be shared by many sites, differences in zeta values for high orders of zeta is only driven by common species. Although conservation actions are mostly orientated towards rare species, common species are getting more attention (McGeoch and Latombe, 2016) as their importance for ecosystem functions is increasingly recognised (Gaston, 2010). Understanding the contribution of common and rare species to species turnover is therefore necessary. In practice, defining the distinction between rare and common species is subjective and must be done for each species individually. By contrast, zeta diversity calculates the contribution of species from rare to common as a continuum, avoiding multiple and largely subjective decisions.

1.3. Aims and novelty of the *zetadiv* R package

149 Here we introduce the *zetadiv* package for R (R CoreTeam, 2013). The *zetadiv*
150 package (available on CRAN; https://CRAN.R-project.org/package=zetadiv) was
151 created to measure and analyse compositional change for occurrence data using zeta
152 diversity. The functions of the *zetadiv* package can be categorised into four kinds of
153 analyses described in detail in the following (Appendix A, Table A1): (i) the analysis
154 of zeta diversity decline explores how the number of species shared by multiple
155 assemblages decreases with increasing number of assemblages within combinations,
156 and what information is contained in the form of this decline; (ii) the analysis of the
157 distance decay of zeta diversity illustrates how zeta diversity for different orders
158 varies with distance between sites; (iii) Multi-Site Generalised Dissimilarity
159 Modelling (MS-GDM, an adaptation of Generalised Dissimilarity Modelling; Ferrier
160 et al., 2007), computes the contribution of different environmental variables and
161 distance to zeta diversity for different orders; (iv) the analysis of the hierarchical
162 scaling of zeta diversity unravels how zeta diversity varies with grain.
163
164 Analysis of the decline in zeta diversity uses an incremental increase in the numbers
165 of assemblages included in the combinations. It is therefore an application unique to
166 zeta diversity as it combines alpha and beta (ζ_1 and ζ_2) with higher orders of zeta in a
167 single analysis, to provide a comprehensive description of species turnover. By
168 contrast, as we detail below, the other three kinds of analyses have in the past been
169 applied using beta diversity, and other R packages exist to compute such analyses.
170 The *vegan* package (Oksanen et al., 2018) enables the computation of a wide range of
171 beta diversity measures and of the hierarchical scaling of beta diversity with sampling
172 grain. The *simba* package (Jurasinski and Retzer, 2012) enables the comparison of
173 different slopes of the distance decay of beta similarity. Generalised Dissimilarity

Modelling can be performed on beta diversity using the *gdm* package (Manion et al., n.d.). As we illustrate in the examples below, the *zetadiv* package extends such analyses and enables their application to zeta diversity for selected numbers of assemblages beyond pairwise beta diversity ($n \geq 2$; see the full R code in Appendix B for fully reproducible examples and figures).

2. Biodiversity data

All the functions of *zetadiv* require at most four types of data (except for the functions that use the outputs of other functions, such as plotting functions). (i) Occurrence data, in the form of sites-by-species (rows-by-columns) data frames, are required by all functions. (ii) When spatial information is needed, a data frame with the projected or geographical coordinates of the sites or assemblages can be used. (iii) Instead of the spatial coordinates, a distance matrix between sites, independently computed, can be provided, when measures of connectivity other than Euclidian or orthodromic (i.e. distance between two points on the globe defined by their geographic coordinates) distance are required (e.g. Manhattan distance or distance accounting for the path of least resistance). (iv) A site-by-variable data frame representing the environmental variables of the sites or assemblages can be provided for MS-GDM analysis.

Two datasets, describing two different ecosystems, and complying with these requirements are included in the package to demonstrate the functions (Appendix A, Table A2). The first dataset is an inventory of resident, terrestrial bird survey data (presence-only) from the BirdLife Australia Atlas of Australian Birds (1998-2013) and covering South-East Australia (Barrett et al., 2003). The species occurrences are

complemented by maps of environmental variables for the same region, including proportion of natural environments, irrigated agriculture and plantations, as well as human density, water features (www.abs.gov.au), temperature and precipitation (www.worldclim.org; Fick and Hijmans, 2017), and elevation (www.gebco.net). The bird and environmental data were arranged into two continuous grids at two different spatial scales (same spatial extent but a fine grain [25 × 25 km grid cells] and a coarse grain [100 × 100 km grid cells], therefore producing two scales. Only cells whose richness was within 10% of estimated asymptotic richness are included in the datasets (Latombe et al., 2017), to limit the occurrence of false absence. The second dataset is an inventory of the presence and absence of springtails and mite species in 12 plots (4 transects and 3 altitudes) on Marion Island, along with the altitude of the sites and the side of the island where they are located (McGeoch et al., 2008; Nyakatya and McGeoch, 2008).

In the following, we use the fine-grain bird data to illustrate the four different kinds of analyses. The fine-grain data, together with the seven environmental variables, can be loaded using the following commands:

```
data(bird.spec.fine)
xy <- bird.spec.fine[,1:2] # geographic coordinates of sites
data.spec <- bird.spec.fine[,3:192] # site-by-species matrix
data(bird.env.fine)
data.env <- bird.env.fine[,3:9] # site-by-environment matrix
```

3. Zeta diversity and zeta decline

224

225 3.1. Description

226 The functions `Zeta.order.ex` and `Zeta.order.mc` compute ζ_i, the number of

227 species shared by any *i* assemblages (the order of zeta) in two alternative ways.

228 `Zeta.order.ex` computes the expected value of zeta diversity for order *i*. Let P_j be

229 the probability of species *j* with occupancy O_j occurring in *i* given sites out of the *N*

230 surveyed sites. The expected value can be calculated as the sum of the probability

231 over all species *S*:

232

233 $$E(\zeta_i) = \sum_j^S E[P_j] = \sum_j^S \frac{C_i^{O_j}}{C_i^N} \quad (1)$$

234

235 where C_i^N and $C_i^{O_j}$ are binomial coefficients giving the total number of possible

236 combinations of *i* sites out of a total of *N* or O_j, respectively. The variance is then

237 given by the summation of the covariance of the probability:

238

239 $$var(\zeta_i) = \frac{C_i^N}{C_i^N - 1} \times \sum_j^S \sum_k^S \left(E\left[P_j P_k\right] - E\left[P_j\right] * E[P_k]\right) \quad (2)$$

240 where

241 $$E[P_j P_k] = \frac{C_i^{O_{jk}}}{C_i^N} \quad (3)$$

242 and O_{jk} is the number of sites in which both species *j* and *k* are present (also referred

243 to as joint occupancy; Hui, 2009). The number O_{ij} corresponds to the element *ij* of the

244 *S*×*S* dimensional matrix $\mathbf{M}^T\mathbf{M}$, where **M** is the site(row)-by-species(column) matrix

245 of occurrence and T matrix transposition. Note that the variance in Equation 2 is

246 corrected for bias using Bessel's correction (Kenney and Keeping, 1951), which

247 corresponds to the default in `Zeta.order.ex`. This is suitable if the assemblages
248 represent a sample of the total study system. In case of a continuous grid sample or in
249 lab experiments, for which the incidence data can be exhaustive, the exact variance
250 can also be computed by setting `sd.correct = FALSE` in the function parameters.
251
252 By contrast, `Zeta.order.mc` (for "Monte Carlo sampling") computes zeta diversity
253 by averaging the number of shared species for *i* assemblages over all possible
254 combinations of the *i* assemblages from *N* total assemblages. The shared species for *i*
255 assemblages is obtained using the dot product of species (1/0) vectors. When all
256 possible combinations are used, `Zeta.order.mc` and `Zeta.order.ex` are
257 equivalent. For large *N* and intermediate *i*, the number of combinations for *i*
258 assemblages, C_i^N, becomes very high, and the computational complexity becomes
259 intractable. The user must therefore provide a value `sam`, representing the number of
260 samples over which ζ_i should be computed. If `sam` > C_i^N, ζ_i is computed exactly, but
261 otherwise approximated over `sam` random combinations. The impact of `sam` on the
262 computation of ζ_i can be assessed using the function `Zeta.sam.sensitivity`
263 (Appendix A, Figure A1).
264
265 In contrary to `Zeta.order.ex`, for each combination *j* of *i* assemblages,
266 `Zeta.order.mc` allows for the computation of a normalised version zeta (i.e. ζ_{ij}/S_j),
267 where S_j is either (i) the total number of species over the assemblages in the specific
268 combination *j* (i.e. the gamma diversity of the combination *j*, therefore equivalent to
269 the Jaccard similarity index), (ii) the average number of species per assemblage in the
270 specific combination *j* (i.e. the alpha diversity of the combination *j*, therefore
271 equivalent to the Sørensen similarity index), or (iii) the minimum number of species

over the assemblages in the specific combination *j* (therefore equivalent to the Simpson similarity index). Normalised zeta may be suitable when richness varies widely across regions or systems being compared.

The formulas described above for `Zeta.order.ex` and `Zeta.order.mc` correspond to combinations of any *i* assemblages over all assemblages. This sub-sampling scheme may nonetheless not be the most appropriate for some data. For example, the turnover of assemblages arranged in a linear fashion along a gradient (e.g. Rivadeneira et al., 2002; Whittaker, 1956) may be better analysed by combining assemblages close to each other, and using a specific assemblage as a reference (Whittaker, 1967). Several sub-sampling schemes are possible in `Zeta.order.mc`. Assemblages can be combined using a nearest-neighbour approach to explore patterns of local turnover. When a nearest-neighbour approach is used, the combinations can be non-directional, or directional, moving away from a fixed-point origin or a fixed-edge origin (for example for ecological systems being invaded from a specific direction) (McGeoch et al., 2017). A focal assemblage plus the closest (*i*-1) assemblages are then be used for calculating ζ_i. The focal assemblage can be the fixed-point origin or any other assemblage. There are therefore 4 possible sub-sampling schemes, whose pertinence depends on the specific study (see McGeoch et al., 2017 for additional details and a comparison of the zeta declines using different sub-sampling schemes for the well-known Smokey Mountain dataset of Whittaker 1956, 1967): the ALL scheme using combinations of any assemblages (the default scheme), the non-directional nearest neighbour (NON) scheme, in which each site is associated to its *i*-1 nearest neighbours to compute ζ_i, the directional nearest neighbour using a specific assemblage or an edge as a reference (DIR), and each site

297 is associated to its *i*-1 nearest neighbours in the opposite direction to the reference to
298 compute ζ_i, and the fixed-point origin (FPO) scheme, in which a specific assemblage
299 is always combined with its *i*-1 nearest neighbours to compute ζ_i (i.e. similar to NON
300 but using one specific assemblage only). When the FPO is located outside of the study
301 area, it corresponds to a fixed-edge origin (FEO) scheme, in which assemblages close
302 to the edge are combined with their nearest neighbours.
303
304 The functions `Zeta.decline.ex` and `Zeta.decline.mc` then compute the values
305 of ζ_i for a range of orders *i* (Figure 2; Appendix A, Figure A2). As the number of
306 assemblages increases, the number of shared species amongst assemblages
307 necessarily decreases, hence a decline in zeta. These functions also compute the ratio
308 ζ_i / ζ_{i-1}, which is called the retention rate and quantifies the proportion of species that
309 are retained in additional samples. The retention rate is especially useful to reveal
310 features of the zeta decline that are indistinguishable from the observation of the
311 decline itself, allowing for highlighting differences in the structure of compositional
312 change between datasets or study areas, and for detecting spatial structure in gradients
313 of vegetation when using different sub-sampling schemes (McGeoch et al., 2017).
314
315 Finally, an exponential and a power law parametric form are fitted to the zeta decline.
316 These are the two most common parametric forms observed in nature (Hui and
317 McGeoch, 2014). The parametric form of the decline may signal the relative roles of
318 stochastic or deterministic assembly processes, although it may also be affected by
319 assemblage richness and sample size. The function `Plot.zeta` plots the outputs of
320 `Zeta.decline.ex` and `Zeta.decline.mc`.
321

Computing zeta diversity for different orders has been used, for example to validate the outputs of self-organising maps used for pest profile analyses, which group together areas with similar profiles of species composition (Roigé et al., 2017). Pairwise comparisons of sites enables the identification of clusters with few shared species and therefore high uncertainty. Using orders of zeta beyond pairwise comparisons enables to further refine the uncertainty level of the remaining clusters by distinguishing between clusters with low (i.e. superficial) and high similarity for higher orders of zeta.

3.2. Example

The zeta decline of bird species over South-East Australia was computed from orders 1 to 50 using the ALL and the NON subsampling schemes across grid cells and plotted using the following commands (the seed is set to 1 for reproducibility):

```
set.seed(1)
dev.new (width = 12, height = 4)
zeta.decline.fine.ex <- Zeta.decline.ex(data.spec, orders =
1:50)
dev.new(width = 12, height = 4)
zeta.decline.fine.NON <- Zeta.decline.mc(data.spec, xy, orders
= 1:50, NON = TRUE, DIR = FALSE, FPO = NULL)
```

The `NON = TRUE` parameter indicates that the NON scheme must be used. If the `FPO` parameter contains coordinates, they take precedence over the `NON` parameter. If `DIR`

346 = `FALSE`, the FPO (or FEO) scheme applies. The DIR scheme requires both `DIR =`
347 `TRUE` and a set of coordinates in `FPO`.
348
349 Comparing outputs of the ALL and the NON subsampling schemes provides
350 information on the effect of the spatial scale on species turnover. When all cells are
351 combined, the zeta decline better fits a power law than an exponential parametric
352 form (Figure 2a; ΔAIC = 270.61), therefore suggesting that species are distributed in
353 a deterministic fashion across South-East Australia. The retention rate (ζ_i / ζ_{i-1})
354 increases steadily, but starts levelling off after 20 assemblages, indicating that, below
355 that value, few species are retained as new assemblages are considered, but many
356 more are, proportionally, beyond 20 assemblages. The asymptote therefore provides
357 an indication of the scale at which species can be considered to be rare and common.
358
359 When the cells are combined using the NON scheme, the retention rate is higher than
360 for the ALL scheme for low orders of zeta, indicating that the zeta values decline at a
361 lower rate. This suggests some level of spatial aggregation of species, with closer
362 cells sharing more rare species (and common species to a lesser extent), as can be
363 expected. The zeta decline computed with the NON scheme is also better fitted by a
364 power law rather than exponential parametric form.
365

366 4. Distance decay of zeta

367

368 4.1. Description

369 The distance decay of similarity is a well-known community descriptor (Morlon et al.,
370 2008; Nekola and White, 1999), i.e. as distance between assemblages increases, two

371 assemblages are expected to become less similar and to share fewer species. Typical
372 research questions that can be addressed by considering the distance decay of zeta-
373 diversity include: (i) the explicit distances over which species assemblages differ; (ii)
374 how do the decay patterns of rare and common species differ, providing insight on the
375 spatial properties of their distributions.
376
377 The function `Zeta.ddecay` generalizes distance decay and enables its computation
378 for any number of assemblages. For many sites, it uses the same Monte Carlo
379 sampling as `Zeta.order.mc`, and can therefore be applied to normalised zeta. For
380 more than two assemblages, distances between assemblages (either computed from
381 sites coordinates or from a custom distance matrix) must be combined for each
382 combination of sites, for example as the mean distance across *n* sites. The function is
383 flexible and enables users to define how they should be combined, using a built-in or
384 a custom function (see Latombe et al., 2017 for a discussion on the impacts of using
385 different functions). `Zeta.ddecay` regresses ζ_i over this measure of distance using
386 three types of regression: (i) a generalized linear model, the default being linear
387 regression, allowing constraints on the signs of the coefficients (ii) a generalized
388 additive model (GAM), to allow for non-linearities and periodicities in the distance
389 decay (Soininen et al., 2007) and (iii) a general additive model under shape constraint,
390 or "shape-constrained additive model" (SCAM; Pya and Wood, 2015), set by default
391 to a monotonically declining GAM. Additional options enable the definition of
392 thresholds for distance which may be desirable, for example, for discarding
393 uninformative long tails that would artificially make the slope of the distance decay in
394 linear models more shallow. It is also possible to specify how to transform spatial
395 distance according to any function. The function `Zeta.ddecays` calls

`Zeta.ddecay` and computes the slope of the distance decay using linear models for different orders of zeta, and plots changes in slope as the order increases (Appendix A, Figure A3).

The distance decay of zeta can also be applied to time series of species composition, using time instead of distance, therefore computing a time decay of zeta diversity. Time decay of zeta diversity has been used to show differences in the response of bird communities of two different river basins to drought (McGeoch et al., 2017).

4.2. Example

The distance decays of ζ_2, ζ_3, ζ_5, and ζ_{10} were assessed using a linear regression and a GAM (set with `reg.type`, whose default is linear regression) using the following commands (for `order = 2`, `order = 3`, `order = 5` and `order = 10`):

```
set.seed(1)
dev.new()
zeta.ddecay.lm.fine <- Zeta.ddecay(xy, data.spec, order = 2,
 confint.level = 0.95) # the default regression is a linear
 model
set.seed(1)
dev.new()
zeta.ddecay.gam.fine <- Zeta.ddecay(xy, data.spec, order = 2,
reg.type="gam") # a generalised additive model is used
insteadn of the default linear model
```

Both methods show a clear distance decay, even for ζ_{10}, although it becomes less pronounced for high orders of zeta (Figure 3). The distance decay is more pronounced for ζ_3 and ζ_5 than for ζ_2 (p-values = 0.001, 0.003; the significance was computed using the `diffslope2` function from the *simba* R package, Jurasinski 2012; see R code in Appendix B for details) suggesting that within the extent of this study, rare species are dispersed relative to the space-filling properties of the species with higher occurrence levels. The GAM shows that the distance decay is not linear for ζ_2 and ζ_3. In particular, ζ_3 allows for the detection of a threshold at ~800 km after which the effect of distance on compositional change mostly disappears (Figure 3).

5. Multi-site generalised dissimilarity modelling

5.1. Description

Multi-Site Generalised Dissimilarity Modelling (MS-GDM; Latombe et al., 2017) is inspired by Generalized Dissimilarity Modelling (GDM; Ferrier et al., 2007), a statistical technique for analysing and predicting changes in beta diversity from pairwise differences in environmental variables and spatial distance between sites using regression techniques. Following the same principles, the function `Zeta.msgdm` enables the regression of rescaled (ζ_i / ζ_1) or normalised ζ_i values (Jaccard, Sørensen or Simpson versions) over environmental differences and distance between assemblages. Since ζ_i is the number of species in common across *i* sites, we call it Multi-site Generalised Dissimilarity Modelling (see Latombe et al., 2017 for details). MS-GDM enables the inclusion of both continuous and categorical environmental variables as predictors. In the latter case, the environmental difference between *i* sites is computed as the number of different values across the *i* sites (and

446 the maximum value is therefore *i*; Latombe et al. *in review*). MS-GDM also enables
447 the inclusion of the zeta values of the same order from another group of species as
448 predictors, when both groups are expected to be related to each other (such as native
449 and alien species; Latombe et al. *in review*). Typical research questions that can be
450 addressed by MS-GDM include: (i) whether variation in the number of shared species
451 (compositional similarity) between assemblages is explained predominantly by either
452 environmental differences or distance; (ii) whether the relative importance of different
453 environmental variables and distance differs for rare and common species (by
454 comparing low and high orders of zeta)
455
456 Four different types of regression techniques have been implemented: generalized
457 linear models (GLM), with possible constraint on the sign of the coefficients, GAMs,
458 SCAMs, and, following Ferrier et al. (2007), a combination of I-spline and GLM with
459 constraints on the signs of the coefficients (see Latombe et al., 2017 for details). I-
460 splines (Ramsay, 1988) are a kind of monotone spline functions that are used to
461 transform the data before applying a generalized linear model with non-negative
462 coefficients. This transformation accommodates non-linear relationships between zeta
463 diversity and changes in environmental variables, but also the fact that the impact of
464 change in an environmental variable may depend on the values of this variable (for
465 example a change of temperature near the limit of the species thermal tolerance may
466 have more impact on species occurrence than the same change in the middle of the
467 range of its thermal tolerance).
468
469 The order of the I-splines and the number of knots (for the GAM, SCAM and I-
470 splines) can be set by the user. The number of knots must be chosen carefully, as too

many knots may result in overfitting (Manion, 2009). Moreover, as for any regression analysis, variables suffering from multicollinearity (e.g. VIF>10) should be removed (Dormann et al., 2013). As for the distance decay, for many sites, `Zeta.msgdm` uses the same Monte Carlo sampling as `Zeta.order.mc`. When $i > 2$, the environmental differences and distances between assemblages must also be combined for each combination, for example using the mean of differences.

A function `Ispline` to transform data using I-splines is also included in the package. Using the output from `Zeta.msgdm`, the function `Predict.msgdm` predicts the zeta values for new environmental data. The function `Plot.spline` is used to plot the I-splines for the different variables. Finally, the function `Zeta.varpart` computes variation partitioning (Legendre, 2008) for a model computed with `Zeta.msgdm`, to determine which part of ζ_i is explained by the environmental or the distance variables. `Zeta.varpart` uses the adjusted R^2, to account for the use of several environmental variables, whereas distance is a single variable. Note that the non-adjusted R^2 is computed as 1 – (residual sum of squares) / (total sum of squares), and makes sense only for linear regression, for which the residual sum of squares is normally distributed. Results of variation partitioning for the other regression techniques should therefore be interpreted with caution. In variation partitioning, some partitions may be negative (Legendre and Legendre, 2012). The function `Pie.neg` therefore considers negative values as 0 to plot the results as a pie diagram.

5.2. Example

Similar to Latombe et al. (2017), MS-GDM was computed for the Sørensen ζ_2 and ζ_{10} using I-splines (and a binomial family with a log link, which requires a negative

intercept, as shown by `cons.inter = -1`), as well as variation partitioning for linear regressions (contrary to MS-GDM, no constraint was applied on the sign of the regression by setting `method.glm = "glm.fit2"` so that the residuals are normally distributed, as explained above), using the following commands (for `order = 2` and `order = 10`):

```
set.seed(1)
zeta.ispline.fine2 <-
Zeta.msgdm(data.spec,data.env,xy,order=2,sam=1000,reg.type="is
pline",normalize="Sorensen",family=binomial(link="log"),cons.i
nter = -1)
Plot.ispline(zeta.ispline.fine2, data.env, distance = TRUE,
legend = FALSE)
set.seed(1)
zeta.varpart.fine2 <-
Zeta.varpart(xy=xy,data.spec=data.spec,data.env=data.env,order
=2,sam=1000,method.glm = "glm.fit2")
dev.new()
pie.neg(zeta.varpart.fine2[4:7,1], density = c(4, 0, 8, -1),
 angle = c(90, 0, 0, 0), labels =
 c("distance","undistinguishable","environment","unexplained"),
 radius = 0.9)
```

In these data, precipitation is the main predictor of bird compositional change for ζ_2, especially for dry environments (as shown by the steep slope of the I-spline for low

521 precipitations), followed by distance (Figure 4). For ζ_{10}, which, contrary to ζ_2,
522 excludes the contribution of the rarest species to turnover, the importance of
523 temperature and area per person increases. The decrease in the relative importance of
524 precipitation may be due to the fact that common species are more likely to find
525 refugia in areas containing water bodies during dry periods, whereas rare species may
526 be more vulnerable to rainfall heterogeneity (discussed in further detail in Latombe et
527 al., 2017). Results are slightly different from Latombe et al. (2017) because the
528 Sorensen version of zeta was used here instead of just rescaling the zeta values by the
529 overall ζ_1, and different indices consider the influence of richness on turnover
530 differently (Baselga, 2010).
531
532 Variation partitioning on ζ_2 and ζ_{10} using I-splines and simple linear regressions
533 shows that variation partitioning explains a larger proportion of variance for low
534 orders of zeta than for high ones, indicating that the spatial distribution of rare species
535 is more predictable than for common species (Figure 5). As expected, the I-splines
536 explain a larger part of variations compared to linear regressions for both ζ_2 and ζ_{10},
537 due to their flexibility. In addition, these results linking environment with species
538 compositional change rather than distance support the interpretation of the fact that
539 the decline of zeta diversity is better fitted by a power law than by an exponential
540 parametric form, suggesting deterministic community assembly.
541

542 6. Hierarchical scaling of zeta

543

544 6.1. Description

545 Like all biodiversity metrics, zeta diversity is sensitive to scale, *i.e.* to grain and extent
546 (Hui and McGeoch, 2014). For compositional change, grain type follows three
547 general sampling schemes (Scheiner et al., 2011): (i) sites arrayed as cells in a
548 contiguous grid, (ii) sites arrayed as cells in a regular but non-contiguous grid and (iii)
549 irregularly distributed sites of potentially varying size, such as islands (Figure 6). For
550 data based on regular grids, the effect of scale can be assessed by grouping
551 assemblages with their immediate neighbours (Figure 6a,b). For irregularly
552 distributed areas, assemblages are grouped based on the distance between them
553 (Figure 6c).
554
555 When defining commonness based on the relative occupancy of a species (the number
556 of sites or cells where the species is present divided by the total number of sites) (see
557 McGeoch and Latombe, 2016), the proportion of rare species necessarily decreases as
558 grain increases, whereas the proportion of common species increases. This is because
559 the relative occupancy of a species necessarily increases (or stays constant) as grain
560 increases. The rate at which rare species become more common with coarser grain
561 depends on their spatial distribution (are species clustered or not) (Hui et al., 2010;
562 Hui and A McGeoch, 2007; McGeoch and Gaston, 2002). For example, a species
563 present in 4 adjacent cells arranged in a square (i.e. highest possible level of
564 clustering) in a $n \times n$ continuous grid (Figure 6a) has an occupancy of $4/n^2$, and an
565 occupancy of $1/(n/2)^2 = 4/n^2$ once the grain is doubled if all four cells are combined
566 into a single one. Any other spatial arrangement of the four cells will therefore
567 generate an occupancy higher than $4/n^2$.
568

569 However, from a community perspective, species commonness and rarity are relative
570 notions (McGeoch and Latombe, 2016). For a given occupancy, a species will be
571 common in a community in which other species have a lower occupancy, and
572 conversely. As we showed in the description of the zeta decline, the shape of the
573 species retention rate across orders of zeta enables defining the threshold at which
574 species can be considered to be rare or common. The distinction between rare and
575 common species can therefore vary differently across scales for communities with
576 different spatial arrangements of their species (McGeoch and Gaston, 2002). Species
577 with different levels of spatial clustering will therefore contribute differently to the
578 various orders of zeta diversity depending on the grain of the study. Species that are
579 spatially dispersed will contribute to higher orders of zeta when the grain becomes
580 coarser than species that are spatially clustered (Figure 7).
581
582 Typical research questions that can be addressed by exploring the hierarchical scaling
583 of zeta diversity therefore include: (i) how the characteristic of being common or rare
584 varies with grain; and (ii) whether the sampling effort is sufficient to comprehensively
585 study species turnover of both common and rare species.
586
587 In the *zetadiv* package, the functions `rescale.regular` and `rescale.min.dist`
588 aggregate the species occurrence data, and combine the environmental data and the
589 coordinates following a user-specified function such as the mean, based on the
590 neighbours and on minimum distance, respectively, for a specific level of
591 aggregation. The functions `Zeta.scale.regular` and `Zeta.scale.min.dist`
592 compute ζ_i for a specific order *i*, for a range of levels of aggregation for the two
593 methods. For `rescale.min.dist` and `Zeta.scale.min.dist`, the assemblages

are aggregated iteratively: given a list of assemblages in a specific order, the first assemblage is combined with the closest ones, then the next available assemblage is combined with the closest available ones, and so on. Since the order of the assemblages in the list can impact the outcome of the algorithm, the function `Zeta.scale.min.dist` performs the analyses several times for each order and returns the mean.

6.2. Example

We assessed the hierarchical scaling of ζ_1 to ζ_{10} by aggregating the 25 × 25 km cells (from 1 to 10 cells and then to 60 cells by steps of 10, as stated by `m = c(1:10,seq(20,60,10))`) based on minimum distance (Figure 6c) using the following commands (for order = 1 to order = 10):

```
set.seed(1)
zeta.scale.irreg <- Zeta.scale.min.dist(xy, data.spec, m =
 c(1:10,seq(20,60,10)), order = 1, reorder = 50, normalize =
 FALSE, plot = FALSE, zeta.type="exact")
```

Since the order in which the cells are aggregated can change the results, the aggregation is performed 50 times (`reorder = 50`) and the average zeta values are computed.

As expected, zeta values increase as grain increases for all orders of zeta (Figure 8a). We also compared (ζ_i- ζ_{i-1}) for each grain, to compare the rates of increase across orders of zeta. Although zeta diversity increases with grain in a similar fashion for all

orders (Figure 8a), the difference between the zeta values of different orders changes with grain and between orders. ζ_1- ζ_2 always decreases as the grain increases (Figure 8b), and the zeta decline becomes more shallow between orders 1 and 2 (Figure 8c). That is, less rare species are lost when increasing grain. By contrast, for higher orders of zeta, differences in the rates of increase between two consecutive orders of zeta increases when grouping 2 or 3 cells, then decreases (Figure 8b). This means that the zeta decline is steeper across orders 2 to 10 when aggregating 2 or 3 cells than for the fine grain data and for aggregating many cells (Figure 8c). These results suggest that a spatial grain of ~1250 km^2 (~35 x 35 km) may be appropriate to study bird communities over Australia, as the sharper and more steady decline of zeta diversity indicates a more gradual distinction between common and rare species than at finer and coarser grains, for which the zeta decline becomes more shallow as the zeta order increases (Figure 8b,c). The ~1250 km^2 grain may therefore be related to the scale at which bird species of different levels of rarity aggregate in South-East Australia.

7. Concluding remarks

By extending the analyses of compositional change to more than pairwise combinations of assemblages, zeta diversity provides a more detailed understanding of species diversity and a more exhaustive description of community assemblages than using alpha and beta diversity alone. In addition to the clear advantages of obtaining accurate descriptions of biodiversity, such as the possibility to better identify the processes that generates it, zeta diversity also enables the differentiation of the role of common species from rare ones in structuring biodiversity patterns. As we have shown in the examples above illustrating the four different types of analyses

currently applicable to zeta diversity applied to bird communities over South-East Australia, considering multiple orders of zeta diversity sheds light on differences in the characteristics and drivers of spatial distribution of common and rare species. It also shows the impact of the spatial resolution at which communities are defined for distinguishing between common and rare species.

The package is also under constant development, and future versions of *zetadiv* will pay special attention to spatially mapping zeta diversity and the parametric form of zeta decline. With increasing recognition of the importance of temporal changes in compositional change (Magurran, 2011) as a consequence of climate change and biotic homogenization (Dornelas et al., 2014), specific functions for temporal decay will be implemented in the future. Current functions can nonetheless already be used to perform such analyses on zeta diversity (e.g. using `Zeta.decline.mc` using the closest assemblages along a temporal gradient). Given the importance of accounting for phylogenetic and functional traits information for the management and conservation of ecological communities (Devictor et al., 2010), phylogenetic and functional measures of zeta diversity will be developed, reflecting similar recent developments for beta diversity (Graham and Fine, 2008; Loiseau et al., 2017). Finally, measures of zeta diversity will be developed for measuring turnover in species interactions.

Software availability

Name of Software: zetadiv (version 1.1.1).

Year of First Release: 2015.

Developers: G. Latombe, Melodie A. McGeoch, David A. Nipperess, Cang Hui

669 Maintainer: G. Latombe

670 E-mail: Latombe.guillaume@gmail.com

671 Available from the CRAN: https://CRAN.R-project.org/package=zetadiv

672

673 ## Acknowledgements

674 We thank Rachel Leihy and Grant Duffy for providing feedback on the functions of
675 the package. We also thank the attendees of the 2015 Ecological Society of Australia
676 and Eco-Stats conferences for fruitful discussions on the concept of zeta diversity. We
677 acknowledge BirdLife Australia as the source of the bird data used here. This research
678 was supported by an Australian Research Council Discovery Project Grant
679 (DP150103017) to MM and CH, and the National Research Foundation of South
680 Africa (grant nos. 81825 and 76912 to CH).

681

682 ## REFERENCES

683

684 Barrett, G., Silcocks, A., Barry, S., Cunningham, R., Poulter, R., 2003. The new atlas
685 of Australian birds, Royal Australasian Ornithologists Union. Melbourne.
686 Baselga, A., 2013. Multiple site dissimilarity quantifies compositional heterogeneity
687 among several sites, while average pairwise dissimilarity may be misleading.
688 Ecography (Cop.). 36, 124–128.
689 Baselga, A., 2010. Partitioning the turnover and nestedness components of beta
690 diversity. Glob. Ecol. Biogeogr. 19, 134–143.
691 Chao, A., Chiu, C.-H., Hsieh, T.C., 2012. Proposing a resolution to debates on
692 diversity partitioning. Ecology 93, 2037–2051.
693 Devictor, V., Mouillot, D., Meynard, C., Jiguet, F., Thuiller, W., Mouquet, N., 2010.
694 Spatial mismatch and congruence between taxonomic, phylogenetic and
695 functional diversity: the need for integrative conservation strategies in a changing
696 world. Ecol. Lett. 13, 1030–1040.
697 Diserud, O.H., Ødegaard, F., 2007. A multiple-site similarity measure. Biol. Lett. 3,
698 20–22.
699 Dormann, C.F., Elith, J., Bacher, S., Buchmann, C., Carl, G., Carré, G., Marquéz,
700 J.R.G., Gruber, B., Lafourcade, B., Leitão, P.J., 2013. Collinearity: a review of
701 methods to deal with it and a simulation study evaluating their performance.
702 Ecography (Cop.). 36, 27–46.
703 Dornelas, M., Connolly, S.R., Hughes, T.P., 2006. Coral reef diversity refutes the
704 neutral theory of biodiversity. Nature 440, 80.

705 Dornelas, M., Gotelli, N.J., McGill, B., Shimadzu, H., Moyes, F., Sievers, C.,
706 Magurran, A.E., 2014. Assemblage time series reveal biodiversity change but not
707 systematic loss. Science (80-.). 344, 296–299.
708 Ferrier, S., Manion, G., Elith, J., Richardson, K., 2007. Using generalized
709 dissimilarity modelling to analyse and predict patterns of beta diversity in
710 regional biodiversity assessment. Divers. Distrib. 13, 252–264.
711 Fick, S.E., Hijmans, R.J., 2017. WorldClim 2: new 1-km spatial resolution climate
712 surfaces for global land areas. Int. J. Climatol. 37, 4302–4315.
713 Gaston, K.J., 2010. Valuing common species. Science (80-.). 327, 154–155.
714 Graham, C.H., Fine, P.V.A., 2008. Phylogenetic beta diversity: linking ecological and
715 evolutionary processes across space in time. Ecol. Lett. 11, 1265–1277.
716 Grimm, V., Railsback, S.F., 2012. Pattern-oriented modelling: a “multi-scope”for
717 predictive systems ecology. Phil. Trans. R. Soc. B 367, 298–310.
718 Grimm, V., Revilla, E., Berger, U., Jeltsch, F., Mooij, W.M., Railsback, S.F., Thulke,
719 H.H., Weiner, J., Wiegand, T., DeAngelis, D.L., 2005. Pattern-Oriented
720 Modeling of Agent-Based Complex Systems: Lessons from Ecology. Science
721 (80-.). 310, 987–991.
722 Hui, C., 2009. On the scaling patterns of species spatial distribution and association. J.
723 Theor. Biol. 261, 481–487.
724 Hui, C., A McGeoch, M., 2007. Modeling species distributions by breaking the
725 assumption of self-similarity. Oikos 116, 2097–2107.
726 Hui, C., McGeoch, M.A., 2014. Zeta diversity as a concept and metric that unifies
727 incidence-based biodiversity patterns. Am. Nat. 184, 684–694.
728 Hui, C., Veldtman, R., McGeoch, M.A., 2010. Measures, perceptions and scaling
729 patterns of aggregated species distributions. Ecography (Cop.). 33, 95–102.
730 Jost, L., 2007. Partitioning diversity into independent alpha and beta components.
731 Ecology 88, 2427–2439.
732 Jost, L., Chao, A., Chazdon, R.L., 2011. Compositional similarity and beta diversity.
733 Biol. Divers. Front. Meas. Assess. 66–84.
734 Jurasinski, G., Retzer, V., 2012. simba: A collection of functions for similarity
735 analysis of vegetation data. https://CRAN.R-project.org/package=simba. R
736 Packag. Version 0.3–5.
737 Kenney, F., Keeping, E.S., 1951. Mathematics of statistics-part two. D. Van Nostrand
738 Company, Inc Princeton,; New Jersey; Toronto; New York; London.
739 Latombe, G., Hui, C., McGeoch, M.A., 2017. Multi-site generalised dissimilarity
740 modelling: using zeta diversity to differentiate drivers of turnover in rare and
741 widespread species. Methods Ecol. Evol. 8, 431–442.
742 Latombe, G., Hui, C., McGeoch, M.A., 2015. Beyond the continuum: a multi-
743 dimensional phase space for neutral–niche community assembly. Proc. R. Soc. B
744 282, 20152417.
745 Latombe, G., Parrott, L., Fortin, D., 2011. Levels of emergence in individual based
746 models: Coping with scarcity of data and pattern redundancy. Ecol. Modell. 222,
747 1557–1568.
748 Legendre, P., 2008. Studying beta diversity: ecological variation partitioning by
749 multiple regression and canonical analysis. J. Plant Ecol. 1, 3–8.
750 Legendre, P., Legendre, L.F.J., 2012. Numerical ecology (3rd ed.). Elsevier,
751 Amsterdam, The Netherlands.
752 Loiseau, N., Legras, G., Gaertner, J., Verley, P., Chabanet, P., Mérigot, B., 2017.
753 Performance of partitioning functional beta-diversity indices: Influence of
754 functional representation and partitioning methods. Glob. Ecol. Biogeogr. 26,

753–762.
Magurran, A.E., 2011. Measuring biological diversity in time (and space), in: Magurran, A.E., McGill, B.J. (Eds.), Biological Diversity: Frontiers in Measurement and Assessment. Oxford University Press, pp. 85–96.
Magurran, A.E., McGill, B.J., 2011. Biological diversity: frontiers in measurement and assessment. Oxford University Press, Oxford.
Manion, G., 2009. A technique for constructing monotonic regression splines to enable non-linear transformation of GIS rasters, in: 18th World IMACS Congress and MODSIM09 International Congress on Modelling and Simulation. pp. 13–17.
Manion, G., Lisk, M., Ferrier, S., Nieto-Lugilde, D., Fitzpatrick, M.C., n.d. gdm: Functions for Generalized Dissimilarity Modeling. R package version 1.2.3. https://github.com/fitzLab-AL/GDM.
McGeoch, M.A., Gaston, K.J., 2002. Occupancy frequency distributions: patterns, artefacts and mechanisms. Biol. Rev. 77, 311–331.
McGeoch, M.A., Latombe, G., 2016. Characterizing common and range expanding species. J. Biogeogr. 43, 217–228.
McGeoch, M.A., Latombe, G., Andrew, N.R., Nakagawa, S., Nipperess, D.A., Roige, M., Marzinelli, E.M., Campbell, A.H., Verges, A., Thomas, T., Steinberg, P.D., Selwood, K.E., Hui, C., 2017. The application of zeta diversity as a continuous measure of compositional change in ecology. bioRxiv 216580.
McGeoch, M.A., Le Roux, P.C., Hugo, A.E., Nyakatya, M.J., 2008. Spatial variation in the terrestrial biotic system, in: Chown, S.L., Froneman, P.W. (Eds.), The Prince Edward Islands: Land-Sea Interactions in a Changing Ecosystem. African SunMedia, Stellenbosch.
Morlon, H., Chuyong, G., Condit, R., Hubbell, S., Kenfack, D., Thomas, D., Valencia, R., Green, J.L., 2008. A general framework for the distance–decay of similarity in ecological communities. Ecol. Lett. 11, 904–917.
Nekola, J.C., White, P.S., 1999. The distance decay of similarity in biogeography and ecology. J. Biogeogr. 26, 867–878.
Nyakatya, M.J., McGeoch, M.A., 2008. Temperature variation across Marion Island associated with a keystone plant species (Azorella selago Hook.(Apiaceae)). Polar Biol. 31, 139–151.
Oksanen, J., Blanchet, F.G., Friendly, M., Kindt, R., Legendre, P., McGlinn, D., Minchin, P.R., O'hara, R.B., Simpson, G.L., Solymos, P., Stevens, M.H.H., Szoecs, E., Wagner, H., 2018. vegan: Community Ecology Package. R package version 2.5–1. https://CRAN.R-project.org/package=vegan.
Pya, N., Wood, S.N., 2015. Shape constrained additive models. Stat. Comput. 25, 543–559.
R CoreTeam, 2013. R: A language and environment for statistical computing.
Ramsay, J.O., 1988. Monotone regression splines in action. Stat. Sci. 425–441.
Ricotta, C., Pavoine, S., 2015. A multiple-site dissimilarity measure for species presence/absence data and its relationship with nestedness and turnover. Ecol. Indic. 54, 203–206.
Rivadeneira, M.M., Fernández, M., Navarrete, S.A., 2002. Latitudinal trends of species diversity in rocky intertidal herbivore assemblages: spatial scale and the relationship between local and regional species richness. Mar. Ecol. Prog. Ser. 245, 123–131.
Roigé, M., McGeoch, M.A., Hui, C., Worner, S.P., 2017. Cluster validity and uncertainty assessment for self-organizing map pest profile analysis. Methods

805 Ecol. Evol. 8, 349–357.
806 Roura-Pascual, N., Sanders, N.J., Hui, C., 2016. The distribution and diversity of
807 insular ants: do exotic species play by different rules? Glob. Ecol. Biogeogr. 25,
808 642–654.
809 Scheiner, S.M., Chiarucci, A., Fox, G.A., Helmus, M.R., McGlinn, D.J., Willig, M.R.,
810 2011. The underpinnings of the relationship of species richness with space and
811 time. Ecol. Monogr. 81, 195–213.
812 Socolar, J.B., Gilroy, J.J., Kunin, W.E., Edwards, D.P., 2016. How should beta-
813 diversity inform biodiversity conservation? Trends Ecol. Evol. 31, 67–80.
814 Soininen, J., McDonald, R., Hillebrand, H., 2007. The distance decay of similarity in
815 ecological communities. Ecography (Cop.). 30, 3–12.
816 Whittaker, R.H., 1967. Gradient analysis of vegetation. Biol. Rev. 42, 207–264.
817 Whittaker, R.H., 1956. Vegetation of the great smoky mountains. Ecol. Monogr. 26,
818 1–80.
819
820

FIGURES

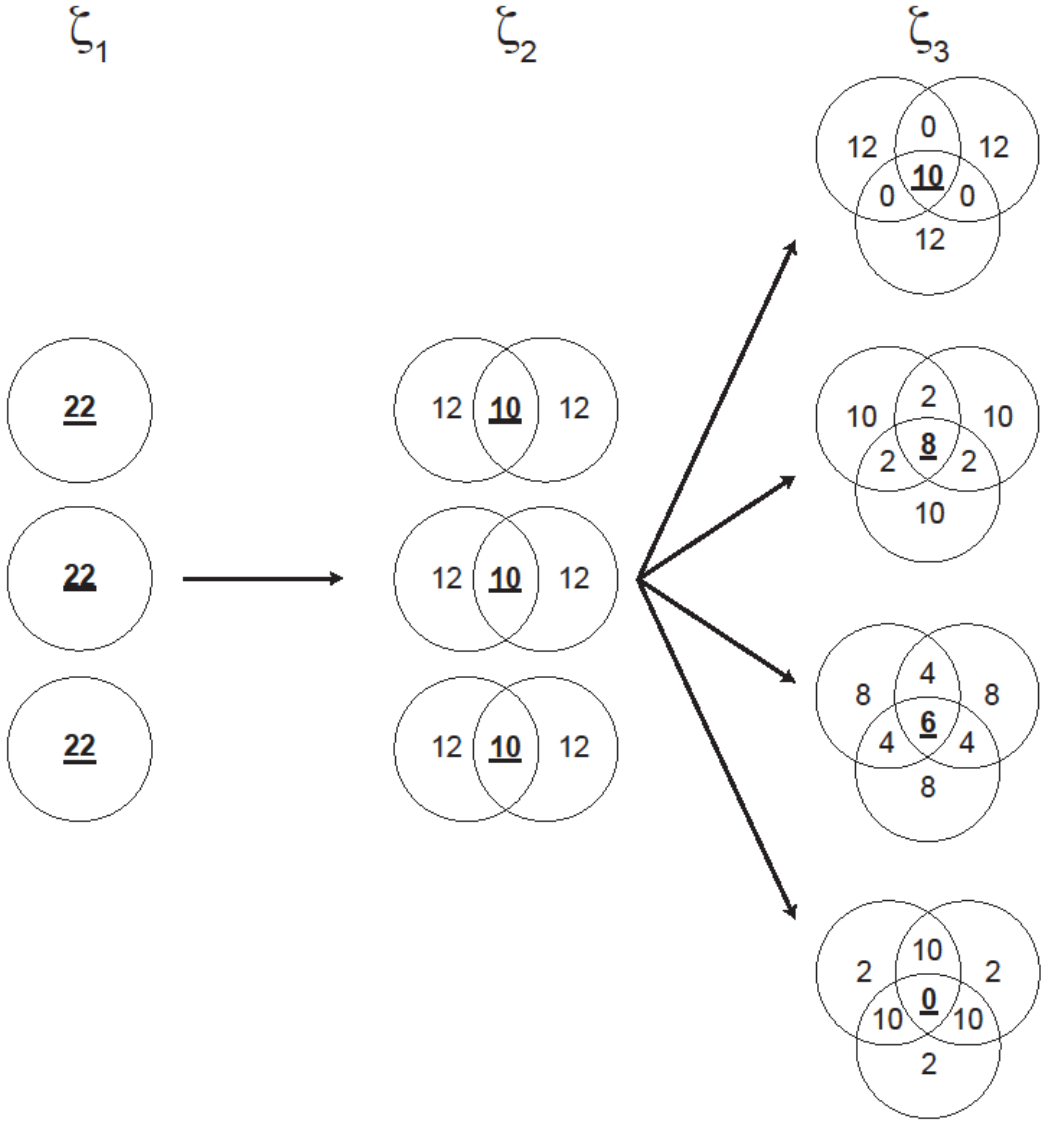


Figure 1. Four (amongst many) different ways to partition species turnover when 3 assemblages (zeta order = 3) are combined. Only considering richness (alpha diversity, corresponding to ζ_1) and pairwise compositional change (beta diversity, corresponding to ζ_2, where zeta order = 2) provides an incomplete description of the community. Numbers in bold and underlined are the values of zeta diversity for orders 1 to 3.

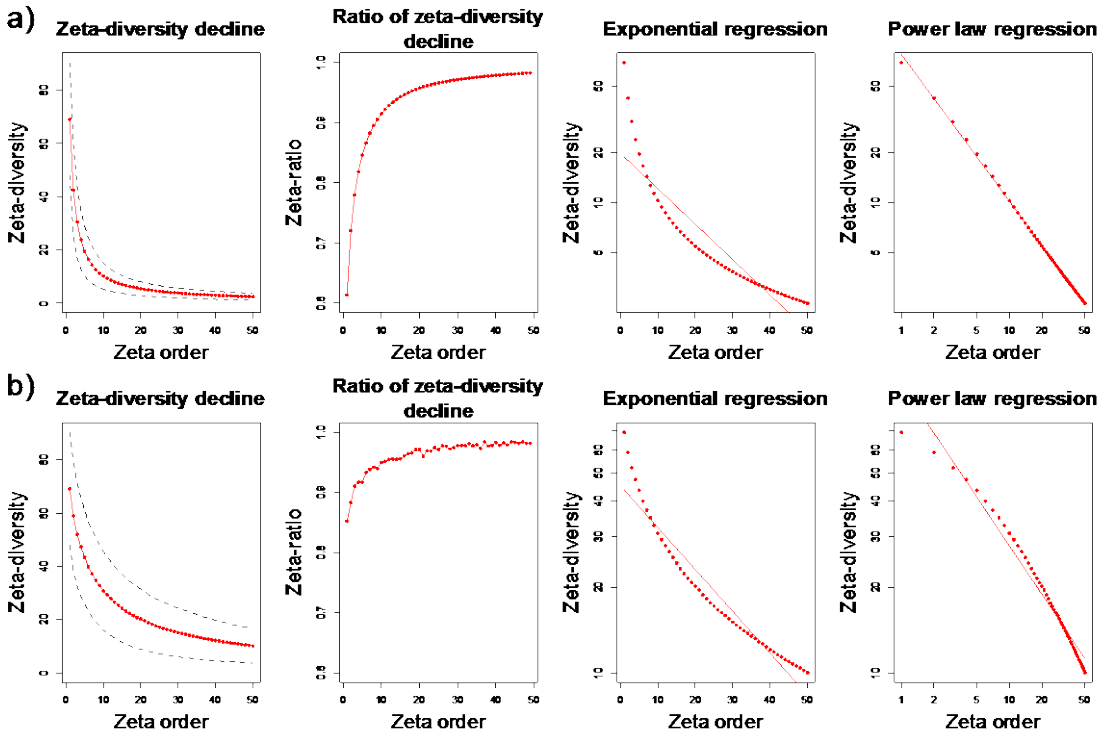


Figure 2. Zeta decline between orders 1 to 50 characterising the bird data for 25 × 25 km cells computed with a) `Zeta.decline.ex` ('expected', i.e. all combinations) and b) `Zeta.decline.mc` with `sam=1000` and using a non-directional nearest-neighbour (NON) subsampling scheme.

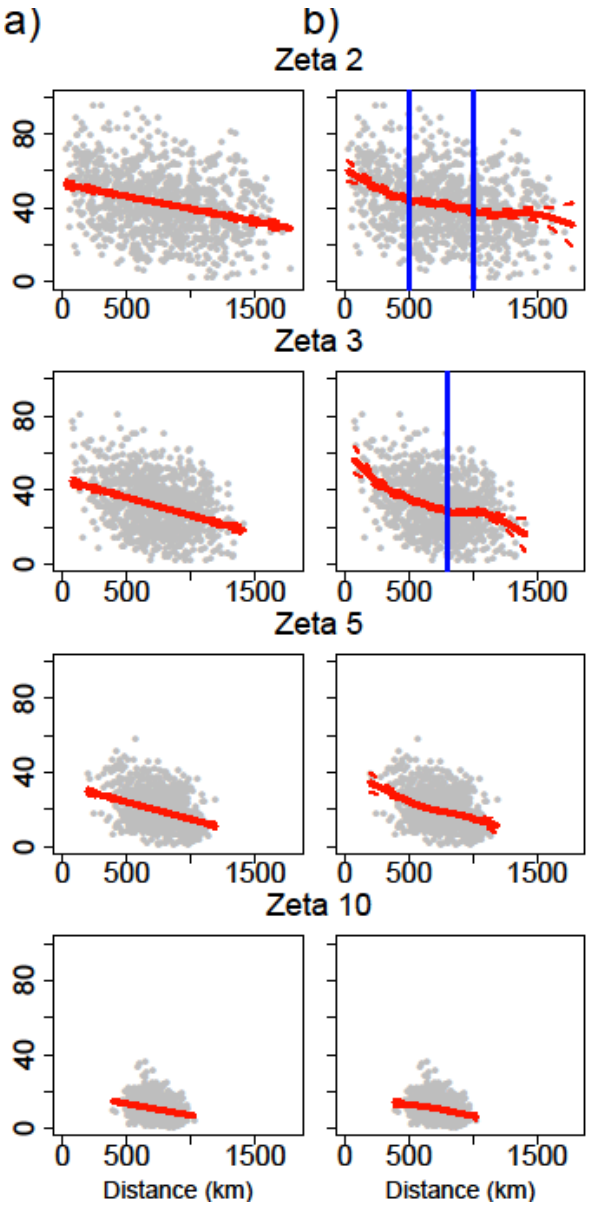
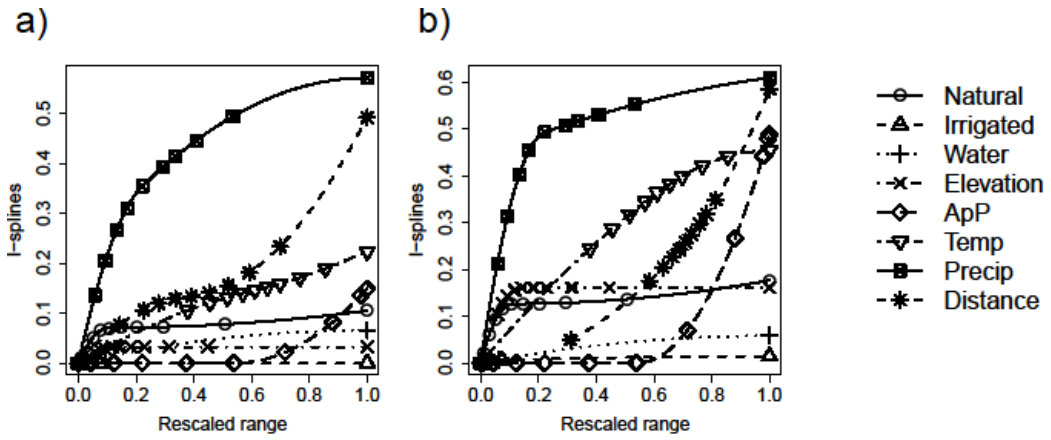


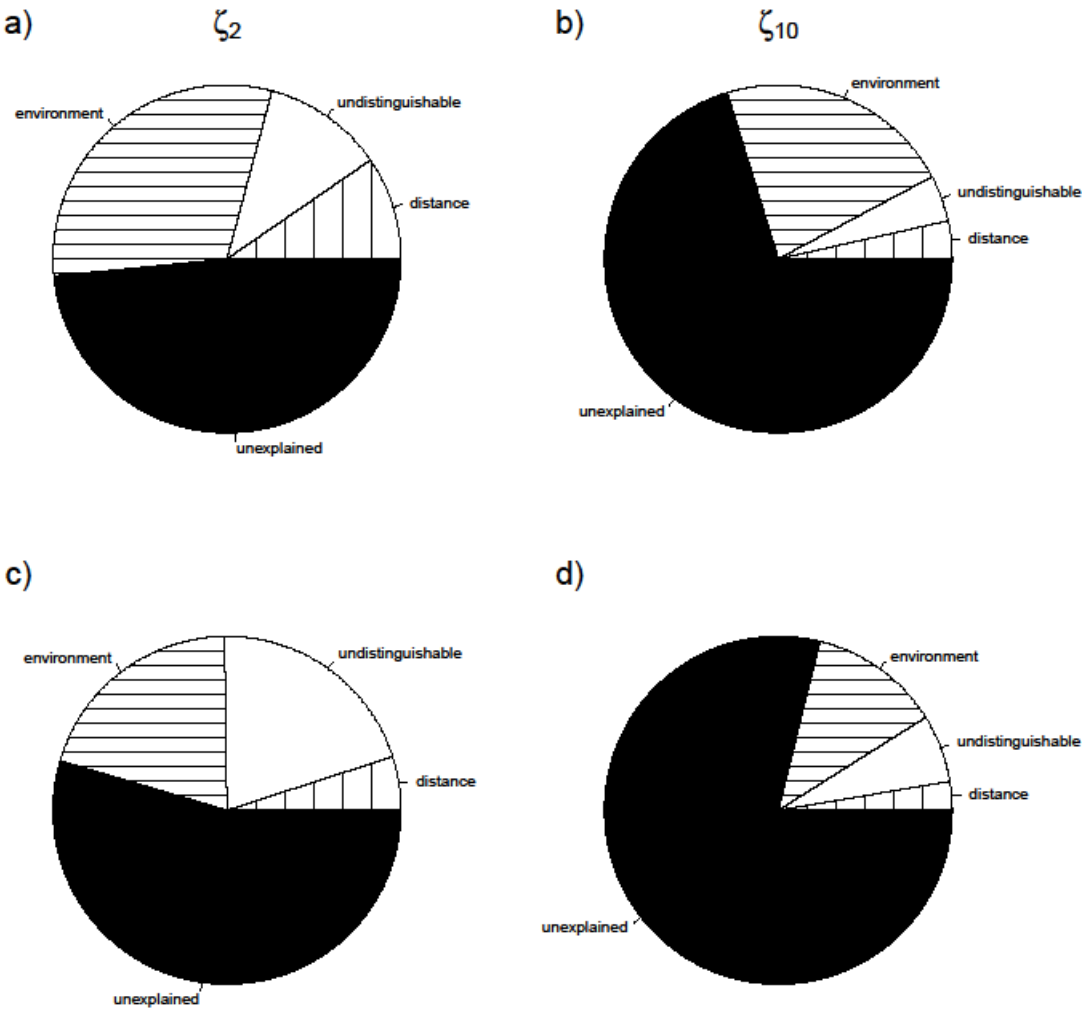
Figure 3. Distance decay of ζ_2 to ζ_{10} for the bird data for 25 × 25 km cells using a) a linear regression and b) a generalised additive model (GAM). The linear regression shows clear distance decay, with a steeper slope for ζ_3 and ζ_5 than for ζ_2, suggesting that rare species are dispersed relative to the space-filling properties of the species with higher occurrence levels. The GAM also reveals three slightly different rates of decline for zeta ζ2 (with thresholds at ~500 km and ~1000km) and two clearer different rates of decline for zeta ζ3 (with a threshold at ~800 km), indicated by the vertical blue lines.



844

845 **Figure 4.** I-splines explaining zeta diversity of bird assemblages over South-East
846 Australia for (a) ζ_2 and (b) ζ_{10}, using 7 environmental variables and spatial distance,
847 for 25 × 25 km cells. The relative maximum values of the splines indicate the relative
848 contribution of each variable to explaining zeta diversity. By contrast, the slope of the
849 splines provide information on how the influence of each variable changes along the
850 gradient of values. For example, changes in precipitation have more influence on
851 compositional change in dry areas (low rescaled range value) than in wet areas (high
852 rescaled range value), especially for ζ_{10} (Latombe et al. 2017).

853



854

855 **Figure 5.** Proportion of variation (variation partitioning) in ζ_2 and ζ_{10} explained by

856 environmental and distance variables for the bird data for 25 × 25 km cells, when I-

857 splines (a,b) and linear regressions (c,d) are used. The larger proportion of variation

858 explained by the environment is consistent with the relative amplitudes of the

859 corresponding I-splines (Figure 4).

860

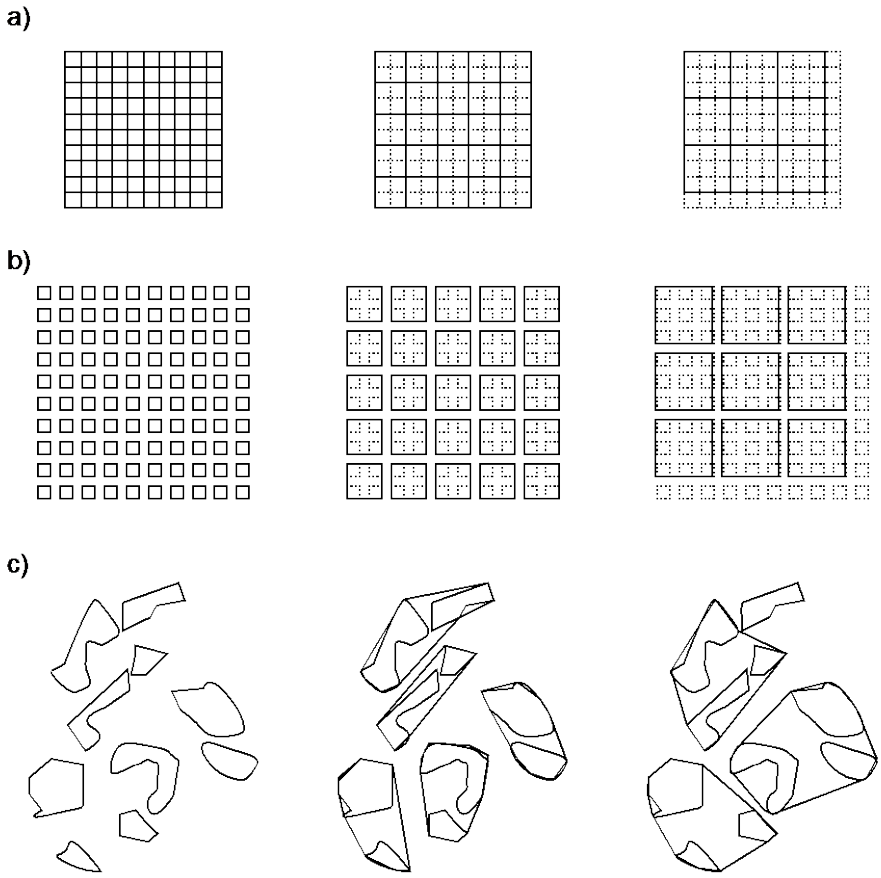


Figure 6. Resampling of data depending on the initial sampling scheme, classified according to three of the four sampling schemes defined by Scheiner et al. (2011) (the fourth type, strictly nested quadrats, is not relevant here) (see also McGeoch et al., 2017). For a) a continuous grid and b) a regular but discontinuous grid, adjacent cells are grouped together. For c) irregularly distributed sites of potentially varying size, such as islands, sites are grouped based on the distance between them. For a) and b), resampling based on minimum distance can also be applied to the grid cells, but grouping adjacent sites is not applicable to c). For a) and b), if the number of cells at fine grain is not a divisor of the number of cells at coarse grains, some cells are lost during aggregation. For c), the order in which sites are grouped can influence the final configuration. The bird datasets can be seen as a) and c), since the original cells are regularly distributed, but only cells with observed richness within 10% of estimated richness are included and the remaining cells are therefore irregularly distributed, but have a constant area.

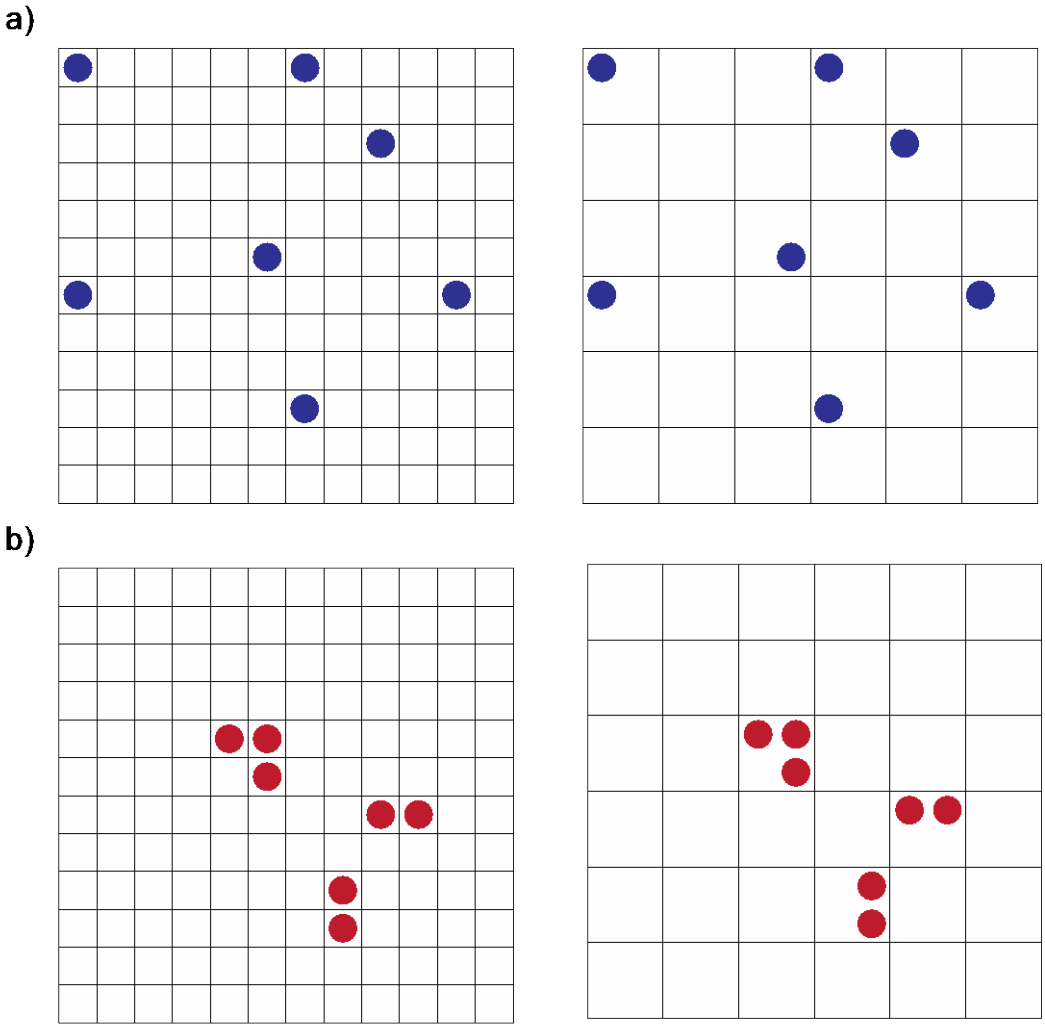
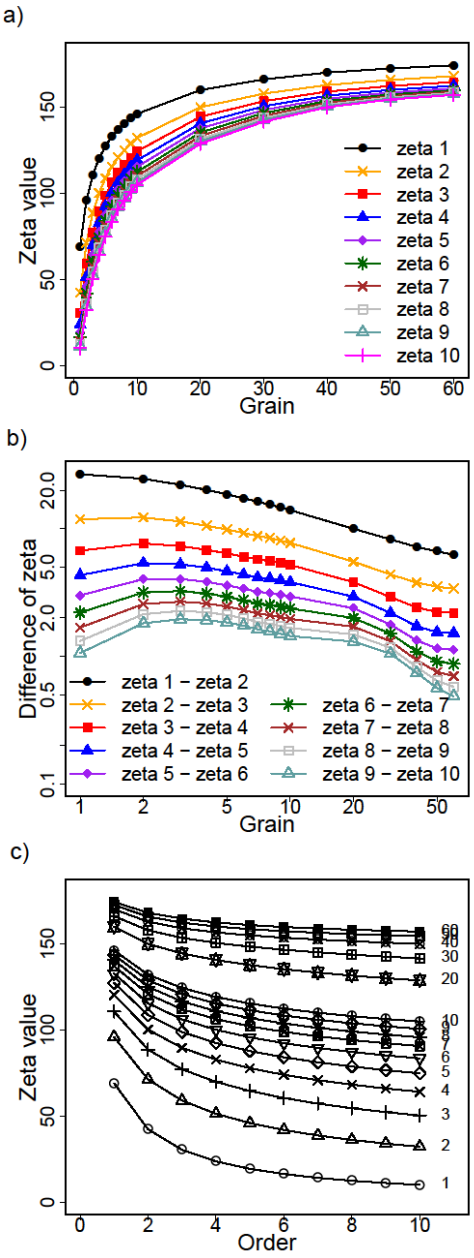


Figure 7. Effect of species spatial aggregation on their contribution to different orders of zeta when the grain changes. a) Highly dispersed species still contribute to high orders of zeta under the ALL sampling scheme (orders 1 to 7 in this example, since the species is present in 7 different grid cells) when the grain becomes coarser. b) At fine grain, spatially aggregated species contribute to higher orders of zeta (orders 1 to 7) than at coarse grain (orders 1 to 3).



884

885 **Figure 8.** Scale dependence of ζ_1 to ζ_6 for the bird data by aggregating 1 to 60 cells

886 based on minimum distance (Figure 6c). a) As the grain increases and cells are

887 aggregated, species share more cells. b) For orders ≥ 2, the difference (ζ_i - ζ_{i+1})

888 initially increases with grain, then decreases (Hui et al., 2010). c) The zeta decline

889 from orders 1 to 10 is slightly sharper when aggregating 2 or 3 cells (~1500 km^2; the

890 grain is indicated on the right) than without aggregating cells (fine grain,

891 corresponding to the '1' zeta decline) or when aggregating more cells (coarse grain,

892 i.e. >4 in this case).

893

Supporting information

Appendix A: Supporting tables and figures

Appendix B: Code for reproducibility of examples