A method of predicting ecological community structure

John Alroy

School of Natural Sciences, Macquarie University, NSW 2109, Australia

Correspondence
John Alroy
john.alroy@mq.edu.au

Abstract

1. Species inventories are the most basic form of ecological data. They provide information both about species richness and about community assembly rules. Fitting species abundance distribution models yields such information. Previous distributions either fit the data badly, assume that all species are equivalent, or ignore sampling processes. A distribution called the compound exponential-geometric series (CEGS) assumes that species vary randomly in their underlying abundances and that inventories are random draws reflecting this variation.
2. The predictive power of CEGS and of four rival distributions is tested in two ways. First, richness estimates for entire inventories are used to predict recomputed estimates after randomly winnowing of individuals. Second, counts for local inventories are used to predict counts for matched samples that represent the same ecological groups and biogeographic realms.
3. CEGS yields the best count predictions and is rarely rejected by the data. Its richness estimates are precise and nearly unbiased, so it outperforms not only other theoretical distributions but the benchmark Chao 1 extrapolation index.
4. Because of its solid performance, simple theoretical basis, and ability to yield absolute species richness estimates that are not lower bounds, CEGS may solve the twin problems of describing abundance distributions and estimating diversity.

KEYWORDS

compound exponential-geometric series distribution, coverage-based rarefaction, Fisher's alpha, log series, Poisson log normal, shareholder quorum subsampling, Weibull distribution

1 | INTRODUCTION

Lists of species found in particular places at particular times are the bedrock foundation of community ecology. When combined with counts of individuals belonging to each species, they provide a powerful tool for understanding ecological structure (Fisher et al., 1943; Preston, 1948; Hurlbert, 1971; Hill, 1973; Hubbell, 2001). Two aspects of structure are considered to be paramount: species richness and the shape of count distributions.

The ecological literature has focused more and more on treating both things as aspects of "diversity" to be measured at once using indices called Hill numbers, such as Shannon's *H* and Simpson's *D* (Hill, 1973; Chao et al., 2014; Moreno et al., 2017; Roswell et al., 2021). This approach is problematic. First, there is no a priori basis for deciding which number is most informative, explaining why workers in this camp tend to present multiple values. Second, the more extreme indices virtually ignore rare species: indeed, singletons are almost entirely zeroed out by the sample-size corrected form of *D* (Hurlbert, 1971). Third, Hill numbers say nothing about the underlying processes that generate the data. Finally, it would seem trivial to argue that any two statistical properties should be measured with two statistics.

A second approach is to ignore the shape problem and simply try to estimate the total richness of a species pool by extrapolation. There are many particular strategies (Colwell & Coddington, 1994), but the most popular method in this class is called Chao 1 (Chao, 1984). Chao 1 is important because it yields values very tightly correlated with those provided by alternatives such as the abundance coverage estimator (ACE: Chao & Lee, 1994) and interpolation and extrapolation with Hill numbers (iNEXT: Chao et al., 2014). The problem with all of these well-known estimators is that they are designed to provide lower-bound values only, so they are literally intended to be inaccurate. Hardly any branch of science favours methods that are intrinsically biased.

Calculating a relative measure of species richness by interpolation, a.k.a. rarefaction, is a third way to estimate diversity – if not other structural properties. The premise is that extrapolation is just too imprecise and inaccurate to be useful, but it is trivial to say how many species would be recovered in a sample with a given number of individuals (Hurlbert, 1971) or with a given level of frequency distribution coverage (Alroy, 2010; Chao & Jost, 2012). Rarefaction is deeply unsatisfying because relative richness says nothing about shape; total species richness is intrinsically interesting; and relative richness can only be understood in the context of a specified sampling quota.

Here I advocate a final strategy: fitting a theoretical model of abundance that yields both a richness estimate and a description of shape. An example is the Poisson log normal distribution (PLN: Bulmer, 1974; Connolly et al., 2005). The log normal in any guise has long been understood to be a reasonable descriptor of count distributions (Preston, 1948; Antão et al., 2021; Callaghan et al., 2023). A second option is the Weibull distribution (see Ulrich et al., 2018), which can provide richness estimates when its discretised version (Nakagawa & Osaki, 1975) is used. Despite having received considerable support (Baldridge et al., 2016), the log series (LS: Fisher et al., 1943) does not specify richness because its equation assumes that the "zero class" of counts is infinitely large. It does, however, yield the powerful statistic called Fisher's α, which is the sole governing parameter of the distribution.

This paper focuses on a novel distribution called the compound exponential-geometric series (CEGS). Like the PLN, it assumes an underlying distribution: exponential instead of log normal. It also assumes a sampling process that yields the actual counts: geometric instead of Poisson. The open question is whether this approach better predicts the two fundamental structural properties of communities.

Predictions of richness are tested here with a validation approach that depends on degrading each species inventory: methods that yield much the same richness estimates for full and subsampled inventories are to be preferred. Predictions of shape are tested by fitting each model to each inventory, and then seeing how these fits project onto count distributions for similar samples representing the same geographic regions and biotic groups. For example, the fit for each Neotropical butterfly inventory is used to predict the counts for another one that is of much the same size.

The results indicate that the only close rival to CEGS is the unrealistic log series. Thus, many or even most real communities might obey its basic model.

2 | MATERIALS AND METHODS

2.1 | Data

The data are species inventories drawn from a global, openly available literature compilation called the Ecological Register (Alroy, 2015, 2017, 2024). This data set is unusual because (1) it is highly and evenly dispersed among taxonomic groups and geographic areas, and (2) every inventory is matched with a list of integer counts, making it possible to fit theoretical

abundance distributions and diversity estimators. The organisms include trees and terrestrial animals. Major groups that dominate particular inventories include ants (161 inventories), bats (276), birds (336), butterflies (211), carnivores (211), dung beetles (182), frogs (162), lizards (91), mosquitoes (185), odonates (134), orthopterans (90), rodents (277), and trees (435). Other groups may be present within a given inventory (e.g., ungulates within carnivore-dominated ones). Additional inventories are dominated by groups such as spiders that are less well-represented. These are included in the diversity analysis but not the analysis that involves predicting distribution shapes.

2.2 | Theoretical abundance distribution

CEGS is derived by assuming that the expected value of a draw from the underlying exponential distribution equals the expected count produced by imposing a geometric sampling process upon it. Let U be a random uniform variate. $-\ln U$ yields an exponential distribution; call this E. The expected value of E is just E divided by a scaling constant called λ. Also conventionally, p is the governing parameter of the geometric series. The expected count is well-known to be $(1 - p)/p = 1/p - 1$. Also assume that p can be varied by taking a root, meaning a shape parameter denoted $1/\gamma$. So we have:

$$1/p^{1/\gamma} - 1 = \int E/\lambda$$
$$p = \int 1/(E/\lambda + 1)^{\gamma} \quad (1)$$

The expression is integrated in this study using the built-in *integrate* function in the R programming language.

The CEGS species abundance distribution (SAD) is just the probability mass function $p_X(x)$ of the log series, which predicts the chance of obtaining a given count x such as the number of singletons or number of doubletons:

$$p_X(x) = (1 - p)^x p \quad (2)$$

The richness estimator is trivial because the chance of obtaining a zero count under the geometric series is just p: $(1 - p)^0 p = p$. Therefore, if R is the total size of the species pool and S is the raw number of observed species, then:

$$R = S/(1 - p) \quad (3)$$

2.3 | Distribution fitting method

The standard maximum likelihood (ML) function for handling SADs (Grøtan & Engen, 2008; Prado et al., 2018) could have been used to fit the model. In this context, the likelihood is just the product of the probabilities of observing the individual counts based on a given SAD. So if the SAD is 0.5, 0.2, 0.1... and the counts are 1, 1, 2, and 3, then the joint likelihood is 0.5^2 x 0.2 x 0.1 = 0.005.

However, ML solutions tend to be unstable. To reduce sampling error, likelihood differencing (LD) was used instead. LD assumes that best parameter estimates are located near regions of rapid changes in likelihoods. After all, when parameters are implausible neighbouring likelihoods are all close to zero, so differences are small. Rates of change should be high near likelihood peaks or high ridges because they should have steep sides.

The LD procedure involves computing a grid of likelihoods representing combinations of λ and γ. In the current analysis, the 250 grid points for each parameter had values of $251/i - 1$ where i was the rank of each point. Therefore, the grid dimensions spanned 250 to 0.004. Separate, orthogonal matrices for each parameter called Λ and Γ were defined. A matrix of sums of the absolute values of the differences in raw likelihoods L between each point i and its neighbours was computed, as in $|L_i - L_{i-1}| + |L_i - L_{i+1}|$. The sums were scaled to their total to produce a difference-based weight matrix D, D was multiplied separately across Λ and Γ, and the summations of the products were used to infer expected values of λ and γ.

LD has the advantage of not being dependent on the grid configuration, unlike Bayesian estimates that assume point sampling. If the grid is already fine, this is because adding points splits the differences in values of neighbouring likelihoods. So if there are x as many points in a region as there were previously, then each one has $1/x$ as much weight, so the total in the region is largely unchanged.

2.4 | Alternative distributions

In addition to CEGS, four interesting and plausible models are considered. (1) As applied to SADs, the geometric series (GS) is structurally related CEGS. It predicts flat distributions,

with relatively few singletons and many species with subequal counts. (2) The log series (Fisher et al., 1943) is not only the oldest one-parameter model in the literature, but a highly popular one (Baldridge et al., 2016). It is problematic because it does not predict richness by itself. Also, it can be interpreted either as a sampling distribution (Fisher et al., 1943) or as the result of a birth-death process (Kendall, 1948; Hubbell, 2001), but not both at once. (3) As mentioned, the Poisson log normal (Bulmer 1974) is biologically realistic and has received much support (Connolly et al., 2005; Antão et al., 2021; Callaghan et al., 2023). (4) The discretised Weibull distribution (Nakagawa & Osaki, 1975) is not compound and has no sampling model, but it does allow for very high and low counts being found in the same inventory.

2.5 | Alternative diversity estimators

In addition to richness estimates provided by CEGS, the PLN, and the Weibull, three other diversity metrics are considered: Fisher's α, the powerful diversity stand-in that governs the log series; Chao 1; and rarefied richness estimated with the method called shareholder quorum subsampling in its algorithmic form (Alroy, 2010) and coverage-based rarefaction (CBR) in its analytical form (Chao & Jost, 2012). A quorum (= target coverage level) of 0.5 was used to guarantee that a large majority of species inventories would yield a CBR estimate. Higher values would exclude many of them.

Hill numbers other than richness itself, such as Simpson's *D* and Shannon's *H* (Hill, 1973), are not considered here because they are not parameters of distributions such as Fisher's α and are not intended as proxies for richness per se. Briefly, *D* and *H* perform poorly when a two-parameter distribution operates because they are very sensitive to the presence of an occasional high count. They are also strongly correlated with the values yielded by CBR, making it redundant to analyse them.

2.6 | Richness prediction test

Returning the same value regardless of the size of a data set is a hallmark feature of a good statistical estimator. To test for this property, species richness estimates were generated using all of the above methods, all of the inventories were harshly degraded, and the estimates were recomputed. The algorithm was to randomly draw two individuals from each inventory for

each species in the full set. For example, when 10 species were present in the raw data 20 individuals were drawn. Not all inventories are this large: 2359 of the 3042 consistently analysable inventories could be used here.

2.7 | Matched inventory prediction test

In many cases, rival distribution models ape each other closely when fitted to the same inventory. Thus, conventional decriptive statistics such as the corrected Akaike information criterion (Hurvich & Tsai, 1993) are of limited utility. A more powerful approach is to predict the shape of each inventory's distribution by first fitting the model to another inventory matching it both in biological terms and in terms of sampling intensity. After all, the goal of science is to test strong predictions.

Here, matches were identified by sorting all inventories into bins based on their dominant ecological groups (see above) and biogeographic realms. For example, all ants from the Afrotropics were considered to be a set. Next, the inventories were ordered from worst to best sampled on the basis of geometric mean counts. For example, counts of 1, 1, 1, and 10 yield a mean of 1.78 and counts of 2, 2, 2, and 2 yield one of 2, so the second inventory is better sampled. Each model was fit to the worst-sampled inventory and projected onto the second-worst, and so on. Therefore, the worst-sampled one in each bin was fitted to predict its neighbour but not predicted itself, and vice versa for the best-sampled. Fit was evaluated by computing the log likelihood with the standard equation.

3 | RESULTS

Degrading the inventories through random subsampling creates large problems for most methods. PLN and Weibull estimates are often nearly random (Figs. 1E, F), while raw counts, GS estimates, and Chao 1 index values are well below the line of unity (Figs. 1A, C, G). CBR yields consistent values (Fig. 1H), but this is no surprise because the same target for frequency distribution coverage (0.5) is used in all cases. The real issues with CBR are that it provides relative estimates only and that it is highly sensitive to random variation in high counts, as could be shown with additional analyses. The basic pattern is much the same for CEGS and Fisher's α (Figs. 1B, D). However, the latter assumes the one-parameter log series model and provides a generic diversity statistic, so it is unrealistic and uninformative about

species richness. In sum, CEGS offers the only consistent and reasonably precise true species richness estimates.

In terms of abundance distribution shapes, samples falling in the same categories with respect to dominant groups and biogeographic realms predict each other consistently better with CEGS (Fig. 2). Although support is split for many small inventories, CEGS comes out well ahead of the other two-parameter models in terms of strong support (defined as a likelihood ratio of 10 or more). Tallies are 732 wins out of 827 comparisons against the PLN (88.5%), and 700 out of 759 against the Weibull (92.2%). CEGS overtakes the GS even more frequently (2015 wins out of 2127 comparisons = 94.7%). Finally, it beats the LS by a substantial margin (484 wins out of 812 comparisons = 59.6%) . Many large differences in log likelihoods favour CEGS over all rivals (Fig. 2), including the LS (Fig. 2B).

4 | DISCUSSION

4.1 | Could CEGS account for a large majority of ecological communities?

CEGS is decisively better than the alternative two-parameter models tested here, and based on this it could describe almost all terrestrial inventories. Conventionally, decisive support for a model equates to a likelihood ratio of about 100. By this standard, CEGS fits 48 matched samples unambiguously better than all of its rivals. Most of the remaining models are much worse: the GS is strongly supported in just 12 cases, the PLN in six, and the Weibull in two. This is surprising because the PLN and Weibull in particular are very good at mimicking a CEGS distribution, since all three include both shape and scale parameters. Meanwhile, the LS comes out well ahead of everything 144 times. This difference is misleading because CEGS support is clear when the similar-seeming PLN and Weibull aren't considered. In head-to-head comparisons, 285 cases favour CEGS and 163 favour the LS.

On top of that, the CEGS prediction is not decisively rejected (ratio < 100) in 2432 out of 2639 cases (92.2%), as opposed to 86.9% for the LS, which is the runner-up. Even with a milder likelihood ratio cutoff of < 10, CEGS is not rejected in 2167 of 2639 cases (82.1%). So there is no real wiggle room here for advocates of the LS, PLN, and Weibull. If the ecological world isn't governed by CEGS, it is either governed by no one model in particular or it is governed by something very similar to CEGS. The former notion is unparsimonious – broad scientific explanations are better than non-explanations. The latter is special pleading.

4.2 | What about the exceptions?

The partial success of the log series probably indicates poor sampling that obscures the underlying abundances, or possibly slow birth-death-immigration processes (Kendall, 1948; Hubbell, 2001) instead of the rapid geometric sampling process assumed to be important by CEGS. Meanwhile, the PLN should result when many factors governing distributions are summed. That's the point of the central limit theorem. The fact that it is a poor model (Figs. 1E, 2C) indicates that most community inventories reflect more than just a summation of many independent and jointly uninteresting processes. And although the poorly-performing Weibull is good at predicting high-variance distributions, it has no real biological basis because it incorporates no sampling process: it assumes an instant transition between underlying abundances and counts.

4.3 | What about overlooked models?

There are many models in the literature (Matthews & Whittaker, 2014), including for example the zero-sum multinomial (ZSM: Hubbell, 2001). In practice, the ZSM makes predictions that very closely track those of the LS, which explains why this paper omits it. Another option is the gambin (Ugland et al., 2007), but this model has not been formulated to fit species abundance distributions sensu strico (= counts of species sharing counts). Finally, a series of mostly one-parameter niche partitioning models has been proposed, but these are not held in high regard by contemporary reviewers (e.g., Baldridge et al., 2016). They tend to make strong assumptions about competition that seem unnecessary. None of the models considered here do such a thing, so arguably all of them are ecologically neutral in the general sense. To put this another way, if another really good, really simple model already exists then I am not aware of it, and if not, then coming up with one would seem like a tall order.

4.4 | What about other systems?

There is some chance that some model other than CEGS might work better for marine organisms or microbes. Finding out whether that is true would be a good path for future research. However, when it come to terrestrial macroscopic organisms, it seems hard to argue

that the current data set omits anything important in terms of geographic and taxonomic scope: all biogeographic realms and 13 of the most commonly-studied groups are each represented by at least scores of inventories compiled with no agenda from the primary literature.

4.5 | Do we still need other methods of estimating species richness?

CEGS has a barely visible sample size bias and perhaps could be more precise (Fig. 1B). However: (1) it has an explicit basis in realistic SAD theory, whereas lower-bound models like Chao 1 generally assume uniform distributions (Alroy, 2017); (2) it is straightforward and minimal in terms of formulation and computation; (3) it does specifically aim to provide an accurate estimate, unlike lower bound methods; (4) it presents absolute values, unlike rarefaction methods; and (5) it presents values in units of species, unlike Fisher's α, which rests on a distribution (the log series) that assumes unlimited richness.

If another robust method of quantifying diversity is left to be found, it would only be universally acceptable if it provided absolute species richness figures. Furthermore, proponents of any rival method would have to provide evidence that it meets two necessary conditions for any good estimator: yielding internally consistent values and depending on a realistic model.

Acknowledgements

I thank Barry Brook, Michael Foote, and other colleagues for helpful feedback on the CEGS distribution.

Data availability

The empirical data used in this study are available via the Dryad Digital Repository at https://datadryad.org/stash/dataset/doi:10.5061/dryad.brv15dvdc.

Conflict of interest

The author declares no conflicts of interest.

REFERENCES

Alroy, J. (2010). The shifting balance of diversity among major marine animal groups. *Science,* 329, 1191-1194. https://doi.org/10.1126/science.1189910

Alroy, J. (2015). The shape of terrestrial abundance distributions. *Science Advances,* 1, e1500082. https://doi.org/10.1126/sciadv.1500082

Alroy, J. (2017). Effects of habitat disturbance on tropical forest biodiversity. *Proceedings of the National Academy of Sciences USA,* 114, 6056-6061. https://doi.org/10.1073/pnas.1611855114

Alroy, J. (2024). Data from: three models of ecological assembly: terrestrial species inventories. *Dryad Digital Repository* https://datadryad.org/stash/dataset/doi:10.5061/dryad.brv15dvdc

Antão, L. H., Magurran, A. E., & Dornelas, M. (2021). The shape of species abundance distributions across spatial scales. *Frontiers in Ecology and Evolution,* 9, 626730. https://doi.org/10.3389/fevo.2021.626730

Bulmer, M. G. (1974). On fitting the Poisson lognormal distribution to species abundance data. *Biometrics,* 30, 651-660. https://doi.org/10.2307/2529621

Callaghan, C. T., Borda-de-Água, L., van Klink, R., Rozzi, R., & Pereira, H. M. (2023). Unveiling global species abundance distributions. *Nature Ecology & Evolution,* 7, 1600-1609. https://doi.org/10.1038/s41559-023-02173-y

Chao, A. (1984). Nonparametric estimation of the number of classes in a population. *Scandinavian Journal of Statistics,* 11, 265-270.

Chao, A., & Jost, L. (2012). Coverage-based rarefaction and extrapolation: standardizing samples by completeness rather than size. *Ecology,* 93, 2533-2547. https://doi.org/10.1890/11-1952.1

Chao, A., & Lee, S.-M. (1992). Estimating the number of classes via sample coverage. *Journal of the American Statistical Association,* 87, 210-217. https://doi.org/10.1080/01621459.1992.10475194

Chao, A., Gotelli, N. J., Hsieh, T. C., Sander, E. L., Ma, K. H., Colwell, R. K., & Ellison, A. M. (2014). Rarefaction and extrapolation with Hill numbers: a framework for sampling and estimation in species diversity studies. *Ecological Monographs,* 84, 45-67. https://doi.org/10.1890/13-0133.1

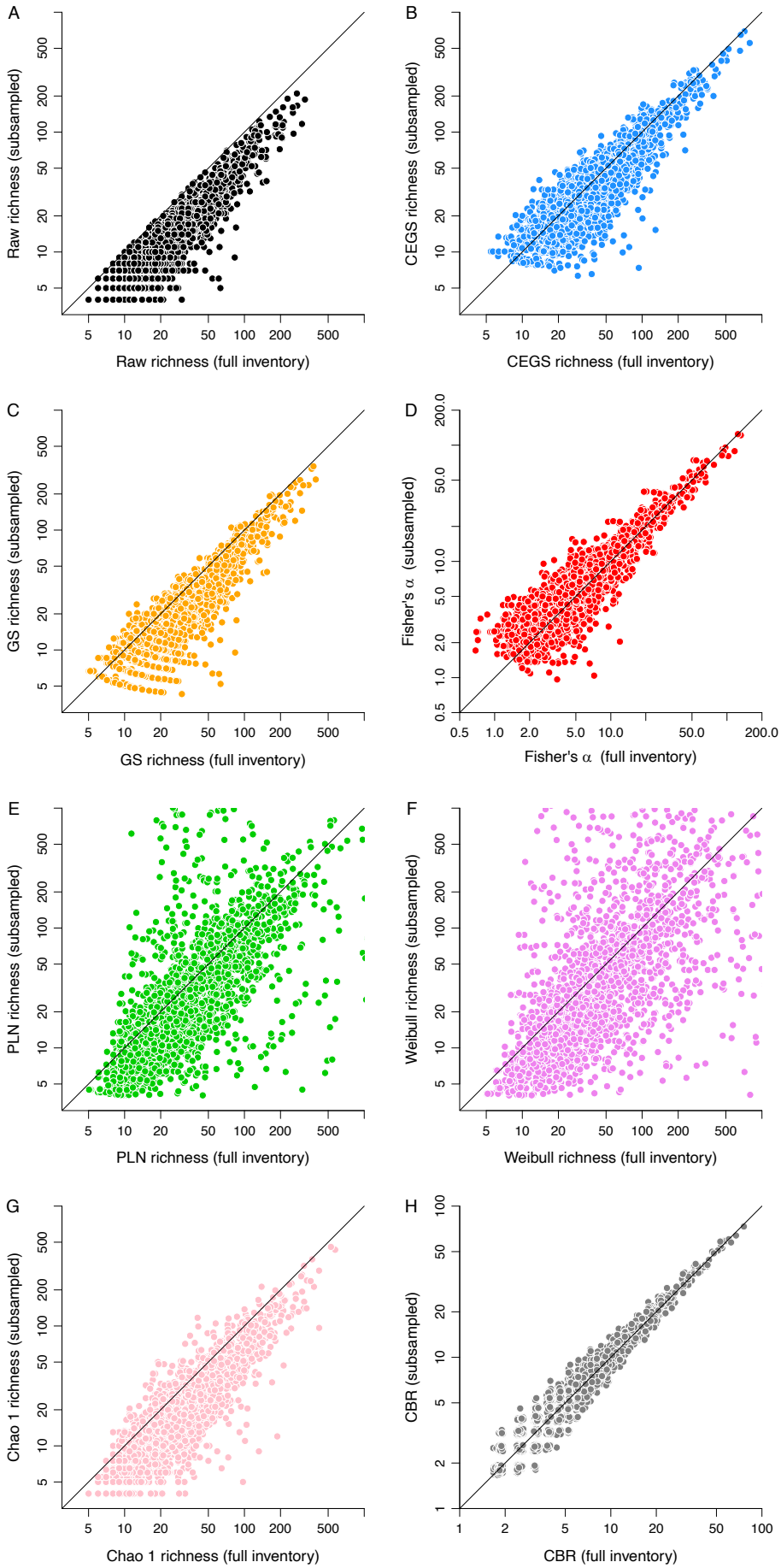
364 Colwell, R. K., & Coddington, J. A. (1994). Estimating terrestrial biodiversity through
365 extrapolation. *Philophical Transactions of the Royal Society B,* 345, 101-118.
366 https://doi.org/10.1098/rstb.1994.0091
367 Connolly, S. R., Hughes, T. P., Bellwood, D. R., & Karlson, R. H. (2005). Community
368 structure of corals and reef fishes. *Science,* 309, 1363-1365.
369 https://doi.org/10.1126/science.1113281
370 Fisher, R. A., Corbet, A. S. & Williams, C. B. (1943). The relation between the number of
371 species and the number of individuals in a random sample of an animal population.
372 *Journal of Animal Ecology,* 12, 42-58. https://doi.org/10.2307/1411
373 Grøtan, V., & Engen, S. (2008). poilog: Poisson lognormal and bivariate Poisson lognormal
374 distribution. R package version 0.4.
375 Hill, M. O. (1973). Diversity and evenness: a unifying notation and its consequences.
376 *Ecology,* 54, 427-432. https://doi.org/10.2307/1934352
377 Hubbell, S. P. (2001). The unified neutral theory of biodiversity and biogeography. Princeton
378 University Press, Princeton, NJ.
379 Hurlbert, S. H. (1971). The nonconcept of species diversity: a critique and alternative
380 parameters. *Ecology,* 52, 577-586. https://doi.org/10.2307/1934145
381 Hurvich, C. M., & Tsai, C. L. (1993). A corrected Akaike information criterion for vector
382 autoregressive model selection. *Journal of Time Series Analysis,* 14, 271-279.
383 https://doi.org/10.1111/j.1467-9892.1993.tb00144.x
384 Kendall, D. G. (1948). On some modes of population growth leading to R. A. Fisher's
385 logarithmic series distribution. *Biometrika,* 35, 6-15. https://doi.org/10.1093/biomet/35.1-
386 2.6
387 Matthews, T .J., & Whittaker, R. J. (2014). Fitting and comparing competing models of the
388 species abundance distribution: assessment and prospect. *Frontiers of Biogeography,* 6,
389 67-82. https://doi.org/10.21425/F5FBG20607
390 Moreno, C. E., et al. (2017). Measuring biodiversity in the Anthropocene: a simple guide to
391 helpful methods. *Biodiversity and Conservation,* 26, 2993-2998.
392 https://doi.org/10.1007/s10531-017-1401-1
393 Nakagawa, T., & Osaki, S. (1975). The discrete Weibull distribution. *IEEE Transactions on*
394 *Reliability,* 24, 300-301. https://doi.org/10.1109/TR.1975.5214915
395 Prado, P. I., Dantas Miranda, M. & Chalom, A. (2018). sads: maximum likelihood models for
396 species abundance distributions. R package version 0.4.2.

Preston, F. W. (1948). The commonness, and rarity, of species. *Ecology,* 29, 254-283. https://doi.org/10.2307/1930989

Roswell, M., Dushoff, J., & Winfree, R. (2021). A conceptual guide to measuring species diversity. *Oikos,* 130, 321-338. https://doi.org/10.1111/oik.07202

Ugland, K. I., Lambshead, P. J. D., McGill, B., Gray, J. S., O'Dea, N., Ladle, R. J., & Whittaker, R. J. (2007). Modelling dimensionality in species abundance distributions: description and evaluation of the Gambin model. *Evolutionary Ecology Research,* 9, 313-324.

Ulrich, W., Nakadai, R., Matthews, T. J., & Kubota, Y. (2018). The two-parameter Weibull distribution as a universal tool to model the variation in species relative abundances. *Ecological Complexity,* 36, 110-116. https://doi.org/10.1016/j.ecocom.2018.07.002


A
Raw richness (subsampled)
Raw richness (full inventory)
B
CEGS richness (subsampled)
CEGS richness (full inventory)
C
GS richness (subsampled)
GS richness (full inventory)
D
Fisher's α (subsampled)
Fisher's α (full inventory)
E
PLN richness (subsampled)
PLN richness (full inventory)
F
Weibull richness (subsampled)
Weibull richness (full inventory)
G
Chao 1 richness (subsampled)
Chao 1 richness (full inventory)
H
CBR (subsampled)
CBR (full inventory)

409 Figure 1. Diversity estimates obtained by analysing complete species inventories (x-axes)
410 and by analysing subsampled inventories (y-axes). Each inventory is randomly sampled so
411 that the number of individuals equals twice the original number of species. (A) Raw richness.
412 Residual standard error (RSE) around the line of unity = 0.656; median distance of the points
413 from the line on a log scale (offset) = –0.551. (B) Estimates based on fitting the CEGS model
414 (RSE 0.420, offset –0.150). (C) Geometric series (GS) model estimates (RSE 0.488, offset –
415 0.307). (D) Fisher's α (RSE 0.368, offset 0.051). (E) Poisson log normal (PLN) model
416 estimates (RSE 1.213, offset –0.310). (F) Weibull model estimates (RSE 1.803, offset –
417 0.254). (G) Chao 1 index values (RSE 0.634, offset –0.440). (H) Coverage-based rarefaction
418 estimates based on a target coverage level (quorum) of 0.5 (RSE 0.184, offset 0.004).
419

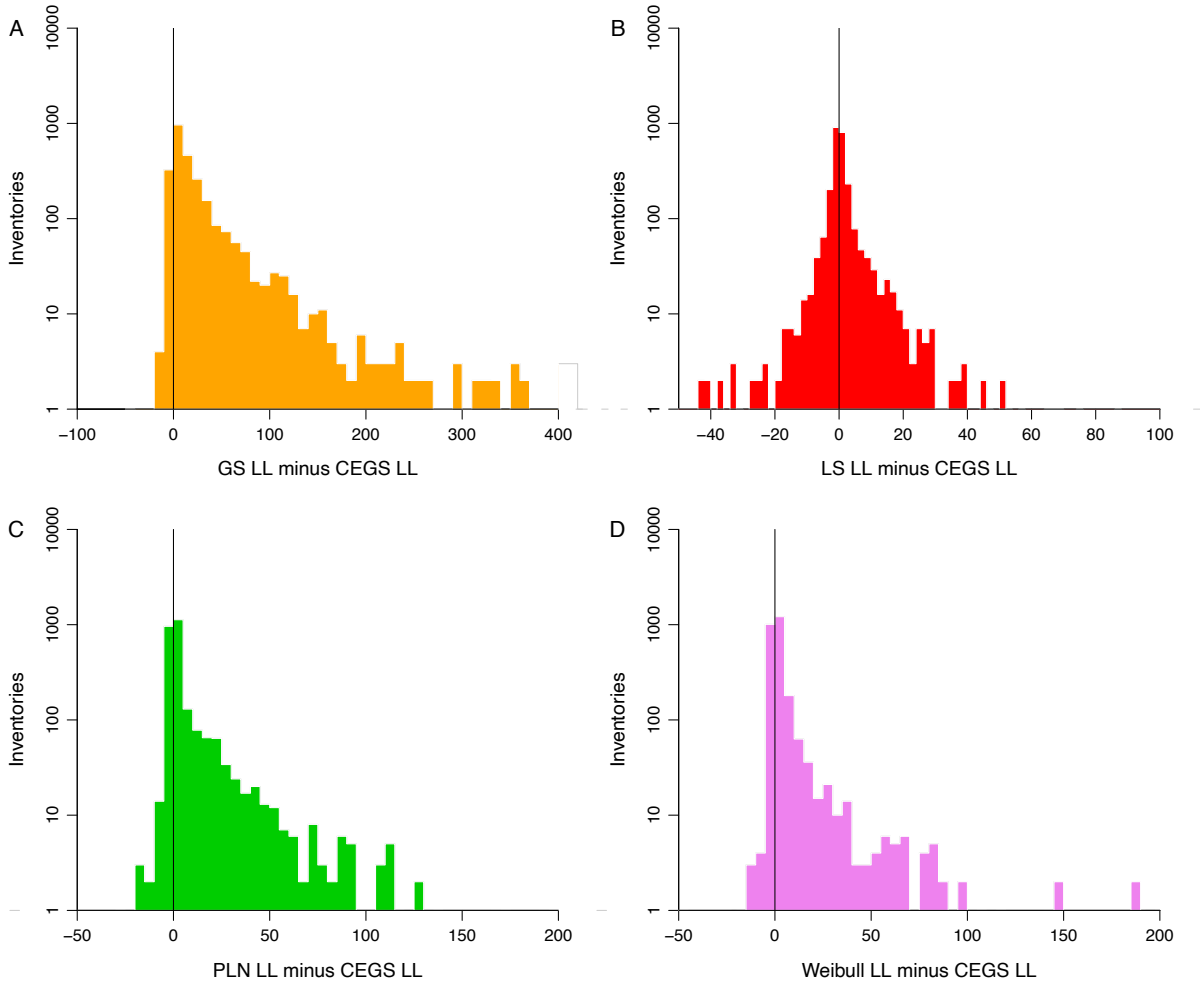


Figure 2. Differences in log likelihoods (LLs) between CEGS and four rival distributions when models are fit to each inventory and projected onto a match. For any given inventory, the match is the next most heavily sampled inventory dominated by the same ecological group and coming from the same biogeographic realm. Ranges of x-axes vary across panels; y-axes are logged to illustrate patterns in the tails of distributions. (A) CEGS LLs minus LLs yielded by the geometric series (GS). (B) CEGS LLs minus log series (LS) LLs. (C) CEGS LLs minus Poisson log normal (PLN) LLs. (D) CEGS LLs minus Weibull LLs.