Received: 26 March 2022 | Revised: 24 June 2022 | Accepted: 15 July 2022

DOI: 10.1111/jbi.14486

Journal of Biogeography WILEY

COMMENTARY

Simple null model analysis subsumes a new species co-occurrence index: A comment on Mainali et al. (2022)

Werner Ulrich[1] | Spyros Sfenthourakis[2] | Giovanni Strona[3] | Nicholas J. Gotelli[4]

[1]Department of Ecology and Biogeography, Nicolaus Copernicus University in Toruń, Toruń, Poland

[2]Department of Biological Sciences, University of Cyprus, Nicosia, Cyprus

[3]Organismal and Evolutionary Biology Research Programme, University of Helsinki, Helsinki, Finland

[4]Department of Biology, University of Vermont, Burlington, Vermont, USA

Correspondence
Werner Ulrich, Department of Ecology and Biogeography, Nicolaus Copernicus University in Toruń, Toruń, Poland.
Email: ulrichw@umk.pl

Funding information
Narodowy Centrum Nauki, Grant/Award Number: 2017/27/B/NZ8/00316; National Science Foundation, Grant/Award Number: grant 2019470

Handling Editor: Fabricio Villalobos

Recently, Mainali et al. (2022, termed MSSF henceforth) proposed a new index of pair-wise species co-occurrence based on the log odds ratio α of conditional occurrence probability. Under the assumption that the total numbers of occurrences of both species are fixed in a number of sites, they demonstrate that the associated random distribution of paired occurrences follows a non-central hypergeometric distribution, with α being the single unknown parameter. MSSF call the maximum likelihood estimator of α the 'affinity' of both species and suggest that it might serve as an index of species co-occurrence. Note that in most of the biodiversity literature, α refers to local (within-site) species richness. The affinity index α is centred around zero (random association), with positive index values indicating increased numbers of joint species occurrences. For pairs of sites that share species, it behaves similarly to Jaccard's index, with positive values indicating a relatively large number of shared species (species aggregation) and negative values indicating a relatively small number shared species (species segregation). As recognized by MSSF, an analogous probabilistic approach was proposed almost a decade ago by Veech (2013). Recently and independently, the hypergeometric distribution was introduced by Carmona and Pärtel (2020) to estimate dark diversity and by Zhou et al. (2022) in medical bibliometrics. Arita (2016) first noted that Veech's (2013) probabilistic approach is identical to Fisher's exact test for a 2×2 matrix of species co-occurrence incidences.

Importantly, the affinity index is defined only for individual pairs of species or pairs of sites. But co-occurrence data are usually organized as a matrix, with the occurrence of multiple species (=rows) recorded at multiple sites (=columns). For such matrices, indices of species co-occurrences quantify the level of β-diversity (between-site differences in species composition). It is unclear how the affinity index would be used with such a matrix, which contains many pairs of species and many pairs of sites. Taking average affinity scores across all possible pairs of species or sites potentially obscures patterns from multiple species (and site) interactions (Chao et al., 2008) and loses information on the ordered loss of species among sites (the degree of nestedness, Baselga, 2010). Like other pairwise indices (e.g. Jaccard, Sørensen and Simpson), the affinity index may be difficult to use or interpret within a biogeographic and ecological context based on multiple sites.

MSSF argue that the known sensitivity of common indices of pair-wise co-occurrence to the number of occurrences invalidates the use of these traditional indices. They also make the bold claim that 'half a century of development in analyses of co-occurrence has been marred by failures' and that this new index will 'resolve all the aforementioned challenges'. However, these sweeping claims are undermined by three problems:

1. To obtain their random expectation, MSSF assume that species have fixed numbers of occurrences among a number of sites. The odds ratio serves as a correction factor in the hypergeometric distribution for the expected number of co-occurrences and is calculated under the assumption that each species has the same chance to occur in any of the sites. The null expectation of the hypergeometric distribution, and consequently, the affinity index, is equivalent to a well-studied null model algorithm: the fixed row—equiprobable column null model (FE) of Gotelli (2000) and Wright et al. (1998). In this algorithm,

row sums (incidences per species) are fixed at observed values and column sums (=species richness per site) are allowed to vary, but are assumed to be equiprobable. We note that the same argument holds for Veech's approach (Arita, 2016).

MSSF do not discuss this equivalence and therefore overestimate the novelty of their affinity index. In Figure 1a, we show that affinity and Veech's (2013) probabilistic occurrence yield very similar results on a large set of empirical species pairs, suggesting that affinity is less innovative than claimed by MSSF. However, Veech's *p* and affinity become independent for higher values of the latter. The red data points in Figure 1 mostly stem from fully nested species pairs, where the occurrences of the less abundant species are a proper subsample of those of the more abundant species. In such situations, the value of affinity, but not of Veech's *p*, is a positive logarithmic function of the total number of sites (Figure 1a).

We further compared the performance of the affinity index with the Jaccard index, using an artificial matrix containing species pairs with differing degrees of species aggregation, and sites with modest variation in the number of species (Figure 1b). The affinity index proved to be equivalent to the standardized effect size of the traditional Jaccard metric used in combination with the FE null model (Figure 1b). Therefore, nothing is gained by using the more complicated calculations of the new affinity index. These results are also consistent with the observation that co-occurrence metrics such as the affinity index that are based on marginal occurrence totals are highly intercorrelated (Keil, 2019).

2. The affinity index and the FE null model share the assumption that all sites are equally suitable for species and that the number of occupied sites per species is fixed. This assumption may be appropriate for standardized samples based on equal sampling effort, such as species counts in quadrats of fixed area (e.g. Dornelas et al., 2019), but it is not valid when sites or samples differ greatly in their carrying capacity, resources, or size.

In the artificial matrix used here (Figure 1b), the total number of sites occupied by at least one species of a given pair varied between 27 and 49. Nevertheless, the affinity index was sensitive to even this modest variation: as the number of occupied sites increased, the average affinity index peaked at intermediate and tended to become negative at higher site-occupancy, indicating species segregation (Figure 1c). For this reason, the affinity index (and simulations based on the FE algorithm) are probably inappropriate for the analysis of island biogeography data, as in MSSF's re-analysis of Chiarucci et al. (2017). In most island systems, species richness per island and numbers of incidences per species are constrained by island size, isolation, and habitat diversity. In these circumstances, a simulation algorithm that preserves both row and column sums of the original matrix is a better choice. This familiar fixed-fixed (FF) null model (Gotelli, 2000) is often conservative (has increased type II error probabilities), but has consistently performed well in benchmark tests with matrices that have heterogeneous row and column totals (Strona et al., 2018; Ulrich & Gotelli, 2007). Moreover, Ulrich et al. (2018) showed that the FF null model is least affected by richness effects.

3. The implementation of the affinity index is not straightforward and might be vulnerable to numerical issues arising from rounding errors and memory allocation. The R code provided by MSSF returns critical error messages when used

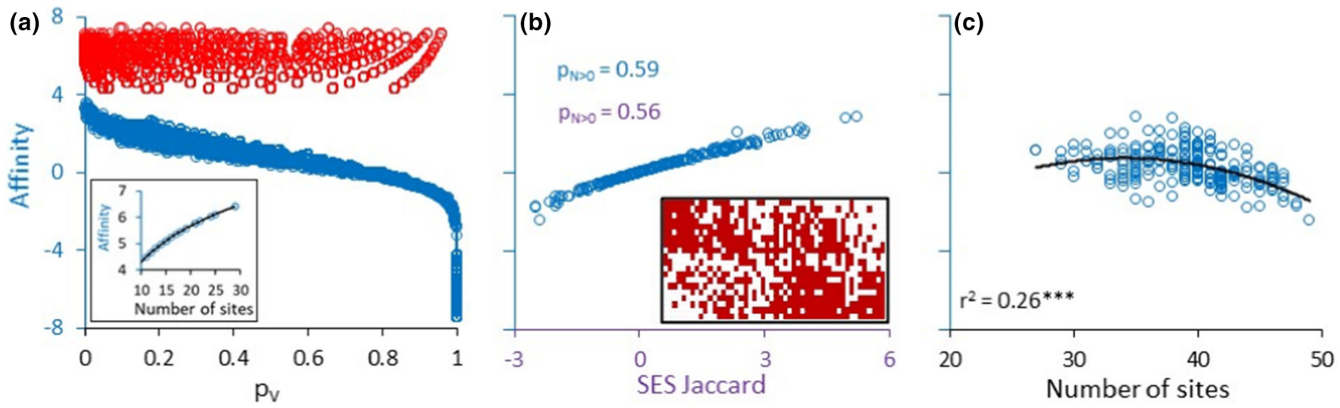

FIGURE 1 (a) Comparison of affinity and Veech probabilistic metric p_V using 47,399 pairwise species occurrence comparisons of study sites from 131 biogeographic presence—absence matrices compiled by Atmar and Patterson (Wright et al., 1998). Marked in red are 21,791 pairs, where the occurrences in one site are nested within the second site. The inlet shows that for these nested pairs affinity becomes a smooth logarithmic function of the number of study sites. (b) Comparison of affinity (blue) and the standardized effect size (SES) of the Jaccard index (violet). $p_{N>0}$ denotes the proportion of comparisons with metric values >0 indicating pairwise positive associations. The inlet shows the artificial 22 species × 50 sites presence—absence matrix used for the plots (b) and (c), where presences where placed in a way that the matrix pattern of co-occurrence was significantly compartmented (SES[C-score] = −6.50, fixed row—equiprobable column constraint null model). (c) Dependence of affinity on the total number of occupied sites. ***Significance of the quadratic regression: $p < .001$.

with many permissible parameter combinations (source codes to reproduce errors available at https://github.com/giovannistrona/co_occurrence). Since we began our re-analysis in February of 2022, the R package provided by MSSF (https://github.com/kpmainali/CooccurrenceAffinity) has already been updated twice. Notably, the three versions of the package yield substantially different results for the same data, which casts doubts on the reliability of the code and the suitability of the underlying analytical approach. In contrast, the FE algorithm is optimized and provided as a base function in many programming languages (e.g. R and Fortran), and it runs reliably on very large species × site matrices. Before its introduction into the literature, the performance of the affinity index should have been compared to established procedures and bench-marked against artificial matrices of very different size and numbers of occurrences (Gotelli & Ulrich, 2012); these quality control procedures would have easily revealed the problems with this new index and its redundancy with established procedures.

In summary, the new affinity index of MSSF does not represent an advance in the analysis of species co-occurrence. Possibly, it might only work reliably on matrices of limited species richness, it assumes sites are equiprobable with respect to occurrence probability, and its results are equivalent to a simple randomization test applied to the standard co-occurrence and β-diversity metrics. We agree with MSSF that raw co-occurrence indices are often correlated with species incidences and species richness per site. However, the solution is not another new metric, but the use of well-performing existing indices in combination with appropriate statistical standards.

Finally, we take the opportunity to caution against an uncritical use of probabilistic distributions in ecology and biogeography, as well as some randomization algorithms, to describe patterns and support inference. Patterns of species occurrences are the product of many interacting processes, both in space and time. These processes constrain the theoretically possible range of occurrences, consequently the patterns of co-occurrence and also the resampling space of null models. Ignoring these constraints in statistical inference can generate both, Type I statistical errors (the Jack Horner effect, Wilson, 1995) and Type II statistical errors (the Narcissus effect, Colwell & Winkler, 1984; Ulrich et al., 2017). In particular, we need to discern between constraints stemming from interactions of the focal species and environmental constraints acting on these species. This may require algorithms and constraints that are tailored to the particular assemblage and set of sites. Over-reliance on analytical solutions, like the hypergeometric distribution, and failure to carefully consider their assumptions and to vet them with benchmark testing will not advance the field.

ACKNOWLEDGEMENTS

WU acknowledges funding from the Polish National Science Centre (UMO-2017/27/B/NZ8/00316). NJG acknowledges funding from the US National Science Foundation (grant 2019470). No permits were needed to carry out this work.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

DATA AVAILABILITY STATEMENT

Code and data ensuring full reproducibility of the analyses presented here are freely available at https://github.com/giovannistrona/co_occurrence.

ORCID

Werner Ulrich https://orcid.org/0000-0002-8715-6619
Spyros Sfenthourakis https://orcid.org/0000-0003-3213-2502
Giovanni Strona https://orcid.org/0000-0003-2294-4013
Nicholas J. Gotelli https://orcid.org/0000-0002-5409-7456

REFERENCES

Arita, H. T. (2016). Species co-occurrence analysis: Pairwise versus matrix-level approaches. *Global Ecology & Biogeography*, *25*, 1397–1400.
Baselga, A. (2010). Partitioning the turnover and nestedness components of beta diversity. *Global Ecology and Biogeography*, *19*, 134–143.
Carmona, C. P., & Pärtel, M. (2020). Estimating probabilistic site-specific species pools and dark diversity from co-occurrence data. *Global Ecology & Biogeography*, *30*, 316–326.
Chao, A., Jost, L., Chiang, S. C., Jiang, Y.-H., & Chazdon, R. L. (2008). A two stage probabilistic approach to multiple community similarity indices. *Biometrics*, *64*, 1178–1186.
Chiarucci, A., Fattorini, S., Foggi, B., Landi, S., Lazzaro, L., Podani, J., & Simberloff, D. (2017). Plant recording across two centuries reveals dramatic changes in species diversity of a Mediterranean archipelago. *Scientific Reports*, *7*, 5415.
Colwell, R. K., & Winkler, D. W. (1984). A null model for null models in biogeography. In D. R. Strong, Jr., D. Simberloff, L. G. Abele, & A. B. Thistle (Eds.), *Ecological communities: Conceptual issues and the evidence* (pp. 344–359). Princeton.
Dornelas, M., Gotelli, N. J., Shimadzu, H., Moyes, F., Magurran, A. E., & McGill, B. J. (2019). A balance of winners and losers in the Anthropocene. *Ecology Letters*, *22*, 847–854.
Gotelli, N. J. (2000). Null model analysis of species cooccurrence patterns. *Ecology*, *81*, 2606–2621.
Gotelli, N. J., & Ulrich, W. (2012). Statistical challenges in null model analysis. *Oikos*, *121*, 171–180.
Keil, P. (2019). Z-scores unite pairwise indices of ecological similarity and association for binary data. *Ecosphere*, *10*, e02933.
Mainali, K. P., Slud, E., Singer, M. C., & Fagan, W. F. (2022). A better index for analysis of co-occurrence and similarity. *Science Advances*, *8*, eabj9204.
Strona, G., Ulrich, W., & Gotelli, N. J. (2018). Bi-dimensional null model analysis of presence-absence binary matrices. *Ecology*, *99*, 103–115.
Ulrich, W., Baselga, A., Kusumoto, B., Shiono, T., Tuomisto, H., & Kubota, Y. (2017). The tangled link between β- and γ-diversity: A narcissus effect weakens statistical inferences in null model analyses of diversity patterns. *Global Ecology & Biogeography*, *26*, 1–5.
Ulrich, W., & Gotelli, N. J. (2007). Null model analysis of species nestedness patterns. *Ecology*, *88*, 1824–1831.
Ulrich, W., Kubota, Y., Kusumoto, B., Baselga, A., Tuomisto, H., & Gotelli, N. J. (2018). Species richness correlates of raw and standardized co-occurrence metrics. *Global Ecology & Biogeography*, *27*, 395–399.

Veech, J. A. (2013). A probabilistic model for analysing species co-occurrence. *Global Ecology & Biogeography*, *22*, 252–260.

Wilson, J. B. (1995). Null models for assembly rules: The Jack Horner effect is more insidious than the narcissus effect. *Oikos*, *72*, 139–144.

Wright, D. H., Patterson, B. D., Mikkelson, G. M., Cutler, A., & Atmar, W. (1998). A comparative analysis of nested subset patterns of species composition. *Oecologia*, *113*, 1–20.

Zhou, X., Zhou, M., Huang, D., & Cui, L. (2022). A probabilistic model for co-occurrence analysis in bibliometrics. *Journal of Biomedical Informatics*, *128*, 104047.

How to cite this article: Ulrich, W., Sfenthourakis, S., Strona, G., & Gotelli, N. J. (2022). Simple null model analysis subsumes a new species co-occurrence index: A comment on Mainali et al. (2022). *Journal of Biogeography*, *49*, 2321–2324. https://doi.org/10.1111/jbi.14486