Stronger Evidence for Trait–Environment Association by Pre-processing of Abundance Tables

Cajo J. F. ter Braak[1]

[1] Mathematical & Statistical Methods group – Biometris, Wageningen University & Research, Wageningen, the Netherlands

Email: cajo.terbraak@wur.nl

ORCID: http://orcid.org/0000-0002-0414-8745

Version 1 : February, 14, 2026; CC-BY
Version 1.2: February, 24, 2026; CC-BY

Abstract

Understanding trait–environment relationships is central to predicting community responses to environmental change, yet statistical evidence for such relationships is often weak in observational datasets. Here, I introduce an N_2-processing method for abundance tables, grounded in the observation that the precision of community-weighted means (CWMs) and species niche centroids (SNCs) is proportional to Hill's effective number N_2. I refine this concept by defining informativeness, which down-weights ubiquitous species whose abundances provide little environmental information. The resulting species and site weights are implemented through iterative proportional fitting (IPF), which preserves the species–site interaction structure while aligning totals with informativeness and effective numbers.

The performance of N_2-processing was evaluated using simulations and 80 published trait–environment datasets from the CESTES global database, analyzed with fourth-corner correlation analysis (FC), RLQ, and double-constrained correspondence analysis (dc-CA). In simulations, N_2-processing substantially increased the power of the max test, especially when abundances were untransformed. In the CESTES analyses, the proportion of datasets exhibiting at least weak trait–environment association increased from 13% when using FC with multiple-testing adjustment to 68% when using dc-CA with N_2-processing. The improvement was largest when associations were weak and diminished when strong prior abundance transformations (such as the logarithmic transformation) had already down-weighted dominant species.

A square-root abundance transformation followed by N_2-processing is recommended as an effective and robust approach to strengthening inference on trait–environment relationships. The N_2-processing is essential because it removes arbitrary, data-analytic choices and ensures that dc-CA operates under conditions that match the assumptions of the model it is testing.

Key words: Hill Number, weighted averaging, community weighted means, trait-environment association, fourth-corner correlation, RLQ, double-constrained correspondence analysis (dc-CA).

1. Introduction

Understanding how biodiversity and the ecosystem services it supports will change under global pressures is a central question in ecology. Predicting shifts in species assemblages along environmental gradients is a major part of this challenge, yet the large number of species involved makes a purely species-based, data-driven approach impractical. A more efficient strategy is to describe species through their functional traits, just as environments are described by sets of environmental variables. Trait-based ecology links species traits, species abundances and environmental conditions, enabling the detection of trait–environment relationships that clarify underlying mechanisms (Garnier *et al.* 2015; Green *et al.* 2022; Westoby 2025). Trait–environment analyses integrate species abundance data, environmental variables, and species traits, each with its own structure and limitations (de Bello *et al.* 2025).

Careful pre-processing of these tables is essential for drawing reliable inferences about community responses to environmental filters (McCune 2015; Broennimann, Collart & Guisan 2026). Here the focus is on pre-processing the abundance table to strengthen evidence for trait–environment associations.

We consider pre-processing approaches tailored to the fourth-corner correlation analysis (FC) (Legendre, Galzin & Harmelin-Vivien 1997; Dray & Legendre 2008; ter Braak 2017), the three-table RLQ method (Dolédec *et al.* 1996; Dray *et al.* 2014), and double-constrained correspondence analysis (dc-CA) (ter Braak, Šmilauer & Dray 2018; ter Braak & van Rossum 2025a). The latter extends community-weighted mean (CWM) regression (Peres-Neto, Dray & ter Braak 2017; Pinho *et al.* 2024) by allowing simultaneous inclusion of multiple traits and environmental variables. Although more advanced methods exist, these approaches, especially dc-CA, provide a practical and robust framework for testing trait–environment relationships in multivariate ecological data (ter Braak & van Rossum 2025a; ter Braak 2026).

Abundance is the weight in a CWM and is often left untransformed in view of Grime's dominance hypothesis (Grime 1973), which suggests that dominant species largely determine ecosystem functioning and thus argues for using raw abundances. Nevertheless, in practice, transforming abundances to presence–absence is often considered as an alternative. This paper proposes a statistical perspective on abundance transformations aimed at strengthening statistical evidence for trait–environment association—an important goal, as such evidence is often weak in observational data (Anderegg 2023). To ensure that models generalize to unseen trait values and trait combinations, site-level CWM regression should be accompanied by species-level Species-Niche-Centroid (SNC) regression. For reliable inference, both regressions must be significant, as first elaborated for FC through the max test (ter Braak, Cormont & Dray 2012; ter Braak 2026).

Ter Braak (2019) proposed N_2-weighting of CWM and SNC regressions, with N_2 the Hill diversity number of order 2 (Hill 1973a). Hill numbers of all orders have the interpretation as effective numbers of occurrences. Here it is shown that order 2 is optimal for use in weighted averaging—a key ingredient of FC, RLQ and dc-CA—as the precision of a weighted average (CWM or SNC) is proportional to N_2. However, a

species occurring with equal abundance in all sites has N_2 equal to the number of sites, yet is not informative about the environment (Lepš & de Bello 2023). A second limitation is that Ter Braak's (2019) N_2-weighting is difficult to extend to dc-CA, as it would require separate species-level and site-level dc-CA analyses.

Both issues are resolved here. First, the effective number N_2 is replaced by a measure of informativeness, $N_2 (1- N_2 /n)$, with n the number of sites, so that a species uniformly distributed across sites is no longer informative. Second, the weighting scheme is reformulated as a transformation of the abundance table that preserves species–site interactions while modifying species and site totals to reflect informativeness and the effective number of informative species. These totals then act as optimal weights in FC, RLQ, and dc-CA. Technically, this is achieved via iterative proportional fitting (IPF).

The effectiveness of this approach was demonstrated in preliminary simulations in ter Braak (2026). Here the simulations are extended, the method is fully documented, and evaluated across 80 published trait–environment datasets from the CESTES database (Jeliazkov *et al.* 2020), by comparing FC, RLQ, and dc-CA with and without N_2-processing and with and without prior abundance transformations. Pre-processing is considered beneficial when it increases statistical power in simulations and tends to strengthen evidence for trait–environment associations (i.e., lowers test P-values) across the CESTES datasets.

The paper is structured as follows. Section 2 describes the method and its evaluation, Section 3 presents the results, and the paper concludes with a discussion. Technical details and additional results are provided in three supplements.

2. Theory and Methods

2.1 N_2-processing

Correspondence analysis (CA) originated in ecology under the name *reciprocal averaging* (Hill 1973b), because species scores in CA are weighted averages of latent site scores, and site scores are in turn weighted averages of species scores. Weighted averages are now most familiar to ecologists through the community-weighted mean trait (CWM). The CWM of a site is the abundance-weighted average of a trait (or trait composite) across species. Conversely, the species niche centroid (SNC) is the abundance-weighted average of an environmental variable (or gradient) across sites for a given species (Peres-Neto et al. 2017). Fourth-corner correlation analysis (FC) can be seen as a combination of CWM- and SNC-regression analysis (Peres-Neto et al. 2017; ter Braak, Peres-Neto & Dray 2018).

Under a null model where abundance is unrelated to the variable being averaged, the precision (inverse variance) of a weighted average is proportional to N_2, the Hill diversity of order 2 (Hill 1973a) (Supplement A). N_2 represents the *effective number* of occurrences. For example, if abundances are 100, 1, and 1, the number of occurrences is three, but the effective number $N_2 \approx 1.04$ because one weight dominates. Although the weighted average is computed over three values, only one effectively contributes to its precision. At the other extreme, a species occurring almost everywhere carries little environmental information (Lepš & de Bello 2023). If a species occurs equally in all sites, its weighted average equals the mean environment, making it uninformative. Thus, informativeness increases with effective

occurrence when rare but decreases again when the species becomes ubiquitous. Under the null model, species informativeness is:

Informativeness= $N_2(1 - N_2/N)$.

where N is the total number of sites (Supplement A).

In CA and its constrained forms—CCA and dc-CA (including FC)—species and site totals act as implicit weights. A simple and effective pre-processing approach is to divide species abundances by their total abundance and multiply by $N_2(1 - N_2/N)$; so species totals reflect informativeness.

This rationale also applies to sites: CWMs from species-poor sites are less precise than those from richer sites, suggesting weighting by the effective number of informative species. Down-weighting extremely species-rich sites (N_2 close to the number of species) was considered but found to offer little advantage in the data sets and simulations presented here. Dividing site abundances by site totals and multiplying by N_2 adjusts site weights, but then species totals no longer reflect informativeness. This catch-22 is resolved by iterative proportional fitting (IPF) (Fienberg 1970; Idel 2016): repeatedly applying both adjustments until species and site totals stabilize (Supplement B). Convergence is typically fast, yielding species and site totals approximately proportional to informativeness and effective numbers while preserving the interaction structure of the table. Informativeness and effective numbers are recomputed at each iteration.

This N_2-processing procedure is implemented in R (R Core Team 2025) as the function *ipf2N2* in the package *douconca* (ter Braak & van Rossum 2025b).

2.2 Evaluation

The effectiveness of N_2-processing was evaluated using fourth-corner correlation analysis, RLQ, and dc-CA on both simulated datasets (ter Braak 2019; ter Braak 2026) and on 80 CESTES datasets (Jeliazkov et al. 2020), summarized in Supplement C. Statistical evidence for trait–environment association was assessed through *P*-values (Muff *et al.* 2021) obtained from the permutational max test (ter Braak et al. 2012; ter Braak 2026) with 999 random permutations. N_2-processing was applied with and without abundance transformations (*e.g.* logarithmic) to determine whether it can replace such transformations and whether it strengthens evidence for trait–environment association in datasets with less extreme abundance variation.

The simulated data sets were used to study power and Type I error rates. These were generated from generalized linear mixed models with random site-dependent trait and species-dependent environmental effects, known as GLMM3 (Niku *et al.* 2021), fitted to the Aravo and Revisit data in ter Braak (2019) and then modified by varying parameters, notably the strength of the trait–environment association(ter Braak 2019).

The 80 CESTES data sets were analyzed with and without pre-processing of abundance tables using FC, RLQ, and dc-CA. FC used the minimum *P*-value of all pairwise fourth-correlations between traits and environmental and was assessed with and without Bonferroni adjustment for multiple testing (FCadj and FC_un). Differences between pre-processing and analysis methods were statistically evaluated using linear regression of evidence, expressed as –log(*P*), on the factors *method of analysis*, *N_2-processing method* (yes, no) and *type of abundance transformation*, with

covariate factor *CESTES data set*. Method of analysis consisted of four levels (FC_un, FCadj, RLQ and dc-CA). The types of abundance transformation were also quantified as 'none' =0, 'square-root'= 1, 'log' = 2 and 'P/A' =4, to more concisely describe the interaction between *N_2-processing method* (yes, no) and *type of abundance transformation*.

The logarithmic transformation of abundance was $\log(y/s+1)$, with y abundance and s the smallest non-zero value, usually $s = 1$ for count data. Non-negative right-skewed traits and environmental variables were log-transformed. When the number of traits or environmental variables was too high relative to species or sites (fewer than 10 residual degrees of freedom), the number of variables used in dc-CA was reduced by selecting variables via CCA. This selection avoided collinearity and discarded variables least related to the pre-processed abundance table.

In this paper, FC and RLQ were fitted to data using the R packages *ade4* v1.7-23 (Thioulouse *et al.* 2018) and dc-CA using *douconca* v1.2.5 (ter Braak & van Rossum 2025a). In RLQ, the Hill-Smith option was used to deal with categorical trait and environmental variables; in dc-CA, abundances were *not* divided by the site totals.

3. Results

N_2-processing of the abundance table increased the power to detect trait–environment associations in both Aravo- and Revisit-simulations (Fig. 1). The increase was largest without any prior transformation and smallest for prior transformation to presence-absence data. With N_2-processing, all prior transformations led to a similar power. Without N_2-processing, there was slight inflation of the Type I error rate when the variance of the site-dependent trait effect was increased relative to its fitted value (Fig. 1). The simulations contained a single trait and single environmental variable only, and therefore apply to both dc-CA and RLQ, as well as fourth-corner correlation analysis.

The median number of IPF iterations was 76, the maximum 8015. Only 0.3% of the data sets and their transformations required more than 1000 iterations.

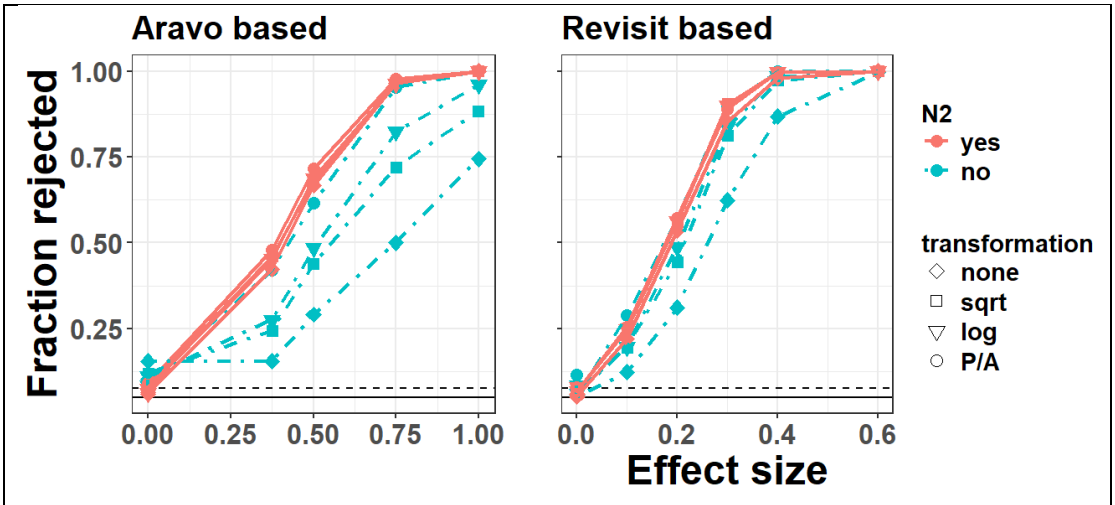


FIG. 1. Rejection rates of pre-processing methods against the size of the trait-environment association (Effect size), based on the simulations of GLMM3 models for the Aravo and Revisit data in ter Braak (2019). Pre-procession methods are transformation of abundance (none, sqrt: square-root, log: logarithmic and P/A: transformation to presence/absence) with and without subsequent N_2-processing

(N_2). Additional simulations with increased trait variance were carried out, as in ter Braak (2019), for effect size zero, and the lines start at this scenario so as to emphasize potential Type I error rate inflation. The horizontal solid line is at the nominal significance threshold; rates above the dashed line (at 0.078) are significantly greater than 0.05.

Evidence for trait-environment association in the 80 CESTES data sets (Fig. 2a) depended more on the analysis method than on the method of pre-processing (abundance transformation with and without subsequent N_2-processing) and whether FC was carried out with or without adjustment for multiple testing. Without adjustment, all data sets had at least one pair of trait and environmental variable that showed at least weak evidence ($P < 0.10$) for trait-environment association, irrespective of the type of pre-processing, as the grey lines in all panels of Figure 2a reach 1 at the vertical line for the 10% significance level. There was one exception, a data set that showed weak significance with N_2-processing, but no evidence without pre-processing. However, adjustment for multiple testing is called for as the data sets typically contained many traits and environmental variables. The adjusted minimum P-value was below 0.10 only in 16-22% of the 80 data sets, as the yellow lines cross the vertical line for the 10% significance level at this range of values.

Without N_2-processing, the percentage of data sets showing at least weak trait-environment association ($P < 0.10$) was much higher using RLQ or dc-CA: between 38 and 51% for RLQ and 50-60 for dc-CA. With N_2-processing these percentages further increased to 52-60% for RLQ and 57-66% for dc-CA, with prior square-root and log transformation yielding the highest percentages at the 5% and 1% significance levels. The benefit of N_2-processing was most pronounced for untransformed data and disappeared once the data were converted to presence–absence (Fig. 2).

The percentage of data sets showing strong trait-environment association ($P < 0.01$) was between 18 and 29% for RLQ and 21-35% for dc-CA without N_2-processing. With N_2-processing, these percentages were similar for both RLQ and dc-CA, except for 'no prior transformation' where N_2-processing increased the percentage data sets with strong evidence in RLQ from 18 to 29% and in dc-CA from 21% to 24% (Fig. 2a).

Within-data-set comparisons using linear regression indicated that the main difference in $-\log(P)$ was between no processing and any pre-processing method; differences among the pre-processing methods themselves were minimal, in RLQ, dc-CA and fourth-corner correlation. Overall, dc-CA showed stronger evidence than RLQ ($P < 0.01$), while fourth-corner analysis with Bonferroni adjustment was much worse than both dc-CA and RLQ. There was no statistical evidence for interaction between *method of analysis* and *method of pre-processing* ($P = 0.68$).

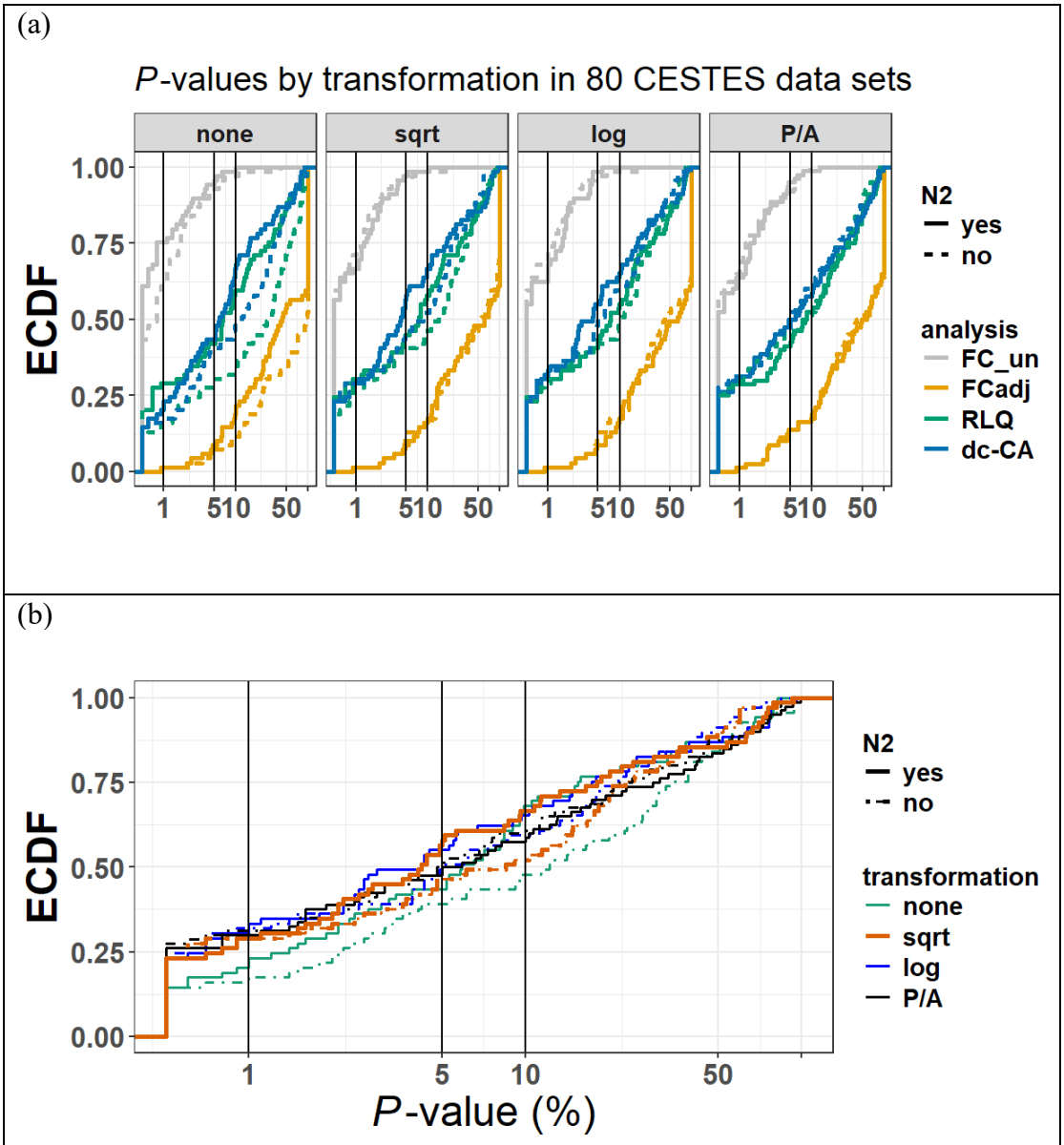


FIG. 2. Empirical cumulative density function (ECDF) of the significance level (*P*-value) for trait-environment association in the 80 data sets of the CESTES database in relation to method of pre-processing and analysis. Pre-procession methods are transformation of abundance (none, sqrt: square-root, log: logarithmic and to P/A: transformation to presence/absence) with and without subsequent N_2-processing (N_2). (a) Analysis methods are fourth corner correlation analysis with and without adjustment for multiple testing (FCadj & FC_un, respectively), RLQ and dc-CA; (b) dc-CA. The abscissa starts at $P = 0.005$ (0.5%). The vertical lines are at the 1, 5 and 10% significance levels.

240

241 Focusing on dc-CA (Fig. 2b), there was strong evidence for effects of both abundance
242 transformation on -log(*P*), with the log transformation performing best and the
243 square-root transformation as close runner-up, and N_2-processing ($P < 0.001$)). A
244 significant interaction ($P < 0.01$) indicated that the effectiveness of N_2-processing
245 diminished as the strength of the abundance transformation increased from none to

square-root, to log, with minor positive or even a small adverse effect after transformation to presence–absence (Fig. 2b; Supplement C).

The median number of IPF iterations was 76, the maximum 5646 and 2% of the data sets and their transformations required more than 1000 iterations.

Supplement C presents results for N_2–variants, explores additional power transformations, and further dissects the CESTES database by type of abundance measurement, ecosystem, and taxonomic group.

4. Discussion

As shown here, the precision of both CWM and SNC is proportional to the Hill's effective number N_2, at least when N_2 is far below it maximum value. This motivates the use N_2 or its modification *informativeness*, as weights in statistical analyses based on CWMs and SNCs, such as FC, RLQ and dc-CA. However, weighting species by their informativeness alters the site totals, and thus the implied site weights. This catch-22 is resolved by iterative proportional fitting (IPF) which conserves the species-site interaction structure of the table while adjusting the relative species and site totals to match, as closely as possible, their target informativeness and effective numbers. The species and site totals of the transformed table represent informativeness and effective number of informative species, respectively. N_2-processing replaces ad-hoc preprocessing choices with a principled, model-based one.

The proposed method increased the power of the max test for detecting trait–environment associations in simulations, and strengthened statistical evidence for such associations in the 80 datasets of the CESTES global database of traits, environments, and abundances. Its effect diminished when extreme abundances are down-weighted, as occurs with log-transformations and, even more strongly, with the transformation to presence–absence.

A key limitation of drawing general conclusions from analyses of simulated and real datasets, as compared with theory, is that it is virtually impossible to encompass all relevant scenarios and data set properties. Having said this, based on the analyses of the simulated and real datasets, a square-root transformation followed by N_2-processing is recommended, offering high effectiveness while avoiding the arbitrary zero replacement required for log transforms. N_2-processing is essential; the square-root step can be replaced by other mild power transforms without loss of efficiency. N_2-processing is essential because it provides the correct model-based variance scaling, reduces the influence of other arbitrary data-analytic choices and maximizes the statistical power and validity of dc-CA across datasets.

Lower P-values in the analysis of the CESTES datasets are desirable, provided that genuine trait–environment associations are present in each dataset. This assumption is reasonable because every dataset in the CESTES compilation originates from a published, peer-reviewed study, implying that the underlying trait–environment relationships were considered scientifically credible by the original authors. Indeed, each dataset contained at least one trait–environment pair showing weak statistical evidence ($P < 0.10$) for association.

However, without prior indication—ideally via pre-registration—of the primary trait–environment pair, multiple-testing correction is required. After such adjustment, only 13% of the datasets retained weak to strong evidence for trait–environment association. Given the typically large number of traits and environmental variables, a

multivariate approach is more appropriate than conducting many univariate tests. Methods such as RLQ and dc-CA are specifically designed to address this issue. When dc-CA was combined with N_2-processing, the proportion of datasets showing at least weak evidence for trait–environment association increased markedly, from 13% using FC to 68% using dc-CA with N_2-processing, demonstrating the substantial gain in sensitivity provided by the proposed method.

In practice, species abundance is often more strongly related to environmental variables than to traits. Sampling designs typically represent one or two dominant environmental gradients, while species occurrence within the chosen taxonomic group cannot be controlled (Gobbi *et al.* 2022). Hence, CWM regression along these gradients often shows stronger evidence for trait–environment associations than SNC regression. Note that a species-level analysis such as SNC-regression is missing in many legacy trait-environment analyses.

Whereas a presence–absence table records occurrences, the entries in the N_2-transformed table can be interpreted as *effective occurrences*. In contrast to presence–absence data, the transformed table retains quantitative information while down-weighting extreme abundances. Predicted effective occurrences can therefore be viewed as analogous to presence probabilities. An advantage of N_2-processing over conversion to presence–absence is that it preserves quantitative information while down-weighting common, uninformative species.

The improvement in statistical evidence for trait–environment association is greatest when the underlying associations are weak—arguably the most common situation in ecological datasets. In such cases, strengthening the analysis through dc-CA combined with N_2-processing becomes particularly important, as it can reveal subtle relationships that might otherwise remain undetected.

Supplement A: N_2 and a species' informativeness

Supplement B: Iterative proportional fitting (IPF)

Supplement C: Real and simulated data sets with additional results

Acknowledgements

I thank xxxxx for their valuable comments that helped to improve the manuscript. By prompting Microsoft Copilot (Version January 6, 2026, https://www.microsoft.com/copilot) with 'Improve: [my English text]', I iteratively refined the language and style of sentences and paragraphs in this paper. All views, insights, and errors are mine.

Data availability

No new data have been generated. The R code is available at

https://doi.org/10.6084/m9.figshare.31398048

CRediT authorship contribution statement

CB is the only author: Writing – review and editing, Writing – original draft, Visualization, Validation, Supervision, Software, Methodology, Formal analysis, Conceptualization.

Competing interests and Funding

The author declares no conflicts of interest. No funds, grants, or other support was received.

5. References

Anderegg, L.D.L. (2023) Why can't we predict traits from the environment? *New Phytologist,* **237,** 1998-2004. https://doi.org/10.1111/nph.18586

Broennimann, O., Collart, F. & Guisan, A. (2026) The ecospat R Package: A Collection of Pre-, Core-, and Post-Modeling Tools to Investigate Species Niches and Distributions. *R Coding for Ecology* (ed. D. Rocchini), pp. 67-98. Springer Nature Switzerland, Cham. https://doi.org/10.1007/978-3-031-99665-8_3

de Bello, F., Fischer, F.M., Puy, J., Shipley, B., Verdú, M., Götzenberger, L., Lavorel, S., Moretti, M., Wright, I.J., Berg, M.P., Carmona, C.P., Cornelissen, J.H.C., Dias, A.T.C., Gibb, H., Lepš, J., Madin, J.S., Majeková, M., Pausas, J.G., Segrestin, J., Sobral, M., Zanne, A.E. & Garnier, E. (2025) Raunkiæran shortfalls: Challenges and perspectives in trait-based ecology. *Ecological Monographs,* **95,** e70018. https://esajournals.onlinelibrary.wiley.com/doi/abs/10.1002/ecm.70018

Dolédec, S., Chessel, D., ter Braak, C.J.F. & Champely, S. (1996) Matching species traits to environmental variables: a new three-table ordination method. *Environmental and Ecological Statistics,* **3,** 143-166. http://dx.doi.org/10.1007/BF02427859

Dray, S., Choler, P., Dolédec, S., Peres-Neto, P.R., Thuiller, W., Pavoine, S. & ter Braak, C.J.F. (2014) Combining the fourth-corner and the RLQ methods for assessing trait responses to environmental variation. *Ecology,* **95,** 14-21. http://dx.doi.org/10.1890/13-0196.1

Dray, S. & Legendre, P. (2008) Testing the species traits environment relationships: The fourth-corner problem revisited. *Ecology,* **89,** 3400-3412. https://doi.org/10.1890/08-0349.1

Fienberg, S.E. (1970) An Iterative Procedure for Estimation in Contingency Tables. *The Annals of Mathematical Statistics,* **41.3,** 907–917. https:://doi.org/10.1214/aoms/1177696968

Garnier, E., Navas, M.-L., Grigulis, K., Garnier, E., Navas, M.-L. & Grigulis, K. (2015) Trait-based ecology: definitions, methods, and a conceptual framework. *Plant Functional Diversity: Organism traits, community structure, and ecosystem properties*, pp. 0. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780198757368.003.0002

Gobbi, M., Corlatti, L., Caccianiga, M., ter Braak, C.J.F. & Pedrotti, L. (2022) Hay meadows' overriding effect shapes ground beetle functional diversity in mountainous landscapes. *Ecosphere,* **e4193**. https://doi.org/10.1002/ecs2.4193

Green, S.J., Brookson, C.B., Hardy, N.A. & Crowder, L.B. (2022) Trait-based approaches to global change ecology: moving from description to prediction. *Proceedings of the Royal Society B: Biological Sciences,* **289,** 20220071. https://doi.org/10.1098/rspb.2022.0071

Grime, J.P. (1973) Competitive Exclusion in Herbaceous Vegetation. *Nature,* **242,** 344-347. https://doi.org/10.1038/242344a0

Hill, M.O. (1973a) Diversity and evenness: a unifying notation and its consequences. *Ecology,* **54,** 427-432. https://doi.org/10.2307/1934352

Hill, M.O. (1973b) Reciprocal averaging: an eigenvector method of ordination. *Journal of Ecology,* **61,** 237-249. https://doi.org/10.2307/2258931

Idel, M. (2016) A review of matrix scaling and Sinkhorn's normal form for matrices and positive maps. https ://arxiv.org/pdf/1609.06349.pdf

Jeliazkov, A., Mijatovic, D., Chantepie, S., Andrew, N., Arlettaz, R., Barbaro, L., Barsoum, N., Bartonova, A., Belskaya, E., Bonada, N., Brind'Amour, A., Carvalho, R., Castro, H., Chmura, D., Choler, P., Chong-Seng, K., Cleary, D., Cormont, A., Cornwell, W., de Campos, R., de Voogd, N., Doledec, S., Drew, J., Dziock, F., Eallonardo, A., Edgar, M.J., Farneda, F., Hernandez, D.F., Frenette-Dussault, C., Fried, G., Gallardo, B., Gibb, H., Gonçalves-Souza, T., Higuti, J., Humbert, J.-Y., Krasnov, B.R., Saux, E.L., Lindo, Z., Lopez-Baucells, A., Lowe, E., Marteinsdottir, B., Martens, K., Meffert, P., Mellado-Díaz, A., Menz, M.H.M., Meyer, C.F.J., Miranda, J.R., Mouillot, D., Ossola, A., Pakeman, R., Pavoine, S., Pekin, B., Pino, J., Pocheville, A., Pomati, F., Poschlod, P., Prentice, H.C., Purschke, O., Raevel, V., Reitalu, T., Renema, W., Ribera, I., Robinson, N., Robroek, B., Rocha, R., Shieh, S.-H., Spake, R., Staniaszek-Kik, M., Stanko, M., Tejerina-Garro, F.L., Braak, C.t., Urban, M.C., Klink, R.v., Villéger, S., Wegman, R., Westgate, M.J., Wolff, J., Żarnowiec, J., Zolotarev, M. & Chase, J.M. (2020) A global database for metacommunity ecology, integrating species, traits, environment and space. *Scientific Data,* **7,** 6. https://doi.org/10.1038/s41597-019-0344-7

Legendre, P., Galzin, R.G. & Harmelin-Vivien, M.L. (1997) Relating behavior to habitat: Solutions to the fourth-corner problem. *Ecology,* **78,** 547-562. https://doi.org/10.2307/2266029

Lepš, J. & de Bello, F. (2023) Differences in trait–environment relationships: Implications for community weighted means tests. *Journal of Ecology,* **111,** 2328-2341. https://doi.org/10.1111/1365-2745.14172

McCune, B. (2015) The front door to the fourth corner: variations on the sample unit × trait matrix in community ecology. *Community Ecology,* **16,** 267-271. https://doi.org/10.1556/168.2015.16.2.14

Muff, S., Nilsen, E.B., O'Hara, R.B. & Nater, C.R. (2021) Rewriting results sections in the language of evidence. *Trends in Ecology & Evolution*. https://doi.org/10.1016/j.tree.2021.10.009

Niku, J., Hui, F.K.C., Taskinen, S. & Warton, D.I. (2021) Analyzing environmental-trait interactions in ecological communities with fourth-corner latent variable models. *Environmetrics,* **n/a,** e2683. https://doi.org/10.1002/env.2683

Peres-Neto, P.R., Dray, S. & ter Braak, C.J.F. (2017) Linking trait variation to the environment: critical issues with community-weighted mean correlation resolved by the fourth-corner approach. *Ecography,* **40,** 806-816. http://dx.doi.org/10.1111/ecog.02302

Pinho, B.X., Melo, F.P.L., ter Braak, C.J.F., Bauman, D., Maréchaux, I., Tabarelli, M., Benchimol, M., Arroyo-Rodriguez, V., Santos, B.A., Hawes, J.E., Berenguer, E., Ferreira, J., Silveira, J.M., Peres, C.A., Rocha-Santos, L., Souza, F.C., Gonçalves-Souza, T., Mariano-Neto, E., Faria, D. & Barlow, J. (2024) Winner–loser plant trait replacements in human-modified tropical forests. *Nature Ecology & Evolution*. https://doi.org/10.1038/s41559-024-02592-5

R Core Team (2025) *R: A language and environment for statistical computing, version 4.1*. R Foundation for Statistical Computing. www.R-project.org, Vienna, Austria

ter Braak, C.J.F. (2017) Fourth-corner correlation is a score test statistic in a log-linear trait–environment model that is useful in permutation testing. *Environmental and Ecological Statistics,* **24,** 219-242. http://dx.doi.org/10.1007/s10651-017-0368-0

ter Braak, C.J.F. (2019) New robust weighted averaging- and model-based methods for assessing trait–environment relationships. *Methods in Ecology and Evolution,* **10,** 1962-1971. https://doi.org/10.1111/2041-210X.13278

ter Braak, C.J.F. (2026) Fourth-corner latent variable models overstate confidence in trait–environment relationships and what to use instead. *Environmental and Ecological Statistics*. https://doi.org/10.1007/s10651-025-00696-0; https://edepot.wur.nl/707477.

ter Braak, C.J.F., Cormont, A. & Dray, S. (2012) Improved testing of species traits–environment relationships in the fourth-corner problem. *Ecology,* **93,** 1525-1526. https://doi.org/10.1890/12-0126.1

ter Braak, C.J.F., Peres-Neto, P.R. & Dray, S. (2018) Simple parametric tests for trait–environment association. *Journal of Vegetation Science,* **29,** 801-811. https://doi.org//10.1111/jvs.12666

ter Braak, C.J.F., Šmilauer, P. & Dray, S. (2018) Algorithms and biplots for double constrained correspondence analysis. *Environmental and Ecological Statistics,* **25,** 171-197. https://doi.org/10.1007/s10651-017-0395-x

ter Braak, C.J.F. & van Rossum, B.-J. (2025a) Linking multivariate trait variation to the environment: the advantages of double constrained correspondence analysis with the R package douconca. *Ecological Informatics,* **88,** 103143. https://doi.org/10.1016/j.ecoinf.2025.103143

ter Braak, C.J.F. & van Rossum, B.-J. (2025b) R package douconca: Double constrained correspondence analysis for multi-trait multi-environment analysis v1.2.5. https://doi.org/10.32614/CRAN.package.douconca

Thioulouse, J., Dray, S., Dufour, A.-B., Siberchicot, A., Jombart, T. & Pavoine, S. (2018) *Multivariate Analysis of Ecological Data with ade4*. Springer New York, New York, NY.978-1-4939-8850-1

Westoby, M. (2025) Trait-based ecology, trait-free ecology, and in between. *The New Phytologist,* **245,** 33-39. http://dx.doi.org/10.1111/nph.20197

6. Supplement A: N_2 and a species' informativeness

This supplement shows the importance of Hill number of order 2, N_2, in weighted averaging of traits (CWM) and environmental variables (SNC). Here it shown that the variance of an SNC or CWM under a permutational null model is inversely proportional to N_2, at least when N_2 is small compared to the number of units (number of species for CWM or number of sites for SNC). N_2 can thus be interpreted the effective number of occurrences or effective number of species in a CWM- and SNC-analysis, respectively A measure of informativeness of species is also defined to cover the case of when its effective number of occurrences, its N_2, is not small, *e.g.* when a species occurs equally in nearly all sites.

The derivation in section 1.1 is also of some interest for winner-loser analysis (Tabarelli, Peres & Melo 2012; Chisté *et al.* 2016; Filgueiras *et al.* 2021; García Criado *et al.* 2023) as it derives the variance of an SNC under permutation of the environmental variables across sites.

6.1 Mean and variance of SNC and CWM under a permutational null model

Being weighted averages, the SNC and the CWM can be expressed as a linear combination $z = \sum_i p_i x_i$, provided p_i is an abundance fraction such that $\sum_i p_i = 1$. In the case of an SNC, x_i is the value of a particular environmental variable or environmental gradient x at site i, and p_i is the abundance of a particular species at site i divided by the total abundance of the species. In the case of a CWM, x_i is the value of a particular trait x of species i, and p_i is the abundance of species i at a particular site divided by the total abundance at the site.

The expectation of z under random permutation of the values $\{x_1, x_2, \dots, x_n\}$ is

$$E(z) = E(\textstyle\sum_i p_i x_i) = \sum_i p_i\, E(x_i) = \sum_i p_i\, \bar{x} = \bar{x} \sum_i p_i = \bar{x}, \qquad \text{(A1)}$$

where $\bar{x} = \frac{1}{n}\sum_i x_i$, the average of the x-values, as $\sum_i p_i = 1$ and $E(x_i) = \bar{x}$ under random permutation. This formula shows that the expected value of SNC under random permutation is the same for all species, irrespective of their total or number of occurrences.

The general formula for the variance of a linear combination is (Rao 1973)

$$var(z) = var(\boldsymbol{p}'\boldsymbol{x}) = \boldsymbol{p}' cov(x) \boldsymbol{p}, \qquad \text{(A2)}$$

where $\boldsymbol{p}' = (p_1, p_2, \dots, p_n)$ and $cov(x)$ is the covariance matrix of $x_1, x_2, \dots, x_n$ under random permutation. The diagonal elements are each s_x^2, the variance of x with denominator n instead of $n-1$ (intuitively, the sample $x_1, x_2, \dots, x_n$ is, in the permutations, the population of interest). The covariance between the elements x_i and x_j $(i \neq j)$ is a constant c, say, because of the symmetry of permutations. The constant can be obtained by noting that the variance of the sum of all x-values is 0 under permutation and can be expressed as

$$var(x_1 + x_2 + \cdots + x_n) = \sum_i var(x_i) + 2\sum_{i<j} cov(x_i, x_j) = n\, s_x^2 + n(n-1)c.$$

Solving for c gives $c = -s_x^2/(n-1)$, which completes the specification of $cov(x)$. Insertion in the general formula (A2), yields

$$var(z) = s_x^2 \sum_i \left(p_i^2 - \frac{1}{n-1} p_i \sum_{j \neq i} p_j \right), \quad \text{(A3)}$$

which can be simplified to

$$var(z) = \frac{s_x^2}{N_2} \left(1 - \frac{N_2 - 1}{n-1} \right), \quad \text{(A4)}$$

where $N_2 = \frac{1}{\sum_i p_i^2}$, the Hill number of order 2.

The first term in (A4) is variance divided by N_2, which resembles the usual formula for the variance of a mean: the population variance divided by the number of units drawn. N_2 can thus be interpreted as the effective number of units in a CWM- or SNC-analysis. When N_2 is small compared to n (*e.g.* sites are relatively species-poor or species are relatively rare in the data set) the second term is small. On neglecting the second term, the precision, *i.e.* the inverse of the variance, is proportional to the effective number N_2.

The variance $s_x^2 \left(1 - \frac{N_2 - 1}{n-1} \right)$ can be interpreted as the variance that is left on average after simple random sampling of N_2 distinct units without replacement from a population of n units (Cochran 1977). Consider $N_2 = 1$ and $N_2 = n$, then the formula says that the variance under permutation is s_x^2 and 0, respectively. In the first case a single x-value is randomly selected so that the variance is that of x (s_x^2), and in the second case all values of x are selected with equal weight ($p_i = 1/n$) so that the SNC or CWM is equal to $\bar{x}$ and is constant under permutation, so that the permutational variance is 0, in agreement with the formula. A species for which $N_2 = n$ is thus uninformative. On replacing n-1 by n in the term $1 - \frac{N_2 - 1}{n-1}$, the term is approximately equal to $1 - \frac{N_2}{n}$, as n is large in practice. A simple measure of a species' informativeness under the null model is therefore $N_2(1 - N_2/n)$.

6.2 Conclusion

Formula (A4) gives a definite interpretation of the Hill number N_2 as the effective number of occurrences or effective number of species for SNC and CWM, respectively. It the Hill number of order 2 (N_2) that plays a special role in weighted averaging, instead of any order, as Lepš and de Bello (2023) appear to suggest.

The simple measure of the informativeness of a species under the null model is $N_2(1 - N_2/n)$.

References

Chisté, M.N., Mody, K., Gossner, M.M., Simons, N.K., Köhler, G., Weisser, W.W. & Blüthgen, N. (2016) Losers, winners, and opportunists: How grassland land-use intensity affects orthopteran communities. *Ecosphere,* **7,** e01545. https://esajournals.onlinelibrary.wiley.com/doi/abs/10.1002/ecs2.1545

Cochran, W.G. (1977) *Sampling Techniques, 3rd ed.* John Wiley & Sons, New York. 978-0-471-16240-7

Filgueiras, B.K.C., Peres, C.A., Melo, F.P.L., Leal, I.R. & Tabarelli, M. (2021) Winner–Loser Species Replacements in Human-Modified Landscapes. *Trends in Ecology & Evolution,* **36,** 545-555. https://www.sciencedirect.com/science/article/pii/S0169534721000562

García Criado, M., Myers-Smith, I.H., Bjorkman, A.D., Normand, S., Blach-Overgaard, A., Thomas, H.J.D., Eskelinen, A., Happonen, K., Alatalo, J.M., Anadon-Rosell, A., Aubin, I., te Beest, M., Betway-May, K.R., Blok, D., Buras, A., Cerabolini, B.E.L., Christie, K., Cornelissen, J.H.C., Forbes, B.C., Frei, E.R., Grogan, P., Hermanutz, L., Hollister, R.D., Hudson, J., Iturrate-Garcia, M., Kaarlejärvi, E., Kleyer, M., Lamarque, L.J., Lembrechts, J.J., Lévesque, E., Luoto, M., Macek, P., May, J.L., Prevéy, J.S., Schaepman-Strub, G., Sheremetiev, S.N., Siegwart Collier, L., Soudzilovskaia, N.A., Trant, A., Venn, S.E. & Virkkala, A.-M. (2023) Plant traits poorly predict winner and loser shrub species in a warming tundra biome. *Nature Communications,* **14,** 3837. https://doi.org/10.1038/s41467-023-39573-4

Lepš, J. & de Bello, F. (2023) Differences in trait–environment relationships: Implications for community weighted means tests. *Journal of Ecology,* **111,** 2328-2341. https://doi.org/10.1111/1365-2745.14172

Rao, C.R. (1973) *Linear statistical inference and its application. 2nd ed.* Wiley, New York. 9780470316436

Tabarelli, M., Peres, C.A. & Melo, F.P.L. (2012) The 'few winners and many losers' paradigm revisited: Emerging prospects for tropical forest biodiversity. *Biological Conservation,* **155,** 136-140. https://www.sciencedirect.com/science/article/pii/S0006320712002893

7. Supplement B: Iterative proportional fitting (IPF)

This supplement described the N_2-processing method in detail. Section 7.1 describe the standard iterative proportional fitting (IPF) procedure. In dc-CA, RLQ and FC, the marginals of the obtained table are the implied weights. A disadvantage of the standard procedure for use in dc-CA, RLQ and FC is, that the marginals of the obtained table and thus the implied weights to not reflect the effective or informative numbers of the obtained table. In order to obtain marginals that reflect effective or informative numbers these numbers need to be recomputed (updated) in each step of the IPF procedure. This is detailed in section 0, while the details of the stopping criterion are described in section 1.3.

Notation: n_{ij} is the abundance of the j^{th} out of m species in the i^{th} out of n sites ($n_{ij} \geq 0$). The $n \times m$ abundance table thus consist of non-negative entries denoted by n_{ij} and after transformation or during or after IPF. A '+' replacing an index denotes summation over that index, e.g. y_{i+} and y_{+j} denote the total for the i^{th} row and j^{th} column, respectively.

7.1 IPF of a two-way table

The original Deming and Stephan IPF procedure (Fienberg 1970) aims to modify a two-way to specified row and column marginals with entries r_i and k_j, respectively. If $r_+ \neq k_+$, calculate $c = k_+ / r_+$ and rescale $r_i \leftarrow$ c × old r_i such that new $r_+ = k_+$.

Initialization: $y_{ij} \leftarrow n_{ij}$

Iterate until to the values stabilize:

Step 1. For each column j execute: $y_{ij} \leftarrow y_{ij}\ (k_j / y_{+j})$

Step 2. For each row i execute: $y_{ij} \leftarrow y_{ij}\ (r_i / y_{i+})$

Step 3. Goto step 1 or stop if the values of the current iteration are sufficiently close to those of the previous iteration.

This IPF procedure can be applied to an abundance table by first calculating the effective numbers $N_{2i} = \frac{1}{\sum_j p_{ij}^2}$, with $p_{ij} = \frac{y_{ij}}{y_{i+}}$ and $\widetilde{N}_{2j} = \frac{1}{\sum_i q_{ij}^2}$, with $q_{ij} = \frac{y_{ij}}{y_{+j}}$, and informative numbers $I_i = N_{2i}\ (1 - N_{2i}/m)$ and $\tilde{I}_j = \widetilde{N}_{2j}\ (1 - \widetilde{N}_{2j}/n)$

then setting the marginals to either the effective number or the informative numbers ($r_i = N_{2i}$ or $r_i = I_i$ and $k_j = \widetilde{N}_{2j}$ or $k_j = \tilde{I}_j$) with subsequent rescaling such that $r_+ = k_+$.

Executing Step 1 only with $k_j = \tilde{I}_j$ without further iteration is the processing method *iter0* of Supplement C in the section N_2-variants. The row totals obtained after Step 1 in this special case do not have an interpretation in terms of effective or informative numbers for sites.

The pre-processing method described so far can be obtained by the function *ipf2N2* in the *douconca* package (ter Braak & van Rossum 2025) by setting the argument *updateN2* to FALSE, while the method denoted *iter0* can be obtained by setting *max_iter* = 0.

Four variants of N_2-processing are:

1. N_2 (in full: N_2iN_2): default N_2-processing, *i.e.* IPF to species marginal that represents informativeness ($N_2(1- N_2/N)$) and site marginal that represents effective numbers of species, in the notation introduced in the following: N_2iN_2i;
2. N_2iN_2i: IPF to marginals that represent informativeness of species and of sites;
3. N_2N_2: IPF to marginals that represent effective numbers (N_2) of species and of sites;
4. iter0: species marginal scaled to informativeness of species without further scaling of the site marginal

For presence–absence data, the N_2N_2 variant yields no transformation because species and site totals in a P/A table coincide with their N_2-values. The default variant $N_2 \equiv N_2iN_2$ similarly produces minimal adjustment when species totals are much smaller than half the number of sites, in which case species informativeness is effectively proportional to their occurrence frequencies.

There is no proof that IPF always returns a table with the desired marginals (Idel 2016).

7.2 IPF while updating N_2 marginals

The IPF procedure of the previous section leads to marginals that are proportional to the effective or informative numbers of the initial data table, but the precision of any CWM or SNC calculated during permutation is proportional to the effective numbers of the final data table, which differ from those of the initial table. For this reason, the effective and informative numbers are recalculated after each step.

The full procedure, with sequence as in the function *ipf2N2* in the *douconca* package, is:

Initialization:

$y_{ij} \leftarrow n_{ij}$

Set, for each row, $r_i = N_{2i}$ or $r_i = I_i$ and, for each column $k_j = \tilde{N}_{2j}$ or $k_j = \tilde{I}_j$ and for subsequent rescaling calculate $c = k_+ / r_+$.

Iterate:

Step 1. For each row i execute: $y_{ij} \leftarrow y_{ij}\ (c\ r_i / y_{i+})$

Step 2: For each column update $\tilde{N}_{2j}$ the effective number of the column of the current table $\{y_{ij}\}$ and set $k_j \leftarrow$ new $\tilde{N}_{2j}$ or new $\tilde{I}_j = \tilde{N}_{2j}\ (1 - \tilde{N}_{2j}/n)$.

Step 3. For each column j execute: $y_{ij} \leftarrow y_{ij}\ (k_j / y_{+j})$

Step 4: For each row update N_{2i} the effective number of the row of the current table $\{y_{ij}\}$ and set $r_i \leftarrow$ new N_{2i} or new $I_i = N_{2i}\ (1 - N_{2i}/m)$.

Step 5: Calculate $c \leftarrow k_+ / r_+$.

Step 6: Stop if criterion is met (section 1.3), otherwise continue with Step 1.

Step 7: Goto step 1.

7.3 Stopping criterion

This section describes the detail of the stopping criterion used in Step 6 of the full IPF procedure of the previous section.

At step 6, each column has the intended sum (because of step 3) , but the row sum (y_{i+}) may differ from the intended value ($c\ r_i$). What counts in dc-CA are the ratios y_{i+}/y_{++}. The criterion is therefore based on the ratios of the row sums and the intended values. As full equality may be impossible to reach and the small row sums have little influence in dc-CA, the criterion is based on the range of the ratios $\Delta_i = (\ c\ r_i + 1\ ) / (y_{i+} + 1)$. The criterion value is the maximum of $\max(\Delta_i)$ and $1/\min(\Delta_i)$. The iteration process stops when the criterion value does not improve since the previous iteration or when the maximum number of iterations has been reached, with a default value of 10000.

7.4 R code of function IPF2N2 in douconca v1.2.5

The R code of function `ipf2N2` implementing the schematic code of sections 1-3. But first the R code for two helper functions, `fN2` for calculating the Hill number N2 and `fN2N_N2` for calculating the effective number and informativeness (if `N2N_N2 =`

TRUE). The functions cater for the limiting cases. An example: When the informativeness of a species is 0, while that species is the only one in a particular site. In this particular case the N_2 of the site is set to a small number, effectively 0. The species and site are not removed as species and sites with very small marginal totals do not influence dc-CA.

```
fN2 <- function(x ) {
  sx<- sum(x)
  sx <- if(is.na(sx) || sx < 1.0e-6) {
    return(.Machine$double.eps)} else {sx}
  x <- x / sx
  return(1 / sum(x * x))
}
```

```
fN2N_N2 <- function(Y,
                    margin,
                    N2N_N2 = TRUE) {
  N2 <- apply(X = Y, MARGIN = margin, FUN = fN2)
  margin1 <- if (margin == 1) 2 else 1
  if (N2N_N2) N2 <- N2 * pmax(1 - N2 / dim(Y)[margin1], .Machine$double.eps
)
  return(N2)
}
```

```
ipf2N2 <- function(Y,
                   max_iter = 10000,
                   updateN2 = TRUE,
                   N2N_N2_species = TRUE,
                   N2N_N2_sites = FALSE
                   ) {

  Y <- as.matrix(Y)
  rownams <- rownames(Y)
  colnams <- colnames(Y)
  R <- rowSums(Y)
  K <- colSums(Y)

  if (any(R == 0)) {
    warning("Some sites do not have species.\n ",
      paste(names(R)[which(R == 0)],  collapse = " "), "\n")
  }
  if (any(K == 0)) {
    warning("Some species are absent in every site.\n",
      paste(names(K)[which(K == 0)],  collapse = " "), "\n")
  }
  K[K < .Machine$double.eps] <- .Machine$double.eps
  R[R< .Machine$double.eps] <- .Machine$double.eps
  N2spp <- N2spp0 <- fN2N_N2(Y, 2, N2N_N2 = N2N_N2_species)
  N2sites <- N2sites0 <- fN2N_N2(Y, 1, N2N_N2 = N2N_N2_sites)
  N <- nrow(Y)
  crit0  <-  1.0e10
  crit1 <- crit0 - 1
  iter <- 0
  ratio <- 1
```

```
  if(max_iter == 0){
    Y <- Y %*% diag(N2spp / K)
  } else {
  while (crit1 < crit0 && iter < max_iter) {
    iter <- iter + 1
    crit0 <- crit1
    Y0 <- Y
    Y <- Y * ((N2sites / R) * ratio)
    K <- colSums(Y)
    if (updateN2) N2spp <- fN2N_N2(Y, 2, N2N_N2 = N2N_N2_species)
    Y <- Y %*% diag(N2spp / K)
    R <- rowSums(Y)
    if (updateN2) N2sites <- fN2N_N2(Y, 1, N2N_N2 = N2N_N2_sites)
    ratio <- sum(N2spp)/sum(N2sites)
    mm <- range((ratio * N2sites+1)/(R+1))
    #mm <- range(c(mm,(N2spp+1)/(K+1)))
    crit1 <- max(c(1/mm[1],mm[2]))
  }
    Y <- Y0
    if (updateN2) {
      N2sites <- fN2N_N2(Y, 1, N2N_N2 = N2N_N2_sites)
      N2spp <- fN2N_N2(Y, 2, N2N_N2 = N2N_N2_species)
    }

  if (iter == max_iter && max_iter > 0) {
    warning(paste0("No convergence in ", max_iter, " iterations."),"\n")
  }
  if (any(Y<0)){
    warning("some values in preprocessed Y negative")
    Y[Y < 0] <- 0 # tiny non-negative values should not occur
  }
  }
  R <- rowSums(Y)
  attr(Y, which = "N2species_original") <- N2spp0
  attr(Y, which = "N2sites_original") <- N2sites0
  attr(Y, which = "N2species") <- N2spp
  attr(Y, which = "N2sites") <- N2sites
  attr(Y, which = "R/N2") <- (R / N2sites) / (sum(R) / sum(N2sites))
  attr(Y, which = "iter") <- iter
  attr(Y, which = "crit") <- crit0
  colnames(Y) <- colnams
  rownames(Y) <- rownams
  if (fN2(R) / length(R) < 0.5) {
    message("Warning: unbalanced site totals in return value:",
            "N2 of row sums less than halve the number of rows.\n")
  }
  if (any(N2sites <= .Machine$double.eps)) {
    warning("After processing: Some sites do not have species.\n ",
            paste(rownames(Y)[which(N2sites <= .Machine$double.eps)],
                  collapse = " "), "\n")
  }
  if (any(N2spp <= .Machine$double.eps)) {
    warning("After processing: Some species are absent in every site.\n ",
            paste(colnames(Y)[which(N2spp <= .Machine$double.eps)],
                  collapse = " "), "\n")
  }
  return(Y)
}
```

7.5 Help text

ipf2N2 {douconca}

Iterative proportional fitting of an abundance table to Hill-N2 marginals

Description

Function for pre-processing/transforming an abundance table by iterative proportional fitting, so that the transformed table has marginals proportional to `N2` or `N2(1-N2/N)` with `N` the number of elements in the margin. Hill-N2 is the effective number of species. It is of intrinsic interest in weighted averaging (CWM and SNC) as their variance is approximately inversely proportional to N2 (ter Braak 2019), and therefore of interest in `dc_CA`.

Usage

```
ipf2N2(
  Y,
  max_iter = 10000,
  updateN2 = TRUE,
  N2N_N2_species = TRUE,
  N2N_N2_sites = FALSE
)
```

Arguments

`Y`	abundance table (matrix or dataframe-like), ideally, with names for rows and columns.
`max_iter`	maximum number of iterative proportional fitting (ipf) iterations. If `max_iter = 0`, the columns are divided by their effective number or informativeness (`N2` or `N2(1-N2/N)`, depending on the setting of `N2N_N2_species`) without further pre-processing and the row sums are then, with `N2N_N2_species = TRUE`, sums of informativeness instead of effective number of informative species.
`updateN2`	logical, default `TRUE`. If `FALSE` the marginal sums are proportional to the N2-marginals of the initial table, but the N2-marginals of the returned matrix may not be equal to their marginal sum. If `updateN2 = TRUE` and `N2N_N2_species = TRUE` (the default), the column marginals are `N2(N-N2)/N` with `N` the number of sites. The row sums are then proportional to, what we term, the effective number of informative species. If `N2N_N2_species = FALSE`, the returned transformed table has N2 columns marginals, *i.e.* `colSums(Y2) = N2species(Y2)` with `Y2` the return value of `ipf2N2`. If converged, N2 row marginals are equal to the row sums, *i.e.* `rowSums(Y2) = approx. const*N2sites(Y2)` and `const` a constant.
`N2N_N2_species`	Set species marginal to the value of `N2(1-N2/N)` for each species. Default `TRUE`. If `FALSE`, the marginal is set to the `N2` value of each species.
`N2N_N2_sites`	Default `FALSE` sets the marginal proportional to the `N2` value of each site. If `TRUE`, the marginal is set to `N2(1-N2/m)`, with `m` the number of species.

Details

Applying `ipf2N2` with `N2N_N2_species=FALSE` to an presence-absence data table returns the same table. However, a species that occurs everywhere (or in most of the sites) is not very informative. This is acknowledged with the default option `N2N_N2_species=TRUE`. Then, with `N2N_N2_species=TRUE`, species that occur in more than halve the number of sites are down-weighted, so that the row sum is no longer equal to the richness of the site (the number of species), but proportional to the number of informative species. The returned matrix has the intended species marginal (column sums), by construction of the algorithm, even without convergence. On convergence, it has the intended site marginal (row sums).

Value

a matrix of the same order as the input `Y`, obtained after ipf to N2-marginals.

References

ter Braak, C.J.F. (2019). New robust weighted averaging- and model-based methods for assessing trait-environment relationships. Methods in Ecology and Evolution, 10 (11), 1962-1971. doi:10.1111/2041-210X.13278

ter Braak, C.J.F. (2026). Fourth-corner latent variable models overstate confidence in trait–environment relationships and what to use instead Environmental and Ecological Statistics. doi:10.1007/s10651-025-00696-0

The arguments `N2N_N2_species` and `N2N_N2_sites` determine the four variants of N_2-processing are follows:

1. N_2 (in full: N_2iN_2): default N_2-processing, *i.e.* IPF to species marginal that represents informativeness ($N_2(1- N_2/N)$) and site marginal that represents effective numbers of species, in the notation introduced in the following: N_2iN_2i; *i.e.* `N2N_N2_species = TRUE` and `N2N_N2_sites = FALSE`.
2. N_2iN_2i.: IPF to marginals that represent informativeness of species and of sites, *i.e.* `N2N_N2_species = TRUE` and `N2N_N2_sites = TRUE`.
3. N_2N_2: IPF to marginals that represent effective numbers (N_2) of species and of sites, *i.e.* `N2N_N2_species = FALSE` and `N2N_N2_sites = FALSE`.
4. iter0: species marginal scaled to informativeness of species without further scaling of the site marginal, *i.e.* `max_iter = 0`.

In the section on N_2 – variants in Supplement C, `updateN2 = TRUE`.

7.6 References

Fienberg, S.E. (1970) An Iterative Procedure for Estimation in Contingency Tables. *The Annals of Mathematical Statistics,* **41.3,** 907–917. https://doi.org/10.1214/aoms/1177696968

Idel, M. (2016) A review of matrix scaling and Sinkhorn's normal form for matrices and positive maps. https ://arxiv.org/pdf/1609.06349.pdf

ter Braak, C.J.F. & van Rossum, B.-J. (2025) R package douconca: Double constrained correspondence analysis for multi-trait multi-environment analysis v1.2.5. https://doi.org/10.32614/CRAN.package.douconca

8. Supplement C: Real and simulated data sets with additional results

A key limitation of drawing general conclusions from analyses of simulated and real datasets, as compared with theory, is that it is virtually impossible to encompass all relevant scenarios. This is reflected below in Figs. S1–S4, where the advantage of updating N_2 during IPF (Appendix B) is not evident in Fig. S1 and only debatable in Fig. S4.

8.1 Real and simulated data sets

Four sets of simulated datasets were used. The first three—Aravo, Revisit, and Beetle—were generated from generalized linear mixed models (GLMMs) fitted to their respective real datasets. The fourth set reproduces the simulations in Fig. 3 of Niku et al. (2021), using the same model specification as in Fig. 3 of ter Braak (2026). Among the four sets, only the Beetle-simulations include multiple traits and multiple environmental variables.

In all sets, the underlying model is a GLM with fixed main effects of the trait, the environmental variables, and their interaction. Each set also includes latent variables associated with the trait or the environmental variables, as in GLMM3 models (ter Braak 2026). In the fourth set, on a GLLVM2(2), the fixed effects are supplemented by an environment-related latent variable and two additional unrelated latent variables to capture residual variation beyond that implied by the response distribution. The response distributions, in order, are Poisson, beta-binomial, and negative binomial (for the third and fourth sets). The link function is log, except in the second set, which uses a logit link.

Aravo-simulations

The Aravo-simulated data sets were taken from Appendix A7 in the supporting information of ter Braak (ter Braak 2019). In that Appendix, a general linear mixed model with random species-dependent environmental effect and site-dependent trait effects (*i.e.* a GLMM3 model, which is termed MLM3 in ter Braak 2019) was fitted using the *aravo* data set in the R package ade4 (Dray & Dufour 2007) with the single trait SLA (Specific Leaf Area) and the single environmental variable Snow (mean snowmelt date in Julian day averaged over 1997-1999) and with Poissonian response distribution.. It has 82 plant species and 75 sites. The fitted model was modified to generate data with varying strength of the trait-environment relationship (250 data sets for each scenario). The model fitted to these data was:

```
library(lme4)
fpol2 <- y~ poly(trait,2) + poly(env,2) + trait:env +
                              (1+trait|site)+ (1+env|species))
MLM3 <- glmer(fpol2, data=dat, family= poisson,
                   nAGQ=0, control = glmerControl(calc.derivs=F))
```

The simulated data are generated using simulate(MLM3) after modification of the regression coefficient of the trait-environment interaction. The data are identical to those used in ter Braak (2019).

Revisit-simulations

The Revisit-simulated data sets were taken from ter Braak (ter Braak 2019), who fitted a GLMM3 model to a subset of the Whittaker Siskiyou Mountains Revisit data(Damschen, Harrison & Grace 2010) with the single trait leaf carbon-to-nitrogen ratio (C:N) and the environmental variable Topographic Moisture Gradient (TMG). Briefly, the data consists of the cover abundance of 75 species in 52 sites, calculated from the number of 100 quadrat corners per site that each species intersected. The model was fitted with a beta-binomial response distribution, which was then modified to generate data with varying strength of the trait-environment relationship (250 data sets for each scenario). The model fitted to these data was:

```
library(glmmTMB)
formula.MLM3 <- y ~ trait*env +(1+trait|site)+(1+env|species)

MLM3 <- glmmTMB(formula.MLM3, family = betabinomial, data= Revisit)
```

The data are generated using simulate(MLM3) after modification of the regression coefficient of the trait-environment interaction. The data are identical to those used in ter Braak (2019).

Beetle-simulations

The Beetle-simulations are obtained by fitting a GLMM3 model to ground beetle data from Ribera et al. (2001) which contains the numbers of individuals of 68 ground beetle species in 87 sites. The traits and environmental variables are those used by Niku et al. (2021). The four traits are total length (LTL), maximum vaulting of the pronotum(LPH), overwintering (OVE) with two classes adults only, and adults and/or larvae and breeding season (BRE) with three classes (spring, summer and autumn or winter), while the four environmental variables are the management intensity index (Management), percentage moisture content (Moist), elevation, and pH, with full definitions in the original paper (Ribera et al. 2001). The GLMM3 model was fitted to the data with negative binomial response distribution, "NB2" in *glmmTMB* v2.0.2 (Brooks *et al.* 2017), and model formula:

```
fGLMM3.H1 <- y ~ Management+ Moist + Elevation+pH +
  LTL + LPH +OVE + BRE +
  (Management+ Moist + Elevation+pH):(LTL + LPH +OVE + BRE)+
  (1 + Management+ Moist + Elevation+pH |species) +
  (1 + LTL + LPH +OVE + BRE |site)
```

and call:

```
GLMM3.fit <- glmmTMB(fGLMM3.H1, data=datalong, family= nbinom2(link =
"log"), se = TRUE)
```

Data sets with varying strength of the trait-environment relationship (250 data sets for each scenario) were generated using simulate(model) after setting the 16 regression coefficients of the trait-environment interaction to the product of the fitted coefficients and effect size (effect size = 0, 0.3, 0.45, 0.6, 0.75, 0.9).

GLLVM2(2) simulations

The set-up the simulations of the GLLVM2(2) model was proposed by Niku et al. (2021) and the actual set used here is the same as that in ter Braak (2026). It has negative binomially distributed abundance of 40 species in 70 sites. The GLLVM2(2) model is a generalized linear mixed model with random species-dependent environmental effects and two latent variables to mimic structured variation beyond the unstructured negative binomial variation. The model differs from the previous GLMM3 models by their latent variables and by *not* including random site-dependent trait effects. The effect size of the trait-environment interactions and the variance of the random species-dependent environmental effects is varied in this set.
The data are generated using simulate(model) after modification of the regression coefficient of the trait-environment interaction. The data are identical to those used in ter Braak (2026).

CESTES data set

The CESTES data base is a set of 80 curated datasets from published trait-studies (Jeliazkov *et al.* 2020; Jeliazkov & Chase 2024). CESTES stands for "metaCommunity Ecology: Species, Traits, Environment and Space". Each dataset includes four matrices: species community abundances or presences/absences across multiple sites, species trait information, environmental variables and spatial coordinates of the sampling sites. Here, the information about space is not used. With number between parentheses, the data sets are from terrestrial (55), fresh water (14) and marine ecosystems (11) all over the globe with many taxonomic group represented and sites varying in extent from small to large. Measures of abundance can be classified as number-of-individuals (45), presence-absence (11), abundance-index (9) and other (15). Table 1 provides a summary of numbers of units and variables in the CESTES data base.

Table 1. Summary of numbers of sites, species and variables in the CESTES data base (nTraitsI and nEnvI are the number of traits and environmental variables, respectively, after transforming categorical variables to indicator variables, as required in dc-CA).

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
nSpecies	6	32	53	71.6	107	208
nTraits	1	5	8	13.8	16	89
nTraitsI	1	8	13	17.4	20	89
nSites	7	27	40	70.9	88	364
nEnv	1	6	9	11.9	13	47
nEnvI	1	7	13	15.1	20	63

References

Brooks, M.E., Kristensen, K., van Benthem, K.J., Magnusson, A., Berg, C.W., Nielsen, A., Skaug, H.J., Maechler, M. & Bolker, B.M. (2017) glmmTMB balances speed and flexibility among packages for zero-inflated generalized linear mixed modeling. *The R Journal,* **9,** 378-400. https://doi.org/10.32614/RJ-2017-066

Damschen, E.I., Harrison, S. & Grace, J.B. (2010) Climate change effects on an endemic-rich edaphic flora: resurveying Robert H. Whittaker's Siskiyou sites (Oregon, USA). *Ecology,* **91,** 3609-3619. https://doi.org/10.1890/09-1057.1

Dray, S. & Dufour, A.B. (2007) The ade4 package: implementing the duality diagram for ecologists. *Journal of Statistical Software,* **22,** 1-20. https://doi.org/10.18637/jss.v022.i04

Jeliazkov, A. & Chase, J.M. (2024) When Do Traits Tell More Than Species about a Metacommunity? A Synthesis across Ecosystems and Scales. *The American Naturalist,* **203,** E1-E18. https://doi.org/10.1086/727471

Jeliazkov, A., Mijatovic, D., Chantepie, S., Andrew, N., Arlettaz, R., Barbaro, L., Barsoum, N., Bartonova, A., Belskaya, E., Bonada, N., Brind'Amour, A., Carvalho, R., Castro, H., Chmura, D., Choler, P., Chong-Seng, K., Cleary, D., Cormont, A., Cornwell, W., de Campos, R., de Voogd, N., Doledec, S., Drew, J., Dziock, F., Eallonardo, A., Edgar, M.J., Farneda, F., Hernandez, D.F., Frenette-Dussault, C., Fried, G., Gallardo, B., Gibb, H., Gonçalves-Souza, T., Higuti, J., Humbert, J.-Y., Krasnov, B.R., Saux, E.L., Lindo, Z., Lopez-Baucells, A., Lowe, E., Marteinsdottir, B., Martens, K., Meffert, P., Mellado-Díaz, A., Menz, M.H.M., Meyer, C.F.J., Miranda, J.R., Mouillot, D., Ossola, A., Pakeman, R., Pavoine, S., Pekin, B., Pino, J., Pocheville, A., Pomati, F., Poschlod, P., Prentice, H.C., Purschke, O., Raevel, V., Reitalu, T., Renema, W., Ribera, I., Robinson, N., Robroek, B., Rocha, R., Shieh, S.-H., Spake, R., Staniaszek-Kik, M., Stanko, M., Tejerina-Garro, F.L., Braak, C.t., Urban, M.C., Klink, R.v., Villéger, S., Wegman, R., Westgate, M.J., Wolff, J., Żarnowiec, J., Zolotarev, M. & Chase, J.M. (2020) A global database for metacommunity ecology, integrating species, traits, environment and space. *Scientific Data,* **7,** 6. https://doi.org/10.1038/s41597-019-0344-7

Niku, J., Hui, F.K.C., Taskinen, S. & Warton, D.I. (2021) Analyzing environmental-trait interactions in ecological communities with fourth-corner latent variable models. *Environmetrics,* **n/a,** e2683. https://doi.org/10.1002/env.2683

Ribera, I., Dolédec, S., Downie, I.S. & Foster, G.N. (2001) Effect of Land Disturbance and Stress on Species Traits of Ground Beetle Assemblages.

Ecology, **82,** 1112-1129. https://doi.org/10.1890/0012-9658(2001)082[1112:EOLDAS]2.0.CO;2

ter Braak, C.J.F. (2019) New robust weighted averaging- and model-methods for assessing trait–environment relationships. *Methods in Ecology and Evolution,* **10,** 1962-1971. https://doi.org/10.1111/2041-210X.13278

ter Braak, C.J.F. (2026) Fourth-corner latent variable models overstate confidence in trait–environment relationships and what to use instead. *Environmental and Ecological Statistics*. https://doi.org/10.1007/s10651-025-00696-0; https://edepot.wur.nl/707477; https://rdcu.be/eXukf.

8.2 N2 updating

Updating N_2 (species informativeness and site effective number of informative species) during IPF (Supplement B) has minor effect on the power in the Aravo- and Revisit-simulations, surely when compared with no N_2-processing (Figure S1).

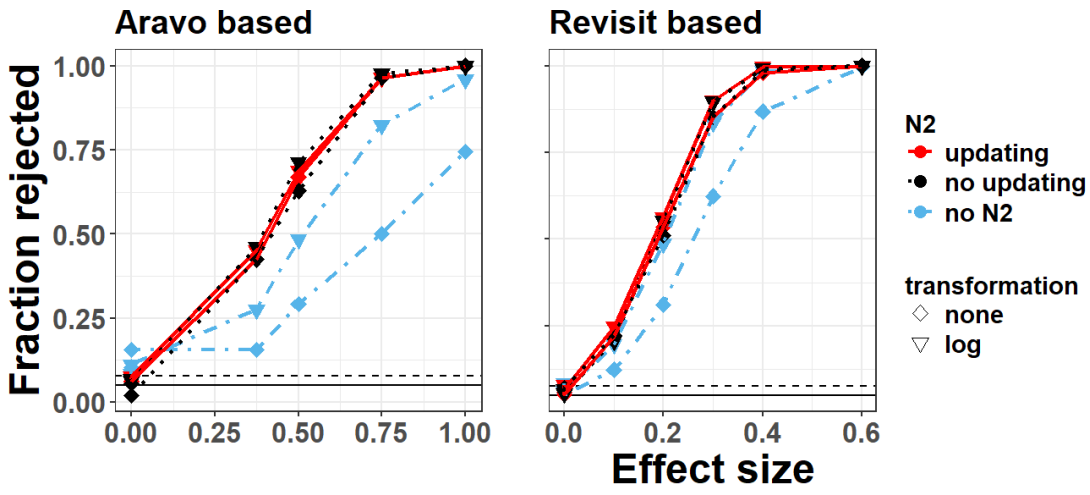


FIG. S1. Effect of updating N_2 during iterative proportional fitting (IPF) with and without prior log-transformation.

In the Beetle-simulations (Fig. S2), updating N_2 led to higher power without prior transformation and, as in Fig. S1, to similar power after log-transformation.

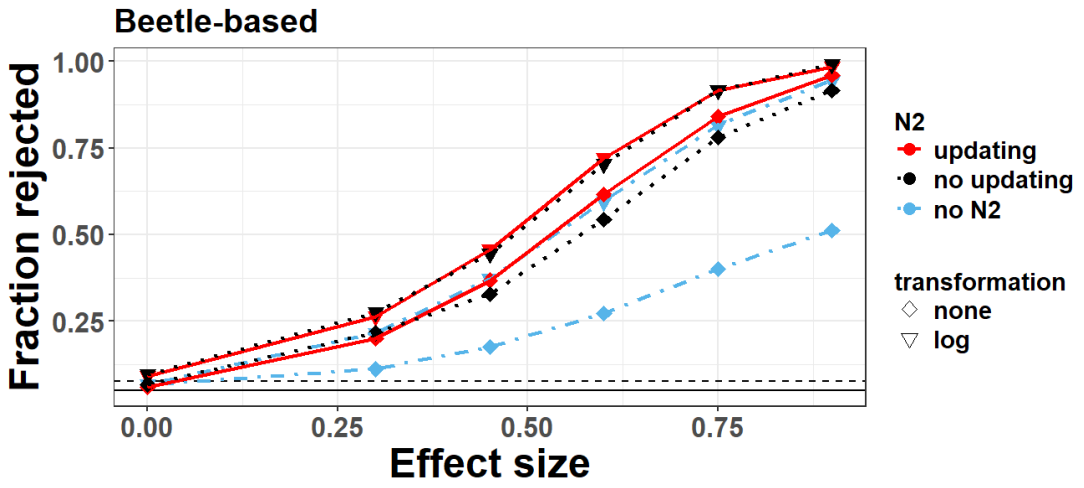


FIG. S2. Effect of updating N_2 during iterative proportional fitting (IPF) with and without prior log-transformation in dc-CA applied to the Beetle-simulations of a multi-trait, multi-environment GLMM3 model (ter Braak 2026).

The same pattern is visible in Figure S3 in the simulations of a GLLVM2(2) model in Figure 3 of both Niku et al. (2017) and ter Braak (2026): improvement without prior transformation but no effect of updating N_2 after log-transformation. Fig. S3a show a relatively rare case in which N_2-processing after a log-transformation led to lower power than N_2-processing alone.

The empirical power in Fig. S3 without pre-processing is much lower than in Fig. 3 of ter Braak (2026) because the latter used abundance data standardized by site totals, whereas Fig. S3 did not. Dividing by site totals is the default in *douconca* when N_2-processing is omitted, making dc-CA align more closely with CWM regressions, which generally avoid differential site weighting. N_2-processing replaces such ad-hoc choices with a principled one.

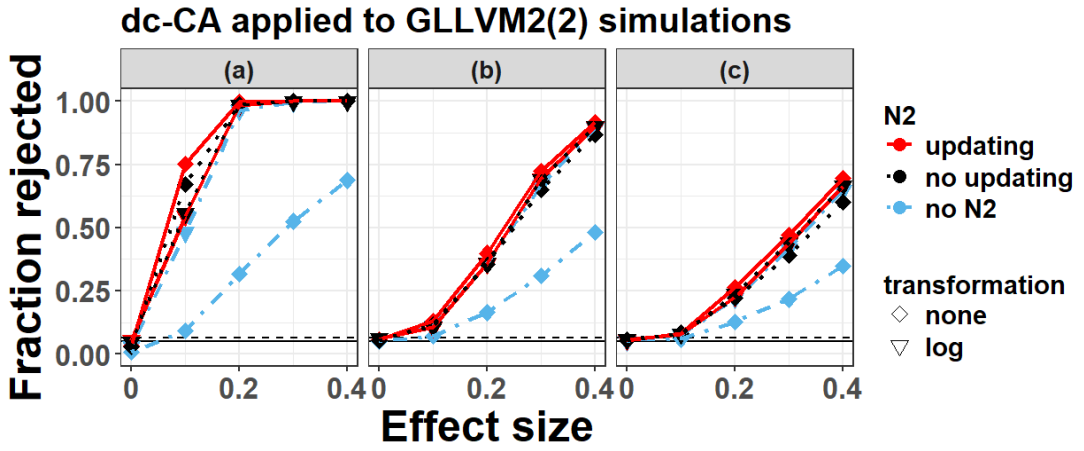


FIG. S3. Effect of updating N_2 during iterative proportional fitting (IPF) with and without prior log-transformation. The 1000 simulated data sets in each scenario are those of the GLLVM2(2) model of Figure 3 in ter Braak (2026), with two independent latent variables and random species-dependent environmental effects with variance $\sigma_b^2 = 0$, 0.4 and 0.8 in (a), (b) and (c), respectively ($n = 70$ sites and m

= 40 species). The scenarios were proposed in the Figures 1 and 3 of Niku et al. (2021).

Updating N_2 during IPF yielded generally very similar *P*-values in dc-CA analysis of the CESTES data sets (Figure S3).

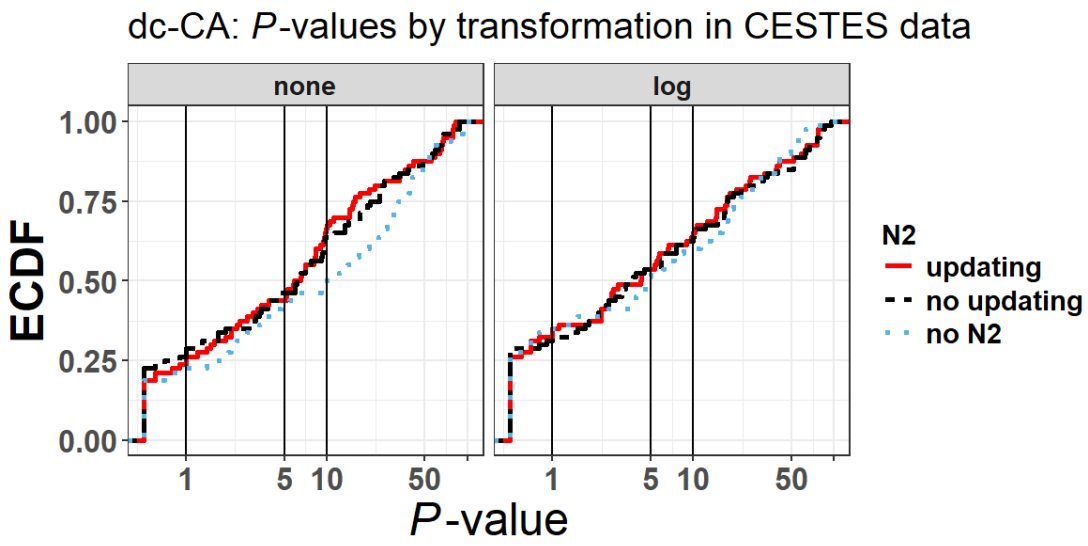


FIG. S3. Effect of N_2 and updating N_2 during iterative proportional fitting (IPF) with and without prior log-transformation: empirical cumulative density function (ECDF) of the significance level (*P*-value) using dc-CA for trait-environment association in the 80 data sets of the CESTES database in relation to updating N_2 and prior transformation (none and log). The abscissa starts at $P = 0.005$ (0.5%). The vertical lines are at the 1, 5 and 10% significance levels.

8.3 N2 variants

The N_2-variant of the main text (N2 in Figure S4), which is the default in the *ipf2N2* function in *douconca*, showed equal or more power in the Aravo- and Revisit-simulations than the other three variants (Fig. S4, Supplement B). With and without log-transformation, the N_2-variant in which both marginals are effective numbers (N2N2 in Figure S4) performs worse than the main text variant in the Aravo-simulations, whereas the simplest variant which is without iteration (iter0 in Figure S1) and divides each species' abundance by the informativeness of the species only, performed worse in the Revisit-simulations. The variant in which both marginals reflect informativeness performed similar to the N_2-variant of the main text.

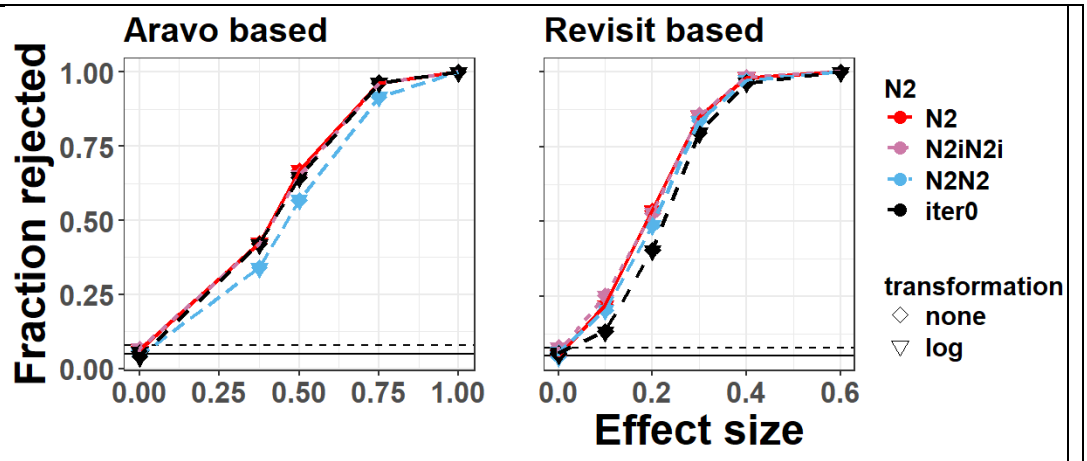


FIG. S4. Variants of N_2-processing with and without prior log-transformation: N2: default N_2-processing, *i.e.* iterative proportional fitting (ipf) to species marginal that represents informativeness ($N_2(1- N_2/N)$) and site marginal that represents effective numbers of species, in the following notation N2iN2; N2iN2i: ipf to marginals that represent informativeness of species and of sites; N2N2, ipf to marginals that represent effective numbers (N_2) of species and of sites;
iter0: species marginal scaled to informativeness of species without further scaling of the site marginal.

The four variants of N_2-processing yielded similar P-value distributions in dc-CA analysis of the CESTES data sets with and without prior log-transformation (Figure S5). Without prior transformation, version 'iter0' did well below $P = 0.07$.

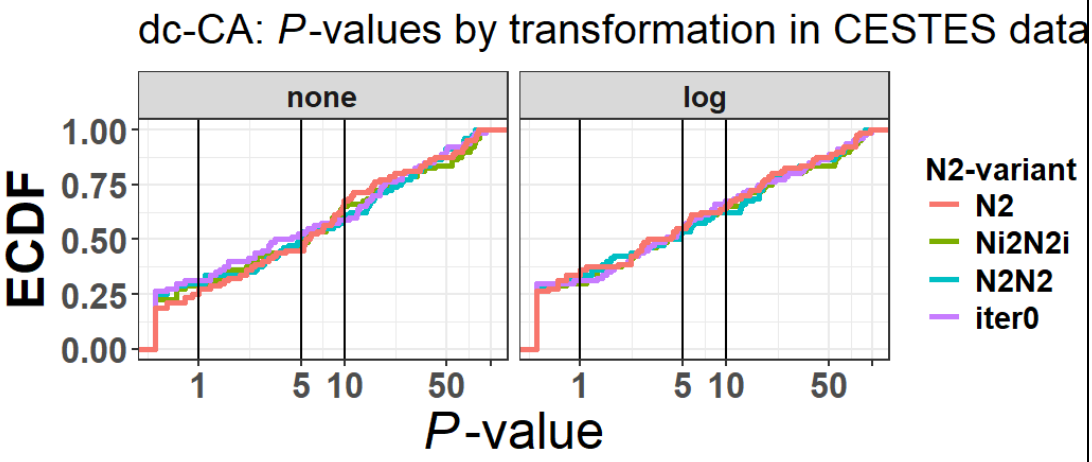


FIG. S5. Variants of N_2-processing with and without prior log-transformation: empirical cumulative density function (ECDF) of the significance level (P-value) using dc-CA for trait-environment association in the 80 data sets of the CESTES database in relation to method of N_2-variants (see legend Fig. S1) and prior transformation (none and log). The abscissa starts at $P = 0.005$ (0.5%). The vertical lines are at the 1, 5 and 10% significance levels.

8.4 Prior power transformations

Figure S6 complements Figure 1 in the main text and shows the square-root transformation with other power transformations with and without subsequent N_2-processing. With N_2-processing, all power transformations gave similar the rejection rates in the Aravo- and Revisit-simulations.

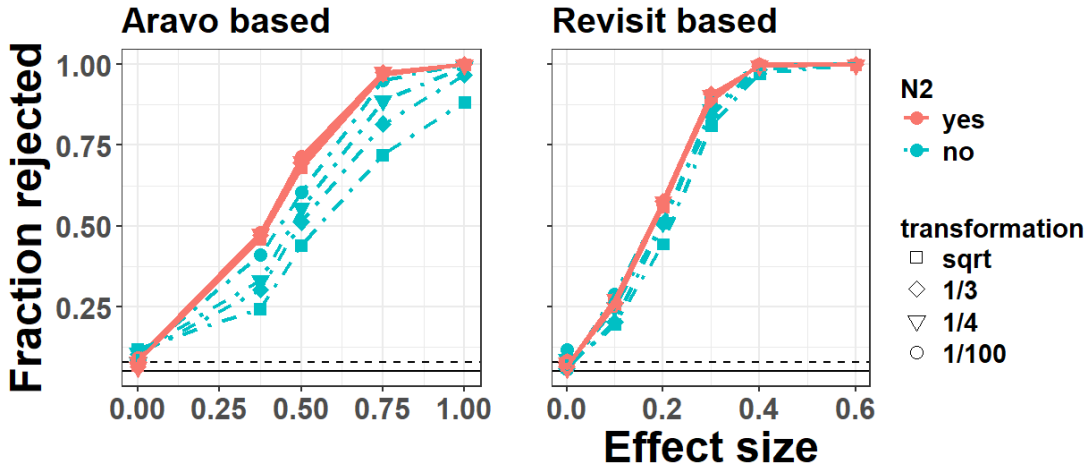


FIG. S6. Rejection rates of preprocessing methods against the size of the trait-environment association (Effect size), on the simulations of GLMM3 models for the Aravo and Revisit data in ter Braak (2019). Pre-procession methods are transformation of abundance (sqrt: square-root, and other fractional powers with and without subsequent N_2-processing (N_2). Additional simulations with increased trait variance were carried out, as in ter Braak (2019), for effect size zero, and the lines start at this scenario so as to emphasize potential Type I error rate inflation. The horizontal solid line is at the nominal significance threshold; rates above the dashed line (at 0.078) are significantly greater than 0.05.

Figure S7 complements Figure 2 in the main text and shows the empirical distributions of the *P*-value obtained with a number of prior power transformations in the CESTES data base. N_2-processing generally increased the evidence but its effect diminished with the smaller powers. With cube-root there are slightly more data sets that give low *P*-values (below 0.025) without subsequent N_2-processing than with such processing (Fig. S7). The lines for power 1/100 are very similar to those for transformation to presence-absence data. The square-root transformation followed by N_2-processing performs particularly well at the significance levels that matter most (5% and 10%) and shows no meaningful loss of power at the 1% level.

.

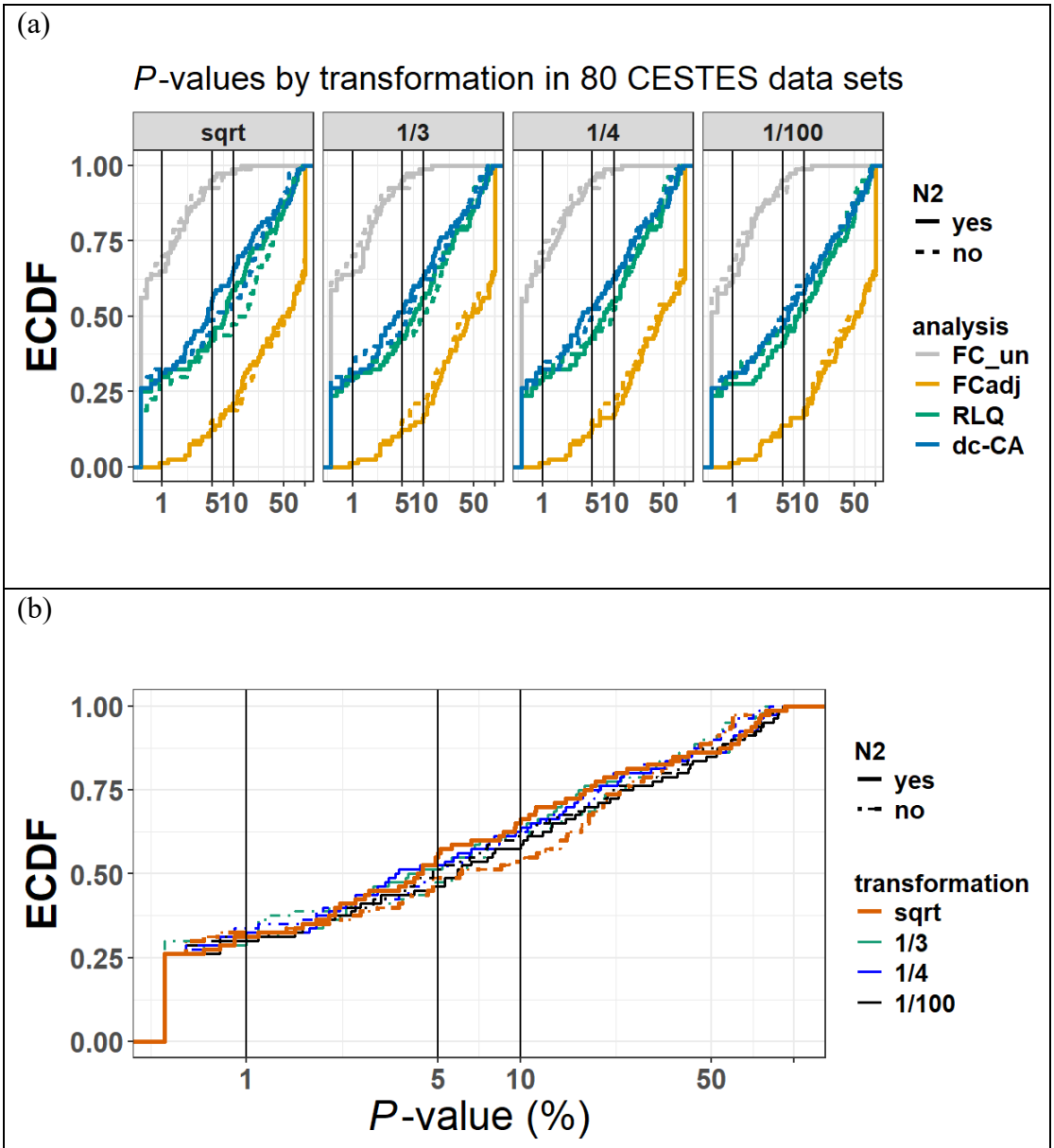


FIG. S7. Empirical cumulative density function (ECDF) of the significance level (*P*-value) for trait-environment association in the 80 data sets of the CESTES database in relation to method of pre-processing and analysis. Pre-procession methods are (power) transformation of abundance (sqrt: square-root and other numerical fractional powers, *e.g.* 1/4 : quarter-root) with and without subsequent N_2-processing (N_2). (a) Analysis methods are fourth corner correlation analysis with and without adjustment for multiple testing (FCadj & FC_un, respectively), RLQ and dc-CA; (b) dc-CA. The abscissa starts at $P = 0.005$ (0.5%). The vertical lines are at the 1, 5 and 10% significance levels.

8.5 Statistical analysis of the N_2-effect in the CESTES data base

This section focusses on the effect of prior-abundance transformation with or without N_2-processing on *P*-values issued by FC, RLQ and dc-CA applied to the 80 CESTES

data sets. The *P*-values are transformed to Evidence with Evidence = $-\log(P)$ to give more emphasis on the small *P*-values. The analysis is by linear regression and Anova of the Evidence on to the factors *method of analysis*, *abundance transformation*, and *N_2-processing*, with 3, 4 and 2 levels, respectively. The four levels of *method of analysis* (*mth*) are FC adjusted for multiple testing (FC-adj), RLQ and dc-CA. The three levels of *abundance transformation* (*pwr*) are none, square-root, log and transformation to presence absence. The two levels of *N_2-processing* (*N2*) are yes and no. As the Evidence varies with data set and is computed for each set, a paired analysis is asked for. This is achieved by adding the factor the *dataset* with 80 levels to each regression and Anova analysis. The Anova analyses use function `LSD.test` from the library `agricolae` (de Mendiburu 2023).

There was no statistical evidence for interaction between *method of analysis* and *method of pre-processing*, here *N2*pwr*, (*P* = 0.68) as judged by the comparison of the models

*Evidence ~ dataset + N2*pwr*mth*

*Evidence ~ dataset + N2*pwr + mth*.

The mean Evidence per *method of analysis* with grouping on least-significant difference (LSD) without adjustment at *P* = 0.05 is

```
##       Evidence groups
## dc-CA 3.305665      a
## RLQ   3.025167      b
## FCadj 1.149048      c
```

The LSD = 0.131. Note that an Evidence of 3 corresponds to an *P* value issued by the method of analysis of 0.05, indicating that the mean Evidence of RLQ and dc-CA is just below *P* = 0.05, the mean Evidence of `FCadj` of 1.149 corresponds to *P* = 0.32. The mean Evidence per *method of pre-processing* with grouping on least-significant difference (LSD) without adjustment at *P* = 0.05, with LSD = 0.213, is

```
##          Evidence groups
## yes.log  2.619430      a
## yes.sqrt 2.616760      a
## no.P/A   2.578651      a
## no.log   2.574581      a
## yes.none 2.535546      a
## yes.P/A  2.523273      a
## no.sqrt  2.464680      a
## no.none  2.033426      b
```

The low values are caused by the low mean Evidence in the `FCadj` analysis. In the following, the *P*- values are restricted to those of dc-CA. The grouping for dc-CA only:

```
## yes.log  3.505116      a
## yes.sqrt 3.462850     ab
## no.log   3.408959     ab
## no.P/A   3.357034     ab
## yes.P/A  3.329233     ab
## no.sqrt  3.297212     ab
```

```
## yes.none 3.225800      b
## no.none  2.859117      c
```

with LSD= 0.270. No pre-processing at all is clearly worst. While N_2-processing gives higher mean Evidence than no preprocessing (`yes.none` compared to `no.none`), additional pre-processing gives further improvement, with log-transformation performing best and the square-root as runner-up.

Quantifying the abundance transformations (none=0, sqrt = 1, log = 2, P/A=4), the model with factors, `Evidence ~ dataset + N2 * pwr` is not statistically better fitting than the quantified model `Evidence ~ dataset+ N2q+pwrq+ I(pwrq^2) + N2qpwrq` ($P = 0.97$). The coefficient and R-squared- and F-statistics of the latter are:

```
## Coefficients:
##                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)                1.96258    0.31738   6.184 1.21e-09 ***
## datasetBarbaro2009a        0.96264    0.43391   2.219 0.026922 *
…
## datasetYates2014          -0.62001    0.43391  -1.429 0.153595
## N2q                        0.33811    0.11480   2.945 0.003363 **
## pwrq                       0.49916    0.11168   4.469 9.51e-06 ***
## I(pwrq^2)                 -0.11287    0.03430  -3.290 0.001064 **
## N2qpwrq                   -0.12529    0.06136  -2.042 0.041643 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8678 on 556 degrees of freedom
## Multiple R-squared:  0.8644, Adjusted R-squared:  0.8441
## F-statistic: 42.68 on 83 and 556 DF,  p-value: < 2.2e-16
```

showing again that N2 has a significant effect ($P < 0.01$) on the Evidence without other preprocessing, but that this effect is diminishing with the stronger prior abundance transformation, as shown by the negative coefficient of the interaction `N2qpwrq`.

With all power transformations tested the group means are (LSD = 0.234):

```
##            Evidence groups
## yes.log    3.505116      a
## yes.1/4    3.475380      a
## yes.sqrt   3.462850      a
## yes.1/3    3.459842     ab
## no.1/4     3.435104     ab
## no.1/3     3.410377     ab
## no.log     3.408959     ab
## no.1/100   3.360803     ab
## no.P/A     3.357034     ab
## no.1/1000  3.356174     ab
## yes.1/100  3.330693     ab
## yes.P/A    3.329233     ab
## yes.1/1000 3.309273     ab
## no.sqrt    3.297212     ab
## yes.none   3.225800      b
## no.none    2.859117      c
```

demonstrating that N2-processing works and that the choice of prior abundance transformation (log, square-, cube- or quarter- root) is of minor importance.

Discussion

While informative about which differences have no statistical meaning, the regression models are, of course, simplistic as they focuss on the mean evidence expressed as -log(P), i.e. on the geometric mean P, only. Quantile regression might also be of interest.

Reference

de Mendiburu F. (2023). agricolae: Statistical Procedures for Agricultural Research, R package version 1.3-7. https://doi.org/10.32614/CRAN.package.agricolae.

8.6 Subdividing the CESTES data base by type

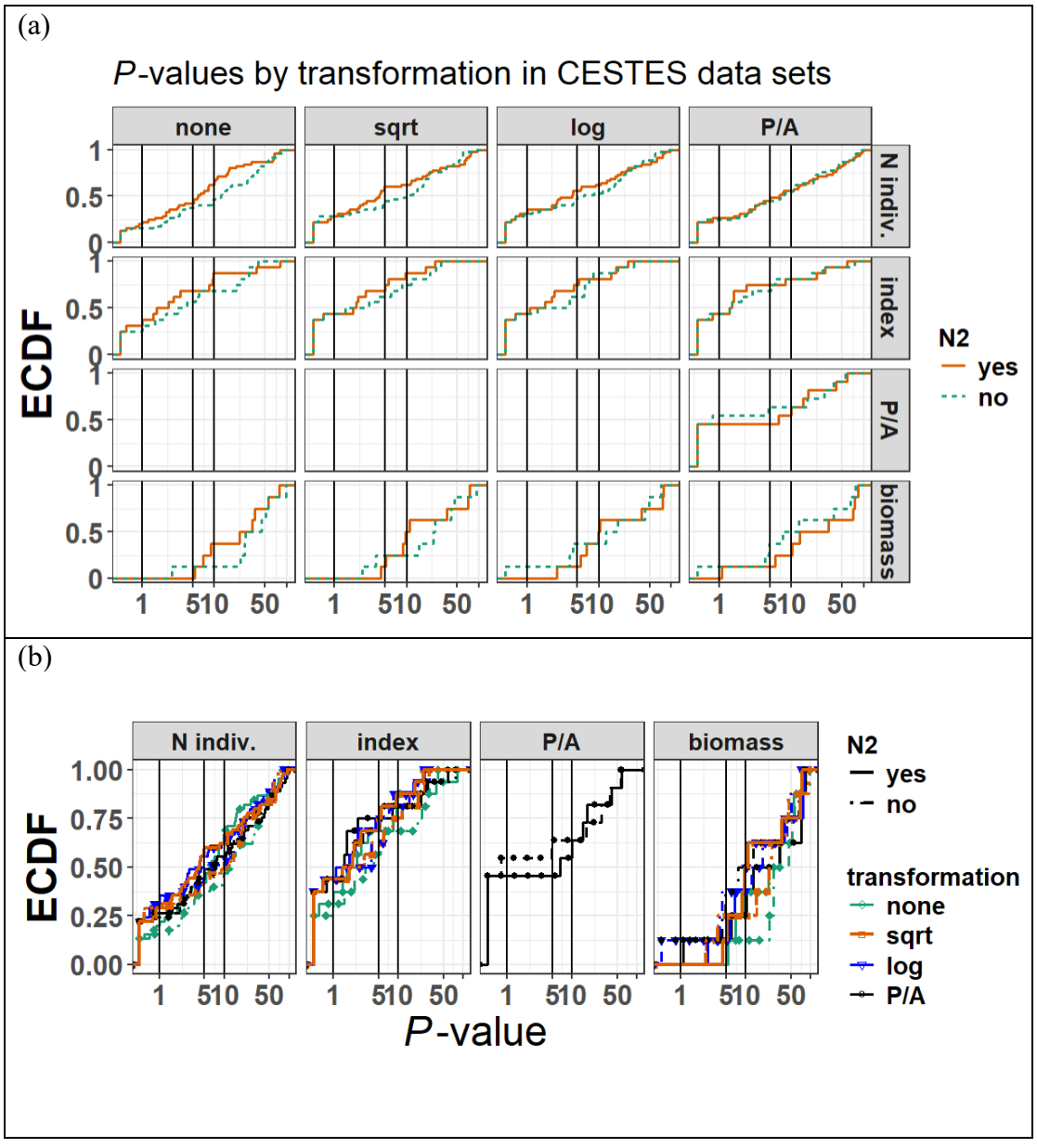
Here in CESTEST data base is subdivided by response type, ecosystem type, and taxonomic group to examine whether the preferred transformation (prior-abundance transformation with or without N_2-processing) varies across these categories. It should be noted that several of these subsets contain relatively few datasets, and the corresponding results should therefore be interpreted with caution and not over-generalized.

Type of abundance measurement

Perhaps the best transformation (prior abundance transformation with or without N_2-processing) differs between the type of response, *i.e.* how abundance was measured. Fig. S8a focusses on the effect of N_2-processing per types of prior transformation and prior transformation. Fig. S8b directly compares the different prior abundance transformations, whereas Fig. S8c compares types of abundance measurement for square-root transformed abundance with and without N_2-processing.

In data sets in which the abundance is expressed as (average) number of individuals or an index, N_2-processing yielded generally more evidence for trait-environment association (lower P-values) with diminishing effect with stronger prior abundance transformation (from none to P/A, Fig. S8a).

Pre-processing resulted in generally lower P values (more evidence for trait-environment association), with prior square-root and log-transformation being best, irrespective of the type of response (Fig. S8b), with the exception of response coded as presence-absence. N_2-processing on the raw abundance, when recorded as number of individuals showed stronger evidence (lower P values) for $P \geq 0.10$ than with prior abundance transformation. For data sets that recorded presence-absence only all pre-processing methods show the same EDCF, except that N_2-processing showed lower P values for $P > 0.10$. Apparently, only in data sets with weak to very weak evidence for trait-environment association contain species that are very common, so that informativeness works better than simply the number of occurrences. Studies using indices of abundance showed generally higher evidence than other measures of abundance. The fraction of data sets with $P < 0.02$ was highest in data sets using presence-absence.



(c)

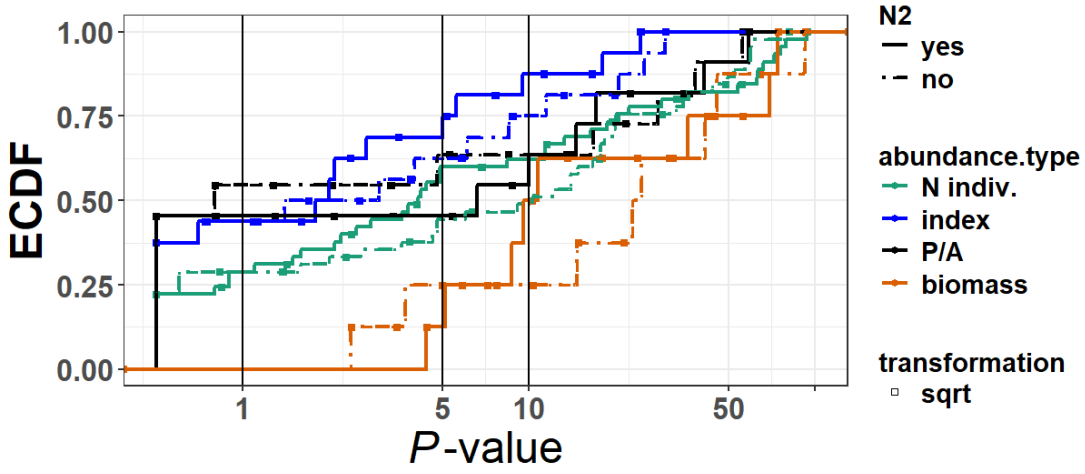
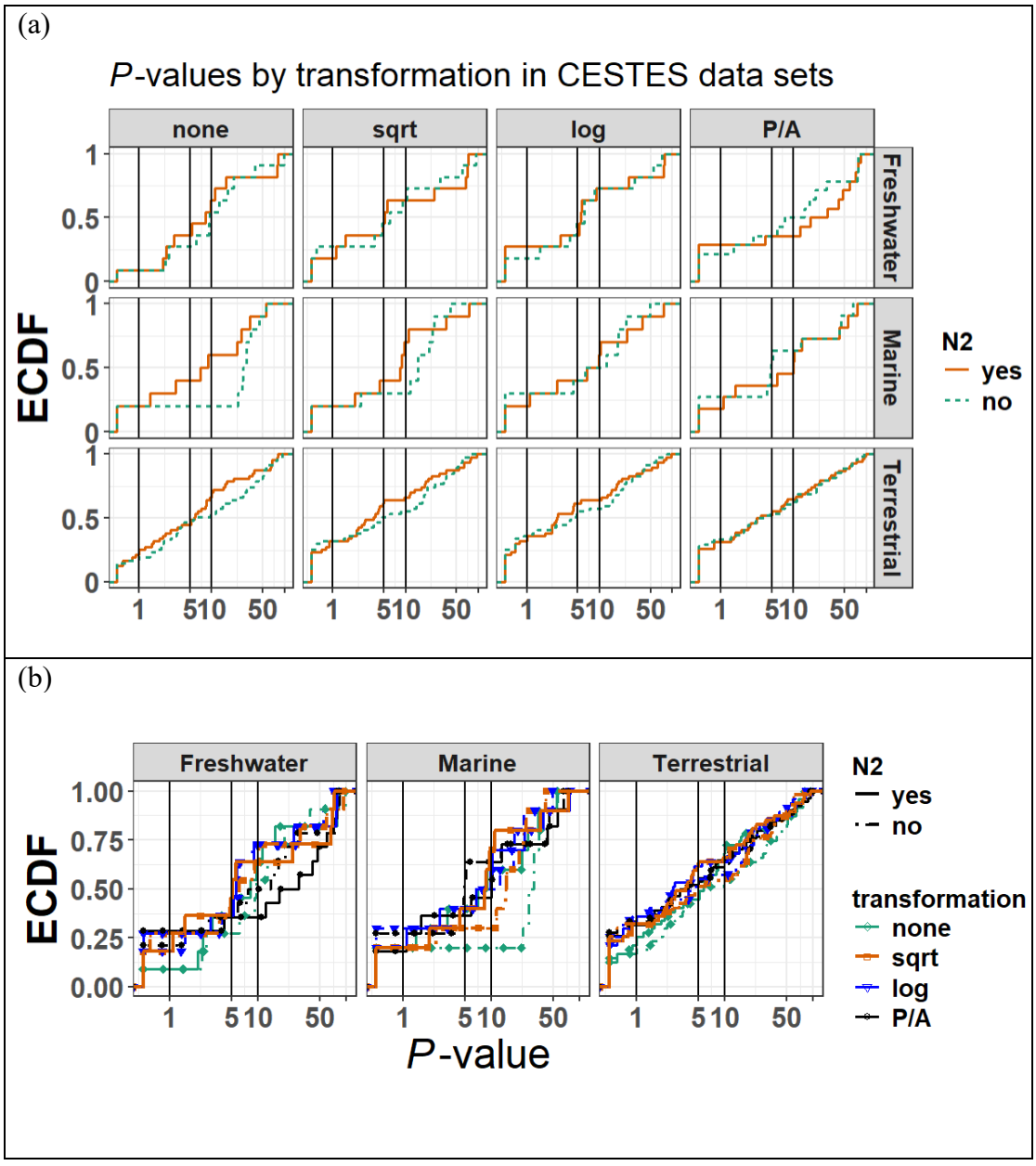


FIG. S8. Empirical cumulative density function (ECDF) of the dc-CA significance (*P*-value) for trait-environment association in the data sets of the CESTES database in relation to methods of pre-processing and type of abundance. Pre-procession methods are transformation of abundance (none, sqrt: square-root, log: logarithmic and to P/A: transformation to presence/absence) with and without subsequent N_2-processing (N_2). Types of abundance are number of individuals (N= 45 datasets) and abundance index (N =16), presence-absence (N =11) and biomass (N =8). (a) Effect of N_2-processing on *P*-values by prior transformation and type of abundance. (b) A *P*-value comparison of prior transformations per type. (c) A *P*-value comparison of types with square-root abundance transformation. The abscissa starts at $P = 0.005$ (0.5%). The vertical lines are at the 1, 5 and 10% significance levels.

Type of ecosystem

Perhaps the best transformation (prior abundance transformation with or without N_2-processing) differs between the type of ecosystem. Fig. S9a focusses on the effect of N_2-processing per type of ecosystem and prior transformation. Fig. S9b directly compares the different prior abundance transformations, whereas Fig. S9c compares ecosystem types for square-root transformed abundance with and without N_2-processing.

The main patterns are similar to those in the main text with an effect of N_2-processing that diminishes with stronger prior abundance transformation and good performance of square-root and log-transformation, with little improvement due to N_2-processing (Fig. 9a). The effect of N_2-processing was largest in Marine data sets, but slightly negative with transformation to presence-absence (Fig. 9a). The fraction of data sets showing strong evidence ($P < 0.01$) for trait-environment association was highest in Terrestrial studies (Fig. 9c).



(c)

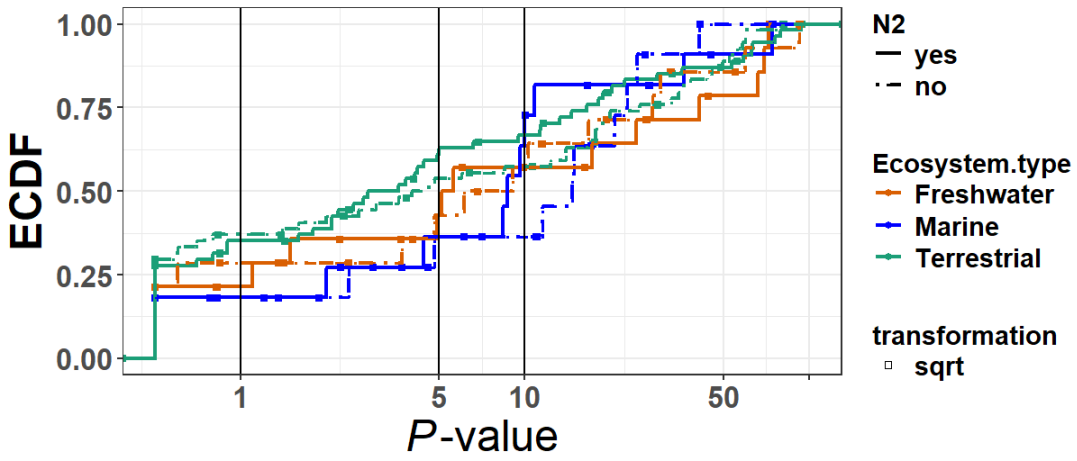


FIG. S9. Empirical cumulative density function (ECDF) of the dc-CA significance (*P*-value) for trait-environment association in the data sets of the CESTES database in relation to methods of pre-processing and type of ecosystem. Pre-procession methods are transformation of abundance (none, sqrt: square-root, log: logarithmic and to P/A: transformation to presence/absence) with and without subsequent N_2-processing (N_2). Types of ecosystems are Freshwater (N= 14 datasets) and marine (N =11) and terrestrial (N =54). (a) Effect of N_2-processing on *P*-values by prior transformation and type of ecosystem. (b) A *P*-value comparison of prior transformations per type. (c) A *P*-value comparison of types. The abscissa starts at $P = 0.005$ (0.5%). The vertical lines are at the 1, 5 and 10% significance levels.

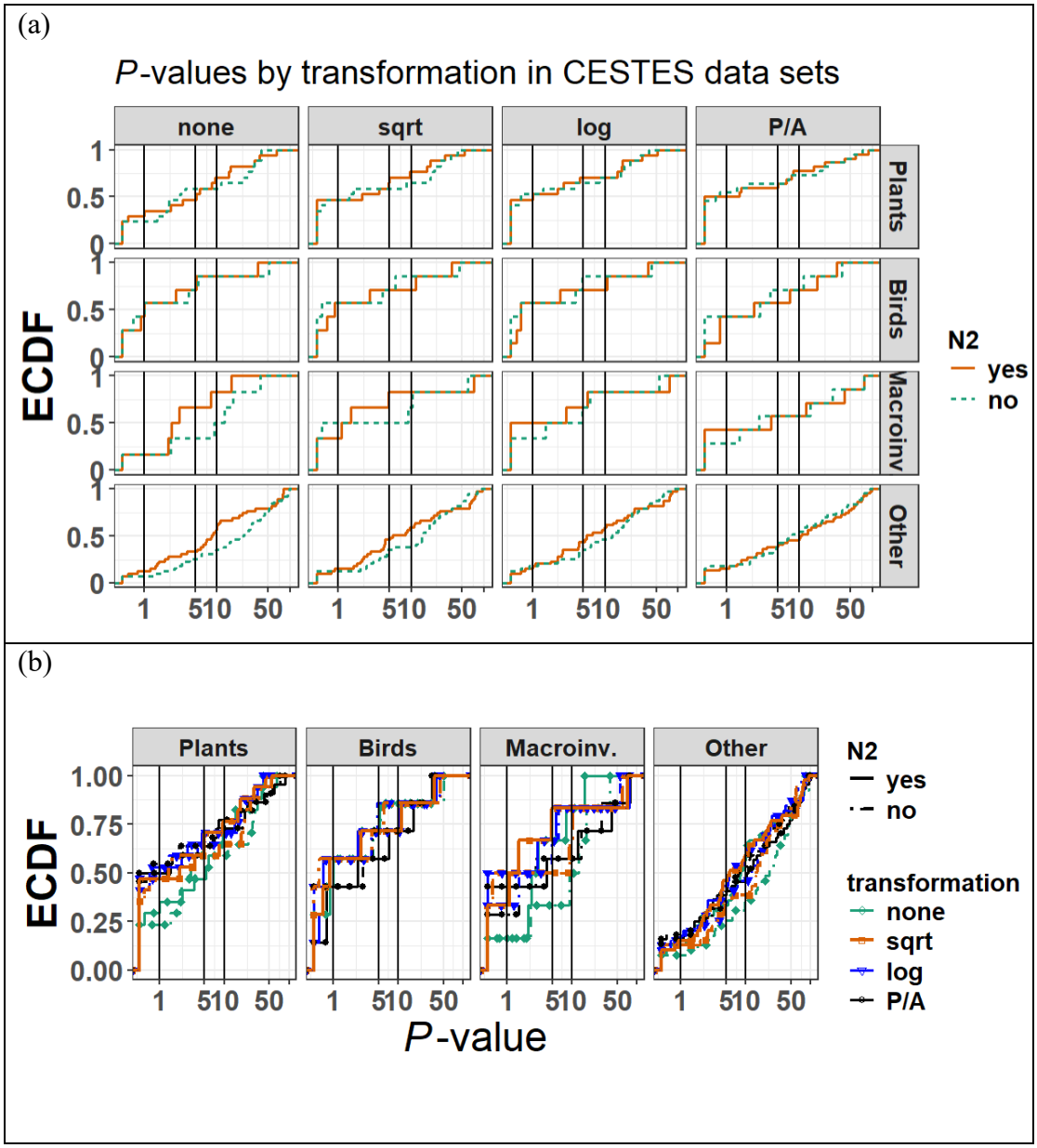
1245

1246

Taxonomic group

Perhaps the best transformation (prior abundance transformation with or without N_2-processing) differs between taxonomic groups. Fig. S10a focusses on the effect of N_2-processing per taxonomic group and prior transformation. Fig. S10b directly compares the different prior abundance transformations, whereas Fig. S10c compares taxonomic groups for square-root transformed abundance with and without N_2-processing.

In the Plant and Bird data sets the effect of N_2-processing is small (Fig. 10a). The evidence for trait-environment association is higher in the Plant and Birds data sets than in the data sets on Macroinvertebrates, with least evidence in 'Other' consisting of 24 different taxonomic groups.



(c)

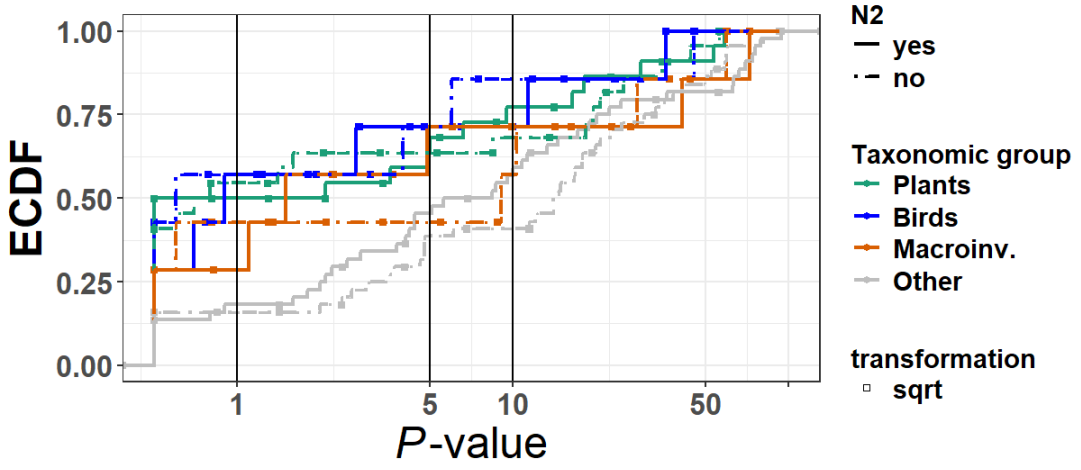


FIG. S10. Empirical cumulative density function (ECDF) of the dc-CA significance (*P*-value) for trait-environment association in the data sets of the CESTES database in relation to methods of pre-processing and taxonomic group. Pre-procession methods are transformation of abundance (none, sqrt: square-root, log: logarithmic and to P/A: transformation to presence/absence) with and without subsequent N_2-processing (N_2). Taxonomic groups: Plants (N=22 datasets) and Birds (N = 7), Macroinvertebrates (N = 7) and Other (N = 44). (a) Effect of N_2-processing on *P*-values by prior transformation and taxonomic group. (b) A *P*-value comparison of prior transformations per group. (c) A *P*-value comparison of taxonomic groups. The abscissa starts at P = 0.005 (0.5%). The vertical lines are at the 1, 5 and 10% significance levels.