ECOGRAPHY

Research article

Use and misuse of trait imputation in ecology: the problem of using out-of-context imputed values

Lucas Damián Gorné[1,2], Jesús Aguirre-Gutiérrez[3,4], Fernanda C. Souza[5], Nathan G. Swenson[6], Nathan Jared Boardman Kraft[7], Beatriz Schwantes Marimon[8], Timothy R. Baker[9], Renato A. Ferreira de Lima[10], Emilio Vilanova[11], Esteban Álvarez-Dávila[12], Abel Monteagudo Mendoza[13,14], Gerardo Rafael Flores Llampazo[15], Rubens Manoel dos Santos[16], Gerhard Boenisch[17], Alejandro Araujo-Murakami[18], Gonzalo Rivas-Torres[19], Hirma Ramírez-Angulo[20], Nayane Cristina dos Santos Prestes[21], Paulo S. Morandi[22], Sabina Cerruto Ribeiro[23], Wesley Jonatar A. da Cruz[24], Mathias Disney[25,26], Anthony Di Fiore[19,27], Ben Hur Marimon-Junior[8], Ted R. Feldpausch[28], Yadvinder Malhi[3], Oliver L. Phillips[29], David Galbraith[29] and Sandra Díaz[1,2]

[1]Universidad Nacional de Córdoba, Facultad de Ciencias Exactas Físicas y Naturales, Córdoba, Argentina
[2]Consejo Nacional de Investigaciones Científicas y Técnicas, CONICET, IMBiV. Córdoba, Argentina
[3]Environmental Change Institute, School of Geography and the Environment, University of Oxford, Oxford, UK
[4]Leverhulme Centre for Nature Recovery, University of Oxford, Oxford, UK
[5]Departamento de Ecologia e Conservação, Instituto de Ciências Naturais, Universidade Federal de Lavras, Lavras, MG, Brazil
[6]Department of Biological Sciences, University of Notre Dame, Notre Dame, IN, USA
[7]Department of Ecology and Evolutionary Biology, University of California, Los Angeles, CA, USA
[8]Universidade do Estado de Mato Grosso (UNEMAT), Programa de Pós-Graduação em Ecologia e Conservação, Nova Xavantina, MT, Brazil
[9]School of Geography, University of Leeds, Leeds, UK
[10]Departamento de Ciências Biológicas, ESALQ, Universidade de São Paulo, Piracicaba, SP, Brazil
[11]Wildlife Conservation Society, Bronx, NY, USA
[12]UNAD-Universidad Nacional Abierta y a Distancia de Colombia, Bogotá, Colombia
[13]Universidad Nacional de San Antonio Abad del Cusco, Cusco, Perú
[14]Jardín Botánico de Missouri, Oxapampa, Perú
[15]Instituto de Investigaciones de la Amazonía Peruana, Iquitos, Perú
[16]Laboratory of Phytogeography and Evolutionary Ecology, Universidade Federal de Lavras, Lavras, MG, Brazil
[17]Max-Planck-Institute for Biogeochemistry, Jena, Germany
[18]Museo de Historia Natural Noel Kempff Mercado, Universidad Autónoma Gabriel Rene Moreno, Santa Cruz de la Sierra, Bolivia
[19]Estación de Biodiversidad Tiputini, Universidad San Francisco de Quito, Quito, Ecuador
[20]Indefor, Facultad de Ciencias Forestales y Ambientales, Universidad de Los Andes, Mérida, Venezuela
[21]Universidade do Estado de Mato Grosso, Caceres, MT, Brazil
[22]Programa de Pós-graduação em Ecologia e Conservação, Universidade do Estado de Mato Grosso, Caceres, MT, Brazil
[23]Centro de Ciências Biológicas e da Natureza, Universidade Federal do Acre, Rio Branco, AC, Brazil
[24]AMAP (botAnique et Modélisation de l'Architecture des Plantes et des Végétations), 8 CIRAD, CNRS, INRAE, IRD, Montpellier Cedex 5, France
[25]Department of Geography, University College London, London, UK
[26]NERC National Centre for Earth Observation (NCEO), London, UK
[27]Department of Anthropology and Primate Molecular Ecology and Evolution Laboratory, The University of Texas at Austin, TX, USA
[28]Geography, Faculty of Environment, Science and Economy, University of Exeter, Exeter, UK
[29]School of Geography, University of Leeds, Leeds, UK

Correspondence: Lucas Damián Gorné (gorneld@gmail.com)

Ecography
2025: e07520
doi: 10.1111/ecog.07520
Subject Editor: Carsten Dormann
Editor-in-Chief: Miguel Araújo
Accepted 18 December 2024

NORDIC SOCIETY OIKOS

www.ecography.org

Despite the progress in the measurement and accessibility of plant trait information, acquiring sufficiently complete data from enough species to answer broad-scale questions in plant functional ecology and biogeography remains challenging. A common way to overcome this challenge is by imputation, or 'gap-filling' of trait values. This has proven appropriate when focusing on the overall patterns emerging from the database being imputed. However, some applications force the imputation procedure out of its original scope, using imputed values independently from the imputation context, and specific trait values for a given species are used as input for computing new variables. We tested the performance of three widely used imputation methods (Bayesian hierarchical probabilistic matrix factorization, multiple imputation by chained equations with predictive mean matching, and Rphylopars) on a database of tropical tree and shrub traits. By applying a leave-one-out procedure, we assessed the accuracy and precision of the imputed values and found that out-of-context use of imputed values may bias the estimation of different variables. We also found that low redundancy (i.e. low predictability of a new value on the basis of existing values) in the dataset, not uncommon for empirical datasets, is likely the main cause of low accuracy and precision in the imputed values. We therefore suggest the use of a leave-one-out procedure to test the quality of the imputed values before any out-of-context application of the imputed values, and make practical recommendations to avoid the misuse of imputation procedures. Furthermore, we recommend not publishing gap-filled datasets, publishing instead only the empirical data, together with the imputation method applied and the corresponding script to reproduce the imputation. This will help avoid the spread of imputed data, whose accuracy, precision, and source are difficult to assess and track, into the public domain.

Keywords: BHPMF, gap-filling, imputation, mice, plant trait, Rphylopars, sparse matrix

Introduction

The use of plant functional trait information to answer questions in ecology and biogeography has greatly increased in recent years. While they have contributed to this, large-scale, open-access data repositories (Kattge et al. 2020) are still rather sparsely populated, so that it is difficult to obtain information for more than a handful of traits for any given species. At the finer scales of local to landscape studies, it is sometimes possible to obtain the missing information from the field, which is also a better practice to use on-site trait data for studies with local questions (Palma et al. 2022). However, this is often unfeasible for regional, continental or global scale studies. One increasingly common way to overcome this constraint is the use of imputation methods to fill the trait information gaps for species orindividual plants (entities in the trait datasets).

A range of methods with different degrees of suitability for different situations exists in the literature (Johnson et al. 2021, Enders 2022). Imputation methods are designed to allocate trait values that are, on average, consistent with the patterns in the observed data. In general, a good imputation method should be useful to keep incomplete cases and thus prevent the loss of information and, at the same time, avoid distorting the information provided by the empirical data. When data are missing completely at random the consequence of removing incomplete observations is a decrease in statistical power, due to decreased sample size (Nakagawa and Freckleton 2008), but no estimation bias is expected in the statistics. However, in real datasets, the missing data for a given variable are often related to the values of other variables. In these cases, deleting incomplete cases can lead to misleading results in comparative studies or biased estimations (Penone et al. 2014). The literature about imputation procedures recommends that the imputation model should contain all variables in the analysis model and any interactions between variables, as well as any auxiliary variables not included in the analysis model to make the missingness mechanism assumptions more plausible and to provide information about the missing values (Madley-Down et al. 2019). The chosen imputation method should be able to take advantage of the redundancies in the data structure (Box 1), e.g. if a phylogenetic signal is expected in a trait an imputation method incorporating a phylogenetic correlation matrix will be preferred. Also, it is recommended that the performance of imputation methods should be based mainly on statistical estimates instead of only on imputation error (Jardim et al. 2021). In summary, imputation methods are solutions for sparse data, but these solutions depend on the statistical model to be applied to the dataset and on the dataset itself. However, other applications are starting to appear, namely imputation performed as a way of predicting specific trait values (i.e. attributes) that will in turn be used as input information to compute new variables and perform further analyses based on them. Two common examples of these 'out-of-context' applications of the imputation procedure are: 1) the computing of trait community weighted means in local plots, which requires taking the attributes of a subset of species from a species by trait matrix to perform a summation of them weighted by the relative abundance of each species in each plot (Garnier et al. 2004); and 2) the projection of new entities (e.g. species, communities) onto pre-defined phenotypic spaces (Segrestin et al. 2021), which also requires taking specific attributes for a given species as an input in a summation of the eigenvector of traits values that will allow the projection. In both cases, even when the imputation preserves the data structure of the species by trait matrix (i.e. covariation and taxonomic/phylogenetic structure), the procedure has no information about the community weighted mean where the species belong or the position of the species onto the pre-defined phenotypic space. Because the imputation method does not have that input, and because any imputation method does not create new information but extracts and distributes information already existing in the database (Box 1), if the isolated (out-of-context) imputed values for specific attributes are incorrect, so will be the arithmetic operations made based on them. Previous works have raised concerns about using imputation methods in

local-scale studies or trait-based community ecology (Swenson 2014, Swenson et al. 2017). Additionally, previous studies have shown that every known imputation approach produces inaccurate values, even with as little as 5% missing data (Johnson et al. 2021). It follows that, when an 'extension' application such as these is planned, a previous testing of the accuracy and precision of the imputed attributes becomes essential.

The present study aims to draw attention to the risk of the out-of-context uses of imputed values. For this purpose, we use an empirical trait database to illustrate the precision and accuracy of the imputed values and the effect of using such values out of the context of the imputation to compute new variables (community weighted means and position onto a pre-defined phenotypic space, in the present case).

Material and methods

Datasets

To produce the pre-defined phenotypic space, we used the dataset released by Díaz et al. (2022) which was the basis of a global spectrum of plant form and function (Díaz et al. 2016). This dataset contains species mean values for six vascular plant traits (plant height, stem specific density, leaf area, leaf mass per area, leaf nitrogen content per dry mass and diaspore mass). From this dataset, we got the subset of species with no missing trait values ('Díaz's dataset').

To test precision and accuracy of the imputed values and its consequences we worked with Baraloto et al. dataset (Baraloto et al. 2010a, b; dataset 269 in the TRY Database https://www.try-db.org/TryWeb/Home.php; hereafter 'Baraloto datasets'). This dataset involves Neotropical woody species, and preserves the trait information at the level of individual plants. We subset it by keeping data from the same traits as for the global spectrum of form and function (seed mass is missing in this dataset). Then, we kept individuals for which more than one trait were measured, and species for which more than one individual were measured. This resulted in a dataset of 7227 individuals from 448 species and 200 genera. The aggregated species-level Baraloto's dataset was 97.32% complete. The individual-level Baraloto's dataset achieved a 67.88% coverage (Supporting information). In this way we obtained the two related databases with the same species but different levels of redundancy, one having repeated observations within each species (at least two individuals), and the other having the mean value for each trait for each species. The individual level is more redundant than the species level dataset because the variability of traits within species is smaller than between species. As a consequence, if we know the trait value for one individual of a given species, the information added by a new individual of the same species is less than the information added by a new individual of a different species.

Box 1.

Information and redundancy. Here we use these terms in the framework of the mathematical theory of information (Shannon and Weaver 1964). **Information** is a measurement of the freedom of all elements in a message. In our case, the 'message' could be the parameter values of a statistical model describing the relationship among variables and the elements could be the elements of the entities by traits matrix. **Redundancy** is the proportion of non-informative elements in the message. An element is considered non-informative when it adds little extra information with respect to the one already provided by the other elements. The stronger the covariation structure in a matrix, because of e.g. convergent evolution into syndromes and/or constraints related to evolutionary history, biomechanics or biochemical pathways, the higher the redundancy will be because each element is less free to vary independently. In other words, in a highly redundant matrix, each new observation will contribute very little new information. The higher the redundancy, the higher the predictability of a new value on the basis of the ones already existing in the matrix. An example of how to operationalize and measure information in the context of imputation was introduced by Jardim et al. (2021). They also showed that the proportion of missing data in a dataset is not always a good proxy of the proportion of missing information in the same dataset.

Leave-one-out procedure

We tested the precision and accuracy of three popular imputation methods on the species-level Baraloto's dataset. First, it was the Bayesian hierarchical probabilistic matrix factorization method (BHPMF, Schrodt et al. 2015). BHPMF was originally designed to preserve the covariation and taxonomic structure of plant attributes, and it suits well the needs of plant trait-based macroecology (Díaz et al. 2016, Bruelheide et al. 2018). Second, it was the multiple imputation by chained equations (MICE) procedure, with the predictive mean matching (PMM) method (Little 1988, van Buuren et al. 1999). MICE-PMM imputes data by matching observed values between traits, then populates missing values in incomplete traits by adopting information from the matched species. Third, it was Rphylopars. Rphylopars is a maximum likelihood frequentist method that uses a phylogeny and a sparse trait matrix to estimate simultaneously the across-species (phylogenetic) and within-species (phenotypic) trait covariance (similar to a phylogenetic mixed model) to reconstruct the ancestral state and impute missing values (Goolsby et al. 2017). BHPMF and Rphylopars methods estimate a mean and a standard deviation (square root of the phenotypic variance in the Rphylopars method) for the imputed values. Because MICE-PMM is a process of multiple imputations we imputed 10 times and computed the mean and standard deviation from these.

Before performing the imputation methods, the traits were $\log_{10}$-transformed and each $\log_{10}$-transformed variable was standardised by centering and scaling ($z = (x_i - \bar{x})/\mathrm{sd}(x)$), as recommended by Schrodt et al. (2015). The quality-check procedure consists of repeating the imputation process while removing each one of the elements in the matrix, one by one (Leave-one-out. Supporting information: script code 1). The process was performed only for those entities (rows) with two or more known attributes (traits). At the end of the process, we obtained the empirically observed values along with the corresponding imputed values when this element was removed. On this basis, we carried out a series of quality-control analyses. First, we controlled for imputed values falling out of the empirically observed range for each trait by constraining all imputed values to that range. Imputed values higher than the maximum or lower than the minimum values of a given trait were replaced by the maximum or minimum observed, respectively. Second, to assess the accuracy of the imputation, we performed an ordinary least squares linear regression analysis with the $\log_{10}$-transformed imputed values as a function of the $\log_{10}$-transformed empirically observed values. Because we aimed to assess the imputed values as predictors of the actual ones, we assumed the empirical values were measured accurately and their error was negligible compared to the error in the imputed values. Here we expected an intercept = 0 and a slope = 1. Third, to assess the precision of the imputation, we computed an index accounting for the proportion of the total variance in a trait which is predicted by the imputed values:

$$Q = 1 - \frac{\sum (I_i - O_i)^2}{\sum (O_i - \bar{O})^2}$$

Where, I_i and O_i are the imputed and empirically observed values for the entity i, respectively, and $\bar{O}$ is the mean empirical value across all entities. Q has an upper limit = 1 (when there is no error in the imputation); the lower value could be < 0 if the errors are larger than the variance in the empiric values for a trait. This would mean that the overall empiric mean is a better estimation than the imputation. We also computed the root mean square error for the z-variables (zRMSE) to assess precision. The lower limit of zRMSE = 0, which would mean that the imputed values are equal to the empiric ones, i.e. the imputation is perfect. Because the z-variables are scaled by the standard deviation, a value of zRMSE>1 would mean that the imputed values are, on average, less precise than using the overall mean as an imputation for the missing values.

The BHPMF procedure was performed in R ver. 3.4.4 (www.r-project.org), which is the version required to run the 'GapFilling' function from the 'BHPMF' R-package (Schrodt et al. 2015, https://github.com/fisw10/BHPMF). The MICE-PMM procedure was performed by running the 'mice' R-package (van Buuren and Groothuis-Oudshoorn 2011) in R ver. 4.4.0 (www.r-project.org). The Rphylopars procedure was performed by running the 'Rphylopars' R-package in R ver. 4.4.0 (www.r-project.org), assuming a Brownian motion evolutionary model. The phylogenetic tree was gotten by running 'rtrees' R-package (Li 2023), based on the Smith and Brown (2018) plants megatree.

Effect of the out-off-context use of the imputed values

From the previous step (section 'Leave-one-out procedure'), we selected the imputation method producing the most precise and accurate imputed values. We assessed the effect of the out-of-context use of the imputed values by comparing: 1) the trait community weighted means estimated from the actual versus imputed values; 2) the position of the species projected onto a pre-defined phenotypic space.

Table 1. Estimation of the accuracy (intercept and slope of the linear regression of imputed values as function of the empiric value, both $\log_{10}$-transformed), and precision (Q and zRMSE) of the imputation with respect to empirical values in the species level Baraloto's datase, for each trait. H: adult plant height; SSD: stem specific density; LA: leaf area, LMA: leaf mass per area; N_{mass}: nitrogen content per unit leaf mass. In bold, are the best values for each trait and statistics.

Trait	Imputation method	Intercept [95% CI]	Slope [95% CI]	Q	zRMSE
H	BHPMF	**0.944 [0.858; 1.029]**	**0.280 [0.214; 0.345]**	−0.006	1.587
	MICE-PMM	1.003 [0.942; 1.064]	0.233 [0.187; 0.280]	0.163	0.914
	Rphylopars	1.089 [1.044; 1.135]	0.163 [0.129; 0.198]	**0.164**	**0.913**
SSD	BHPMF	**−0.117 [−0.132; −0.102]**	**0.427 [0.360; 0.493]**	0.162	1.491
	MICE-PMM	−0.177 [−0.188; −0.167]	0.160 [0.115; 0.206]	0.059	0.969
	Rphylopars	−0.129 [−0.139; −0.119]	0.391 [0.348; 0.435]	**0.414**	**0.765**
LA	BHPMF	**1.875 [1.620; 2.130]**	**0.495 [0.428; 0.563]**	0.214	1.679
	MICE-PMM	3.087 [2.917; 3.258]	0.169 [0.123; 0.214]	0.070	0.963
	Rphylopars	2.282 [2.119; 2.445]	0.388 [0.345; 0.431]	**0.408**	**0.769**
LMA	BHPMF	**1.180 [1.053; 1.308]**	**0.402 [0.337; 0.467]**	0.158	1.338
	MICE-PMM	1.244 [1.145; 1.343]	0.367 [0.317; 0.417]	**0.308**	**0.831**
	Rphylopars	1.595 [1.530; 1.661]	0.189 [0.156; 0.222]	0.215	0.885
N_{mass}	BHPMF	**0.626 [0.537; 0.715]**	**0.528 [0.461; 0.594]**	0.313	1.484
	MICE-PMM	0.913 [0.841; 0.985]	0.314 [0.260; 0.368]	0.225	0.879
	Rphylopars	0.775 [0.716; 0.834]	0.415 [0.371; 0.460]	**0.453**	**0.738**

We computed the trait community weighted means from imputed versus the actual species trait values for a set of 1000 simulated communities. Communities were simulated by sampling the species list of Baraloto's dataset. The number of species in each community ranges from 10 to 100 with a random uniform distribution. The relative abundance of each species in each community follows a power law distribution with an alpha parameter belonging to a random uniform distribution ranging from 1.5 to 4. This results in a set of communities with richness ranging from 10 to 100 and Shannon evenness ranging from 0.009 to 0.999. The power law distribution was simulated by running the*rpldis* function of the 'poweRlaw' R-package (Gillespie 2015). We fitted an ordinary least squares linear regression of the community weighted means from imputed versus actual species trait values to asses the effect of imputation on the variable. We assumed the empirical values were accurately measured and their error was negligible compared to the error introduced by the imputed values in the community weighted means calculus.

We defined a phenotypic space by performing a principal component analysis with Díaz's dataset. We kept the same traits as in Baralotos's dataset, this is plant height, stem specific density, leaf area, leaf mass per area and mass based leaf nitrogen content. These variables were $\log_{10}$-transformed and then standardized before running the *prcomp* function of the 'stats' R-package (www.r-project.org). Then we compute the position of each species in Baraloto's dataset onto phenotypic space defined by the first and second axes of the previous ordination analysis. To do so, we $\log_{10}$-transformed the imputed and the actual trait values, then we subtracted the mean and divided by the standard deviation of each $\log_{10}$-transformed trait from Díaz's dataset. Once we have scaled the trait values according to the pre-defined phenotypic space we compute the actual and imputed position of Baraloto's dataset species by a matrix multiplication of the vectors of scaled trait values by the eigenvector of principal components one and two. Finally, we fitted an ordinary least squares linear regression of the imputed versus actual PC1 and PC2 position of the species to assess the effect of imputation on these variables. We assumed the empirical values were accurately measured and their error was negligible compared to the error introduced by the imputed values in the PC1 and PC2 position.

Effect of redundancy on the imputed values

Finally, we illustrated how the dataset redundancy affected the imputation in our case. The Bayesian part of the method BHPMF sets the '*prior*' according to the mean value of the traits in the hierarchical grouping levels of the cases in the matrix. So, for each trait, we computed the relative range amplitude (i.e. the proportion of the total range amplitude of the dataset represented in each grouping level). The larger the relative range amplitude, the less precise the imputed values will be because of a less informative '*prior*'. To perform these analyses, we computed the relative error of the imputed values with respect to the actual values ([imputed – empiric] × empiric^{-1}) and constructed a linear model to analyse whether the relative range amplitude for a given trait in the smallest grouping level (genus or species) has an effect on the relative error of the imputed values.

To performe these analyses we applied the leave-one-out procedure with the BHPMF imputation method on both, the individual- and the species-level Baraloto's datasets. At the level of individuals we followed two strategies to vary the level of redundancy. The first one proceeded exactly as described

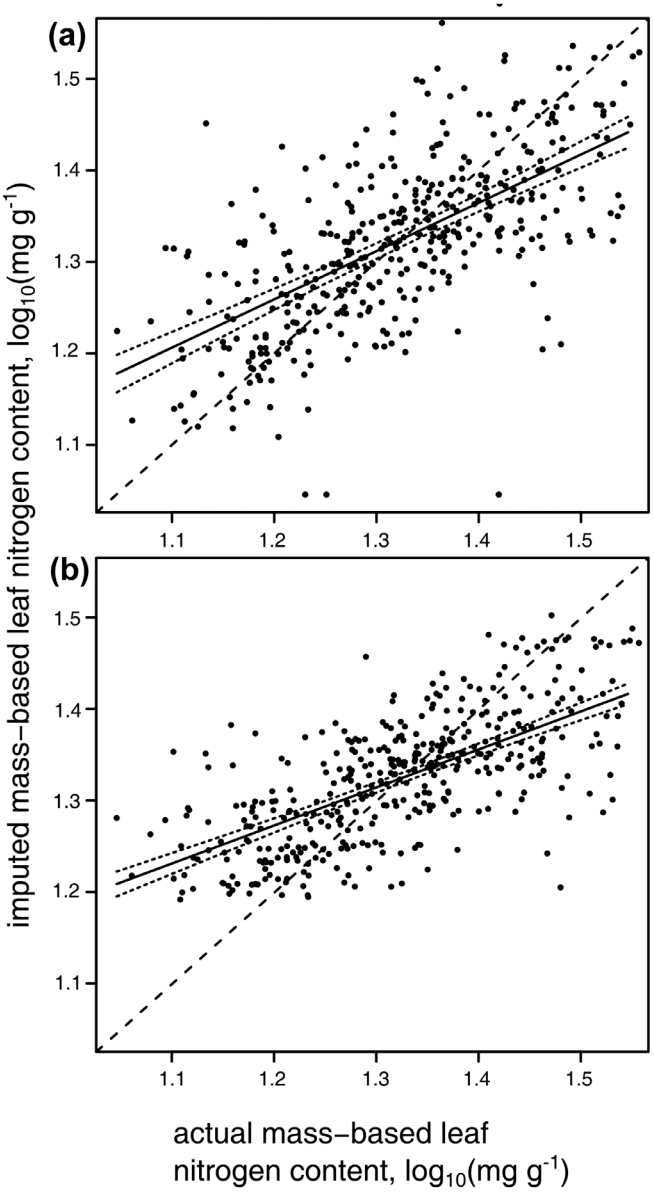

Figure 1. Relationship between imputed and actual mass-based leaf nitrogen content values for the species level Baraloto's dataset. Imputed values were obtained by performing (a) the BHPMF method, and (b) the Rphylopars method. The dashed line represents the identity line (intercept = 0; slope = 1). The continuous and dotted lines represent the fitted ordinary least-squares linear regression line and its 95% confidence interval.

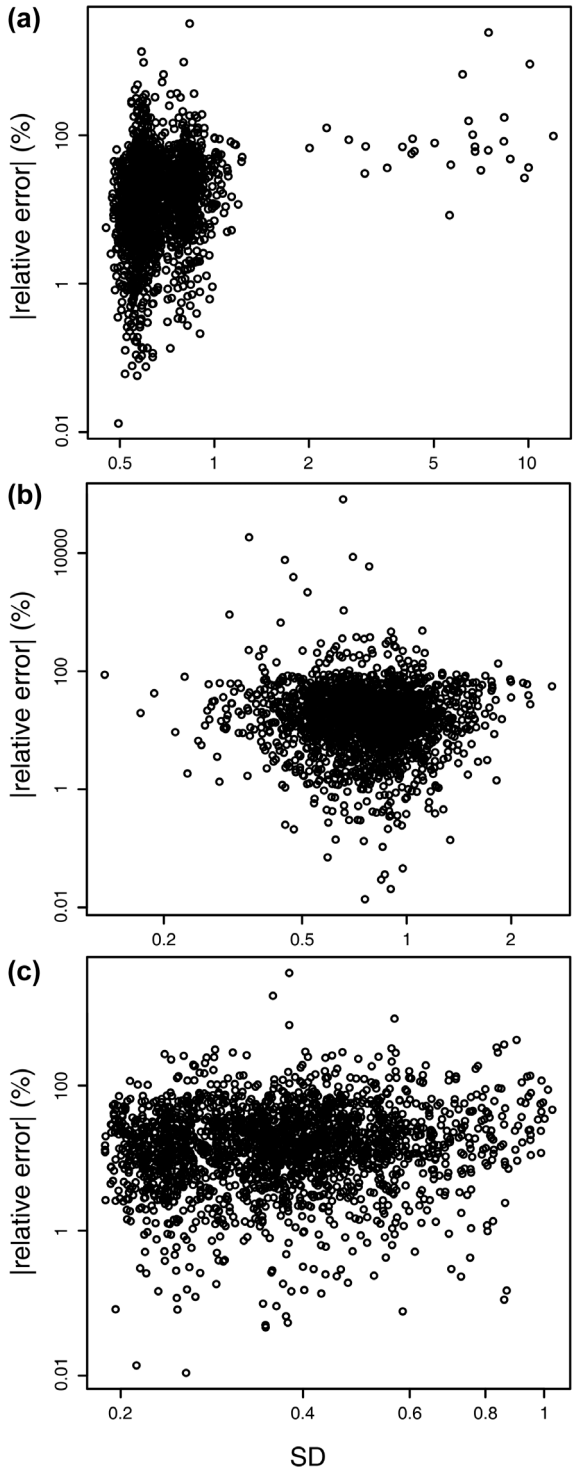

Figure 2. From the species level Baraloto's dataset, bivariate distribution of the absolute value of the relative error in the imputed values and the related uncertainty. (a) imputed values according to the mean value obtained by performing the BHPMF method and SD according to the standard deviation predicted by the same procedure. (b) imputed values and SD obtained by performing the mice-PMM method 10 times and computing the mean and standard deviation from these values. (c) imputed values obtained by performing the Rphylopars method and SD computed as the squared root of the estimated phenotypic variance.

above in the section 'Leave-one-out procedure' (this is the individual level). In the second strategy we removed not only the measurement we aim to test but also all the additional information for the particular combination of species and trait being imputed in each step (this is the individual2 level). That is, we kept the within-species information (replicates) for all traits except the one being tested. When applying the imputation once to each full database (species- and individual- level), the outliers were constrained as previously described.

Results and discussion

Precision and accuracy of the trait imputed values

For the species level Baraloto's dataset, the three imputation methods tested here produced imputed values accounting for 0–45.3% of the proportion of the total variance in each trait (Table 1). The BHPMF imputation method exhibited the lowest precision across all traits, with zRMSE values exceeding one for every trait, indicating that the imputed values were, on average, less precise than the overall mean. Conversely, the Rphylopars method showed the highest precision for all traits except LMA, where mice-PMM proved to be more precise (Table 1). Furthermore, BHPMF was the only method that generated imputed values out of the observed empirical range of each trait in the dataset (Supporting information).

The three imputation methods tested here produced biased trait values. In four (H, LA, LMA and N_{mass}) of the five traits, the intercept of the regression of imputed as a function of empirical values was > 0 and the slope was < 1 (Table 1). This means that the imputed values at both extremes of the range were biased, overestimating trait values at the lower extreme and underestimating them at the higher extreme. For SSD, the intercept was < 0 and the slope was < 1 (Table 1), meaning that this trait was systematically underestimated by the three imputation methods. The less biased method was BHPMF, this is the intercepts were closer to zero and slopes closer to one for the five traits (Table 1).

The imputed values for leaf nitrogen content were the most precisely and accurately imputed in the species level Baraloto's dataset (Table 1). Figure 1 shows the dispersion and bias of the imputed values in respect to the actual values.

We found that none of the estimations of variability/uncertainty associated with every single imputed value, according to the three methods tested here, was related to the relative error of each observation ($100 \times$ [|imputed value-actual value|/actual value]) (Fig. 2). As a consequence, nor the SD reported for every single value when performing the BHPMF method, nor the phenotypic variance estimated by Rphylopars, nor the variance associated with the multiple imputation by chained equations procedure are informative criteria to know the accuracy and precision of any particular imputed value.

Effect of the out-of-context use of imputed values

This section aims to illustrate our concerns about the out-of-context use of the imputed values. The results introduced

here are not an exhaustive assessment of the factors affecting the accuracy or precision of the computed variables. Because the method producing the most precise imputed values for most of the traits was Rphylopars, but BHPMF was the method producing the least biased imputed values, we explored the effect of the imputed values from both methods on the community-weighted trait means (CWM) calculus and on the projection of species onto a pre-defined phenotypic space (PCs).

In both sets of variables (i.e. the CWM and PCs) and for both imputation methods, the imputed values led to biased estimations (Fig. 3, Fig. 4, Supporting information). The intercepts of the regressions of imputed-derived values as a function of empirical values were > 0 and the slopes were < 1. This means that the imputed-derived values at both extremes of the range were biased, overestimating trait values at the lower extreme and underestimating them at the higher extreme. In line with Johnson et al. (2021),

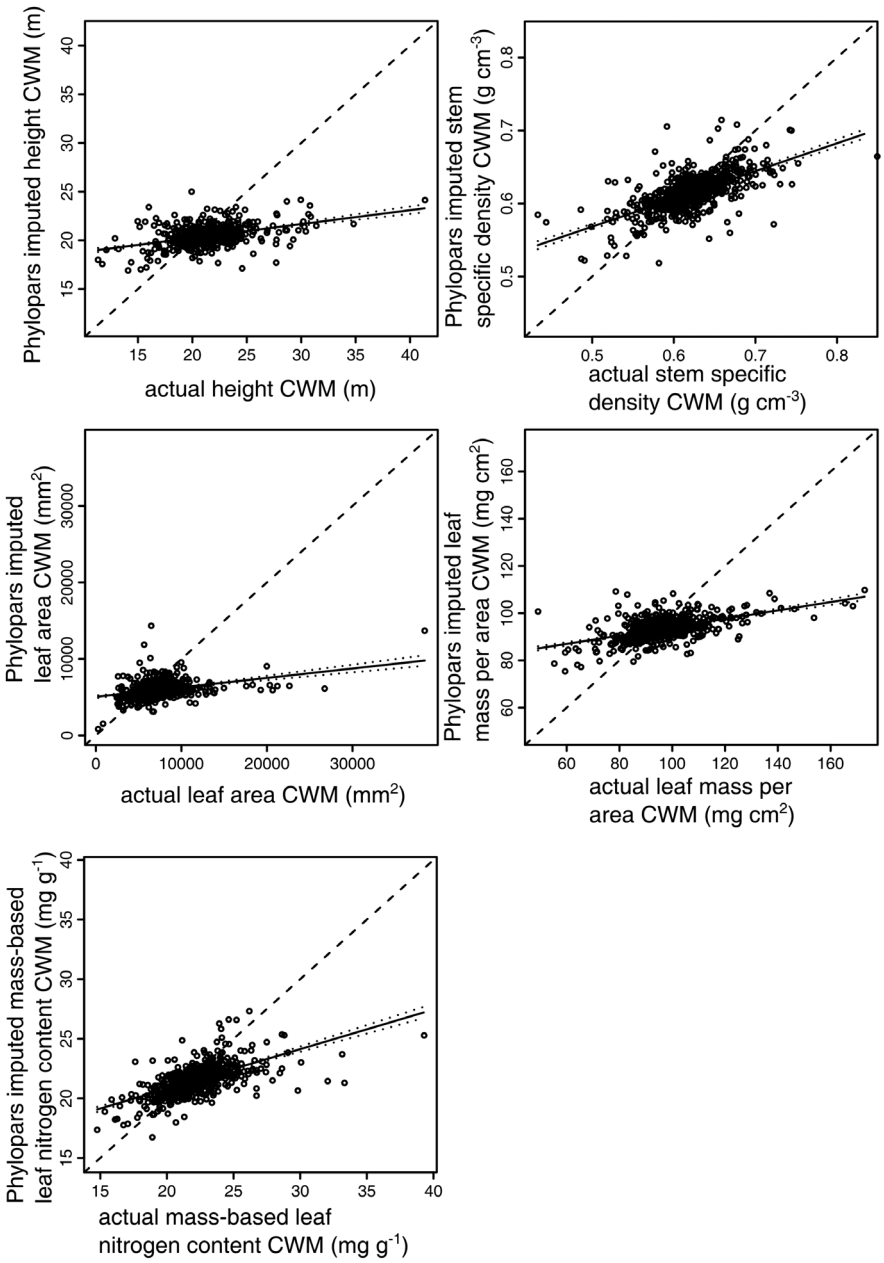

Figure 3. Relationship between community weighted means for each trait in the species-level Baraloto's dataset computed, for simulated communities, with the actual value of each trait versus the imputed values by the Rphilopars method. The dashed line represents the identity line (intercept = 0; slope = 1). The continuous and dotted lines represent the fitted ordinary least-squares linear regression line and its 95% confidence interval.

we found that imputation is not accurate enough to estimate trait values for individual species or records. As such, imputed values should be interpreted cautiously when any out-of-context use for these values is planned. Also, in agreement with Johnson et al. (2021), in our view the threshold for deciding on whether imputation is accurate depends on the research question.

The examples presented here are illustrative, but a more nuanced assessment is advised when using out-of-context imputed values. An imprecise or biased estimation could be tolerable in some contexts, depending on the magnitude of such errors. The leave-one-out procedure we used here is a tool that may help in that assessment (Fig. 5). In some cases, a picture of the imputed values' overall accuracy and precision could be enough to decide whether to use them. For a deeper evaluation, the frequency distribution of errors of the imputed values, obtained in the leave-one-out procedure, could be used in a randomization process specifically designed to assess the error propagation in each situation.

Redundancy and error of the imputed values

We suspected that the poor performance of the imputation may be due to limited redundancy (as defined in Box 1) in the matrix, due to the fact that the traits chosen to compute the global spectrum of plant form and function were deliberately selected to minimize redundancy (Díaz et al. 2016) and are thus weakly correlated overall.

To test how redundancy affects the accuracy and precision of the imputed values, we compare the outcome of the BHPMF method between the species-level and the individual-level Baraloto's dataset. For all traits, the imputation at the species level was less precise than simply using the global average for each trait (zRMSE >1). Additionally, the sum of square errors for H was larger than the actual variance of the trait ($Q < 0$). The precision and accuracy of the imputed values improved in the individual-level dataset (Table 2 – level: individual) but decreased sharply, becoming similar to the one at the species level, when information on the same trait for additional individuals of the same species was removed (Table 2 – level: individual2). This strongly suggests that indeed the degree of redundancy in the empirical matrix is key to obtain imputed values close to the empirical ones. In this case, there are at least two individuals for each species, and therefore there is a higher redundancy in the individual level dataset than in the species level dataset.

The BHPMF method uses the average value for the grouping levels (species in this case) as priors for the imputation. As a consequence, 1) the procedure applied in individual2 results in a situation similar to that of the species-level dataset (that is, we removed most of the redundant information relevant for the imputation); 2) the narrower the variability within a grouping level, the more informative the prior will be. As expected, we found that, for all traits, the relative error of the imputed values in the individual-level dataset was positively impacted by the relative range amplitude of the trait within each species. In other words, the wider the relative range of the trait within the species, the larger the relative error of the imputed values (Supporting information).

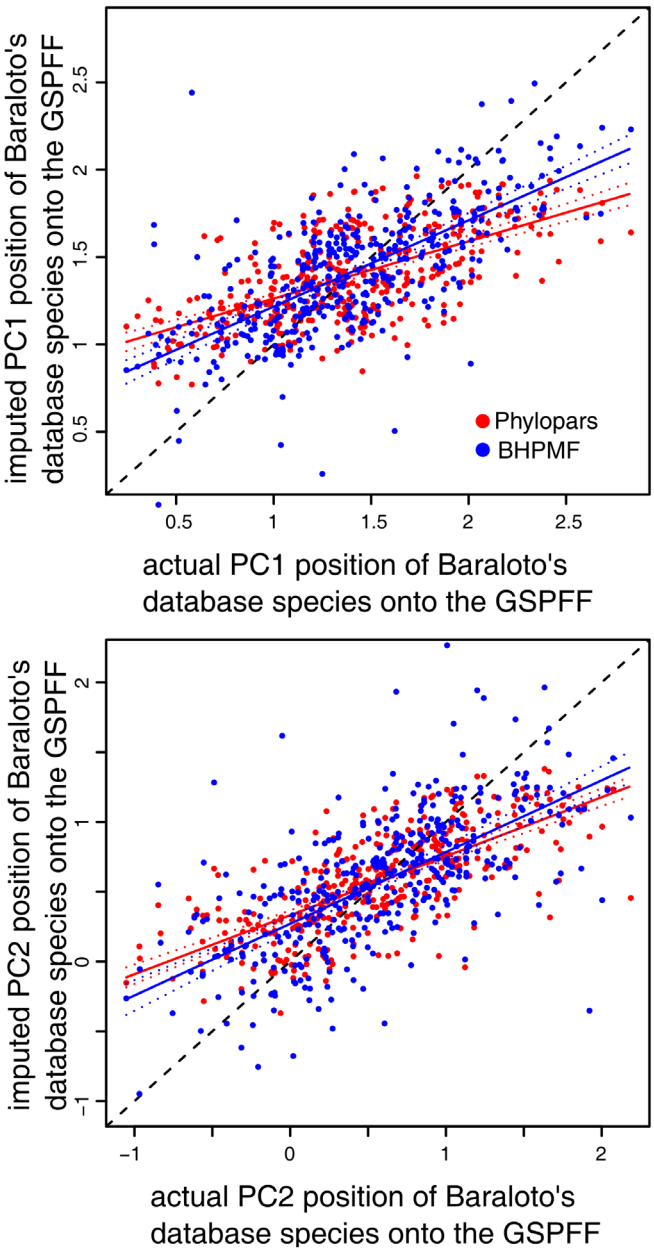

Figure 4. Relationship between the computed position of each species in the species-level Baraloto's dataset onto a pre-defined phenotipic space. (a) position onto the principal component 1 computed with the actual value of each trait versus the imputed values by the Rphilopars (red) and BHPMF method (blue). (b) position onto the principal component 2 computed with the actual value of each trait versus the imputed values by the Rphilopars (red) and BHPMF method (blue). The dashed line represents the identity line (intercept = 0; slope = 1). The continuous and dotted lines represent the fitted ordinary least-squares linear regression line and its 95% confidence interval.

It is important to highlight that the aggregated species-level Baraloto's dataset is almost complete (significantly less

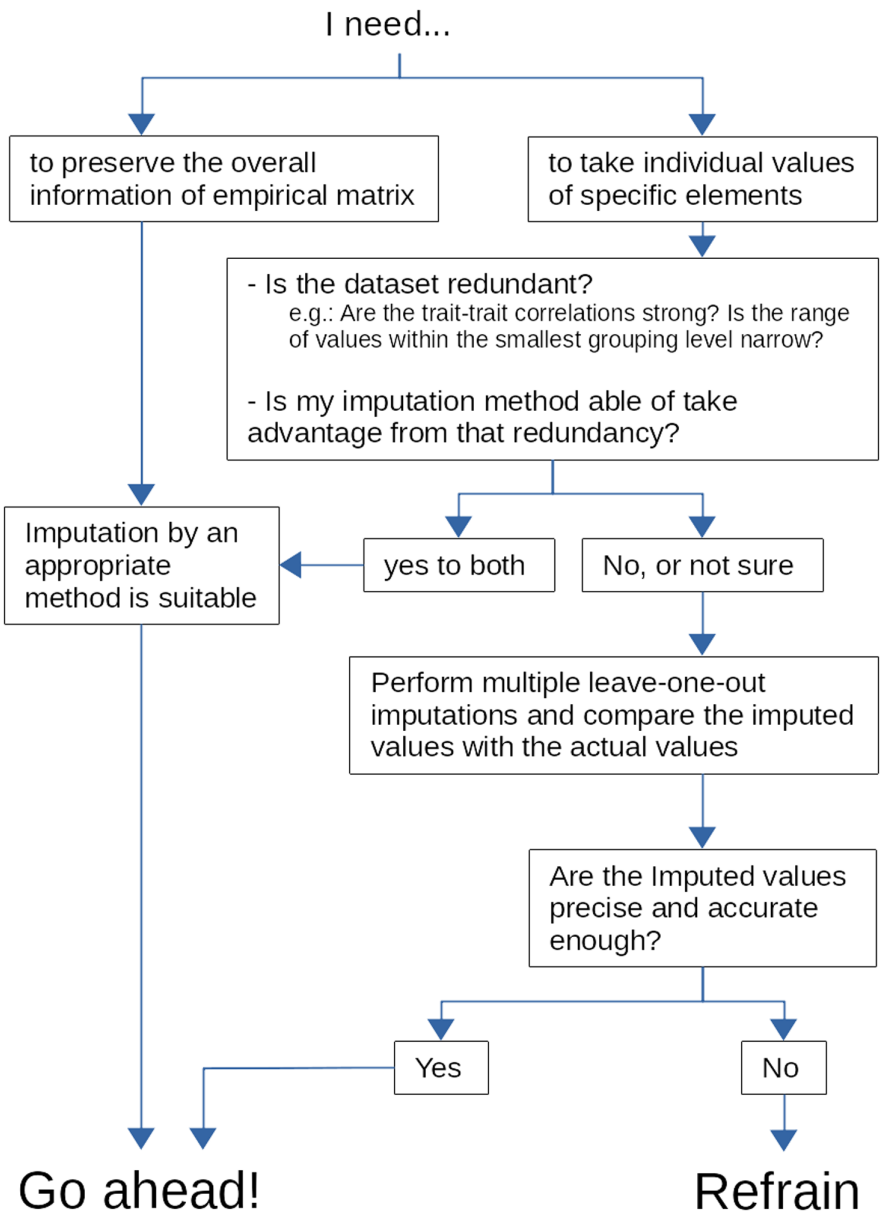

Figure 5. Decision tree. Some criteria to decide when we should be more cautious about imputing data in an entity by traits matrix.

sparse than what most people use for their imputation work in recent macroecological studies). As a consequence, the performance of the imputation is not the result of a particularly sparse matrix.

Final remarks

When combining datasets from incomplete sources, imputation helps to increase the number of species and traits to be considered when testing ecological and biogeographical questions. However, no imputation method produces new information; they all work with the information already implicit in the empirical matrix. Our study illustrates how the lower the redundancy in the empirical trait matrix, the less reliable the imputed values will be, and what consequences may derive from an out-of-context use of such imputed values. The accuracy and precision of the imputed values will be high only if the dataset is highly redundant in information, and the chosen method is able to take advantage of this redundancy. Providing external information (i.e. data from additional entities belonging to the same hierarchical groups, or information on other traits correlated with those of interest) may improve the quality of the imputation (Joswig et al. 2023).

Rather than undermining or supporting a particular method of imputation, our procedure provides a useful test of the quality of the imputation achieved in specific cases (Fig. 5). In line with this, we offer a practical procedure to check the precision and accuracy of imputed values in contexts in which the use of imputation goes beyond the scope

Table 2. Estimation of the accuracy (intercept and slope of the linear regression of imputed values as a function of the empiric value, both log_{10}-transformed), and precision (Q and zRMSE) of the imputed values in the species and individual-level Baraloto's dataset, for each trait. Individual2 is the individual-level Baraloto's dataset with a variation in the leave-one-out procedure. Here, we removed the measurement we aim to test and all the additional information for the particular combination of species and traits being imputed in each step. That is, we kept the within-species information (replicates) for all traits except the one being tested. H: adult plant height; SSD: stem specific density; LA: leaf area, LMA: leaf mass per area; N_{mass}: nitrogen content per unit leaf mass.

Trait	Level	Intercept [95% CI]	Slope [95% CI]	Q	zRMSE
H	Species	0.944 [0.858; 1.029]	0.280 [0.214; 0.345]	−0.006	1.587
	Individual	**0.674 [0.651; 0.696]**	**0.471 [0.454; 0.489]**	**0.375**	**0.864**
	Individual2	1.067 [1.039; 1.095]	0.165 [0.143; 0.187]	−0.230	1.269
SSD	Species	−0.117 [−0.132; −0.102]	0.427 [0.360; 0.493]	0.162	1.491
	Individual	**−0.081 [−0.086; −0.077]**	**0.613 [0.593; 0.633]**	**0.518**	**0.728**
	Individual2	−0.141 [−0.147; −0.136]	0.333 [0.310; 0.357]	0.108	1.137
LA	Species	1.875 [1.620; 2.130]	0.495 [0.428; 0.563]	0.214	1.679
	Individual	**0.887 [0.850; 0.923]**	**0.763 [0.753; 0.773]**	**0.766**	**0.490**
	Individual2	2.201 [2.139; 2.263]	0.405 [0.388; 0.422]	0.134	0.956
LMA	Species	1.180 [1.053; 1.308]	0.402 [0.337; 0.467]	0.158	1.338
	Individual	**0.847 [0.823; 0.873]**	**0.570 [0.557; 0.583]**	**0.520**	**0.735**
	Individual2	1.311 [1.278; 1.344]	0.338 [0.321; 0.355]	0.028	1.048
N_{mass}	Species	0.626 [0.537; 0.715]	0.528 [0.461; 0.594]	0.313	1.484
	Individual	**0.372 [0.348; 0.396]**	**0.714 [0.696; 0.732]**	**0.679**	**0.566**
	Individual2	0.748 [0.714; 0.781]	0.429 [0.403; 0.454]	0.194	0.944

for which it was initially designed (i.e. 'extension' applications). We suggest this is a key tool for researchers to decide, in the light of their specific objectives, whether such imputation meets the standards needed and thus can be used as a basis for further analysis. For example, if the whole matrix is the main focus of interest, as when performing multivariate analyses to detect trait syndromes on their own or to correlate the main axes of such space with environmental or ecosystem-level variables, it would be better to use imputation than work only with complete cases. This is because a properly specified imputation method, even if biased or imprecise in some specific values, will prevent the large loss of information involved in excluding all incomplete cases and will lead to less biased conclusions (Madley-Dowd et al. 2019). However, much more caution is in order when using imputation in cases where the focus is on individual elements of those matrices, as in the examples we provided above (i.e. using an imputation method to predict specific values).

Finally, we argue that, in order to prevent the uncritical use of imputed values and their propagation and potential misuse in the scientific literature, it would be good practice to avoid publishing gap-filled datasets. Much better would be to provide only the empirical data, together with the imputation method applied and the corresponding script to reproduce the imputation. In this way, any user should be able to reproduce the analyses but in full knowledge of which data were actually measured, and which ones imputed. Crucially, such practice would avoid the spread of imputed data, whose accuracy or precision might be poor or in any case difficult to assess, into the public domain.

Acknowledgements – We want to thank: to Jens Kattge for his valuable, thorough and selfless contribution to this manuscript; to Jon Lloyd, Sandra Patiño, Hans ter Steege, Bruno X. Pinho, Christopher Baraloto, Jos Barlow, Sandra Muller, Nancy Garwood, Ian Wright, Frans Bongers, Joseph Wright, Adriana Prieto, Armando Torres-Lezama, David Neill, Eurídice Honorio Coronado, Luzmila Arroyo, Nelson Miranda, Rodolfo Vasquez Martinez, Lina Mercado, Ima Célia Guimarães Vieira, Julio Serrano and Nigel Pitman for their contribution of empirical trait information; to Aurora Levesley for contributing to the development of this work; and to the Subject Editor Dr Carsten Dormann for valuable comments.

Funding – The present work is part of the ARBOLES project, funded by NERC (NE/S011811/1). This project was supported by ForestPlots.net approved Research Project no. 139: "How are Neotropical Forests placed in the global spectrum of plant form and function". Some data used in the analyses were collected with the support of the Brazilian National Council for Scientific and Technological Development (CNPq)/Long Term Ecological Research Project (PELD), Proc. 441244/2016-5 and 441572/2020-0. TreeCo trait data compilation was supported by grant 2013/08722-5, São Paulo Research Foundation (FAPESP). JAG was funded by the Natural Environment Research Council (NERC; NE/T011084/1) and the Oxford University John Fell Fund (10667). He is part of the National Systems of Researchers (SNI level 1) from the CONAHCYT, Mexico.

Author contributions

Lucas Damián Gorné: Conceptualization (lead); Data curation (lead); Formal analysis (lead); Investigation (lead); Methodology (lead); Validation (lead); Visualization (lead); Writing – original draft (lead). **Jesús Aguirre-Gutiérrez**: Data curation (equal); Writing – review and editing (equal). **Fernanda C. Souza**: Data curation (equal); Writing – review and editing (equal). **Nathan G. Swenson**: Data curation (equal); Writing – review and editing (equal). **Nathan Jared Boardman Kraft**: Data curation (equal); Writing – review and editing (equal). **Beatriz Schwantes Marimon**: Data curation (equal); Writing – review and editing (equal). **Timothy R. Baker**: Data curation (equal); Writing – review and editing (equal). **Renato A. Ferreira de Lima**: Data curation (equal); Writing – review and editing (equal). **Emilio Vilanova**: Data curation (equal); Writing – review and editing

(equal). **Esteban Álvarez Dávila**: Data curation (equal); Writing – review and editing (equal). **Abel Monteagudo Mendoza**: Data curation (equal); Writing – review and editing (equal). **Gerardo Rafael Flores Llampazo**: Data curation (equal); Writing – review and editing (equal). **Rubens Manoel dos Santos**: Data curation (equal); Writing – review and editing (equal). **Gerhard Boenisch**: Data curation (equal); Writing – review and editing (equal). **Alejandro Araujo-Murakami**: Data curation (equal); Writing – review and editing (equal). **Gonzalo Rivas-Torres**: Data curation (equal); Writing – review and editing (equal). **Hirma Ramírez-Angulo**: Data curation (equal); Writing – review and editing (equal). **Nayane Cristina dos Santos Prestes**: Data curation (equal); Writing – review and editing (equal). **Paulo S. Morandi**: Data curation (equal); Writing – review and editing (equal). **Sabina Cerruto Ribeiro**: Data curation (equal); Writing – review and editing (equal). **Wesley Jonatar A. da Cruz**: Data curation (equal); Writing – review and editing (equal). **Mathias Disney**: Data curation (equal); Writing – review and editing (equal). **Anthony Di Fiore**: Data curation (equal); Writing – review and editing (equal). **Ben Hur Marimon-Junior**: Data curation (equal); Writing – review and editing (equal). **Ted R. Feldpausch**: Data curation (equal); Writing – review and editing (equal). **Yadvinder Malhi**: Funding acquisition (equal); Project administration (equal); Resources (equal); **Oliver L. Phillips**: Data curation (equal); Funding acquisition (equal); Project administration (equal); Resources (equal); Writing – review and editing (equal). **David Galbraith**: Funding acquisition (equal); Project administration (lead); Resources (equal); Writing – review and editing (equal). **Sandra Díaz**: Data curation (equal); Funding acquisition (equal); Project administration (equal); Resources (equal); Supervision (lead); Writing – review and editing (lead).

Transparent peer review

The peer review history for this article is available at https://www.webofscience.com/api/gateway/wos/peer-review/10.1111/ecog.07520.

Data availability statement

The present work was not built on new data. The sources of the used data are referenced in the Methods section.

Supporting information

The Supporting information associated with this article is available with the online version.

References

Baraloto, C., Paine, C. E. T., Patino, S., Bonal, D., Herault, B. and Chave, J. 2010a. Functional trait variation and sampling strategies in species-rich plant communities. – Funct. Ecol. 24: 208–216.

Baraloto, C., Paine, C. E. T., Poorter, L., Beauchene, J., Bonal, D., Domenach, A.-M., Hérault, B., Patiño, S., Roggy, J.-C. and Chave, J. 2010b. Decoupled leaf and stem economics in rainforest trees. – Ecol. Lett. 13: 1338–1347.

Bruelheide, H. et al. 2018. Global trait–environment relationships of plant communities. – Nat. Ecol. Evol. 2: 1906–1917.

Díaz, S. et al. 2016. The global spectrum of plant form and function. – Nature 529: 167–171.

Díaz, S. et al. 2022. The global spectrum of plant form and function: enhanced species-level trait dataset. – Sci. Data 9: 755.

Enders, C. K. 2022. Applied missing data analysis, 2nd edn. – Guildford Press.

Garnier, É., Cortez, J., Billès, G., Navas, M., Roumet, C., Debussche, M., Laurent, G., Blanchard, A., Aubry, D., Bellmann, A., Neill, C. and Toussaint, J.-P. 2004. Plant functional markers capture ecosystem properties during secondary succession. – Ecology 85: 2630–2637.

Gillespie, C. S. 2015. Fitting heavy tailed distributions: the poweRlaw package. – J. Stat. Softw. 64: 1–16.

Goolsby, E. W., Bruggeman, J. and Ané, C. 2017. Rphylopars: fast multivariate phylogenetic comparative methods for missing data and within-species variation. – Methods Ecol. Evol. 8: 22–27.

Jardim, L., Bini, L. M., Diniz-Filho, J. A. F. and Villalobos, F. 2021. A cautionary note on phylogenetic signal estimation from imputed databases. – Evol. Biol. 48: 246–258.

Johnson, T. F., Isaac, N. J. B., Paviolo, A. and González-Suárez, M. 2021. Handling missing values in trait data. – Global Ecol. Biogeogr. 30: 51–62.

Joswig, J. S., Kattge, J., Kraemer, G., Mahecha, M. D., Rüger, N., Schaepman, M. E., Schrodt, F. and Schuman, M. C. 2023. Imputing missing data in plant traits: a guide to improve gap-filling. – Global Ecol. Biogeogr. 32: 1395–1408.

Kattge, J. et al. 2020. TRY plant trait database – enhanced coverage and open access. – Global Change Biol. 26: 119–188.

Li, D. 2023. rtrees: an R package to assemble phylogenetic trees from megatrees. – Ecography 2023: e06643.

Little, R. J. A. 1988. Missing data adjustments in large surveys (with discussion). – J. Bus. Econ. Stat. 6: 287–296.

Madley-Dowd, P., Hughes, R., Tilling, K. and Heron, J. 2019. The proportion of missing data should not be used to guide decisions on multiple imputation. – J. Clin. Epidemiol. 110: 63–73.

Nakagawa, S. and Freckleton, R. P. 2008. Missing inaction: the dangers of ignoring missing data. – Trends Ecol. Evol. 23: 592–596.

Palma, E., Vesk, P. A. and Catford, J. A. 2022. Building trait datasets: effect of methodological choice on a study of invasion. – Oecologia 199: 919–935.

Penone, C., Davidson, A. D., Shoemaker, K. T., Di Marco, M., Rondinini, C., Brooks, T. M., Young, B. E., Graham, C. H. and Costa, G. C. 2014. Imputation of missing data in life-history trait datasets: which approach performs the best? – Methods Ecol. Evol. 5: 961–970.

Schrodt, F., Kattge, J., Shan, H., Fazayeli, F., Joswig, J., Banerjee, A., Reichstein, M., Bönisch, G., Diaz, S., Dickie, J., Gillison, A., Karpatne, A., Lavorel, S., Leadley, P., Wirth, C. B., Wright, I. J., Wright, S. J., Reich, P. B. and Reich, P. B. 2015. BHPMF – a hierarchical Bayesian approach to gap-filling and trait prediction for macroecology and functional biogeography. – Global Ecol. Biogeogr. 24: 1510–1521.

Segrestin, J., Sartori, K., Navas, M. L., Kattge, J., Díaz, S. and Garnier, É. 2021. PhenoSpace: a Shiny application to visualize

trait data in the phenotypic space of the global spectrum of plant form and function. – Ecol. Evol. 11: 1526–1534.
Shannon, C. E. and Weaver, W. 1964. The theory of mathematical communication, 10th edn. – URBANA, the Univ. of Illinois Press.
Smith, S. A. and Brown, J. W. 2018. Constructing a broadly inclusive seed plant phylogeny. – Am. J. Bot. 105: 302–314.
Swenson, N. G. 2014. Phylogenetic imputation of plant functional trait databases. – Ecography 37: 105–110.
Swenson, N. G., Weiser, M. D., Mao, L., Araújo, M. B., Diniz-Filho, J. A. F., Kollmann, J., Nogués-Bravo, D., Normand, S., Rodríguez, M. A., García-Valdés, R., Valladares, F., Zavala, M. A. and Svenning, J. C. 2017. Phylogeny and the prediction of tree functional diversity across novel continental settings. – Global Ecol. Biogeogr. 26: 553–562.
van Buuren, S. and Groothuis-Oudshoorn, K. 2011. mice: multivariate imputation by chained equations in R. – J. Stat. Softw. 45: 1–67.
van Buuren, S., Boshuizen, H. C. and Knook, D. L. 1999. Multiple imputation of missing blood pressure covariates in survival analysis. – Stat. Med. 18: 681–694.