Received: 8 March 2022 | Revised: 20 March 2023 | Accepted: 14 April 2023

DOI: 10.1111/geb.13695

Global Ecology and Biogeography | A Journal of Macroecology | WILEY

METHOD

Imputing missing data in plant traits: A guide to improve gap-filling

Julia S. Joswig[1,2] | Jens Kattge[2,3] | Guido Kraemer[2,3,4,5] | Miguel D. Mahecha[2,3,4,5] | Nadja Rüger[3,6,7] | Michael E. Schaepman[1] | Franziska Schrodt[8] | Meredith C. Schuman[1,9]

[1]Department of Geography, University of Zurich, Zürich, Switzerland

[2]Max Planck Institute for Biogeochemistry, Jena, Germany

[3]German Centre for Integrative Biodiversity Research (iDiv) Halle-Jena-Leipzig, Leipzig, Germany

[4]Remote Sensing Centre for Earth System Research, University of Leipzig, Leipzig, Germany

[5]Helmholtz Centre for Environmental Research, Leipzig, Germany

[6]Department of Economics, University of Leipzig, Leipzig, Germany

[7]Smithsonian Tropical Research Institute, Ancón, Panama

[8]School of Geography, University of Nottingham, Nottingham, UK

[9]Department of Chemistry, University of Zürich, Zürich, Switzerland

Correspondence
Julia S. Joswig, Department of Geography, University of Zurich, Zürich, Switzerland.
Email: juliajoswigjj@gmail.com

Funding information
NSF national science foundation, Grant/Award Number: US NSF 20-508; NOMIS; Deutsche Forschungsgemeinschaft DFG, Grant/Award Number: RU1536/3-1

Handling Editor: Pedro Peres-Neto

Abstract

Aim: Globally distributed plant trait data are increasingly used to understand relationships between biodiversity and ecosystem processes. However, global trait databases are sparse because they are compiled from many, mostly small databases. This sparsity in both trait space completeness and geographical distribution limits the potential for both multivariate and global analyses. Thus, 'gap-filling' approaches are often used to impute missing trait data. Recent methods, like Bayesian hierarchical probabilistic matrix factorization (BHPMF), can impute large and sparse data sets using side information. We investigate whether BHPMF imputation leads to biases in trait space and identify aspects influencing bias to provide guidance for its usage.

Innovation: We use a fully observed trait data set from which entries are randomly removed, along with extensive but sparse additional data. We use BHPMF for imputation and evaluate bias by: (1) accuracy (residuals, RMSE, trait means), (2) correlations (bi- and multivariate) and (3) taxonomic and functional clustering (valuewise, uni- and multivariate). BHPMF preserves general patterns of trait distributions but induces taxonomic clustering. Data set–external trait data had little effect on induced taxonomic clustering and stabilized trait–trait correlations.

Main Conclusions: Our study extends the criteria for the evaluation of gap-filling beyond RMSE, providing insight into statistical data structure and allowing better informed use of imputed trait data, with improved practice for imputation. We expect our findings to be valuable beyond applications in plant ecology, for any study using hierarchical side information for imputation.

KEYWORDS
Bayesian hierarchical model, gap-filling, imputation, induced pattern, machine learning, matrix factorization, plant functional trait, sensitivity analysis, sparse matrix, TRY

1 | INTRODUCTION

Plant traits are characteristics of plants whose expression is influenced by their phylogeny, biotic (Navarro-Cano et al., 2021) and abiotic (Joswig et al., 2022) environmental factors that vary in space and time (Jetz et al., 2016) and trait–trait relationships (Díaz et al., 2016; Joswig et al., 2022; Thomas et al., 2020; Wright et al., 2004). They can be related to ecosystem functioning (Musavi

et al., 2015), bringing about understanding of how ecosystems may evolve under global change scenarios (Myers-Smith et al., 2019), about biosphere–atmosphere feedback mechanisms and on the influence of biodiversity on ecosystem processes (Díaz et al., 2007; Jetz et al., 2016; Reichstein et al., 2014; Shipley & Keddy, 2016). Many such studies rely on databases of in situ-collected trait data (Mohanraj et al., 2018; Tavşanoğlu & Pausas, 2018). The largest global meta-collection is the TRY database (Kattge et al., 2020).

The TRY database (version 6.0) contains 15,409,681 trait records of 2661 trait variables, mounting a large matrix. For each individual plant (represented by a row in the matrix), however, TRY has only usually few measured traits, leaving most cells in the matrix empty, covering only 0.1% (Kattge et al., 2020). This sparseness of observations limits the statistical power of analyses (Nakagawa & Freckleton, 2008), as well as the non-random nature of missingness. Using only measured traits for all individuals results in fewer individuals which can be included in multivariate analyses (Kattge et al., 2020; Shan et al., 2012) and deletion of non-randomly missing data (Johnson et al., 2020). In addition, sparse data may be biased when compared to fully observed data sets (Johnson et al., 2020; Kim et al., 2018; Nakagawa & Freckleton, 2008; Sandel et al., 2015).

Gap-filling, better termed imputation, is becoming a promising approach to handle sparse data (Johnson et al., 2020). Trait side information usually increases power, that is, reduces the error when imputing missing trait values (Poyatos et al., 2018). Trait data are currently filled by making use of three types of side information: (1) the species' relationships described by their taxonomy or phylogeny (e.g. species mean, Bayesian hierarchical probabilistic matrix factorization, BHPMF, Fazayeli et al., 2014; Schrodt et al., 2015, Rphylopars, Johnson et al., 2020); (2) plant trait–trait correlation structure (e.g. multiple imputation by chained equation, MICE, van Buuren & Groothuis-Oudshoorn, 2011, *k*-Nearest Neighbour, kNN, Dudani, 1976, BHPMF, Schrodt et al., 2015); or (3) trait–environment correlations (e.g. advanced hierarchical probabilistic matrix factorization, aHPMF; Schrodt et al., 2015).

BHPMF serves as one example of an upcoming paradigm of theory-guided initialization in the field of data science (Karpatne et al., 2017), generating data-based models which incorporate prior knowledge as side information (Barredo Arrieta et al., 2020). Probabilistic matrix factorization (PMF) is based on the assumption that the original matrix has correlated columns ('is of low rank') and can therefore be approximated and imputed by the product of two lower dimensionality matrices (Mnih & Salakhutdinov, 2007; Udell & Townsend, 2019). In case of BHPMF, the plant taxonomic hierarchy is added to the data matrix, based on the prior knowledge that plant traits cluster within taxonomic and functional groups (Kattge et al., 2011). While PMF is based on the distribution and correlation of data within the incomplete matrix, the added taxonomic hierarchy substantially improves the accuracy of imputation by BHPMF (Schrodt et al., 2015; Shan et al., 2012). The implementation of imputation by BHPMF has therefore become integral to the vision of linking different trait data streams for spatio-temporal monitoring of plant functional biodiversity (Jetz et al., 2016). BHPMF often performs well in comparison to other commonly used techniques (Fazayeli et al., 2014; Schrodt et al., 2015), particularly in large and very sparse data sets (Johnson et al., 2020) like the TRY database (Kattge et al., 2020). Therefore, BHPMF imputation and its data have been used in a wide array of studies: BHPMF-imputed data have been shown to produce comparable multivariate results to observed data in case studies (Díaz et al., 2016; Schrodt et al., 2015) and when used in an advanced form of trait–trait correlations: trait connectivity (Flores-Moreno et al., 2019). BHPMF has also been used to support the development of process-based range models for many species at different spatial scales (Evans et al., 2016). An initiative collecting species abundance data across vegetation plots (sPlot) aims at broadening its applicability by linking to BHPMF-imputed trait data, derived from TRY (Bruelheide et al., 2019). This large data set is used to analyse community trait–environment relationships (Bruelheide et al., 2018).

However, the limitations of BHPMF-imputed data are still not well understood (Poyatos et al., 2018). Because BHPMF learns from taxonomic side information and trait–trait correlation patterns (matrix factorization, Schrodt et al., 2015), it may introduce biases. Any additional (side) information that reduces the error during imputation is likely to also introduce bias because of the assumptions incorporated into the imputation algorithm. Whereas PMF-based imputation may strengthen trait–trait correlation patterns. In detail, imputed values may systematically deviate according to taxonomic or correlation patterns, or both. If introduced into data during imputation, these artificial patterns of taxonomic and trait–trait correlation biases could then lead to false conclusions, for example, in studies testing for taxonomic differences. Such biases cannot be detected in the current evaluation of BHPMF. This is because BHPMF, like most imputation techniques, optimizes against observed data using measures of imputation accuracy (here the root mean squared error, RMSE), but does not evaluate induced versus observed patterns of taxonomy or trait–trait relationships. Any additional (side) information that reduces the error during imputation is likely to also introduce bias because of the assumptions incorporated into the imputation algorithm.

Evidence of bias emerges as systematic patterns in residuals, the distance of imputed to observed values ($y_{\text{imputed}} - y_{\text{observed}}$), are evidence of bias. Residuals may be non-randomly distributed due to information added during imputation. Molina-Venegas et al. (2018) analysed variability in plant trait imputation accuracy in relation to phylogeny using both Brownian Motion and Monte Carlo approaches. They showed that predictive accuracy depends on the strength of the phylogenetic signal in traits, producing variable levels of imputation accuracy among phylogenetic taxa. This suggests that well-conserved species (and traits) are well explained, while trait samples, which are outliers within their taxon, are less accurately predicted. This may be because imputation moves all samples within a taxon towards the taxon mean. The second potential source of bias stems from the other

side information, that is, trait–trait correlation patterns that are retrieved by BHPMF directly from the data.

Functional groups—coarse approximations of plants' traits—are often used to draw broad-scale conclusions, for example, for trait mapping (Moreno-Martínez et al., 2018). Despite their importance in research, bias in plant functional types and growth forms are not explicitly considered in bias detection analyses. Therefore, implications of induced patterns during gap-filling specifically for plant functional types and growth forms may be of importance for these analyses. Functional groups can sometimes be related to taxonomy (e.g. pteridophytes and ferns), but may also represent different categories (e.g. Fabaceae, including trees and herbs).

The imputation accuracy may be increased by adding an external larger plant trait data set during imputation. Schrodt et al. (2015) point to this potential when imputing a (geographically constrained) subset with and without external data. Still both approaches resulted in similar imputation errors (SD and RMSE) and were not tested for induced patterns. A greater available data set during imputation may help to stabilize trait predictions and ameliorate biases when a local data set has been collected as a focal study with sparse traits recorded. However, the extended data can also bias imputation, which may be reflected in terms of imputation error, trait distributions, taxonomic clustering and trait–trait correlations.

Here, we investigate whether data imputation by BHPMF leads to biases, and which aspects influence the bias. We aim to demonstrate bias in imputed data using BHPMF. We expect (H1) that BHPMF-filled matrices may be of lower rank than the original matrices: Trait–trait correlations would be increased after imputation; (H2) that the taxonomic side information added within BHPMF may strengthen or even introduce taxonomic patterns; and (H3) trait data which were imputed with external data to have smaller residuals and a reduced bias.

We used an observational data set drawn from the TRY database (https://try-db.org, Kattge et al., 2020), and randomly removed different numbers of samples from it to achieve several levels of missingness. After imputations, we compared observed and imputed data sets in terms of (1) error (RMSE, distribution, trait means, residuals of individual values and species means), (2) trait–trait correlations (Pearson correlation coefficient, principal component analysis [PCA], Procrustes test) and (3) taxonomic and functional clustering (silhouette index, distance to species mean). We describe the differences between observed and imputed trait data as 'deviations'.

2 | MATERIALS AND METHODS

2.1 | Data

The data used in this study are based on fully observed trait data and BHPMF-imputed trait data with different sparsity on the largest completely observed in situ trait data collection that could be derived from the TRY database. The traits were selected based on their number of entries. We chose to use observed rather than synthetic trait data to represent the caveats of trait data as realistically as possible. This comes at the cost of using available observations only as well as relying on the inherent bias from data sets. Hence, the completely observed trait data comprise a set of non-randomly selected traits. In order to test for coherence, and influence of different properties, we selected two completely observed trait data sets (OBS, OBS2) which are part of the same global data set (TRY17), but do not share any overlapping entries. Results of OBS are presented in the main text, and those from OBS2 are in the supplement (for a summary of data used, see Figure 1, and Tables S1 and S2).

To obtain the trait data sets, we first extracted records for the 17 most frequently observed continuous traits from the TRY database (try-db.org, 8.10.2016 and TRY version 3). In total, we retrieved traits from 241,653 individual plants. The resulting trait–individual matrix has a sparsity of 93.3%, that is, only 6.7[%] of cells contain a trait record (Table S3). Information on genus, family and phylogenetic group, growth form and plant phenology was added from the TRY—Categorical Traits Data set (https://www.try-db.org/TryWeb/Data.php#3; Table S4). The individuals were attributed to taxonomic (species, genera, families and seed plant clades: Angiosperm–Eudicotyl, Angiosperm–Magnoliid, Angiosperm–Monocotyl Pteridophytes, Gymnosperm) and functional groups (plant functional type [PFT] and growth form Table S4). Species and genera are completely nested within the functional groups.

In order to build the two test trait data sets (OBS, OBS2), we extracted from the total trait data (extended data set) two subsets with fully observed data only. These observed data comprise two collections each having the maximum available number of individual observations and traits (Table S3). OBS is a mainly tropical data set dominated by trees (OBS: n(trees) = 806 n(herbs) = 119, n(grasses) = 118). In contrast, OBS2 consists mainly of data from temperate regions, in which herbs and grasses are better represented than trees (OBS2: n(trees) = 28 n(herbs) = 390, n(grasses) = 119). The respective data stemming from the extended data are n(trees) = 65,474, n(herbs) = 78,561, n(graminoids) = 25,049. OBS and OBS2 have approximately the same number of observations per species (n(OBS) = 2.4; n(OBS2) = 2.2). The observations added to OBS and OBS2 by external data (excluding OBS, OBS2, respectively) vary, for example, per species present in OBS, EXT adds four observations on average, in comparison to OBS2, to which species EXT adds five to six observations on average (Table S3). The retained total data set including OBS and OBS2, together with the external trait data (EXT, EXT2), is called the extended trait data from TRY (TRY17). For a summary of all data sets, see Figure 1 and Table S1.

2.1.1 | Data transformation

Before running BHPMF, all trait data sets were normalized by log- and z–transformation. Specifically, each value y of a trait k was first

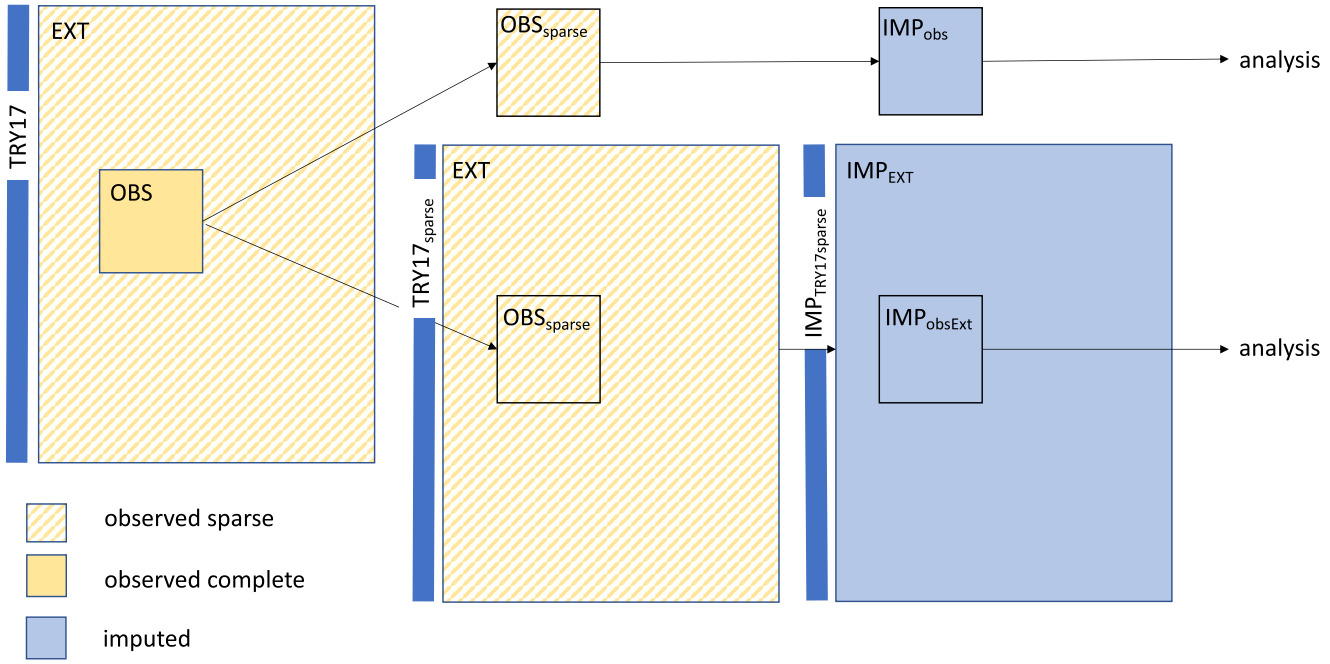

FIGURE 1 Data sets used and produced in this study. TRY17, a sparse global data set of 17 plant traits derived from TRY including OBS, a completely observed data set (for OBS2, the naming is accordingly with added '2'). OBS is then added with gaps (missingness from 1% to 80%), resulting in OBS_{sparse}. The same OBS_{sparse} is replacing OBS in TRY17, making it $TRY17_{sparse}$. OBS_{sparse} and $TRY17_{sparse}$ are being BHPMF imputed. This results in IMP_{obs} from OBS_{sparse} as well as $IMP_{TRYsparse}$, including IMP_{obsExt} and IMP_{EXT}. The imputations IMP_{obs} IMP_{obsExt} are then further analysed and compared to OBS.

log-transformed, then in a second step, the trait mean $\mu(\log(k))$ was subtracted, and the resulting value was divided by its standard deviation $\sigma(\log(k))$.

$$zlog(y) = \frac{\log(y) - \mu(\log(k))}{\sigma(\log(k))}$$

Log transformation was chosen to achieve a closer to normal distribution of values per trait (Kattge et al., 2011). We additionally conducted *z* transformation because a given difference for small trait values (absolute value) is likely to be physiologically more relevant than the same difference (absolute value) for large trait values. The *z* -log transformed data were used for all analyses of BHPMF imputations (unless specifically mentioned).

2.1.2 | Data preparation for BHPMF imputation

The gap-filling procedure was prepared by perforation (see below), then perforated data were BHPMF-imputed and analysed. The observed trait data set (OBS or OBS2) was perforated with varying numbers of missing entries: 1%, 5%, 10%, 20%, 30%, 40%, 50%, 60%, 70% and 80% of observed data were randomly deleted across the whole data set, with the constraint to keep at least one trait record per individual plant (i.e. row) and at least one observation per trait (i.e. column). In total, these minimum values represent 14% of the original, observed data set. For each level of missingness, we repeated the random gap setting three times. In preliminary analyses, repetitions of BHPMF imputations did not show significant differences in total RMSE and were thus not considered. For the imputation with external data, the perforated trait data (OBS_{sparse}) were complemented with external data (EXT).

2.1.3 | BHPMF imputation

BHPMF decomposes or factorizes the trait matrix probabilistically (probabilistic matrix factorization, PMF Salakhutdinov & Mnih, 2009) at different hierarchical levels (here: taxonomy) within a Bayesian framework (Schrodt et al., 2015). PMF probabilistically factorizes the trait matrix by latent vectors for each row and for each column of the matrix. The trait values are imputed as the inner products of the latent vectors. Using a Gibbs sampler (a Markov Chain Monte Carlo algorithm), BHPMF sequentially performs PMF at the different hierarchical levels, using the latent vectors of the PMF at the current hierarchical level as prior information for the next hierarchical level. BHPMF thus samples the higher level-constrained probability density distributions of the latent vectors at the level of the individuals. Eventually these iterations are used to derive imputation means as well as imputation confidence in the form of standard deviations (SD), which are per-value estimates of uncertainty in trait imputations (Fazayeli et al., 2014; Schrodt et al., 2015). The underlying premise of BHPMF is therefore to impute traits of the individual plants using PMF to account for trait–trait correlations as well as intra- and interspecific trait variability and using the taxonomic hierarchy to constrain the sampling of the spars individual-based trait

matrix by well-covered trait matrices at the higher levels of the taxonomic hierarchy (Fazayeli et al., 2014; Schrodt et al., 2015).

BHPMF internally splits the data sets randomly into a training data set for parameter setting (80%), a validation data set for parameter adjustment by optimizing performance and (10%) and a test data set for independently testing the performance after parameter adjustment and learning (10%, Schrodt et al., 2015). The training data set is used for the training of latent vectors, while the validation data set is used to evaluate and stop the optimization process of the latent vectors after five consecutive iterations with stable RMSE, and finally, the test data set serves as the basis for independent performance testing after parameter adjustment and learning (Schrodt et al., 2015).

The R package BHPMF was run with a maximum of 1000 total iterations, where the first 200 were discarded during the 'burn-in' phase. In order to avoid autocorrelation, only every 20th iteration was used to calculate the resulting trait values. The mean of these maximum 40 imputations resulted in the final trait values used as the output.

To determine the effect of adding the external trait data, we imputed the perforated observed data with and without the external trait data (Table S3). This resulted in four main data sets (OBS is used here to indicate both OBS and OBS2, lookup table for both approaches here: Figure 1, Table S1, and for OBS2, Table S2): (1) OBS, observed trait data (OBS_{sparse} if OBS is perforated), (2) IMP_{obs}, trait data imputed based on perforated trait data OBS_{sparse}, (3) TRY17, the extended trait data, including the trait data (OBS) and external data (EXT; TRY17 = OBS + EXT). $TRY17_{sparse}$, if included trait data are sparse (OBS_{sparse}), and (4) IMP_{obsExt} trait data imputed based on OBS_{sparse} with the extended trait data (EXT).

2.2 | Analysis

We computed the error (RMSE, mean, distribution, residuals), the trait–trait correlations (Pearson, PCA) and taxonomic and functional clustering (silhouette index, distance to cluster mean, coefficient of variation). The distance between the observed data (OBS) and the imputed data sets (IMP_{obs} and IMP_{obsExt}) are defined here as deviations.

2.2.1 | Error

The error was calculated for individual observations as residuals, and the distance of individual observations to cluster mean, whereas clusters are taxonomic (species, genus, family) or phylogenetic (clade) and functional (growth form [GF] and plant functional type [PFT]). Furthermore, the root mean square error (RMSE) was calculated per trait and for the whole data set, and distributions were plotted.

Residuals are calculated as the distance between (back-transformed) imputed value $y_{imputed}$ and observed value $y_{observed}$. $y_{residual} = y_{imputed} - y_{observed}$.

RMSE was calculated from transformed data.

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(y_{\mathrm{imputed}} - y_{\mathrm{observed}}\right)^2}$$

The trait distributions across individual values were calculated with the density function of the 'stats' package in R.

2.2.2 | Trait–trait correlations

Trait–trait correlation patterns were calculated for trait pairs and for multivariate trait sets.

For trait pairs, we calculated Pearson correlation coefficients of both the back-transformed and the respective $z - \log$ transformed data set. The Pearson correlation coefficients were calculated for all three imputation repetitions, and aggregated to mean and SD. For multivariate correlation patterns, we performed principal component analysis (PCA) with the R package 'princomp' for one random repetition of imputations. The kernel density was calculated with the 'kernel' package. For quantitative comparison of the PCA results of imputed (IMP_{obs}, IMP_{obsExt}) with those of the PCA result from the observed data set (OBS), we applied a Procrustes test (Peres-Neto & Jackson, 2001) using the 'procrustes' and 'protest' functions in R package 'vegan' (Oksanen et al., 2020). The procrustes analysis rotates the two PCA axes to maximize similarity with the other PCA axes to be tested against (function procrustes). Second, we test the non-randomness ('significance') between two PCA results (function protest). Non-random results (e.g. <0.05) for the first two PCA axes indicate non-randomness, thus similarity.

2.2.3 | Clustering

Clustering was analysed for single values, clusters of a trait and clusters of all traits (multivariate clusters). Clusters were defined by taxonomic and functional groups.

For the value-wise analysis of clustering, we calculated the distance to the respective cluster mean for each value y of a cluster A, ($\bar{y} - y$, z-log transformation). Clusters were species, genera, families, clades, growth forms (GFs), and plant functional types (PFTs).

For the traitwise (univariate) cluster analysis, the coefficient of variation (CV) was calculated per trait and cluster A (e.g. species) consisting of its single values $y_1, y_2, \ldots y_n$. The CV is calculated as the division of standard deviation σ of all observations of one cluster ($y_1, y_2, \ldots y_n$), n = number of values of cluster A, and the mean μ of all observations of the same cluster ($y_1, y_2, \ldots y_n$), n = number of values of cluster A. The CV of clusters with $n > 1$ was calculated. Values with missing functional group cluster attributions were excluded from analyses. The CV is based on back-transformed data.

$$\mathrm{CV} = \frac{\sigma}{\mu}$$

For calculating the multivariate cluster analysis, we used the silhouette index (Rousseeuw, 1987) Rpackage: 'clues', function: get_Silhouette. The silhouette index (S) calculates the dissimilarity of one element of a group to its cluster. First, we calculated for each object *y*, within any cluster *A* (e.g. individual within any taxon or functional group), its distance to all other objects y_n, independent of their cluster attribution. Second, we averaged the distances of all elements per cluster in two ways. On the one hand, we averaged the distances of all elements to the elements within the same cluster, resulting in the dissimilarity *a*(*y*). We also averaged the distances of the same cluster to all elements of the closest neighbouring cluster *b*(*y*). Finally, the dissimilarity *a*(*y*) within the cluster *A* is compared to the dissimilarity *b*(y) of the closest neighbouring cluster, so *a*(*y*) = average distance to objects of group A, while *d*(*y*) = average distance to objects of cluster ≠ A, for example, *C*, and *b*(*y*) = minimum (*d*(*y*, all clusters)).

$$S = \frac{b(y) - a(y)}{\max(a(y), b(y))}$$

A value of $S = -1$ indicates great dissimilarity (little clustering), $S = 0$ indicates that items are situated evenly between two clusters, and *S* close to 1 indicates little dissimilarity, such that items are taken to be part of the attributed cluster. Silhouette indices of clusters with one individual must result in $S = 0$ and thus were not calculated. Values with missing functional group cluster attributions were excluded from analyses. We extracted median values of all clusters of more than one individual per group (species, genera, families, seed plant clades, growth forms, PFTs). Only fully observed clusters can be used to calculate S, thus sparse and perforated data cannot be calculated.

3 | RESULTS

We compared the observed trait data (OBS) to the imputed trait data (IMP_{obs}, IMP_{obsExt}) either from OBS_{sparse} alone or on the basis of extended trait data ($TRY17_{sparse}$). All analyses were run in parallel for a second data set (OBS2, see supplementary material).

3.1 | Imputation error and residuals

In a first step, we analysed the patterns of error and tested whether missingness was associated with error for IMP_{obs} and IMP_{obsExt}, measured as RMSE and residuals per trait and per individual plant value. We found RMSE to increase with missingness (Figure 2) and to be higher for gaps than for available samples (Figure 2 top vs. bottom). This translates into non-gap data being better imputed than absent data (compare also OBS and OBS_{sparse} in Figure S2e).

Including external data during imputation did not change the RMSE much for IMP_{obs} to IMP_{obsExt} (Figure 2). However, $IMP2_{obsExt}$ showed a greater RMSE-reduction in comparison to $IMP2_{obs}$ (Figure S3). The addition of external data did not increase residuals in comparison to IMP_{obs} (Figure S4a) and reduced them for OBS2 (Figure S5a). Across taxa, addition of EXT kept the correlations of samples similar (OBS) or ameliorated them (OBS2, Figure S2e). For species mean, adding external data reduced the fit for OBS,

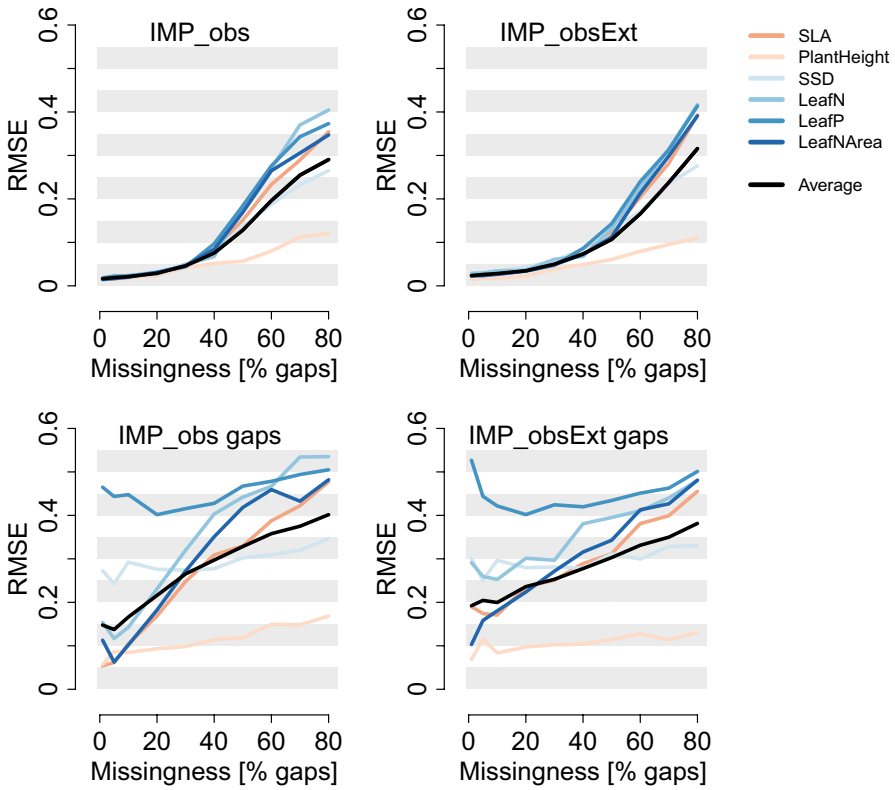

FIGURE 2 Percentage of missingness in trait data (OBS) impacts the classical estimator of imputation error: RMSE; while the extent of the input data set (with or without external data) does not. Lines connect the points at missingness of 0%, 1%, 5%, 10%, 20%, 30%, 40%, 50%, 60%, 70% and 80%. The points are average values of three repetitions while the missingness per trait is approximate, since gaps were set randomly. Top figures refer to the RMSE calculated from all values, that is, imputed values for both types of samples: the observed values as well as the introduced gaps. Both types of samples have an imputed equivalent. Bottom figures refer to the imputed gaps only. Left figures refer to input being trait data only, while figures on the right show results when external data are included during imputation. For illustration purposes, we show one repetition. For trait data 2, see Figure S3.

while for OBS2, it improved it (Figure S6), this external data effect depended on the trait (Figures S7 and S8). To better understand the implications for ecological analyses, we converted the imputations back to original units (Table S5). Single imputed trait values deviated substantially in comparison to the original observations. For example, converted into original units, residuals translate for plant height into a median deviation of 2.3 m (7.72 for 75th quantile), which may be as large as 43.4 m (Extended trait data: median = 2.54 m; 75th quantile = 6.9 m; max = 47.6 m, Table S5). This is interesting given plant height shows the lowest RMSE, when transformed. The extreme (original unit) residuals of specific leaf area (SLA), leaf nitrogen (leafN), leaf phosphorus (leafP) and leaf nitrogen per leaf area (LeafNArea) were reduced for IMP_{obsExt} in comparison to IMP_{obs} (Table S5).

3.2 | Distributions

To determine whether imputed trait data represented the trait distributions well, we compared these for trait means (Figure 3a,b), for single values (Figure S4a, Table S5) and for taxon means (Figures S6a and S4b–d).

The trait distributions of IMP_{obs} and IMP_{obsExt} (of both trait data sets) reproduced the observed distributions in OBS (OBS2) well, but with reduced variance (Figure 3a, OBS2 Figure S9a). IMP_{obsExt} distribution resembled the distribution of OBS more than the distribution of the TRY17 (i.e. OBS with external data, Figures 3a). Consistently, trait means were similar among the imputations IMP_{obs} and IMP_{obsExt}, and OBS than to the added external data (Figure 3b). However, some trait means were shifted: for example, tall plants become shorter than observed, and stem density (SSD) greater than observed.

When assessing the residuals at all taxonomic and phylogenetic mean levels (species to clades), we found adding supplement trait data did usually not change the residuals for OBS (Figure S4, but it did reduce them for OBS2, Figure S5). This was also reflected in the similar correlation coefficients between IMP_{obs} or IMP_{obsExt} with OBS (Figures S1, S2 and S6), but an amelioration for OBS2. It is worth mentioning that a perfect reproduction of the trait distribution and mean is theoretically possible even with a large RMSE.

3.3 | Trait–trait correlations

To determine how imputation affected trait–trait correlations, we analysed the trait–trait correlations of OBS, IMP_{obs} and IMP_{obsExt} on the level of pairwise (Figure 4a,b) and multivariate trait combinations (Figure 4c–e).

The mean Pearson correlation coefficients of pairwise correlation coefficients are reduced for IMP_{obs} and $IMP2_{obs}$ in comparison

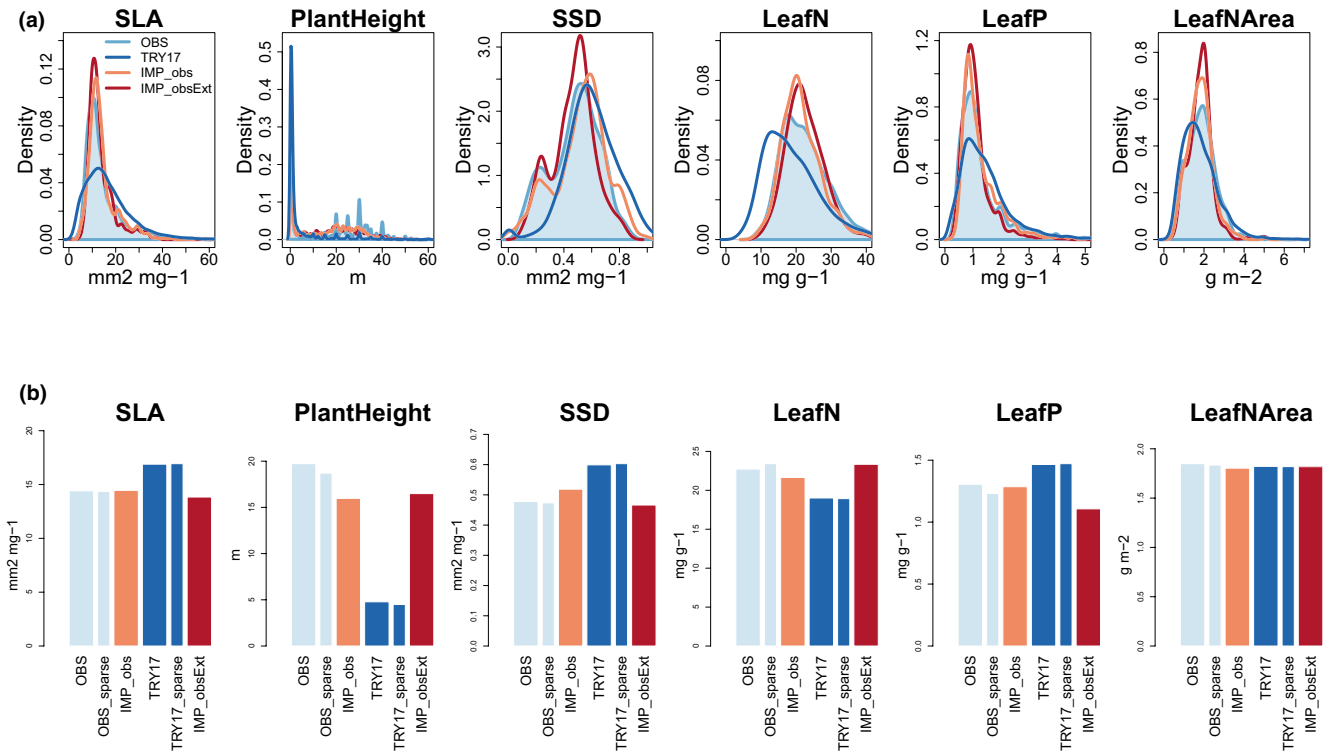

FIGURE 3 Trait distribution deviations from observed to BHPMF-imputed values (missingness: 80% gaps, $n = 1136$; values back-transformed one BHPMF repetition). (a) Trait value distributions of observed trait data (OBS, filled light blue), extended trait data (TRY17, dark blue) and imputed values without (IMP_{obs}, orange) and with external data (IMP_{obsExt}, red). (b) Bar plot of mean trait value of trait data (OBS, light blue), IMP_{obs} (orange) and IMP_{obsExt} (red) imputed including extended trait data (EXT, blue). For OBS2, see Figure S9.

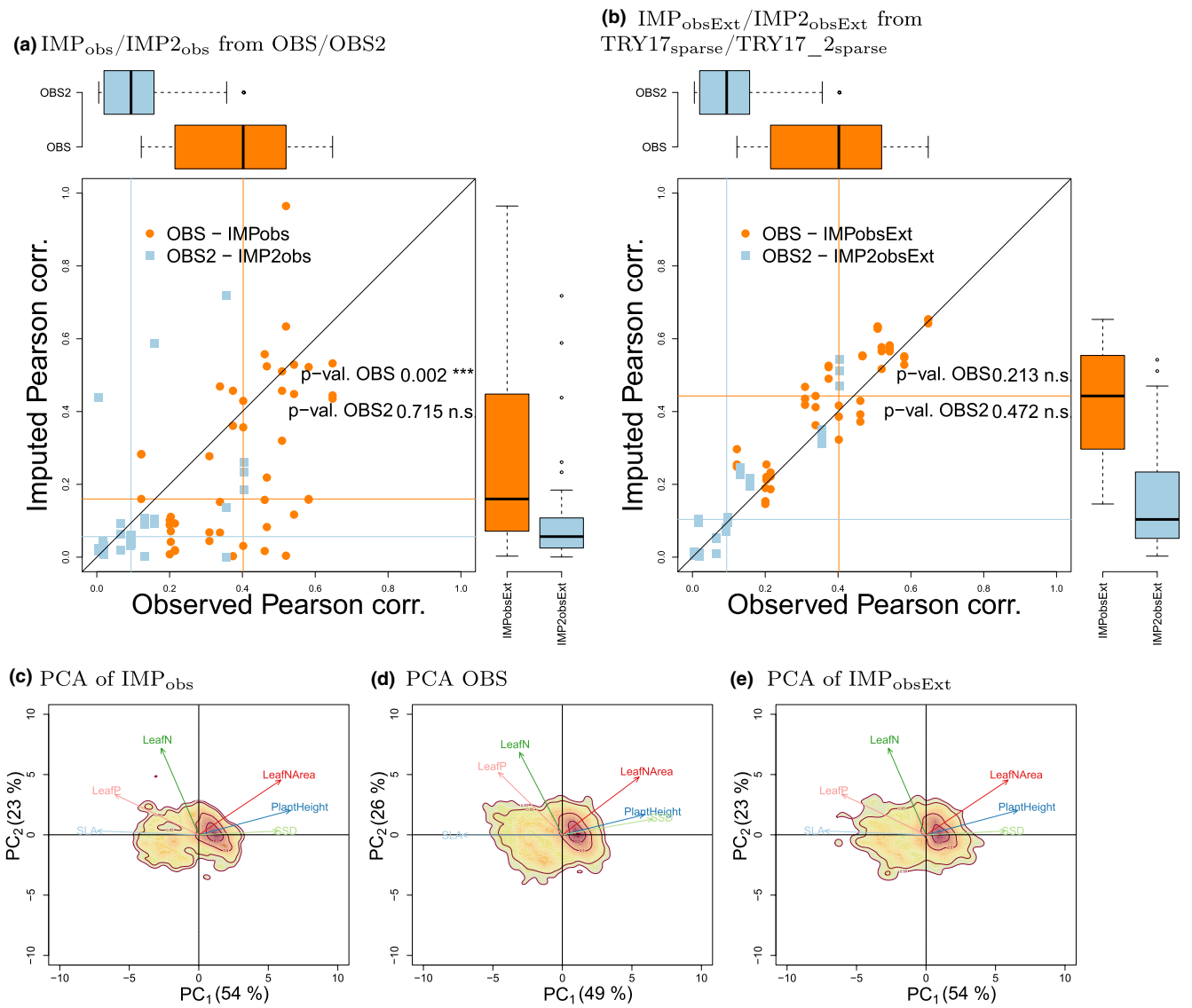

FIGURE 4 Comparison of observed versus imputed pairwise trait–trait correlations. (a) Pearson coefficients for OBS and OBS2, including *t*–test result, comparing trait–trait relationships of observed trait data (OBS, OBS2) and data imputed with 80% missingness (IMP$_{obs}$, IMP2$_{obs}$) (for other transformations and missingness levels: see Figure S10, for OBS2: Figure S11). Imputed without external trait data (repetition $n = 3$). (b) Pearson coefficients for OBS and OBS2, imputed with external trait data, including t–test result, comparing trait–trait relationships of observed trait data (OBS, OBS2) and data imputed with 80% missingness (IMP$_{obsExt}$, IMP2$_{obsExt}$) (IMP$_{obsExt}$, IMP2$_{obsExt}$, repetition $n = 3$). (b–d) Principal component analyses for log-transformed traits (b) trait data imputed with trait data, IMP$_{obs}$ (c) trait data observed, OBS (c) trait data imputed from extended trait data IMP$_{obsExt}$. For bivariate trait–trait relationships and different transformations and missingness levels of OBS, see Figure S10, and for OBS2, see Figure S11. For multivariate relationships of OBS2, see Figure S19. Significance levels: * < 0.05, ** < 0.01; *** < 0.005.

to OBS, OBS2 (Figure 4a although not significantly for IMP2$_{obs}$). The correlation coefficients varied along missingness levels, less so when supplemented with external data (Figures S10 and S11 and Tables S6 and S12). Adding external data improved the match with observed correlation coefficients for both OBS and OBS2, accompanied by an insignificant increase in correlation coefficient values (Figure 4a vs. b). The perforated trait data (OBS$_{sparse}$) had few entries, sometimes only 17, to deduce trait–trait relationships (Table S6). Multivariate trait–trait relationships were similar for imputed and observed data (Figure 4c–e) according to a Procrustes test (Table S7).

3.4 | Clustering

In order to detect possible taxonomic patterns introduced during BHPMF imputation, we analysed clustering at different levels for observed and imputed data arranged in groups (Figure 6), for individual traits (Figure S13) and for single values (Figure 5, and separated according to trait Figures S14a–f). On all three of these 'levels', we observed changes in clustering due to imputation.

First, based on the Silhouette index, we tested whether originally observed clustering was changed by imputation. We found increased clustering after imputation for all groups (Figure 6).

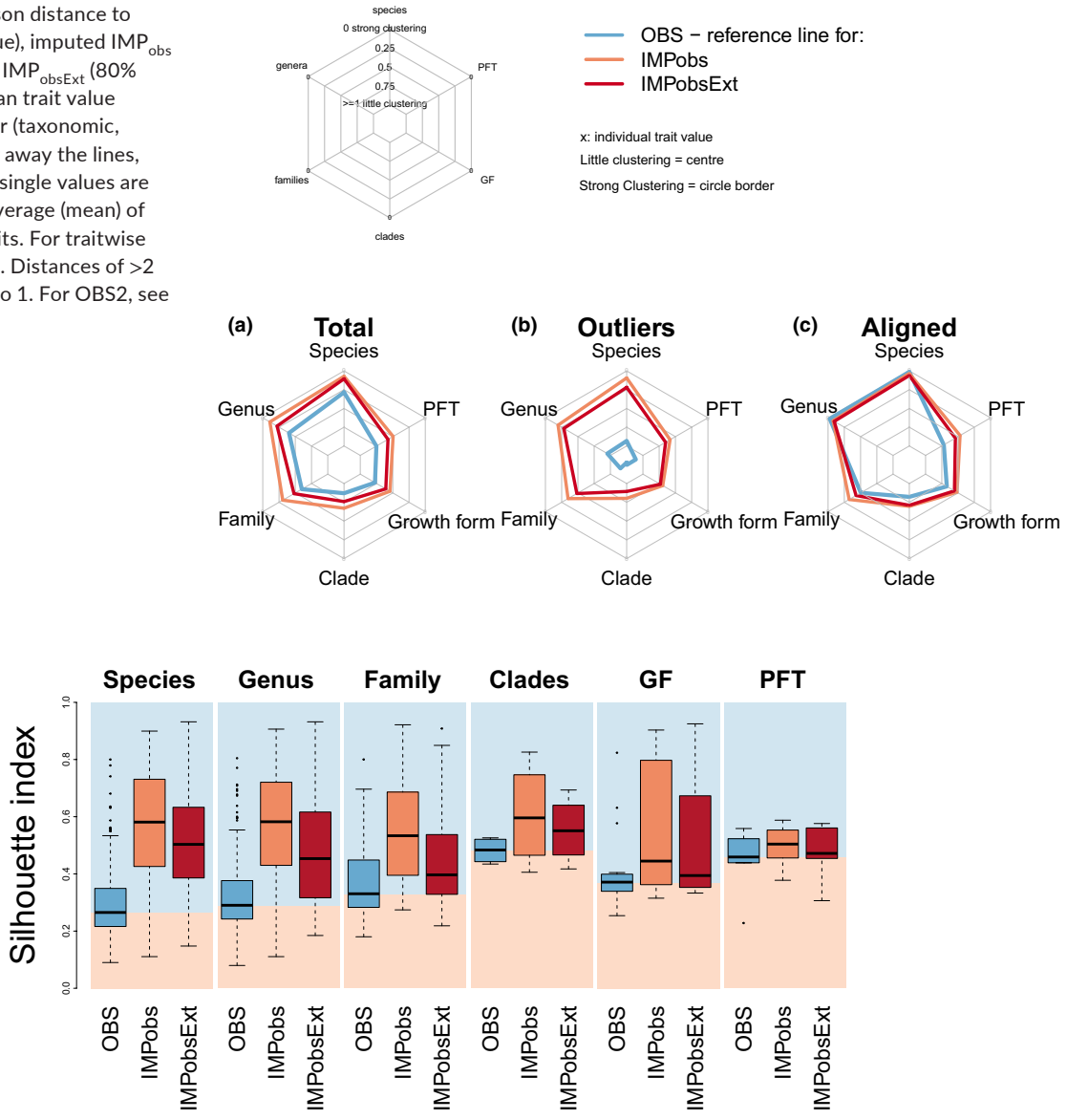

FIGURE 5 Comparison distance to group mean of OBS (blue), imputed IMP_{obs} (80% gaps, orange) and IMP_{obsExt} (80% gaps, red). Outside: mean trait value of the respective cluster (taxonomic, functional). The further away the lines, the more dissimilar the single values are to it. Lines depict the average (mean) of all single values and traits. For traitwise analysis, see Figure S14. Distances of >2 are not shown and set to 1. For OBS2, see Figures S16 and S21.

FIGURE 6 Comparison of OBS and IMP_{obs} and IMP_{obsExt} in their degree of clustering. Degree of multivariate clustering for different groups (taxonomic and functional) compared for observed trait data (OBS), imputed trait data (IMP_{obs}) and trait data imputed using external data (IMP_{obsExt}). The light orange region indicates tighter within-group clusters than for the observed data (greater similarity), and the light blue region indicates looser within-group clusters than for the observed data. Analysis on the basis of *z*-log transformed data (reference of observed data). For OBS2, see Figure S15.

Species and genera increased clustering most, clades and functional groups least. Second, we found that intraspecific trait diversity was also reduced by imputation as measured by the coefficient of variation (CV), especially for heterogeneous and observation-poor species (Figure S13). The reduction in CV occurred regardless of the addition of external data, but external data seemed to reduce the effect size for observation-rich species (Figure S13). Third, we found that single values became closer to the group mean with imputation (Figure 5). We analysed the distance to taxon mean (as well as clade and functional group mean) for OBS, IMP_{obs}, and IMP_{obsExt}. We found an overall reduced distance to taxon mean for IMP_{obs} and IMP_{obsExt} in comparison to OBS, with values which were originally further from the mean shifted more than values which were originally closer to the mean (Figure 5, and separated according to trait Figure S14a–f). This was most pronounced for leaf nitrogen concentration and leaf phosphorus concentration, which clustered only loosely in observed data, and least pronounced for plant height, which is more similar within observed groups.

OBS2, in comparison to OBS, clustered more tightly in the observed data set, and also clustered more tightly following imputation. For OBS2, the effect of imputation on clustering was, however, eliminated by the addition of external data during imputation (Figure S15). Single values for OBS2 also became closer to the group mean with imputation, with outlier values shifting more than values which were initially close to the mean (Figure S16).

4 | DISCUSSION

We investigated whether data imputation with BHPMF leads to inaccuracies and biases which could cause problems in downstream analyses, and aimed to identify aspects influencing the bias. We found that the error of imputed data in terms of RMSE increased with missingness up to an RMSE of $OBS_{mean} = 0.3$ ($OBS2_{mean} = 0.25$; Figure 2a). The individual error per value depended on the trait identity, as well as presence during imputation (if a missing or an observed value was predicted; Figure 2a,c; see summary Table 1). However, the imputations of trait means (IMP_{obs},IMP_{obsExt}) as well as taxa means were usually close to the ones of the fully observed data set, even with high missingness, and were not much affected by imputation in the presence of sparse external data (Figures 3b; Figure S2e).

The PMF approach and the side information used by BHPMF suggest two aspects of potential bias: trait–trait relationships depicted from the sparse data set by BHPMF, and taxonomic relationships which are added to the data set as clustering side information in terms of the taxonomic hierarchy. A third aspect potentially introducing bias is given by using a larger data set for imputation than is actually needed for the downstream analyses. We did not find evidence for strengthened trait–trait relationships after imputation (H1). Instead, we found reduced trait–trait relationships, also linked to stronger variability with increasing missingness (IMP_{obs}, $IMP2_{obs}$; Figure 4e, Figures S11 and S12). In contrast, the use of external data during imputation improved and stabilized the predicted trait–trait correlations across all levels of missingness. Imputation with external data revealed an insignificant trend for increasing correlation strength with increasing missingness, which might provide very weak support for H1 (but see discussion of H3). The relatively unbiased prediction of trait–trait relationships with BHPMF may help explain why the observed and imputed trait data published in Díaz et al. (2015) and Joswig et al. (2022, suppl. section 7.2) revealed very similar trait–trait correlations, and it should be noted that both studies used very large data sets (for Díaz et al., 2015, 45,507 species; and for Joswig et al., 2022, 652,957 individuals). Imputation generally did not modify trait–trait relationship patterns. This supports the use of imputed data for analysing trait connectivity as in Flores-Moreno et al. (2019), as well as for trait–trait relationships (Joswig et al., 2022). Our method only included linear trait–trait relationships, as these are most commonly investigated (Díaz et al., 2015; Joswig et al., 2022), and may miss non-linear relationships. Future work should investigate to what extent non-linear relationships may be affected by BHPMF imputation.

In contrast, Hypothesis 2 (H2) was mainly supported because imputed data clustered more tightly into taxonomic groups (taxonomic hierarchy) and consequently in functional groups (growth forms and PFTs), in comparison to observed data (Figures 5 and 6, Figure S13). Functional groups comprise sets of entire taxa, that is, members of a species belong to the same functional group and therefore patterns of functional groups follow the patterns observed for taxonomy (Figure 5). However, the strength of the effect depended on the taxonomic level (strong for species and genera, Figure 6), data set heterogeneity (more bias when heterogeneous; Figure 5) and sample size per cluster (less bias with more samples; Figure S6a). This was also influenced by the presence of external data during imputation (Figure 6, Figures S1 and S2). For the second data set, $IMP2_{obsExt}$ was less (taxonomically) biased than $IMP2_{obs}$ (Figure S15). Observed values which were far from their respective cluster means became—following imputation—more similar to their cluster mean (Figure 5). Therefore, plant traits are likely to be differently well imputed depending on their homogeneity within species and genera. In ecological terms, the intraspecific diversity influences the amount of clustering bias introduced during imputation. Traits that are taxonomically well conserved thus show a smaller error and bias in comparison to less-conserved traits. This points into the same direction as the finding by Molina-Venegas et al. (2018), who used two imputation methods (PEM and pGLM), and showed a negative relationship between accuracy and phylogenetic tip length (i.e. evolutionary time) depending on the strength of the phylogenetic signal. In the future, the guiding aspect of each trait for its variation should be investigated (i.e. taxonomy, environment, etc.), as it is of importance for decisions regarding scale, aggregation and input during imputation. Compared to OBS, OBS2 is of low functional diversity and high taxonomic conservatism, in part because of the low number of individuals per taxon in OBS2 (Table S3). Thus for OBS2, BHPMF even reduced clustering for some species (Figure S13). External data likely increased accuracy for OBS2 more than for OBS because, per trait and species, the external data provided more new observations for OBS2 (average number of observations per species $n(\text{mean}) = 5.4$) than for OBS (average number of observations per species, $n(\text{mean}) = 4$; Table S3).

The effect of adding external data during imputation (H3) depended on the analysis and the data set. The addition of external data resulted in more consistent predictions of trait–trait correlations which, however, showed a very weak tendency to increase with increasing missingness in the trait data. The addition

TABLE 1 General rules for the factors and their direction on accuracy.

Accuracy
• The average accuracy of imputations depends on the input data. The relationships with accuracy is: – missingness (Figure 2, Figure S3) + taxonomic clustering/ phylogenetic conservatism (Figure 5) – functional diversity/environmental plasticity (Figure 5, Figures S13 and S21) + / – trait–trait relationships (Figure 4, Figures S10 and S12)
Accuracy of single samples/entries
• The accuracy of single samples/entries on: – gap during imputation (compare top Figure 2 vs. bottom Figure 2) (+) external data (Figures S2e, S4 and S5) + similarity to taxon mean (outlier/aligns, Figure 5) ? alignment with trait–trait relationships + addition of samples in the external trait data, which are same as in your local target data set (see difference between OBS and OBS2 in Table S3, Figures 6, and S15)

of external data slightly ameliorated taxonomic clustering. This buffering effect of external data (IMP_{obsExt})—where present—is likely to be the consequence of having a greater trait sample size including trait distribution and trait pairs to draw from (Table S3). Yet, not all predictions were improved by the use of an extended trait data (Figures S7a–S8c). This may be due in part to the environmental plasticity of traits, leading to high intraspecific variability. As BHPMF learns from all individuals of a taxon, imputation of local trait data from widely distributed species could deviate towards trait values originating in other environmental conditions. Consequently, external data may shift taxa means and cause large errors, as seen for single values (Table S5). For the second trait data set ($IMP2_{obsExt}$), wider clustering (Figures S16 and S15) may stem from greater variance introduced by the addition of external data. The addition of external data during imputation has little effect on RMSE, although it tends to increase RMSE for low missingness, and to decrease RMSE for high missingness. It may ameliorate the tendency of BHPMF to increase taxonomic clustering, and stabilizes the prediction of trait–trait relationships when data are sparse. Thus, it is generally recommended to include external data during imputation of sparse data sets (e.g. data sets where more than 10% of entries are missing). We recommend comparing taxon means of external data with the target trait data, and recommend against including external trait and taxon data with large differences in taxon means, as these are likely to result in biases (recommendations Table 2).

Accuracy and taxonomic side information were interlinked. For example, $IMP2_{obs}$ was more accurate (Figure S3) with smaller residuals (Figure S5) and was less biased or even unbiased from taxonomy than IMP_{obs} (Figure S15). The accuracy of OBS2 was likely due to tighter taxonomic clusters than OBS (Figure 5, Figure S16), and thus, its samples could not shift much when deviating to cluster mean after imputation (see also summary Table 1).

This also gives rise to recommendations when external data are useful (Table 2). External data buffered the taxonomic bias slightly more for OBS2 than for OBS (Figure S15), also species mean was better predicted with external data for OBS2 (Figure S6). This may have been due to the fact that external data provide 5.4 additional samples per OBS2 species, while for OBS only added four per sample (Table S3). Further OBS2 may have represented the average environmental condition in the external data better than OBS, since OBS2 was a tropical tree data set, while OBS consists of growth forms dominating temperate, well-sampled areas (Kattge et al., 2020).

For single traits, bias resulted from information in taxonomy and trait–trait relationships. Plant height, for example, was imputed with lowest error for OBS (Figure 2a). This was likely due to the high taxonomic information content: plant height was, for this data set, highly conserved (Figures S14 and S13a). Despite the poor usage of information from trait–trait relationships (Figure S20), the phylogenetic conservatism on low taxonomic levels was sufficient for high prediction accuracy. For plant height, the effect of adding external data (EXT) was not visible in terms of error (Figure 2) or in terms of distributions (Figure 3) or of taxonomic clustering (Figure S14). Yet, the trait–trait relationships were much better imputed for IMP_{obsExt} than for IMP_{obs} in terms of stability of repetitions and similarity to OBS (Figure S20).

Leaf nitrogen values in OBS (Figure S17) were the worst imputed. Leaf nitrogen neither used much information from correlations (Figure S17) nor was it well conserved (Figures S14 and S13a), being strongly dependent on the environment and particularly soil composition. Leaf phosphorus was also poorly conserved, but appeared to use more correlation information (Figure S18).

The z- and log-transformation of the imputed data may have hidden some relevant deviations when back-transformed to the original scale. It is important to note that, while central tendencies and trait–trait relationships can be successfully predicted from imputed data with little bias, there can be large deviations of individual values from observations, causing a reduction in the apparent

TABLE 2 Choices for gap-filling with BHPMF and its usage summarized.

Gap-filling with BHPMF
- We recommend to include additional external trait data to stabilize trait–trait relationships (Figure 4).
  Disclaimer The trait–trait relationships are likely to become stable, as the additional external trait data show rather unaltered trait–trait relationship itself. For varying trait–trait relationships, for example, due to scales (small, big, variable), BHPMF gap-filled data are yet to be tested.
- We recommend to include additional data (e.g. from TRY) if the external taxa add entries to the target data sets species and genera. With additional entries for taxa (see difference OBS and OBS2 Table S3), the imputations are likely to become more accurate (compare OBS in Figure 6, and OBS2 in Figure S15), while still keeping distribution and mean (Figure 3).
  Disclaimer If the taxa replicated in the external data are different from the target data set, for example, because of coming from different environments etc., the accuracy may be reduced. This is because the output taxa (not trait mean) resemble the input taxa mean (Figures S13 and S6b).

Trait–trait relationships with BHPMF imputations
- BHPMF gap-filled data can be used to analyse trait–trait relationships (Figure 4). We recommend including additional external trait data to stabilize the relationships.

Disclaimer for local and sparse data sets: trait–trait relationships may be reduced

Taxonomic analyses with BHPMF imputations
- Do not use BHPMF gap-filled trait data for taxonomic analyses (Figures 6 and 5, Figures S14a,d,e,f and S21d). Especially intraspecific analyses cannot be made with BHPMF gap-filled data, samples within taxa are inherently reduced in trait–space, thus biased.

Disclaimer Exceptions may exist for specific data. Originally strongly clustered trait data, that is, data that have either few entries per taxon or phylogenetically very conserved species, will likely show high accuracy and little deviation after BHPMF imputation (e.g. OBS2 in, Figures S15, S16 and S21b,c, or some traits in OBS Figure S14b).

Difference of imputed and actual gap-filled data
- Always use the observed value. BHPMF generates imputations independently if the value was available. Replace imputations with observations.

biological variation of data sets. In most applications, imputation was likely to produce differently distributed bias than in the example we present here, with larger bias in chunks missing and smaller bias where chunks are available. In our approach, trait samples are missing completely at random (MCAR), while most sparse data sets, trait values are not missing at random (NMAR), but rather the more difficult to observe traits (taxa, regions, etc.) are systematically missing (Jetz et al., 2016). NMAR missingness is likely to show a more variable taxonomic bias than in our approach, because missingness will be distributed unevenly. This results in stronger clustering in large and sparse taxa, and less clustering in small and observed ones. Especially for these large and sparse taxa, additional sparse NMAR big data can strongly influence the imputation result. Either way, it is not recommended to use imputed data for the analysis of taxonomic diversity or within-taxa diversity.

CONCLUSIONS

This study identified the potential and limitations of BHPMF-imputed data sets and deduced guidelines for appropriate use. We found that imputation either reduced the Pearson correlation coefficients for trait–trait relationships or let them be little altered. External data stabilized imputation and improved trait–trait correlation accuracy. The accuracy of imputed values depended on their distance to taxon mean, as BHPMF tends to systematically shift imputed values to the means of the taxonomic groups. Imputed values, which are closer to the genus or species mean are therefore likely to be more accurately predicted than outliers. Following imputation, taxonomic groups clustered more closely together than in the observed data, as did functional groups. We conclude that trait–trait relationship patterns may be inferred from BHPMF-imputed data, but taxonomic patterns should not be as they are strongly biased and have reduced variance. Our study extends criteria for the evaluation of gap–filling beyond RMSE, providing insight into statistical data structure and allowing better-informed use of imputed trait data, with improved practices for imputation. The accuracy and bias of BHPMF is now well characterized, which may make it more useful in comparison to other hierarchical imputation techniques. Moreover, the bias testing suggested here can be beneficial for developing tailor-made gap-filling methods for specific problems. For taxonomic analyses, taxonomic information could be excluded during imputation. It may be replaced by other meaningful information—that is, environmental, temporal ones—and finalized with bias testing. This study's approach, results and conclusions may thus further be useful beyond applications in plant ecology, for any study using hierarchical side information for imputation (see summary of recommendations Tables 1 and 2).

AUTHOR CONTRIBUTIONS

Julia S. Joswig, Jens Kattge and Miguel D. Mahecha were involved in conceptualization—ideas. Julia S. Joswig was involved in data curation—management. Julia S. Joswig was involved in formal analysis. Michael E. Schaepman was involved in funding acquisition. Julia S. Joswig was involved in investigation. Julia S. Joswig, Miguel D. Mahecha and Guido Kraemer were involved in methodology—development or design of methodology; creation of models. Jens Kattge and Michael E. Schaepman were involved in project administration—management and coordination responsibility for the research activity planning and execution. Jens Kattge, Michael E. Schaepman and Meredith C. Schuman were involved in resources. Julia S. Joswig and Guido Kraemer were involved in software. Jens Kattge, Miguel D. Mahecha, Nadja Rüger, Michael E. Schaepman and Meredith C. Schuman were involved in supervision. Julia S. Joswig was involved in visualization and writing—original draft. Julia S. Joswig, Jens Kattge, Guido Kraemer, Miguel D. Mahecha, Nadja Rüger, Michael E. Schaepman, Franziska Schrodt and Meredith C. Schuman were involved in writing—review & editing.

ACKNOWLEDGEMENTS

This work was supported by the University Research Priority Program Global Change and Biodiversity (URPP GCB) of the University of Zurich. J.S.J. acknowledges the International Max Planck Research School for global biogeochemical cycles. J.S.J., M.E.S. and M.C.S. acknowledge support from the University of Zurich University Research Priority Program on Global Change and Biodiversity. P.B.R., M.E.S. and M.C.S. acknowledge membership in the US NSF 20-508 BII-Implementation project, 'The causes and consequences of plant biodiversity across scales in a rapidly changing world'. M.E.S. acknowledges the NOMIS grant of Remotely Sensing Ecological Genomics that funds J.S.J. and M.C.S. C.W. acknowledges the support of the Max Planck Society via its fellowship programme. N.R. was funded by a research grant from Deutsche Forschungsgemeinschaft DFG (RU 1536/3-1). The study was supported by the TRY initiative on plant traits (http://www.try-db.org). The TRY database is hosted at the Max Planck Institute for Biogeochemistry (MPI BGC, Germany) and supported by Future Earth and the German Centre for Integrative Biodiversity Research (iDiv) Halle-Jena-Leipzig. We would like to thank all PIs contributing to the TRY database, whose efforts allowed this analysis. We appreciate the discussions at the MPI BGC. The authors affiliated with the MPI BGC acknowledge funding by the European Union's Horizon 2020 project BACI under grant agreement no. 640176.We thank Farideh Fazayeli for the introduction into installing and running the BHPMF manuscript. Open access funding provided by Universitat Zurich.

CONFLICT OF INTEREST STATEMENT

The authors declare no conflict of interest.

DATA AVAILABILITY STATEMENT

The code is available on Github https://github.com/juliajoswig/Repo_BHPMF_bias. The data that support the findings of this study are available in the TRY archive at [https://www.try-db.org/TryWeb/Data.php#96 with DOI: 10.17871/TRY.96]. These data were derived from the following resources available in the public domain: TRY-db.org.

ORCID

Julia S. Joswig https://orcid.org/0000-0002-7786-1728

Miguel D. Mahecha https://orcid.org/0000-0003-3031-613X

REFERENCES

Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., Garcia, S., Gil-Lopez, S., Molina, D., Benjamins, R., Chatila, R., & Herrera, F. (2020). Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, *58*, 82–115. https://doi.org/10.1016/j.inffus.2019.12.012 ISSN 15662535.

Bruelheide, H., Dengler, J., Jiménez-Alfaro, B., Purschke, O., Hennekens, S. M., Chytrý, M., Pillar, V. D., Jansen, F., Kattge, J., Sandel, B., Aubin, I., Biurrun, I., Field, R., Haider, S., Jandt, U., Lenoir, J., Peet, R. K., Peyre, G., Sabatini, F. M., … Zizka, G. (2019). sPlot—A new tool for global vegetation analyses. *Journal of Vegetation Science*, *30*, 161–186. https://doi.org/10.1111/jvs.12710

Bruelheide, H., Dengler, J., Purschke, O., Lenoir, J., Jiménez-Alfaro, B., Hennekens, S. M., Botta-Dukát, Z., Chytrý, M., Field, R., Jansen, F., Kattge, J., Pillar, V. D., Schrodt, F., Mahecha, M. D., Peet, R. K., Sandel, B., van Bodegom, P., Altman, J., Alvarez-Dávila, E., … Jandt, U. (2018). Global trait–environment relationships of plant communities. *Nature Ecology & Evolution*, *2*(12), 1906–1917. https://doi.org/10.1038/s41559-018-0699-8 ISSN 2397-334X. http://www.nature.com/articles/s41559-018-0699-8

Díaz, S., Kattge, J., Cornelissen, J. H. C., Wright, I. J., Lavorel, S., Dray, S., Reu, B., Kleyer, M., Wirth, C., Prentice, I. C., Garnier, E., Bönisch, G., Westoby, M., Poorter, H., Reich, P. B., Moles, A. T., Dickie, J., Gillison, A. N., Zanne, A. E., … Gorné, L. D. (2016). The global spectrum of plant form and function. *Nature*, *529*(7585), 167–171. https://doi.org/10.1038/nature16489 ISSN 14764687.

Díaz, S., Kattge, J., Cornelissen, J. H. C., Wright, I. J., Lavorel, S., Dray, S., Reu, B., Kleyer, M., Wirth, C., Prentice, I. C., Garnier, E., Bönisch, G., Westoby, M., Poorter, H., Reich, P. B., Moles, A. T., Dickie, J., Gillison, A. N., Zanne, A. E., … Gorné, L. D. (2015). The global spectrum of plant form and function. *Nature*, *529*(7585), 1–17. https://doi.org/10.1038/nature16489 ISSN 0028-0836.

Díaz, S., Lavorel, S., Stuart Chapin, F., III, Tecco, P. a., Gurvich, D. E., & Grigulis, K. (2007). Functional diversity—At the crossroads between ecosystem functioning and environmental filters. *Terrestrial Ecosystems in a Changing World*, *1*, 81–91. https://doi.org/10.1007/978-3-540-32730-1_7 ISSN 00314056.

Dudani, S. A. (1976). The distance-weighted k-nearest-neighbor rule. *IEEE Transactions on Systems, Man and Cybernetics*, *4*, 325–327. https://doi.org/10.1109/TSMC.1976.5408784 ISSN 21682909.

Evans, M. E. K., Merow, C., Record, S., McMahon, S. M., & Enquist, B. J. (2016). Towards process-based range modeling of many species. *Trends in Ecology & Evolution*, *31*(11), 860–871. https://doi.org/10.1016/j.tree.2016.08.005 ISSN 01695347.

Fazayeli, F., Banerjee, A., Kattge, J., Schrodt, F., & Reich, P. B. (2014). Uncertainty quantified matrix completion using Bayesian Hierarchical Matrix factorization. *International Conference on Machine Learning and Applications (ICMLA)*. https://ieeexplore.ieee.org/document/7033133

Flores-Moreno, H., Fazayeli, F., Banerjee, A., Datta, A., Kattge, J., Butler, E. E., Atkin, O. K., Wythers, K., Chen, M., Anand, M., Bahn, M., Byun, C., Cornelissen, J. H. C., Craine, J., Gonzalez-Melo, A., Hattingh, W. N., Jansen, S., Kraft, N. J. B., Kramer, K., … Reich, P. B. (2019). Robustness of trait connections across environmental gradients and growth forms. *Global Ecology and Biogeography*, *28*(12), 1806–1826. https://doi.org/10.1111/geb.12996 ISSN 1466-822X.

Jari Oksanen, F., Blanchet, G., Friendly, M., Kindt, R., Legendre, P., Mcglinn, D., Minchin, P. R., O'hara, R. B., Simpson, G. L., Solymos, P., Henry, M., Stevens, H., Szoecs, E., & Maintainer, H. W. (2020). Package 'vegan' title community ecology package version 2.5-7. R, 2.5.

Jetz, W., Cavender-Bares, J., Pavlick, R., Schimel, D., Davis, F. W., Asner, G. P., Guralnick, R., Kattge, J., Latimer, A. M., Moorcroft, P., Schaepman, M. E., Schildhauer, M. P., Schneider, F. D., Schrodt, F., Stahl, U., & Ustin, S. L. (2016). Monitoring plant functional diversity from space. *Nature Plants*, *2*(3), 16024. https://doi.org/10.1038/nplants.2016.24 ISSN 2055-0278. http://www.nature.com/articles/nplants201624

Johnson, T. F., Isaac, N. J. B., Paviolo, A., & González-Suárez, M. (2020). Handling missing values in trait data. *Global Ecology and Biogeography*, *10*, 51–62. https://doi.org/10.1111/geb.13185 ISSN 1466-822X.

Joswig, J. S., Wirth, C., Schuman, M. C., Kattge, J., Reu, B., Wright, I. J., Sippel, S. D., Rüger, N., Richter, R., Schaepman, M. E., van Bodegom, P. M., Cornelissen, J. H. C., Díaz, S., Hattingh, W. N., Kramer, K., Lens, F., Niinemets, Ü., Reich, P. B., Reichstein, M., … Mahecha, M. D. (2022). Climatic and soil factors explain the two-dimensional spectrum of global trait variation. *Nature Ecology and Evolution*, *6*, 36–50. https://doi.org/10.1038/s41559-021-01616-8

Karpatne, A., Atluri, G., Faghmous, J. H., Steinbach, M., Banerjee, A., Ganguly, A., Shekhar, S., Samatova, N., & Kumar, V. (2017). Theory-guided data science: A new paradigm for scientific discovery from data. *IEEE Transactions on Knowledge and Data Engineering*, *29*, 2318–2331. https://doi.org/10.1109/TKDE.2017.2720168 ISSN 10414347.

Kattge, J., Bönisch, G., Díaz, S., Lavorel, S., Prentice, I. C., Leadley, P., Tautenhahn, S., Werner, G. D. A., Aakala, T., Abedi, M., Acosta, A. T. R., Adamidis, G. C., Adamson, K., Aiba, M., Albert, C. H., Alcántara, J. M., Carolina Alcázar, C., Aleixo, I., Ali, H., … Wirth, C. (2020). TRY plant trait database—Enhanced coverage and open access. *Global Change Biology*, *26*(1), 119–188. https://doi.org/10.1111/gcb.14904 ISSN 1354-1013.

Jens Kattge, S. Díaz, S. Lavorel, I. C. Prentice, P. Leadley, G. Bönisch, E. Garnier, M. Westoby, P. B. Reich, I. J. Wright, J. H C Cornelissen, C. Violle, S. P. Harrison, P. M. Van Bodegom, M. Reichstein, B. J. Enquist, N. A. Soudzilovskaia, D. D. Ackerly, M. Anand, O. Atkin, M. Bahn, T. R. Baker, D. Baldocchi, R. Bekker, C. C. Blanco, B. Blonder, W. J. Bond, R. Bradstock, D. E. Bunker, F. Casanoves, J. Cavender-Bares, J. Q. Chambers, F. S. Chapin, J. Chave, D. Coomes, W. K. Cornwell, J. M. Craine, B. H. Dobrin, L. Duarte, W. Durka, J. Elser, G. Esser, M. Estiarte, W. F. Fagan, J. Fang, F. Fernández-Méndez, A. Fidelis, B. Finegan, O. Flores, H. Ford, D. Frank, G. T. Freschet, N. M. Fyllas, R. V. Gallagher, W. A. Green, A. G. Gutierrez, T. Hickler, S. I. Higgins, J. G. Hodgson, A. Jalili, S. Jansen, C. A. Joly, A. J. Kerkhoff, D. Kirkup, K. Kitajima, M. Kleyer, S. Klotz, J. M H Knops, K. Kramer, I. Kühn, H. Kurokawa, D. Laughlin, T. D. Lee, M. Leishman, F. Lens, T. Lenz, S. L. Lewis, J. Lloyd, J. Llusià, F. Louault, S. Ma, M. D. Mahecha, P. Manning, T. Massad, B. E. Medlyn, J. Messier, A. T. Moles, S. C. Müller, K. Nadrowski, S. Naeem, Ü Niinemets, S. Nöllert, A. Nüske, R. Ogaya, J. Oleksyn, V. G. Onipchenko, Y. Onoda, J. Ordoñez, G. Overbeck, W. A. Ozinga, S. Patiño, S. Paula, J. G. Pausas, J. Peñuelas, O. L. Phillips, V. Pillar, H. Poorter, L. Poorter, P. Poschlod, A. Prinzing, R. Proulx, A. Rammig, S. Reinsch, B. Reu, L. Sack, B. Salgado-Negret, J. Sardans, S. Shiodera, B. Shipley, A. Siefert, E. Sosinski, J. F. Soussana, E. Swaine, N. Swenson, K. Thompson, P. Thornton, M. Waldram, E. Weiher, M. White, S. White, S. J. Wright, B. Yguel, S. Zaehle, A. E. Zanne, and C. Wirth. TRY—A global database of plant traits. *Global Change Biology*, 2011, *17*:2905-2935. ISSN 1354-1013.

Kim, S. W., Blomberg, S. P., & Pandolfi, J. M. (2018). Transcending data gaps: A framework to reduce inferential errors in ecological analyses. *Ecology Letters*, *21*(8), 1200–1210. https://doi.org/10.1111/ele.13089 ISSN 1461023X.

Mnih, A., & Salakhutdinov, R. R. (2007). Probabilistic matrix factorization. In J. Platt, D. Koller, Y. Singer, & S. Roweis (Eds.), *Advances in*

neural information processing systems (Vol. *20*). Curran Associates, Inc https://proceedings.neurips.cc/paper/2007/file/d7322ed717dedf1eb4e6e52a37ea7bcd-Paper.pdf

Mohanraj, K., Karthikeyan, B. S., Vivek-Ananth, R. P., Chand, R. P. B., Aparna, S. R., Mangalapandi, P., & Samal, A. (2018). IMPPAT: A curated database of Indian medicinal plants, phytochemistry and therapeutics. *Scientific Reports*, *8*(1), 4329. https://doi.org/10.1038/s41598-018-22631-z ISSN 2045-2322. http://www.nature.com/articles/s41598-018-22631-z

Molina-Venegas, R., Moreno-Saiz, J. C., Parga, I. C., Davies, T. J., Peres-Neto, P. R., & Rodríguez, M. Á. (2018). Assessing among-lineage variability in phylogenetic imputation of functional trait datasets. *Ecography*, *41*(10), 1740–1749. https://doi.org/10.1111/ecog.03480 ISSN 09067590.

Moreno-Martínez, Á., Camps-Valls, G., Kattge, J., Robinson, N., Reichstein, M., van Bodegom, P., Kramer, K., Cornelissen, J. H. C., Reich, P., Bahn, M., Niinemets, Ü., Peñuelas, J., Craine, J. M., Cerabolini, B. E. L., Minden, V., Laughlin, D. C., Sack, L., Allred, B., Baraloto, C., … Running, S. W. (2018). A methodology to derive global maps of leaf traits using remote sensing and climate data. *Remote Sensing of Environment*, *218*, 69–88. https://doi.org/10.1016/j.rse.2018.09.006 ISSN 00344257.

Musavi, T., Mahecha, M. D., Migliavacca, M., Reichstein, M., van de Weg, M. J., van Bodegom, P. M., Bahn, M., Wirth, C., Reich, P. B., Schrodt, F., & Kattge, J. (2015). The imprint of plants on ecosystem functioning: A data-driven approach. *International Journal of Applied Earth Observation and Geoinformation*, *43*, 119–131. https://doi.org/10.1016/j.jag.2015.05.009 ISSN 1872826X.

Myers-Smith, I. H., Thomas, H. J. D., & Bjorkman, A. D. (2019). Plant traits inform predictions of tundra responses to global change. *New Phytologist*, *221*, 1742–1748 ISSN 14698137.

Nakagawa, S., & Freckleton, R. P. (2008). Missing inaction: The dangers of ignoring missing data. *Trends in Ecology & Evolution*, *23*(11), 592–596. https://doi.org/10.1016/j.tree.2008.06.014 ISSN 01695347. https://linkinghub.elsevier.com/retrieve/pii/S0169534708002772

Navarro-Cano, J. A., Goberna, M., Valiente-Banuet, A., & Verdú, M. (2021). Phenotypic structure of plant facilitation networks. *Ecology Letters*, *24*, 509–519. https://doi.org/10.1111/ele.13669 ISSN 14610248.

Peres-Neto, P. R., & Jackson, D. A. (2001). How well do multivariate data sets match? The advantages of a procrustean superimposition approach over the mantel test. *Oecologia*, *129*, 169–178. https://doi.org/10.1007/s004420100720 ISSN 00298549.

Poyatos, R., Sus, O., Badiella, L., Mencuccini, M., & Martínez-Vilalta, J. (2018). Gap-filling a spatially explicit plant trait database: Comparing imputation methods and different levels of environmental information. *Biogeosciences*, *15*(9), 2601–2617. https://doi.org/10.5194/bg-15-2601-2018 ISSN 1726-4189.

Reichstein, M., Bahn, M., Mahecha, M. D., Kattge, J., Dennis, D., & Baldocchi, D. D. (2014). Linking plant and ecosystem functional biogeography. *Proceedings of the National Academy of Sciences*, *111*(38), 1091–6490. https://doi.org/10.1073/pnas.1216065111 ISSN 0027-8424.

Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, *200*(C), 53–65. https://doi.org/10.1016/0377-0427(87)90125-7 ISSN 03770427.

Salakhutdinov, R., & Mnih, A. (2009). Probabilistic matrix factorization. *Advances in neural information processing systems 20—proceedings of the 2007 conference*.

Sandel, B., Gutiérrez, A. G., Reich, P. B., Schrodt, F., Dickie, J., & Kattge, J. (2015). Estimating the missing species bias in plant trait measurements. *Journal of Vegetation Science*, *26*(5), 828–838. https://doi.org/10.1111/jvs.12292 ISSN 11009233.

Schrodt, F., Kattge, J., Shan, H., Fazayeli, F., Joswig, J., Banerjee, A., Reichstein, M., Bönisch, G., Díaz, S., Dickie, J., Gillison, A., Karpatne, A., Lavorel, S., Leadley, P., Wirth, C. B., Wright, I. J., Joseph Wright, S., & Reich, P. B. (2015). BHPMF—A hierarchical Bayesian approach to gap-filling and trait prediction for macroecology and functional biogeography. *Global Ecology and Biogeography*, *24*(12), 1510–1521.

Shan, H., Kattge, J., Reich, P., Banerjee, A., Schrodt, F., & Reichstein, M. (2012). Gap filling in the plant kingdom—Trait prediction using hierarchical probabilistic matrix factorization. *Icml*, 1303–1310 http://arxiv.org/abs/1206.6439

Bill Shipley and Paul A Keddy. *Nordic society oikos evaluating the evidence for competitive hierarchies in plant communities* Wiley on behalf of Nordic Society Oikos Stable *69*(2):340–345, 2016. http://www.jstor.org/stable/3546158 Linked references are available on JSTOR for this article.

Tavşanoğlu, Ç., & Pausas, J. G. (2018). A functional trait database for Mediterranean Basin plants. *Scientific Data*, *5*(1), 180135. https://doi.org/10.1038/sdata.2018.135 ISSN 2052-4463. http://www.nature.com/articles/sdata2018135

Thomas, H. J. D., Bjorkman, A. D., Myers-Smith, I. H., Elmendorf, S. C., Kattge, J., Diaz, S., Vellend, M., Blok, D., Cornelissen, J. H. C., Forbes, B. C., Henry, G. H. R., Hollister, R. D., Normand, S., Prevéy, J. S., Rixen, C., Schaepman-Strub, G., Wilmking, M., Wipf, S., Cornwell, W. K., … de Vries, F. T. (2020). Global plant trait relationships extend to the climatic extremes of the tundra biome. *Nature Communications*, *11*(1), 1351. https://doi.org/10.1038/s41467-020-15014-4 ISSN 20411723.

Udell, M., & Townsend, A. (2019). Why are big data matrices approximately low rank? *SIAM Journal on Mathematics of Data Science*, *1*, 144–160. https://doi.org/10.1137/18m1183480

van Buuren, S., & Groothuis-Oudshoorn, K. (2011). mice: Multivariate imputation by chained equations in R. *Journal of Statistical Software*, *45*, 1–67. https://doi.org/10.18637/jss.v045.i03 ISSN 15487660.

Wright, I. J., Westoby, M., Reich, P. B., Oleksyn, J., Ackerly, D. D., Baruch, Z., Bongers, F., Cavender-Bares, J., Chapin, T., Cornellissen, J. H. C., Diemer, M., Flexas, J., Gulias, J., Garnier, E., Navas, M. L., Roumet, C., Groom, P. K., Lamont, B. B., Hikosaka, K., … Villar, R. (2004). The worldwide leaf economics spectrum. *Nature*, *428*, 821–827. https://doi.org/10.1038/nature02403 ISSN 0028-0836.

BIOSKETCH

The research interests of the first author relate to aspects of global plant trait variation, including spatio-temporal variation as well as implications of global change on global plant traits. The authors of this manuscript have expertise in functional biogeography, macroecology, biogeochemistry, genetics, remote sensing and empirical inference.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Joswig, J. S., Kattge, J., Kraemer, G., Mahecha, M. D., Rüger, N., Schaepman, M. E., Schrodt, F., & Schuman, M. C. (2023). Imputing missing data in plant traits: A guide to improve gap-filling. *Global Ecology and Biogeography*, *32*, 1395–1408. https://doi.org/10.1111/geb.13695