Biogeosciences Discuss., https://doi.org/10.5194/bg-2017-503
Manuscript under review for journal Biogeosciences
Discussion started: 27 November 2017
© Author(s) 2017. CC BY 4.0 License.

Gap-filling a spatially-explicit plant trait database: comparing imputation methods and different levels of ecological information.

Rafael Poyatos[1,2], Oliver Sus[3], Llorenç Badiella[4], Maurizio Mencuccini[1,5], Jordi Martínez-Vilalta[1,6]

[1] CREAF, E08193 Bellaterra (Cerdanyola del Vallès), Catalonia, Spain
[2] Laboratory of Plant Ecology, Faculty of Bioscience Engineering, Ghent University, Coupure links 653, 9000 Gent, Belgium
[3] EUMETSAT, Eumetsat Allee 1, 64295 Darmstad, Germany
[4] Servei d'Estadística Aplicada, Universitat Autònoma de Barcelona, Cerdanyola del Vallès 08193, Barcelona, Spain
[5] ICREA, Barcelona, Spain
[6] Universitat Autònoma de Barcelona, E08193 Bellaterra (Cerdanyola del Vallès), Catalonia, Spain

Correspondence to: Rafael Poyatos (r.poyatos@creaf.uab.es)

Abstract. The ubiquity of missing data in plant trait databases may hinder trait-based analyses of ecological patterns and processes. Spatially-explicit datasets with information on intraspecific trait variability are rare but offer great promise in improving our understanding of functional biogeography. At the same time, they offer specific challenges in terms of data imputation. Here we compare statistical imputation approaches, using varying levels of ecological information, for five plant traits (leaf biomass to sapwood area ratio, leaf nitrogen content, maximum tree height, leaf mass per area and wood density) in a spatially-explicit plant trait dataset of temperate and Mediterranean tree species (Catalonia, north-east Iberian Peninsula, 31900 km^2). We simulated gaps at different missingness levels (10% – 80%) in a complete trait matrix, and we used overall trait means, species means, k-nearest neighbours (kNN), ordinary and regression kriging and multivariate imputation using chained equations (MICE) to impute missing trait values. We assessed these methods in terms of their accuracy and of their ability to preserve trait distributions, multi-trait correlation structure and bivariate trait relationships. The relatively good performance of mean and species mean imputations in terms of accuracy masked a poor representation of trait distributions and multivariate trait structure. Species identity improved MICE imputations for all traits, whereas forest structure and topography improved imputations for some traits. No method performed best consistently for the five studied traits, but, considering all traits and performance metrics, MICE informed by relevant ecological variables produced globally more plausible datasets. However, at higher missingness (> 30%) species mean imputations and regression kriging tended to outperform MICE for some traits. Spatial patterns of the studied traits in Catalan forests (5495 plots) were largely similar when using species means, regression kriging or the best-performing MICE application. Our results highlight the need to assess imputation quality beyond just imputation accuracy, and show that including ecological information in statistical imputation approaches yields more plausible imputations in spatially-explicit plant trait datasets.

1 Introduction

Trait-based ecology has emerged in recent years as one of the most active ecological sub-disciplines, specially in plant ecology (Westoby & Wright, 2006; Violle et al. 2007). The move from a taxonomic perspective of biodiversity towards a focus on continuous axes of functional variation holds promise for greater generalisation, synthesis and predictive ability in ecology (Funk et al. 2016; Shipley et al. 2016). As a result, plant ecologists have increasingly embraced trait-based approaches because they may be specially suited to study plant strategies (Reich, 2014), community assembly and dynamics (McGill et al. 2006) or ecosystem functioning, particularly in the context of global environmental change (Reichstein et al. 2014). But trait-based ecology is also unquestionably thriving because of the increasing availability and reliability of plant trait data (Kattge et al. 2011).

Plant trait databases compiled from primary sources and multiple individual contributions lack a common design and inevitably result in sparse data matrices (e.g. Jetz et al. 2016). Complete-case analyses (i.e., data analyses using only sampling units with complete data availability) entail a reduced sampling size, which complicates community-level studies (Pakeman, 2014) and limit the spatial coverage of trait maps usable in trait-based models of ecosystem function. Data deletion may also bias parameter estimates (e.g., in trait relationships) if the data are not missing completeley at random (MCAR; Little & Rubin, 2002; Nakagawa & Freckleton, 2008). Imputation (i.e., gap-filling) of missing data with plausible values has only relatively recently started to be widely advocated in ecology (Nakagawa & Freckleton, 2008), although imputation may not be recommended in certain studies (Blonder 2016).

Single imputation methods replace a missing datum by one value and proceed with the analysis as if the imputed data had been observed (Nakagawa & Freckleton, 2008); species mean or median imputation are probably the most widely used methods in ecology, but they ignore the variance of the imputed variables. Other single imputation approaches such as model-based imputation (Gelman & Hill, 2007), *k*-nearest neighbour (kNN) or machine-learning methods (Stekhoven & Bühlmann, 2012) may be more appropriate for multivariate data, but they all alter, to different degrees, the univariate trait distributions and the covariance structure of the dataset. In a multiple imputation framework, *m* imputed datasets are obtained through simulation and may be jointly analysed to provide parameter estimates that take into account the uncertainty introduced by the imputations themselves (e.g. Fisher et al. 2003). Some multiple imputation techniques, such as multivariate imputation using chained equations (MICE) may be specially well-suited to preserve the original structure and distribution of multivariate datasets (van Buuren & Groothuis-Oudshoorn, 2011; van Buuren, 2012).

While forest inventories have adopted sophisticated imputation methods for some time, as for example the kNN methods (Eskelson et al. 2009 and references therein), imputation methods have only recently been started to be used in trait-based ecology (Baraloto et al. 2010; Pyšek et al. 2015). Complex imputation methods such as kNN, MICE or random forests

© Author(s) 2017. CC BY 4.0 License.

generally outperform overall mean or species mean imputations (Penone et al. 2014; Taugourdeau et al. 2014). In earlier applications of these methods, it has been common to assume that interspecific, compared to intraspecific trait variability, was dominant. The strong phylogenetic signal may then be sufficient to impute species-averaged trait values using taxonomic information (Swenson, 2014). However, intraspecific variability in plant traits may be substantial (Siefert et al. 2015; Vilà-Cabrera et al. 2015) and imputation methods that use environmental information may be more appropriate when assessing trait relationships and trait-environment covariance in a spatially explicit context. Biotic or abiotic variables other than the trait matrix of interest can be included in imputation algorithms as auxiliary variables to reduce imputation bias (Azur et al. 2011; Rezvan et al. 2015). Geostatistical methods of spatial interpolation can also be used with (e.g. regression kriging) or without (e.g. ordinary kriging) auxiliary variables (e.g. Hengl et al. 2007).

Additional challenges occur in the imputation of traits in large databases. The expected declining performance of imputation methods with increasing missingness levels, may be trait- and dataset-dependent (Penone et al. 2014; Taugourdeau et al. 2014). Moreover, the impact of imputations on altering bivariate trait relationships has only been tested for single relationships (Penone et al. 2014; Schrodt et al. 2015) and not for the multiple relevant relationships within a plant trait dataset. Likewise, there are few studies quantifying how different imputation methods alter the multivariate structure of plant trait datasets (Schrodt et al. 2015).

Our overarching aim here is to assess the performance of different imputation methods to fill simulated gaps at different missingness levels in a spatially-explicit plant trait dataset. We imputed these missing data using single imputation (kNN), multiple imputation (MICE) and geostatistical approaches (ordinary and regression kriging, OrdKrig and RegKrig, respectively), and compared the imputations with baseline scenarios of overall mean and species mean imputation. Imputation performance was assessed in terms of accuracy, univariate trait distributions, multivariate trait structure and deviations in trait relationships. Our specific objectives are: (i) to test which imputation method (overall mean imputation, kNN, MICE, OrdKrig) performed best when relying only on plant trait data; (ii) to assess the impact of including additional predictors (i.e. ecological information such as species identity, climate, forest structure, topography, lithology and sampling date) in MICE and kNN imputations; (iii) to compare the performance of kNN, MICE and RegKrig using optimum levels of ecological information; and, finally, (iv) to apply the best performing method to fill the gaps in a major subset of the IEFC database to obtain 'continuous' maps of plant traits for the main forest species across a relatively large Mediterranean region.

2. Methods

2. 1 Study area

The study area is the entire territory of Catalonia (31900 km^2), in the north-east Iberian peninsula. Catalonia has 38% forest cover (1.2 x 10^6 ha) and forests are largely dominated by species belonging to the Pinaceae and Fagaceae families. We

© Author(s) 2017. CC BY 4.0 License.

selected 13 tree species, including 6 *Pinus* spp., 5 deciduous and evergreen *Quercus* spp., *Abies alba* and *Fagus sylvatica*, which altogether cover >90% of the forested area in Catalonia (see Supplement S1).

2. 2 Data

Plant trait and forest data were retrieved from the Ecological and Forest Inventory of Catalonia (IEFC), carried out between 1988 and 1998 (Gracia et al. 2000–2004). A complete description of the sampling scheme and methods used to measure plant traits in the IEFC can be found in the Supplement S1. The subset of the IEFC limited to the 13 study species, hereby called 'IEFC incomplete dataset', included 5495 plots. Stand structure, lithology and sampling information for each plot were retrieved from the IEFC database, whereas climate data were obtained from the Climatic Digital Atlas of Catalonia, with a spatial resolution of 180 m (Ninyerola et al. 2000).

We selected five plant traits (leaf mass per area, LMA, mg cm^{-2}; nitrogen per unit mass, N_{mass}, %mass; maximum tree height, H_{max}, m; wood density, WD, gm cm^{-3}; leaf biomass to sapwood area ratio, B_L:A_S, t m^{-2}) that are used to describe major plant functional strategies (Westoby et al. 2002; Wright et al. 2004; Chave et al. 2009; Laforest-Lapointe et al. 2014). In Catalan forests, four of these traits (LMA, N_{mass}, H_{max}, WD) mostly vary across families (Pinaceae and Fagaceae) and within species (Vilà-Cabrera et al. 2015). The missing data patterns in this trait data matrix shows a much higher percentage of missing data (hereafter, 'missingness) for foliar traits, corresponding to a less intense sampling of these traits (Supplement S1). These *intentional* missing data (van Buuren 2012) would correspond to a *planned missing data design*, where missingness at random (MAR) is deliberately applied (Nagakawa 2015).

2.3 Experimental design

All data manipulations, imputations and statistical analyses were performed with the R programming language (R Core Team, 2015). We created a subset of the IEFC incomplete dataset only including those plots (N = 630) where all 5 traits had been measured on the dominant species ('IEFC complete dataset'). In this dataset, we randomly deleted measured values at different probability levels (10%, 20%, 30%, 50% and 80%) and independently for each trait, thus the missing data artificially introduced are MCAR. This data deletion was replicated, to yield 30 simulated datasets for each missingness level (Fig. 1). Hence, the different imputation methods were assessed on 150 datasets (5 missingness levels x 30 replicates).

We ran different single and multiple imputation algorithms (see *2.3 Imputation methods*) to fill the gaps in the trait data of the simulated incomplete datasets. Single imputation methods yield $m = 1$ imputed dataset per simulated dataset and here we set the multiple imputation methods to yield $m = 5$ datasets per simulated dataset to incorporate imputation uncertainty. Prior to the calculation of different performance metrics for each dataset, trait values in multiply imputed datasets were averaged (Penone et al. 2014). Performance metrics were assessed using the measured values of each trait in the IEFC complete

© Author(s) 2017. CC BY 4.0 License.

dataset (see *2.5. Statistical evaluation of the imputations)*. Note that each imputed *dataset* contains both measured and gap-filled data, but the expression 'imputed *values'* refers only to gap-filled data.

2.3 Imputation methods

We used two simple approaches to provide baseline imputations. Mean imputation ('Mean') filled missing data using the overall mean value for each trait and species mean imputation ('Spmean') replaced missing values with trait means computed for each species.

k-nearest neighbour imputation ('kNN') calculates a multivariate distance using only non-missing variables, selects the *k*-nearest plots with measured values for the target missing trait and aggregates these *k* neighbouring values to replace the missing value (R package VIM; Templ et al. 2013). We selected $k = 7$ and median aggregation (Supplement S2). We also analysed how the inclusion of auxiliary variables in the distance calculation affected imputation performance (see *2.4 Comparative assessment of imputation methods*).

We also used two geostatistical approaches. Ordinary kriging calculates a weighted average of nearby observations to predict values of a target variable in an unmeasured location, with weights that minimize prediction error and depend on spatial structure of the target variable via a variogram model (Hengl et al. 2007). Regression kriging combines a deterministic model of the target variable as a function of auxiliary variables with kriging applied to fit the residuals (Hengl et al., 2007). We included climate and forest structure variables in the model used for regression kriging imputations (cf. '*Comparative assessment of imputation methods*'), but not species identity, because there were not enough data to generate the experimental variograms for some of the less common species for all the simulations. We performed all kriging imputations with the 'autoKrige' function in the automap R package. This function tests different variogram models and applies the best-fit variogram model for kriging (Hiemstra et al. 2009).

The MICE (Multivariate Imputation using Chained Equations) algorithm (van Buuren & Groothuis-Oudshoorn, 2011; van Buuren, 2012) sequentially and iteratively imputes incomplete data, variable by variable, using individual imputation models conditionally specified by the user. One cycle through all the imputed variables is one iteration and MICE performs *t* iterations in *m* parallel streams, generating *m* multiple imputations (Supplement S3). Stochasticity is introduced in the imputation process because the parameters of the univariate imputation models are drawn from their posterior distributions, obtained using a Gibbs sampler (van Buuren, 2012). Here we use predictive mean matching (PMM) as the univariate imputation model, as it is robust to non-normality and preserves non-linear relationships between variables (Supplement S3). Several parameters must be tuned to specify the imputation models in the R implementation of MICE (mice package) to yield reliable imputations (van Buuren & Groothuis-Oudshoorn, 2011). The specific settings used in this study are assessed in Supplement S4. Please note that we will use the uppercase acronym 'MICE' to refer to the technique in general and the lowercase acronym 'mice' to refer to a particular application in this study.

2.4 Comparative assessment of imputation methods

We conducted three methodological comparisons of imputation performance. A first exercise compared 'Mean', 'OrdKrig', 'kNN' and 'mice' imputations. 'Mean' imputations used only the information on the target trait, 'OrdKrig' additionally used the spatial coordinates and 'mice' included only the information in the trait matrix.

A second exercise assessed in detail the impact on trait imputation of including additional ecological information as auxiliary predictors in MICE and kNN. We focused our detailed analysis on MICE only but we also made a simplified comparison between kNN and MICE (cf. next paragraph). The auxiliary variables we considered were species identity, a set of climatic variables (mean annual temperature, annual thermal amplitude, both in °C), a set of forest structure variables (total aboveground biomass [T ha^{-1}] and stem density [stems ha^{-1}]), a set of topographical variables (county, elevation [m.a.s.l.], slope [°] and aspect), lithology and sampling month. These predictors were complete and they did not need to be imputed themselves. The selection of the specific variables describing climate and forest structure was based on a recent analysis of trait variation in the same IEFC dataset (Vilà-Cabrera et al. 2015). We further added topographical variables, lithology and sampling month given that they may influence some trait values (Niinemets, 2015; Simpson et al. 2016). Species identity ('s'), climate ('c') and forest structure ('t') were introduced in a factorial design. Topography ('p'), lithology ('l') and sampling month '(m') were sequentially added to MICE and kNN imputations using species, climate and forest structure. Topography included spatial structure through the 'county' variable; preliminary tests using coordinates instead of 'county' did not show better results. Thus, 'mice_ctsplm' was the MICE application with the highest level of ecological information (Fig. 1).

The third exercise compared species mean imputations ('Spmean') with MICE and kNN using two different levels of auxiliary variables: (i) only species identity ('mice_s' and 'kNN_s') and (ii) the level of auxiliary information which performed best overall in the second exercise. In this same exercise, we also compared the previous approaches with 'OrdKrig' and regression kriging ('RegKrig') imputations. This third exercise thus compares a baseline scenario of 'Spmean' with imputation approaches informed either by species identity only or by an optimum level of ecological information.

2.5 Statistical evaluation of the imputations

Imputation performance was evaluated by comparing the imputed datasets with the complete, original dataset. A first set of metrics, Normalised Root Mean Square Error (NRMSE) and Kling-Gupta Efficiency (KGE), was calculated only for those

© Author(s) 2017. CC BY 4.0 License.

values that had been randomly deleted and subsequently gap-filled. We tested whether the distribution of imputed and original trait values differed using a two-sample Kolmogorov-Smirnov test, which tests the null hypothesis that two samples are identically distributed.

For each simulated dataset and trait, we calculated the Normalised Root Mean Square Error (NRMSE) as:

$$NRMSE = \sqrt{\frac{mean\left[\left(y_{imp} - y_{obs}\right)^2\right]}{var\left(y_{obs}\right)}} \text{(Eq. 1)}$$

where y_{imp} and y_{obs} represent the vectors of imputed and observed values for a given trait, respectively. Values of NRMSE approaching zero denote a better performance of the imputation method. We also calculated a dataset-averaged NRMSE by averaging the values of NRMSE for all the traits.

We further assessed imputation performance for each trait by using KGE, a goodness-of-fit measure originally developed for hydrological models, as implemented in the R package hydroGOF (Zambrano-Bigiarini, 2014):

$$KGE = 1 - \sqrt{\left(r-1\right)^2 + \left(vr-1\right)^2 + \left(\beta-1\right)^2} \text{(Eq. 2)}$$

where r is the Pearson correlation coefficient between observed and imputed values, vr is the ratio of the standard deviations between imputed and observed values and β is the ratio of imputed and observed means. The KGE range is [-∞,1], with higher values indicating better imputation performance. KGE jointly assesses correlation, bias and variability between imputed and observed values, and it is therefore a powerful, synthetic indicator of imputation quality in spatially-explicit datasets. We also calculated alternative metrics such as the R^2 between observed and imputed data, but results did not differ from those obtained using NRMSE and KGE (Supplement S5, S6, S7).

A second set of metrics compared the whole complete trait dataset Y_{obs} with the whole imputed dataset Y_{imp} (i.e. including observed and gap-filled trait values). The deviations from the original multivariate structure of the trait dataset were quantified by comparing the correlation matrices of the original and imputed datasets using the following index:

$$\Delta cormat = \sum \left| L\left[cor\left(Y_{obs}\right)\right] - L\left[cor\left(Y_{imp}\right)\right] \right| \text{(Eq. 3)}$$

Where $L\left[cor\left(Y_{obs}\right)\right]$denotes the lower triangular part of the correlation matrix of the observed dataset and $L\left[cor\left(Y_{imp}\right)\right]$ denotes the lower triangular part of the correlation matrix of the imputed dataset. Δcormat is indicative of the aggregated absolute difference between correlation matrices. Note that some traits were log-transformed before the calculation of the corresponding correlation matrix, following Vilà-Cabrera et al. (2015).

We also tested the impact of the imputation algorithms on selected bivariate trait relationships: H_{max}–WD and N_{mass}–LMA (log-transformed when necessary); as the correlation coefficients (r) of these relationships were >0.3 in absolute value and

© Author(s) 2017. CC BY 4.0 License.

were highly significant in the complete dataset. We quantified the relative difference between the complete and the imputed datasets by calculating:

$$\% \Delta r = 100 \cdot \left| r_{obs} - r_{imp} \right| / \left| r_{obs} \right| \text{(Eq. 4)}$$

Throughout the paper, we show violin plots representing the median and the distribution of each performance metric as a function of missingness levels, but we only graphically display the 10%, 30%, 50% and 80% levels, for ease of visualisation.

2.6 Imputing traits for the main forest species in Catalonia

Finally, we applied three imputation methods to gap-fill and map the five traits across all the plots in the 'IEFC incomplete dataset'. We chose 'Spmean', as the most widely used imputation method in trait-based studies 'RegKrig, as a reference geostatistical approach including auxiliary variables and 'mice_ctsp', as the best method globally, considering all traits and performance metrics (see *3. Results and discussion*). We ran 'mice_ctsp' setting *m = 50* (i.e. 50 imputations per missing value), a value closer to the missingness rate, as recommended for final MICE applications (van Buuren, 2012).

3. Results and discussion

3.1 Mean imputations compared to MICE and kNN imputations using only trait information

In general, 'mice' and 'kNN' imputations resulted in more accurate imputations in terms of NRMSE than 'Mean' at low missingness rates (10%). However, at moderate and high missingness both 'mice' and 'kNN' were comparable or outperformed by 'Mean', and specially by 'OrdKrig' (Fig. 2, Supplement S5). Recent assessments also report that the performance of MICE and kNN notably declines when missingness is ≥ 30% (Penone et al. 2014; Taugourdeau et al. 2014). Even if 'Mean' imputations imply the rather naive assumption that species identity may be unknown in a given dataset, it is nonetheless useful to compare 'Mean' imputations against 'mice' and 'kNN', which use the full trait matrix for prediction, because this comparison shows how trait covariation can be used to improve imputations. However, our results for 'OrdKrig' show that spatial structure, rather than trait covariation, may provide more accurate trait imputations when gaps are frequent (Fig. 2, Supplement S5).

As expected (Gelman and Hill, 2007), however, 'Mean' imputation severely altered trait distributions (Supplement S5), and introduced larger errors in selected trait correlations (Fig. 3). 'Mean' imputations tended to cause larger deviations in the correlation matrix, and, although kNN showed the lowest *Δcormat* below 50% missingness, its performance declined at high missingness (Supplement S5). In contrast, 'mice' closely tracked observed trait distributions (Supplement S5), introduced the least error in trait correlations under high missingness levels (Fig. 3) and yielded low *Δcormat* at extreme missingness levels (Supplement S5). Recent results also show that kNN tends to introduce larger bias in bivariate trait relationships compared to MICE (Penone et al. 2014). Despite showing the lowest *Δcormat* at 80% missingness, 'OrdKrig' imputations altered distributions and trait correlations more than 'mice' (Fig. 3, Supplement S5).

3. 2 MICE imputations using different levels of ecological information

Introducing auxiliary variables as predictors improved MICE performance substantially (Fig. 4, Supplement S6). However, these improvements were dependent on the specific predictor set and trait. Species identity increased KGE for all traits (Fig. 4) and it was the major predictor for N_{mass}, LMA and WD. Forest structure notably improved imputations for H_{max}, and only slightly for B_L:A_S. Climate only produced minor increases in KGE, but in a consistent way across traits (i.e., compare 'mice' with 'mice_c' in Fig. 4). These results are in line with the distinct role of phylogeny and environmental variables as drivers of trait variability recently observed for the same tree species in the IEFC (Laforest-Lapointe et al. 2015; Vilà-Cabrera et al. 2015). Including topography in MICE imputations only substantially improved B_L:A_S imputations, probably because the leaf area used in B_L:A_S calculations are obtained from county-level allometries, and county is one of the variables included in the topography predictor set (see *2. Methods* and Supplement S1).

In general, high levels of auxiliary information greatly improved MICE imputations. In terms of dataset-averaged NRMSE, *Δcormat* (data not shown) and preservation of trait distributions, 'mice_ctsp' imputations (including climate, forest structure, species and topography) tended to show the best performance (Supplement S6). Nevertheless, introducing sampling month in the predictor sets did not appreciably improve MICE imputations in terms of KGE (Fig. 4, Supplement S6), despite that phenological variation has been reported for some foliar traits (Niinemets, 2015; but see Fajardo & Siefert, 2016). Lithology did not appreciably improve MICE imputations, in contrast with the reported influence of soil pH on some foliar traits (Maire et al. 2015; Simpson et al. 2016).

Including auxiliary variables as predictors also decreased %Δr for selected trait relationships. In this case, lithology and sampling month slightly reduced %Δr for H_{max} – WD and N_{mass} – LMA relationships (Fig. 5). However, these reductions were not consistent across missingness levels. For example, sampling month reduced %Δr for the N_{mass} – LMA relationships at 50% but not at 80% missingness (Fig. 5).

Our results collectively suggest that, apart from species identity, different types of ecological information, particularly stand structure and topography, may improve statistical imputation schemes. In contrast, the role of climate, lithology and sampling month in improving imputations was comparatively minor. Similarly, including climate and soil data as predictors did not improve imputations of the TRY database obtained with the recently proposed Bayesian Hierarchical Product Matrix Factorisation method, (BHPMF; Schrodt et al. 2015). It is unclear, however, to what extent these results simply reflect the relatively poor quality of the climate and soil data generally available at regional scales.

3.3 Species mean imputations compared to MICE and kNN using optimum levels of ecological information

Adding auxiliary variables to calculate the distance matrix also improved kNN imputations. Values of KGE for ‘kNN_s’ and ‘kNN_ctsp’ were much higher than those observed for ‘kNN’ imputations, which only included the trait data in the distance matrix (Supplementary material, compare Fig. A7 and A12). Likewise, adding climate and forest structure as auxiliary variables improved ‘RegKrig’ performance compared to ‘OrdKrig’ (Fig. 6). For both, kNN and kriging methods, WD and H_{max} were the traits for which these improvements were largest.

In general, ‘mice_ctsp’ and ‘Spmean’ tended to be the best methods in terms of KGE, except for H_{max}, for which ‘mice_ctsp’ and ‘RegKrig’ performed best. At highest missingness levels, ‘Spmean’ showed the highest KGE for N_{mass}, LMA and WD (Fig. 6, Supplement S7). Compared to ‘Spmean’ and ‘RegKrig’, performance of MICE and kNN declined more with increasing missingness (Supplement S8), but MICE generally outperformed kNN (Fig. 2, 3, Supplement S7), as already observed in a recent imputation assessment of species-level, life-history traits (Penone et al. 2014).

MICE imputations, especially ‘mice_ctsp’ showed the lowest %Δr for the studied trait correlations, although ‘Spmean’ also showed low values for for the N_{mass}–LMA relationship (Fig. 7). Neither kNN nor kriging imputations succeeded in minimising changes in trait correlations; for kriging imputations, only ‘RegKrig’ showed a relatively high value of %Δr for the H_{max}–WD relationship (Fig. 7). Kernel density plots and Kolmogorov-Smirnov tests showed that MICE produced imputations (especially ‘mice_ctsp’) most consistent with observed distributions at all missingness levels (Supplement S7). ‘Spmean’ and ‘OrdKrig’ imputations modified trait distributions substantially, while’kNN_ctsp’ and ‘RegKrig’ showed an intermediate performance, but generally far from that of ‘mice_ctsp’ (Supplement S7). ‘Spmean’ and kriging imputations also yielded larger *Δcormat* values (data not shown).

The application of ‘mice_ctsp’ successfully filled the gaps in the IEFC incomplete dataset and quantified the variation among the multiple imputations, providing an estimation of the level of confidence in the imputed values for specific traits (Fig. 8). ‘Spmean’ and ‘RegKrig’ show a largely similar spatial pattern of trait variation compared to ‘mice_ctsp’, although some discrepancies between ‘Spmean’ and ‘mice_ctsp’ can be observed in the north-eastern pre-litoral and coastal area for LMA (Fig. S23). Here, ‘Spmean’ imputations tend to predict lower values compared to ‘mice_ctsp’ imputations. These areas are mostly dominated by *Quercus suber forests* (Supplement S1), and LMA was only measured in 5 out of the 149 plots of this species present in the IEFC incomplete dataset. Therefore, as there is little information on trait covariation for the imputation of LMA in *Q. suber* plots, MICE imputations are largely based on the auxiliary variables and they yield a distinct spatial pattern of trait variation, compared to ‘Spmean’.

4 Implications

The problem of missing data is ubiquitous in plant trait datasets of regional to global scope. Nevertheless, ecologists have recently made substantial progress in (i) the assessment of the best imputation methods in trait-based applications, (ii) how these methods perform with increasing missingness , (iii) which ecological covariates aid to improve imputations and (iv) how different imputation methods impact the results of trait-based analyses (Pakeman, 2014, Taugourdeau et al. 2014, Penone et al. 2014, Schrodt et al. 2015). Most effort thus far, however, has been directed at imputing species-level trait means and all the abovementioned questions have rarely been assessed on the same dataset. Here we deal here with all the previous issues simultaneously and focusing on the spatial component of trait variability, where the intra-specific component cannot be neglected.

Our results show that, in terms of trait prediction error, no imputation method performs best consistently for the five studied traits. However, when all performance metrics are jointly considered (i.e. errors in trait prediction, multivariate trait distribution and trait correlations), MICE informed by relevant ecological variables outperforms approaches based on trait averaging, geostatistical models and kNN methods, albeit this superiority of MICE tends to vanish at higher missingness. For kNN, MICE and kriging imputations we have highlighted the key role of auxiliary variables as necessary covariates to yield reliable imputations in spatially explicit settings. The importance of covariates differed across traits, but, in addition to the expected influence of species, climate and topography in predicting trait values, we also showed a prominent role of stand structure for some traits. The ongoing development of global databases of vegetation structure (e.g. Dengler *et al.* 2014) will likely enable the incorporation of stand variables in future imputation exercises.

Given the limited number of species in our study, reflecting the relatively low richness of the studied communities, taxonomic information introduced as species identity was enough to improve imputations of all studied traits. However, in studies coping with a larger set of species, phylogeny may need to be considered in the imputation models (Schrodt et al. 2015, Swenson et al. 2017). For global trait datasets, a combination of imputation with data augmentation approaches (e.g. Nakagawa & Freckleton, 2008) has been proposed to minimise potential errors in trait-driven analyses caused by incomplete and biased species sampling (Sandel et al. 2015).

Compared to other imputation approaches, MICE is well-suited to deal with multivariate missing data (i.e. MICE produce imputations when some predictors are also missing) and provides information to quantify the uncertainty associated with the imputed data (Fig. 8). MICE also provides a full framework for inference from incomplete datasets. Because our comparative assessment of imputation methods is already complex, here we have only dealt with *imputation*, the first step of the full process (e.g. Nakagawa & Freckleton 2008), but MICE produces multiple datasets that are individually combined in the *analysis* and *pooling* steps. The analysis step refers to the estimation of the parameters of scientific interest (e.g. a regression coefficient) for each dataset. In MICE, parameters are then pooled across datasets to produce unbiased estimates and standard errors, providing a natural way to take into account the additional uncertainty introduced in the analysis by the presence of missing data (van Buuren 2012). However, ecological studies using multiple imputation approaches usually only apply the imputation step (Baraloto et al. 2010, Paine et al. 2011, Pyšek et al 2015, Díaz et al. 2016) and do not take advantage of the multiple imputation framework to quantify the uncertainty resulting from the presence of missing data (but see Fisher et al. 2003).

Our results have important implications given that the demand for spatially explicit datasets is increasing rapidly and that species mean imputation and casewise data deletion are still widespread practices in trait-based ecology. We show that species mean imputation may result in substantial information loss that may hinder research development on important topics in functional biogeography, such as the ecological significance of intraspecific trait variability (e.g. Siefert et al. 2015), the synthesis of plant form and function (Díaz et al. 2016) or the development of trait-driven modelling approaches (Yang et al. 2015). We also show that spatially-distributed layers of ecological information for trait imputation, as shown here for MICE, can thus be used to obtain trait maps to inform trait-driven ecosystem process models (Christoffersen et al. 2016).

Data availability

The IEFC complete trait dataset will be made available in a public repository.

Author contributions

RP, OS, JMV and MM conceived the study and all authors contributed to design the simulations and statistiscal analyses. RP, OS and LB carried out the simulations and analyses, assisted by the rest of the authors. RP wrote the paper with the contribution of all the coauthors.

Competing interests

The authors declare that they have no conflict of interest

Acknowledgements

We thank J. Vayreda, A. Vilà-Cabrera and S. Saura-Mas for their help with the IEFC dataset. This study was funded by the Spanish Ministry of Economy and Competitiveness (MINECO) through the grants CGL2013-46808-R, CGL2014-55883-JIN and MTM2015-69493-R.

References

Azur, M. J., Stuart, E. A., Frangakis, C. and Leaf, P. J.: Multiple imputation by chained equations: what is it and how does it work?, Int. J. Methods Psychiatr. Res., 20(1), 40–49, doi:10.1002/mpr.329, 2011.

Baraloto, C., Timothy Paine, C. E., Poorter, L., Beauchene, J., Bonal, D., Domenach, A.-M., Hérault, B., Patiño, S., Roggy, J.-C. and Chave, J.: Decoupled leaf and stem economics in rain forest trees, Ecology Letters, 13(11), 1338–1347, doi:10.1111/j.1461-0248.2010.01517.x, 2010.

Blonder, B.: Do Hypervolumes Have Holes?, The American Naturalist, 187(4), E93–E105, doi:10.1086/685444, 2016.

van Buuren, S.: Flexible Imputation of Missing Data, CRC Press., 2012.

van Buuren, S. and Groothuis-Oudshoorn, K.: mice: Multivariate Imputation by Chained Equations in R, Journal of Statistical Software, 45(3) [online] Available from: http://www.jstatsoft.org/v45/i03 (Accessed 9 April 2015), 2011.

Chave, J., Coomes, D., Jansen, S., Lewis, S. L., Swenson, N. G. and Zanne, A. E.: Towards a worldwide wood economics spectrum, Ecology Letters, 12(4), 351–366, doi:10.1111/j.1461-0248.2009.01285.x, 2009.

Christoffersen, B. O., Gloor, M., Fauset, S., Fyllas, N. M., Galbraith, D. R., Baker, T. R., Kruijt, B., Rowland, L., Fisher, R. A., Binks, O. J., Sevanto, S., Xu, C., Jansen, S., Choat, B., Mencuccini, M., McDowell, N. G. and Meir, P.: Linking hydraulic traits to tropical forest function in a size-structured and trait-driven model (TFS v.1-Hydro), Geosci. Model Dev., 9(11), 4227–4255, doi:10.5194/gmd-9-4227-2016, 2016.

Dengler, J., Bruelheide, H., Purschke, O., Chytrỳ, M., Jansen, F., Hennekens, S. M., Jandt, U., Jiménez-Alfaro, B., Kattge, J., Pillar, V. D., Sandel, B., Winter, M. and sPlot Consortium: sPlot–the new global vegetation-plot database for addressing trait-environment relationships across the world's biomes, in Biodiversity and vegetation: patterns, processes, conservation, edited by L. Mucina, J. N. Price, and Kalwi, J.M., p. 90, Kwongan Foundation, Perth, AU. [online] Available from: https://www.bayceer.uni-bayreuth.de/pfloek/de/pub/pub/130464/JD196_Dengler_et_al_2014_IAVS_Proceedings.pdf (Accessed 21 November 2016), 2014.

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., Chave, J., Wright, S. J., Sheremet'ev, S. N., Jactel, H., Baraloto, C., Cerabolini, B., Pierce, S., Shipley, B., Kirkup, D., Casanoves, F., Joswig, J. S., Günther, A., Falczuk, V., Rüger, N., Mahecha, M. D. and Gorné, L. D.: The global spectrum of plant form and function, Nature, 529, 161–171, doi:10.1038/nature16489, 2015.

Doove, L. L., Van Buuren, S. and Dusseldorp, E.: Recursive partitioning for missing data imputation in the presence of interaction effects, Computational Statistics & Data Analysis, 72, 92–104, doi:10.1016/j.csda.2013.10.025, 2014.

Eskelson, B. N. I., Temesgen, H., Lemay, V., Barrett, T. M., Crookston, N. L. and Hudak, A. T.: The roles of nearest neighbor methods in imputing missing data in forest inventory and monitoring databases, Scandinavian Journal of Forest Research, 24(3), 235–246, doi:10.1080/02827580902870490, 2009.

© Author(s) 2017. CC BY 4.0 License.

Fajardo, A. and Siefert, A.: Phenological variation of leaf functional traits within species, Oecologia, 180(4), 951–959, doi:10.1007/s00442-016-3545-1, 2016.

Fisher, D. O., Blomberg, S. P. and Owens, I. P. F.: Extrinsic versus intrinsic factors in the decline and extinction of Australian marsupials., Proc Biol Sci, 270(1526), 1801–1808, doi:10.1098/rspb.2003.2447, 2003.

Funk, J. L., Larson, J. E., Ames, G. M., Butterfield, B. J., Cavender-Bares, J., Firn, J., Laughlin, D. C., Sutton-Grier, A. E., Williams, L. and Wright, J.: Revisiting the Holy Grail: using plant functional traits to understand ecological processes, Biol Rev, 92, 1156–1173, doi:10.1111/brv.12275, 2016.

Gelman, A. and Hill, J.: Data Analysis Using Regression and Multilevel/Hierarchical Models, Cambridge University Press., 2007.

Gracia, C., Burriel, J., Mata, T. and Vayreda, J.: Inventari Ecològic i Forestal de Catalunya: Centre de Recerca Ecològica i Aplicacions Forestals. 10 volumes., CREAF, Bellaterra, Spain., 2000.

Hengl, T., Heuvelink, G. B. M. and Rossiter, D. G.: About regression-kriging: From equations to case studies, Computers & Geosciences, 33(10), 1301–1315, doi:10.1016/j.cageo.2007.05.001, 2007.

Hiemstra, P. H., Pebesma, E. J., Twenhöfel, C. J. W. and Heuvelink, G. B. M.: Real-time automatic interpolation of ambient gamma dose rates from the Dutch radioactivity monitoring network, Computers & Geosciences, 35(8), 1711–1721, doi:10.1016/j.cageo.2008.10.011, 2009.

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. and Ustin, S. L.: Monitoring plant functional diversity from space, Nature Plants, 2(3), 16024, doi:10.1038/nplants.2016.24, 2016.

Kattge, J., Díaz, S., Lavorel, S., Prentice, I. C., Leadley, P., Bönisch, G., Garnier, E., Westoby, M., Reich, P. B., Wright, I. J., Cornelissen, J. H. C., Violle, C., Harrison, S. P., Van Bodegom, P. M., Reichstein, M., Enquist, B. J., Soudzilovskaia, N. A., Ackerly, D. D., Anand, M., Atkin, O., Bahn, M., Baker, T. R., Baldocchi, D., Bekker, R., Blanco, C. C., Blonder, B., Bond, W. J., Bradstock, R., Bunker, D. E., Casanoves, F., Cavender-Bares, J., Chambers, J. Q., Chapin Iii, F. S., Chave, J., Coomes, D., Cornwell, W. K., Craine, J. M., Dobrin, B. H., Duarte, L., Durka, W., Elser, J., Esser, G., Estiarte, M., Fagan, W. F., Fang, J., Fernández-Méndez, F., Fidelis, A., Finegan, B., Flores, O., Ford, H., Frank, D., Freschet, G. T., Fyllas, N. M., Gallagher, R. V., Green, W. A., Gutierrez, A. G., Hickler, T., Higgins, S. I., Hodgson, J. G., Jalili, A., Jansen, S., Joly, C. A., Kerkhoff, A. J., Kirkup, D., Kitajima, K., Kleyer, M., Klotz, S., Knops, J. M. H., Kramer, K., Kühn, I., Kurokawa, H., Laughlin, D., Lee, T. D., Leishman, M., Lens, F., Lenz, T., Lewis, S. L., Lloyd, J., Llusià, J., Louault, F., Ma, S., Mahecha, M. D., Manning, P., Massad, T., Medlyn, B. E., Messier, J., Moles, A. T., Müller, S. C., Nadrowski, K., Naeem, S., Niinemets, Ü., Nöllert, S., Nüske, A., Ogaya, R., Oleksyn, J., Onipchenko, V. G., Onoda, Y., Ordoñez, J., Overbeck, G., et al.: TRY – a global database of plant traits, Global Change Biology, 17(9), 2905–2935, doi:10.1111/j.1365-2486.2011.02451.x, 2011.

Laforest-Lapointe, I., Martínez-Vilalta, J. and Retana, J.: Intraspecific variability in functional traits matters: case study of Scots pine, Oecologia, 175(4), 1337–1348, doi:10.1007/s00442-014-2967-x, 2014.

Little, R. J. A. and Rubin, D. B.: Statistical Analysis with Missing Data, Wiley., 2002.

Maire, V., Wright, I. J., Prentice, I. C., Batjes, N. H., Bhaskar, R., van Bodegom, P. M., Cornwell, W. K., Ellsworth, D., Niinemets, Ü., Ordonez, A., Reich, P. B. and Santiago, L. S.: Global effects of soil and climate on leaf photosynthetic traits and rates, Global Ecology and Biogeography, 24(6), 706–717, doi:10.1111/geb.12296, 2015.

McGill, B. J., Enquist, B. J., Weiher, E. and Westoby, M.: Rebuilding community ecology from functional traits, Trends in Ecology & Evolution, 21(4), 178–185, doi:10.1016/j.tree.2006.02.002, 2006.

Nakagawa, S.: Missing data: mechanisms, methods, and messages, in Ecological statistics: contemporary theory and application, edited by G. Fox, S. Negrete-Yankelevich, and V. J. Sosa, pp. 81–105, Oxford University Press., 2015.

Nakagawa, S. and Freckleton, R. P.: Missing inaction: the dangers of ignoring missing data, Trends in Ecology & Evolution, 23(11), 592–596, doi:10.1016/j.tree.2008.06.014, 2008.

Niinemets, Ü.: Is there a species spectrum within the world-wide leaf economics spectrum? Major variations in leaf functional traits in the Mediterranean sclerophyll Quercus ilex, New Phytol, 205(1), 79–96, doi:10.1111/nph.13001, 2015.

Ninyerola, M., Pons, X., Roure, J. M., Ninyerola, M., Pons, X. and Roure, J. M.: A methodological approach of climatological modelling of air temperature and precipitation through GIS techniques, A methodological approach of climatological modelling of air temperature and precipitation through GIS techniques, International Journal of Climatology, International Journal of Climatology, 20(14, 14), 1823–1841, doi:10.1002/1097-0088(20001130)20:14<1823::AID-JOC566>3.0.CO;2-B, 10.1002/1097-0088(20001130)20:14<1823::AID-JOC566>3.0.CO;2-B, 2000.

Paine, C. E. T., Baraloto, C., Chave, J. and Hérault, B.: Functional traits of individual trees reveal ecological constraints on community assembly in tropical rain forests, Oikos, 120(5), 720–727, doi:10.1111/j.1600-0706.2010.19110.x, 2011.

Pakeman, R. J.: Functional trait metrics are sensitive to the completeness of the species' trait data?, Methods Ecol Evol, 5(1), 9–15, doi:10.1111/2041-210X.12136, 2014.

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.: Imputation of missing data in life-history trait datasets: which approach performs the best?, Methods Ecol Evol, 5(9), 961–970, doi:10.1111/2041-210X.12232, 2014.

Pyšek, P., Manceur, A. M., Alba, C., McGregor, K. F., Pergl, J., Štajerová, K., Chytrý, M., Danihelka, J., Kartesz, J., Klimešová, J., Lučanová, M., Moravcová, L., Nishino, M., Sádlo, J., Suda, J., Tichý, L. and Kühn, I.: Naturalization of central European plants in North America: species traits, habitats, propagule pressure, residence time, Ecology, 96(3), 762–774, doi:10.1890/14-1005.1, 2015.

© Author(s) 2017. CC BY 4.0 License.

R Core Team: R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria. [online] Available from: https://www.R-project.org/, 2016.

Reich, P. B.: The world-wide 'fast–slow' plant economics spectrum: a traits manifesto, J Ecol, 102(2), 275–301, doi:10.1111/1365-2745.12211, 2014.

Reichstein, M., Bahn, M., Mahecha, M. D., Kattge, J. and Baldocchi, D. D.: Linking plant and ecosystem functional biogeography, PNAS, 111(38), 13697–13702, doi:10.1073/pnas.1216065111, 2014.

Rezvan, P. H., Lee, K. J. and Simpson, J. A.: The rise of multiple imputation: a review of the reporting and implementation of the method in medical research, BMC Medical Research Methodology, 15(1), 30, doi:10.1186/s12874-015-0022-1, 2015.

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., Wright, S. J. and Reich, P. B.: BHPMF – a hierarchical Bayesian approach to gap-filling and trait prediction for macroecology and functional biogeography, Global Ecology and Biogeography, 24, 1510–1521, doi:10.1111/geb.12335, 2015.

Shipley, B., Bello, F. D., Cornelissen, J. H. C., Laliberté, E., Laughlin, D. C. and Reich, P. B.: Reinforcing loose foundation stones in trait-based plant ecology, Oecologia, 180(4), 923–931, doi:10.1007/s00442-016-3549-x, 2016.

Siefert, A., Violle, C., Chalmandrier, L., Albert, C. H., Taudiere, A., Fajardo, A., Aarssen, L. W., Baraloto, C., Carlucci, M. B., Cianciaruso, M. V., de L. Dantas, V., de Bello, F., Duarte, L. D. S., Fonseca, C. R., Freschet, G. T., Gaucherand, S., Gross, N., Hikosaka, K., Jackson, B., Jung, V., Kamiyama, C., Katabuchi, M., Kembel, S. W., Kichenin, E., Kraft, N. J. B., Lagerström, A., Bagousse-Pinguet, Y. L., Li, Y., Mason, N., Messier, J., Nakashizuka, T., Overton, J. M., Peltzer, D. A., Pérez-Ramos, I. M., Pillar, V. D., Prentice, H. C., Richardson, S., Sasaki, T., Schamp, B. S., Schöb, C., Shipley, B., Sundqvist, M., Sykes, M. T., Vandewalle, M. and Wardle, D. A.: A global meta-analysis of the relative extent of intraspecific trait variation in plant communities, Ecol Lett, 18(12), 1406–1419, doi:10.1111/ele.12508, 2015.

Simpson, A. H., Richardson, S. J. and Laughlin, D. C.: Soil–climate interactions explain variation in foliar, stem, root and reproductive traits across temperate forests, Global Ecol. Biogeogr., 25, 964–978, doi:10.1111/geb.12457, 2016.

Stekhoven, D. J. and Bühlmann, P.: MissForest—non-parametric missing value imputation for mixed-type data, Bioinformatics, 28(1), 112–118, doi:10.1093/bioinformatics/btr597, 2012.

Swenson, N. G.: Phylogenetic imputation of plant functional trait databases, Ecography, 37(2), 105–110, doi:10.1111/j.1600-0587.2013.00528.x, 2014.

Taugourdeau, S., Villerd, J., Plantureux, S., Huguenin-Elie, O. and Amiaud, B.: Filling the gap in functional trait databases: use of ecological hypotheses to replace missing data, Ecol Evol, 4(7), 944–958, doi:10.1002/ece3.989, 2014.

Templ, M., Alfons, A., Kowarik, A. and Prantner, B.: VIM: Visualization and Imputation of Missing Values. R package version 3.0. 3.1., 2013.

Troyanskaya, O., Cantor, M., Sherlock, G., Brown, P., Hastie, T., Tibshirani, R., Botstein, D. and Altman, R. B.: Missing value estimation methods for DNA microarrays, Bioinformatics, 17(6), 520–525, doi:10.1093/bioinformatics/17.6.520, 2001.

Vilà-Cabrera, A., Martínez-Vilalta, J. and Retana, J.: Functional trait variation along environmental gradients in temperate and Mediterranean trees, Global Ecology and Biogeography, 24(12), 1377–1389, doi:10.1111/geb.12379, 2015.

Violle, C., Navas, M.-L., Vile, D., Kazakou, E., Fortunel, C., Hummel, I. and Garnier, E.: Let the concept of trait be functional!, Oikos, 116(5), 882–892, doi:10.1111/j.0030-1299.2007.15559.x, 2007.

Westoby, M. and Wright, I. J.: Land-plant ecology on the basis of functional traits, Trends in Ecology & Evolution, 21(5), 261–268, doi:10.1016/j.tree.2006.02.004, 2006.

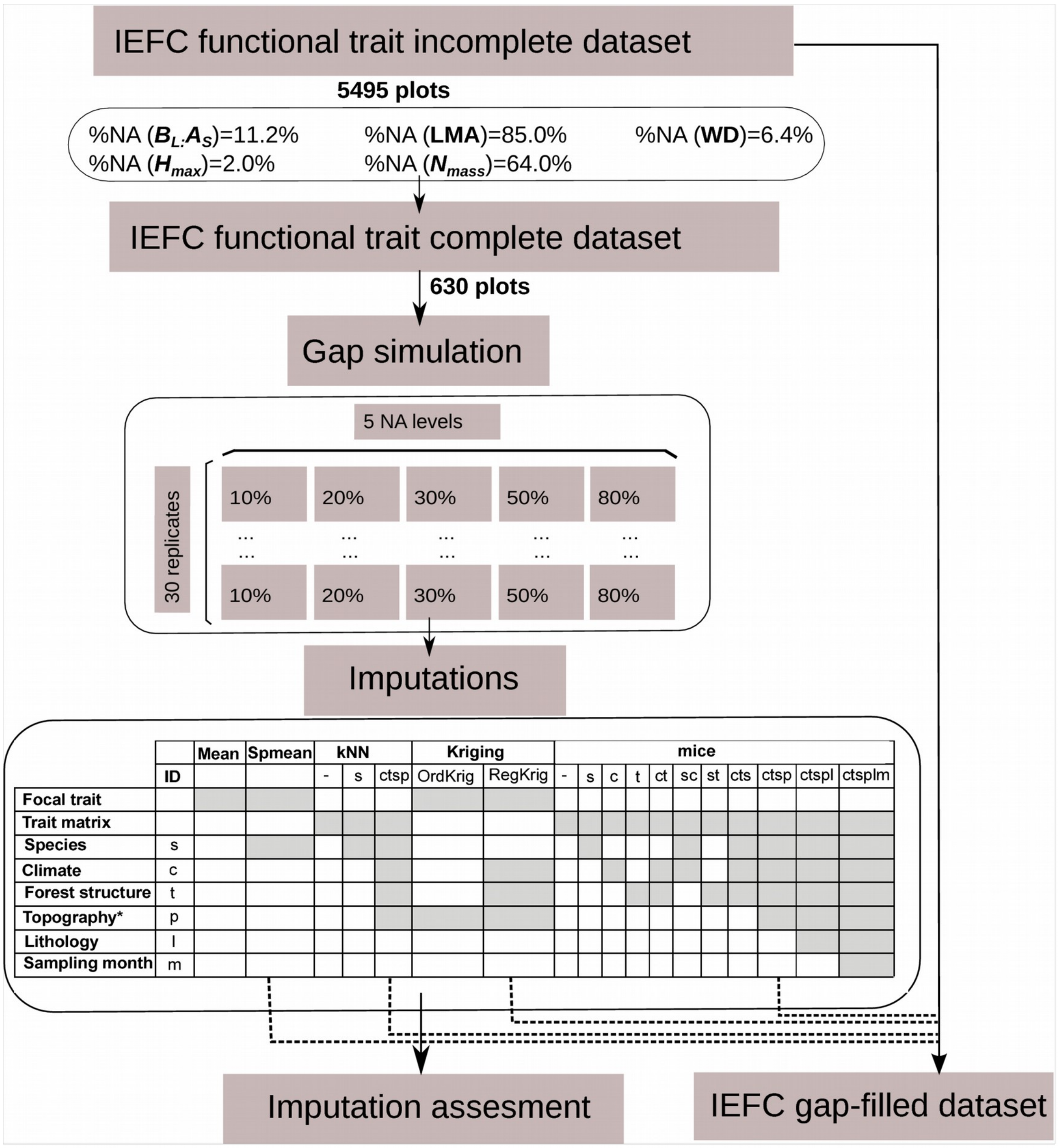
Westoby, M., Falster, D. S., Moles, A. T., Vesk, P. A. and Wright, I. J.: Plant Ecological Strategies: Some Leading Dimensions of Variation between Species, Annual Review of Ecology and Systematics, 33, 125–159, doi:10.2307/3069259, 2002.

Wright, I. J., Reich, P. B., Westoby, M., Ackerly, D. D., Baruch, Z., Bongers, F., Cavender-Bares, J., Chapin, T., Cornelissen, J. H. and Diemer, M.: The worldwide leaf economics spectrum, Nature, 428(6985), 821–827, 2004.

Zambrano-Bigiarini, M.: hydroGOF: Goodness-of-fit functions for comparison of simulated and observed hydrological time series. [online] Available from: https://cran.r-project.org/web/packages/hydroGOF/index.html (Accessed 15 July 2016), 2014.

© Author(s) 2017. CC BY 4.0 License.

520



		Mean	Spmean	kNN			Kriging		mice										
	ID			-	s	ctsp	OrdKrig	RegKrig	-	s	c	t	ct	sc	st	cts	ctsp	ctspl	ctsplm
Focal trait		▨	▨				▨	▨											
Trait matrix				▨	▨	▨			▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨
Species	s		▨		▨	▨				▨				▨		▨	▨	▨	▨
Climate	c					▨		▨			▨		▨	▨		▨	▨	▨	▨
Forest structure	t					▨		▨				▨	▨		▨	▨	▨	▨	▨
Topography*	p					▨	▨	▨									▨	▨	▨
Lithology	l																	▨	▨
Sampling month	m																		▨

Figure 1. Description of the experimental design. A subset was obtained from the incomplete IEFC trait dataset containing only plots where all functional traits had been measured (complete dataset) to perform the gap simulations and the imputations. Imputation methods are described in terms of the input information used. The selected methods for the final application of imputation methods to obtain a gap-filled IEFC trait dataset are also shown. 525

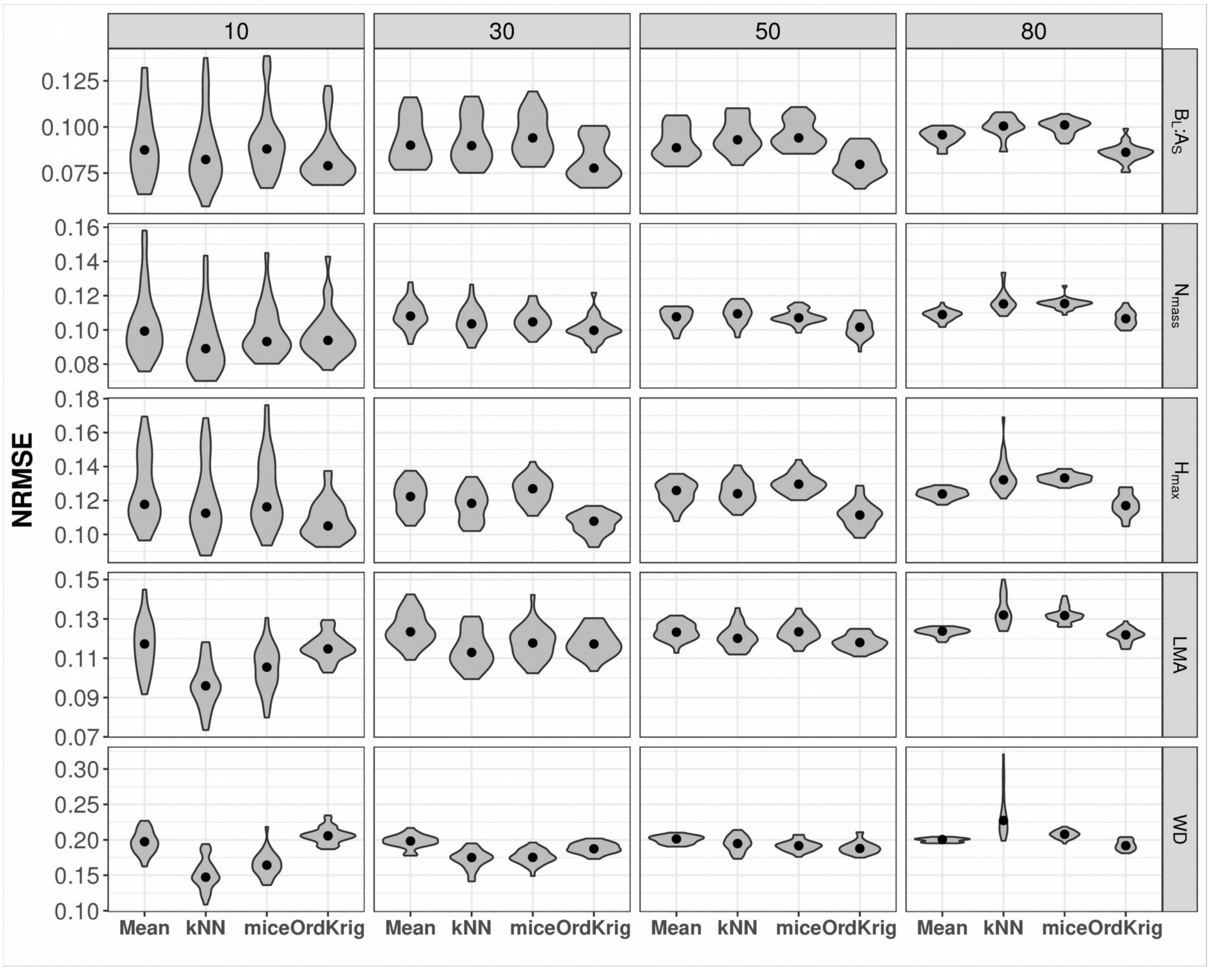


Figure 2. Trait-specific NRMSE at increasing missingness levels (10% to 80%) for different imputation methods: overall trait mean (Mean), mice (using only the trait matrix in the predictor set), kNN (using only the trait matrix for the distance calculation) and ordinary kriging (OrdKrig).

© Author(s) 2017. CC BY 4.0 License.

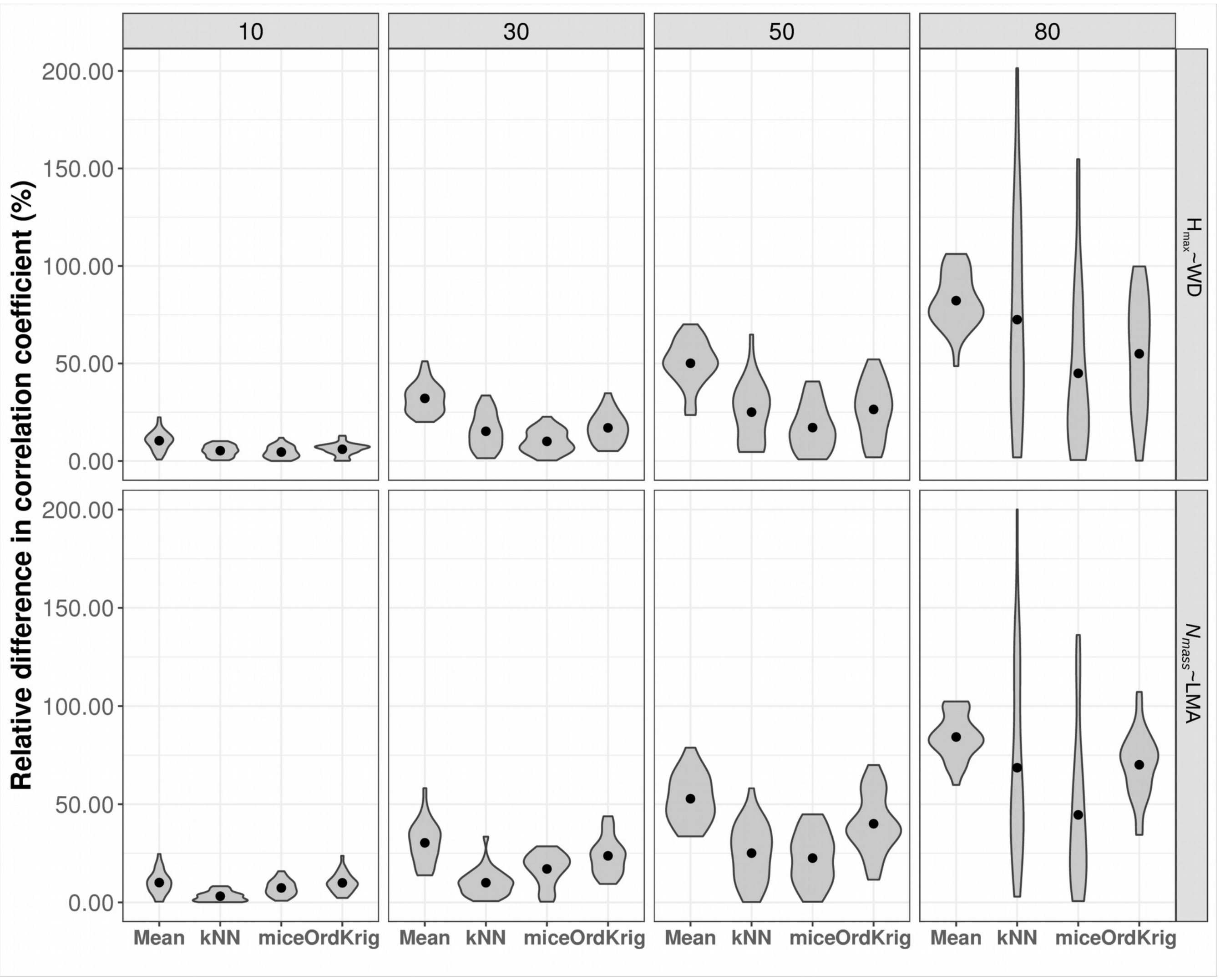


Figure 3. Errors in the correlation coefficient for two selected trait relationships, at increasing missingness levels (10% to 80%) and for different imputation methods: overall trait mean (Mean), mice (using only the trait matrix in the predictor set), kNN (using only the trait matrix for the distance calculation) and ordinary kriging (OrdKrig).

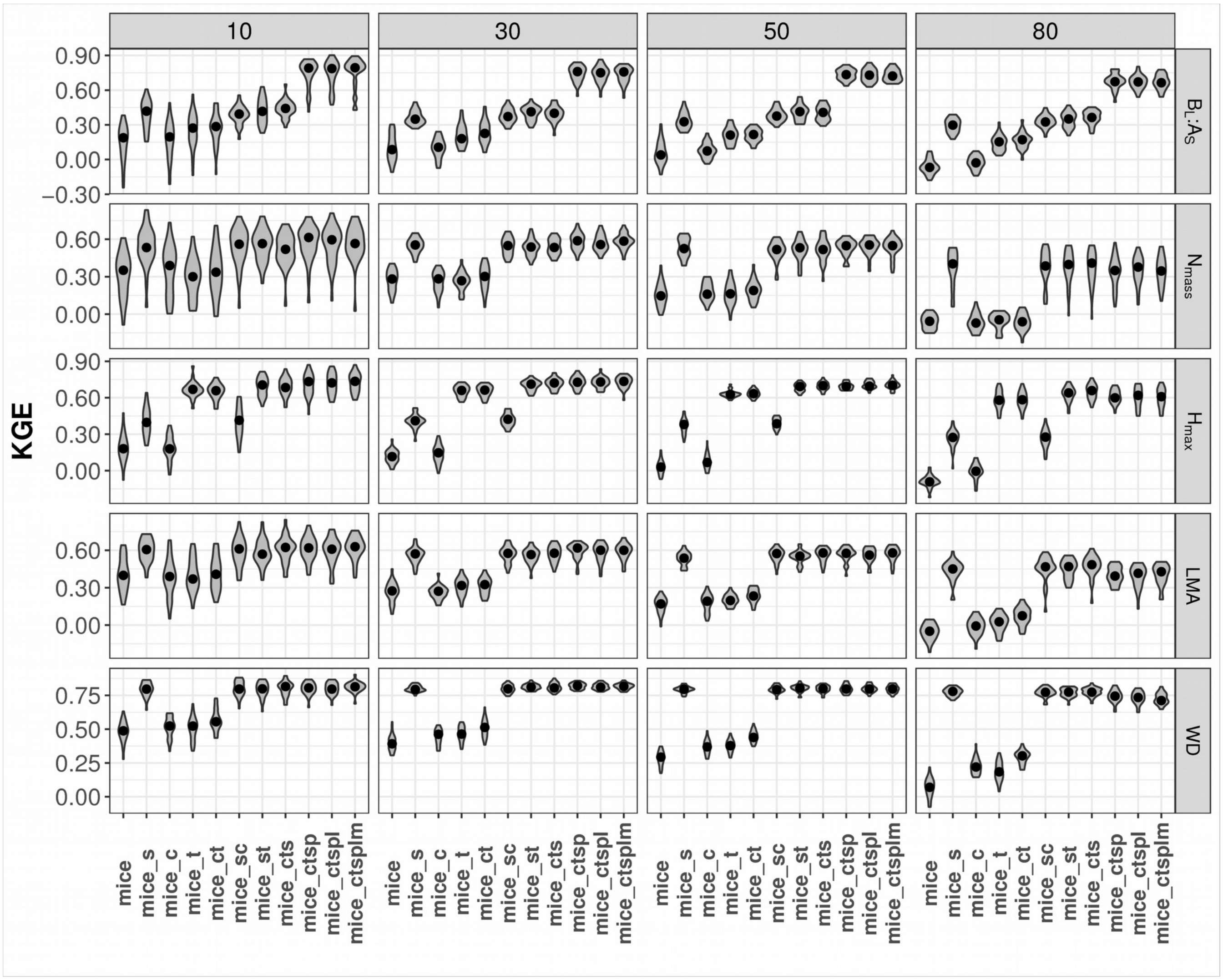


Figure 4. Trait-specific KGE at increasing missingness levels (10% to 80%) and for different MICE imputations using different combinations of additional predictor sets: species identity (s), climate (c), forest structure (t), sptatial structure (p), lithology (l) and sampling month (m). See Fig. 1 for an overall view of the experimental design and the Methods section for a detailed description of the variables employed in each predictor set.

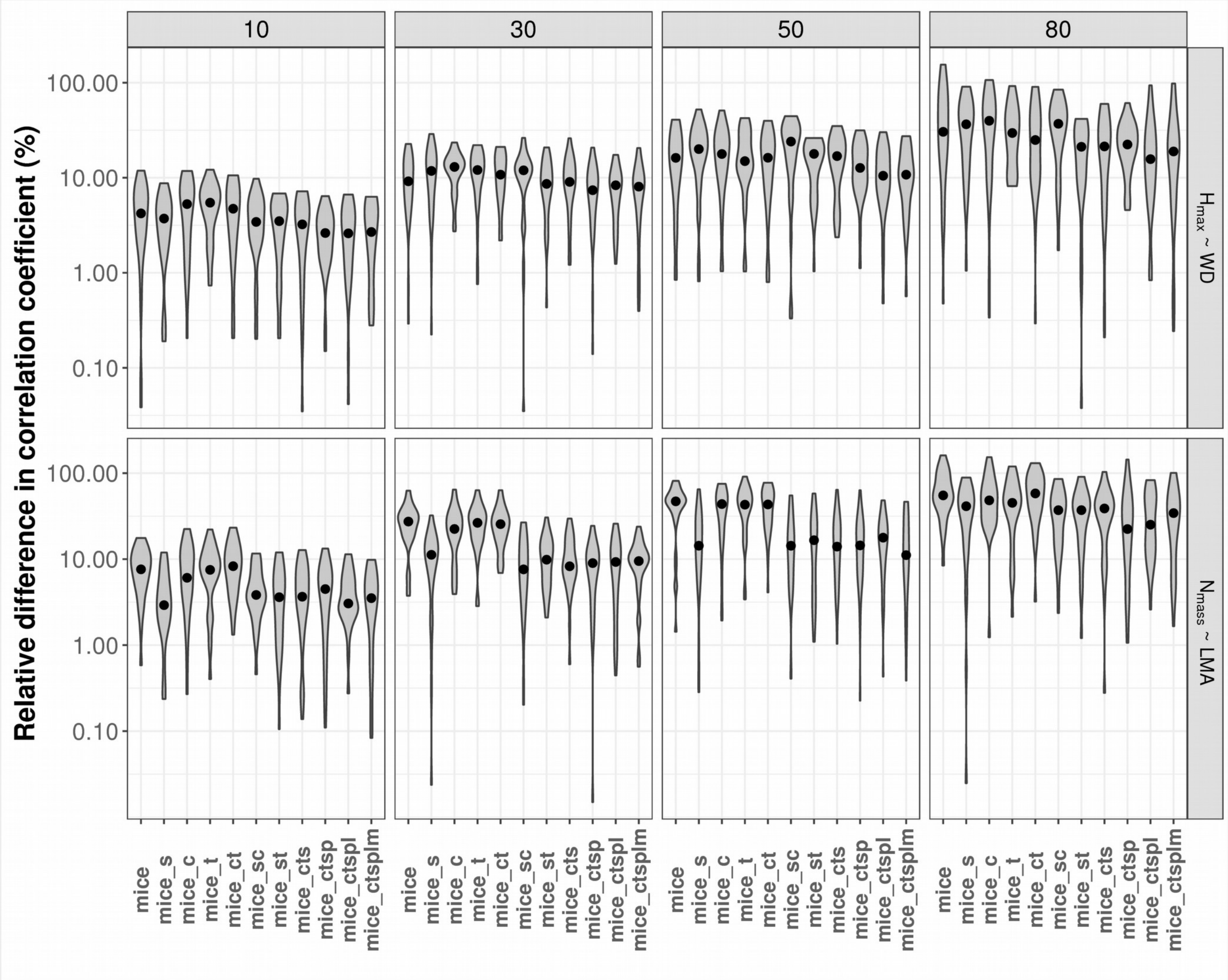
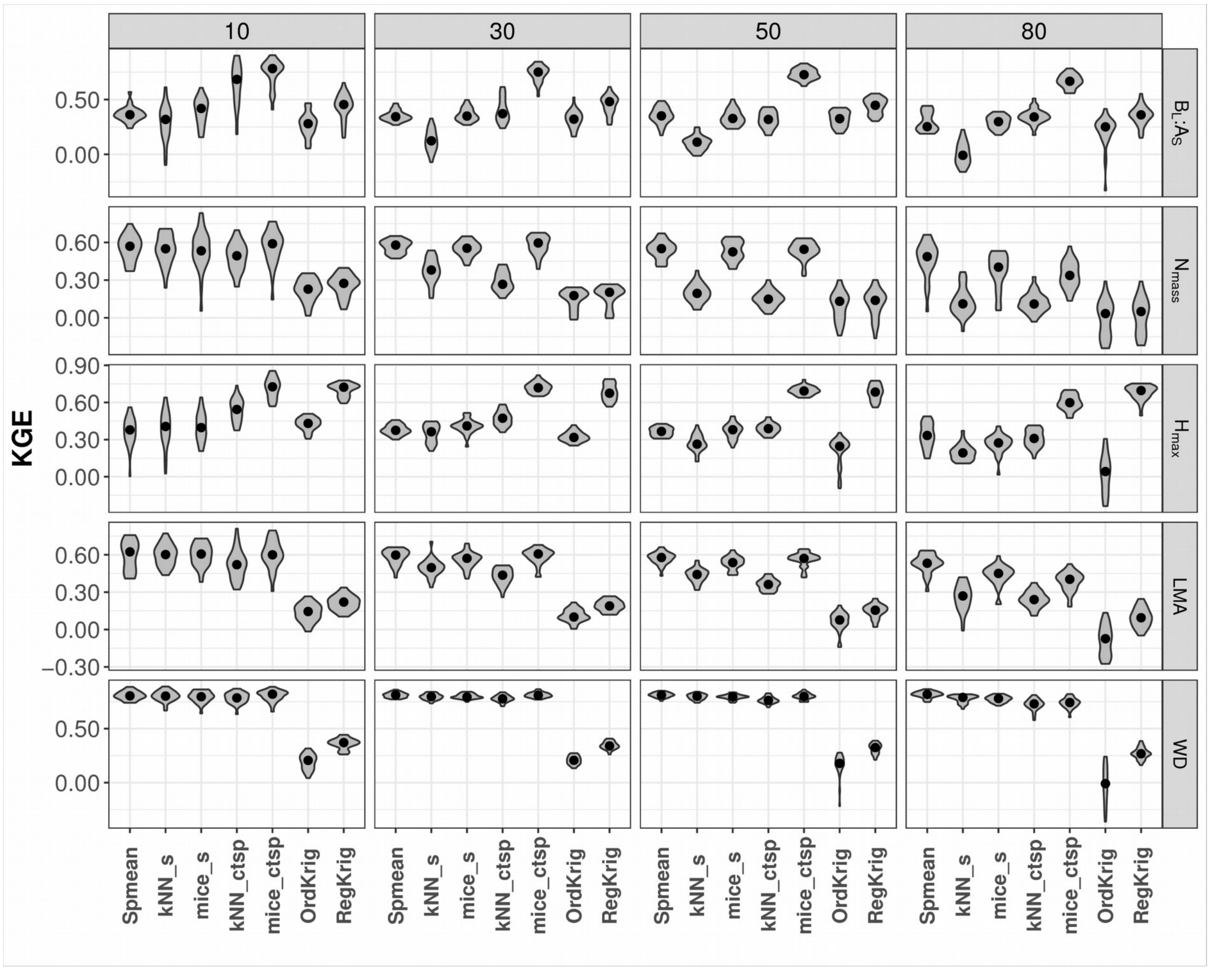


Figure 5. Errors in the correlation coefficient for two selected trait relationships, at increasing missingness levels (10% to 80%) and for different MICE imputations using different combinations of additional predictor sets: species identity (s), climate (c), forest structure (t), topography (p), lithology (l) and sampling month (m). See Fig. 1 for an overall view of the experimental design and the Methods section for a detailed description of the variables employed in each predictor set. Note that the y-axis is in the logarithmic scale.

© Author(s) 2017. CC BY 4.0 License.



545 **Figure 6.** Trait-specific KGE at increasing missingness levels (10% to 80%) for different imputation methods: species mean (Spmean), mice and kNN with species as predictor (mice_s and kNN_s, respectively), mice and kNN with species, climate, forest structure and spatial variables as predictors (mice_ctsp and kNN_ctsp, respectively) and universal krigingregression kriging (RegKrig).

© Author(s) 2017. CC BY 4.0 License.

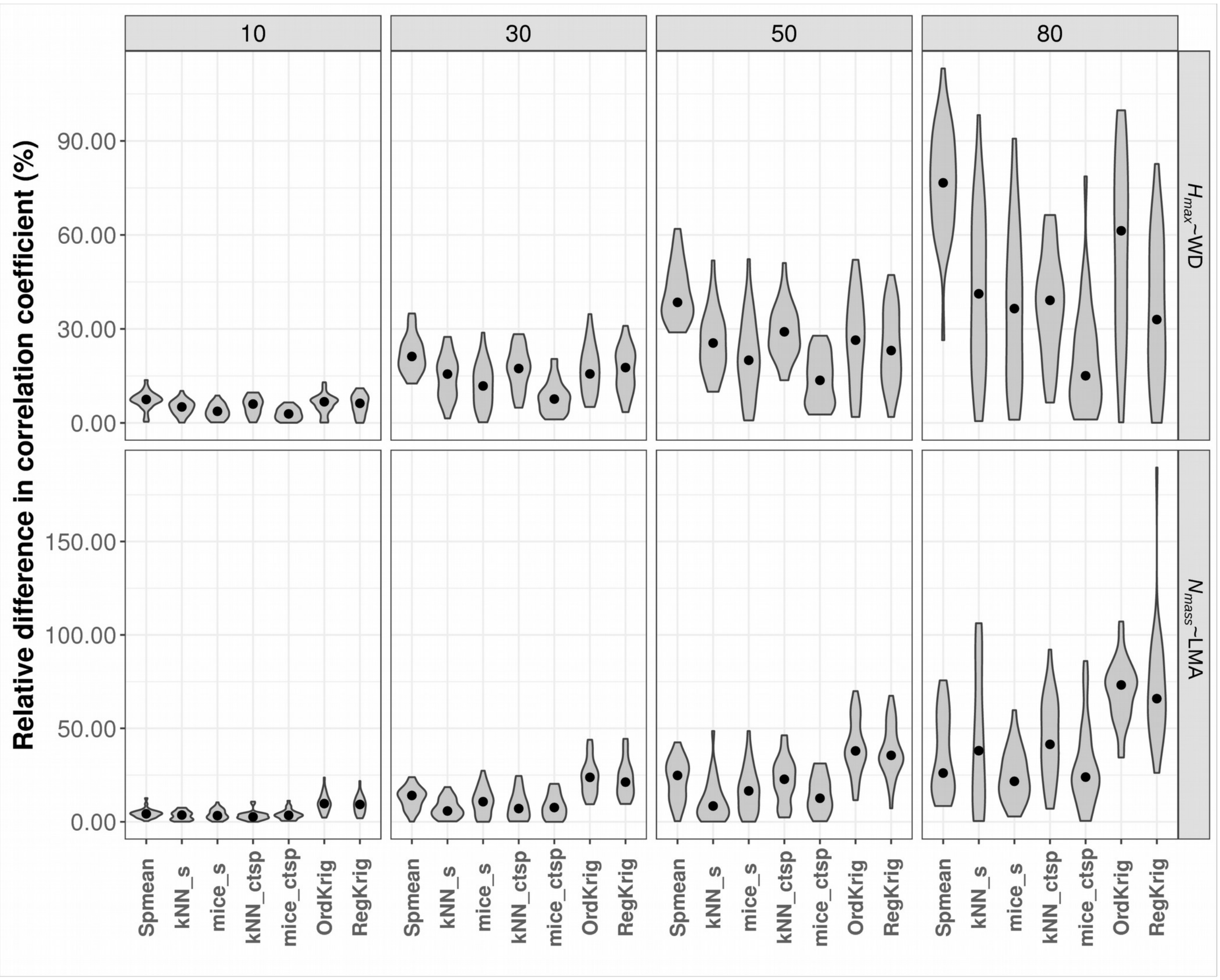


Figure 7. Errors in the correlation coefficient for two selected trait relationships, at increasing missingness levels (10% to 80%) and for different imputation methods: species mean (Spmean), mice and kNN with species as predictor (mice_s and kNN_s, respectively), mice and kNN with species, climate, forest structure and spatial variables as predictors (mice_ctsp and kNN_ctsp, respectively) and universal krigingregression kriging (RegKrig).

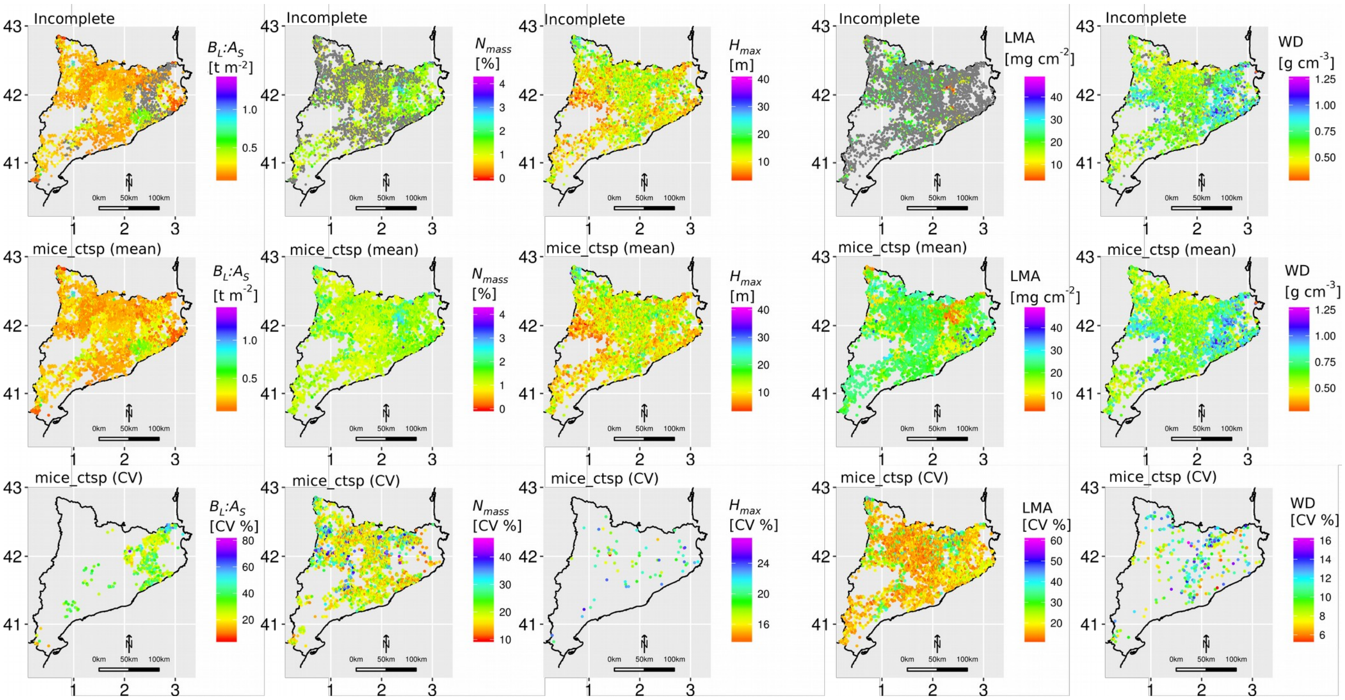


Figure 8. Maps with the distribution of functional traits across the selected plots in the IEFC. The first row shows the incomplete dataset, with missing values in grey. The second row shows the mean of 50 multiple imputations for each missing value using the 'mice_ctsp' approach (MICE imputation using species identity, climate, forest structure and topography as predictors). The third row shows the corresponding coefficient of variation (CV) for these multiple imputations. Note that, for the third row, only imputed values are shown and that the colour scale varies across different traits. 555