Supplement of Biogeosciences, 15, 2601–2617, 2018
https://doi.org/10.5194/bg-15-2601-2018-supplement

Biogeosciences Open Access EGU

Supplement of

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

Rafael Poyatos et al.

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

Supplement

S1. Ecological and Forestry Inventory of Catalonia (IEFC): description and methods

The IEFC (Gracia *et al.*, 2000 - 2004) covers the entire forested territory of Catalonia (31900 km^2), in the North-east Iberian Peninsula. Catalonia is bound by the Mediterranean Sea to the East, the Pyrenees to the North and the continental Ebro depression to the West. The region is hydroclimatically diverse, as a result of the wide altitudinal range (0-3000 m.a.s.l.) and a strong continentality gradient. The dominant climate type is Mediterranean, gradually shifting towards Continental types inland and towards mountainous climates associated to mountain ranges, including an alpine belt in the Pyrenees. Mean annual temperature ranges from 18°C (on the southern coast) to 3°C (in the alpine belt) and annual rainfall varies from 400 mm to more than 1,500 mm (Climatic Digital Atlas of Catalonia, Ninyerola *et al.*, 2000).

Forested land covers 38% of the total land area of Catalonia (1.2 x 10^6 ha), and the 13 species selected in this study dominate > 90% of the forested area. The IEFC sampled the whole forested area of Catalonia between 1988 and 1998, totalling 10638 plots of 10 m radius, at a density of approximately 1 plot/km^2 of forest. A detailed methodological description of the sampling and measurements can be found in http://www.creaf.uab.es/iefc/pub/metodes/Index.htm (in Catalan). In all plots, all trees > 5 cm in diameter at breast height were tagged, identified to species and measured for diamaeter at breast height (dbh). For one or two individuals from each dbh class (5 cm intervals) of the dominant tree species, tree height was measured with a hypsometer. Maximum tree height was estimated as the highest value for a given species at a given plot. Wood samples were taken either from cores or from large first-order branches, and sapwood depth was visually estimated. Wood density (WD) was calculated as the ratio of dry weight to fresh wood volume. Fresh volume was established either by Archimedes' principle or measuring precisely the dimensions of the sample (for wood cores), and dry weight was obtained by weighting samples to a precision of 0.001 g after oven-drying for 48 h at 75ºC.

In a random subsample of all plots (20%), one or two individuals from each dbh class of the dominant tree species were selected to estimate branch size distribution and to measure foliar and wood traits. Leaves from all cohorts in exposed branches were sampled proportionally to their abundance. Leaf mass per area was calculated as the ratio of dry weight to fresh leaf area. One-sided, projected leaf area was measured either using a leaf area meter (LiCor 3100 AM; LiCor, Lincoln, NE) for broadleaves or measuring precisely the length and width of needles. Leaves were then weighted to a precision of 0.001 g after oven-drying for 48 h at 75ºC. A subsample of the dried leaves was ground with a Cyclotec Foss Tecator 1093-001 grinder (Foss Analytical, Hilleroed, Denmark) and sent for chemical analyses to the Scientific-Technical Service of the University of Barcelona. Nitrogen content was measured with an elemental analyzser (C.E. Instruments, Wigan, UK). For the dominant species in each plot, first-order branches were collected, and their branch diameter and supported leaf biomass was measured. Allometries between branch diameter and leaf biomass were built by aggregating branch measurements for each species at the county level (N=41 counties). Tree-level leaf biomass was obtained from individual trees' branch size distributions and branch-level B_L:A_S allometries, as described in Laforest-Lapointe *et al.*, (2014).

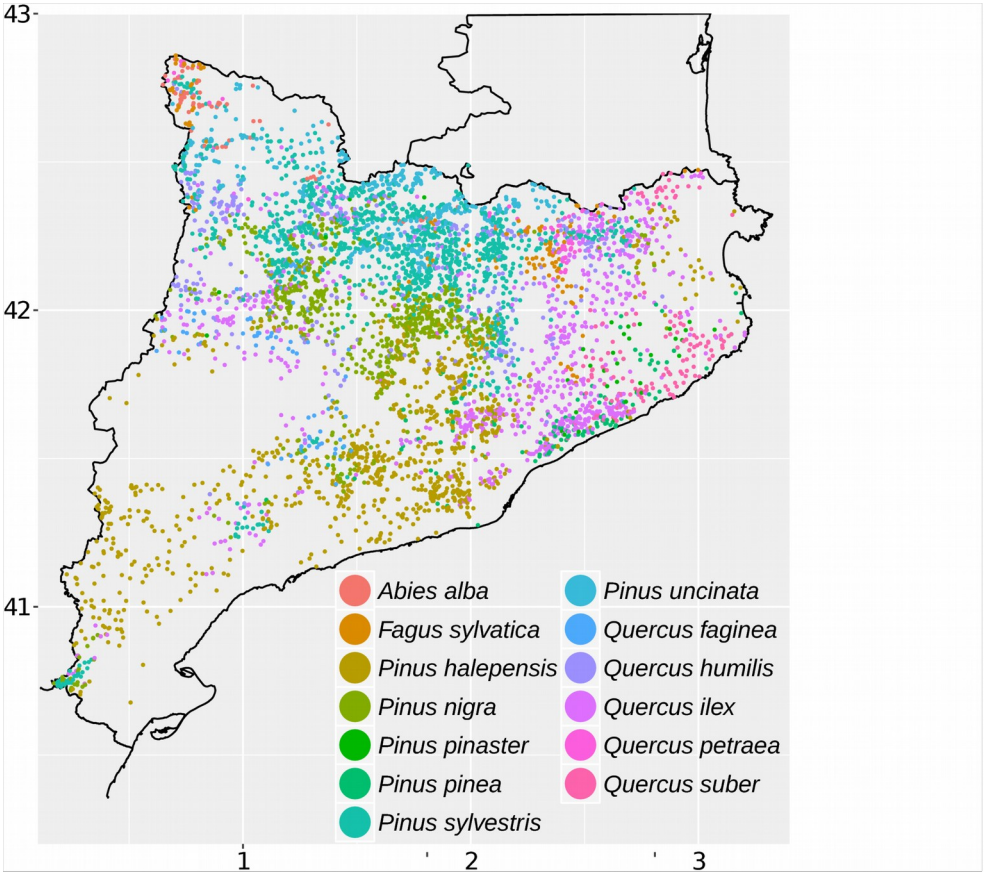

Figure S1. Map of Catalonia with the plots in the IEFC incomplete trait dataset used in this study, categorised by the dominant species.

S2. Description of the kNN algorithm and optimisations of settings

The implementation of the kNN algorithm in the 'VIM' package (Templ *et al.*, 2013) uses the Gower distance (Gower, 1971). The choice of the distance metric, the covariates selected to compute such distance, the *k* value and the aggregation method may impact the performance of the kNN method. A value of $k = 1$ yields imputations that preserve the dependence structure of the variables and increasing *k* improves the accuracy of predictions up to an optimum, beyond which the probability of finding poor matches increases (Eskelson *et al.*, 2009). The choice of an optimum *k* also depends on the number of observations, and we followed McRoberts (2012) to optimise the value of *k* that minimises NRMSE. We explored the performance of the kNN method for $k = 1, 2, 3, 5, 7, 10, 15$ and 20 and observed that the lowest prediction error was observed for $k = 7$ (Fig. S2), similar to results reported elsewhere (Troyanskaya *et al.*, 2001; McRoberts *et al.*, 2002; Penone *et al.*, 2014). Median aggregation of the *k*-neighbouring values performed better than mean aggregation (data not shown), as also found by Penone *et al.* (2014) .

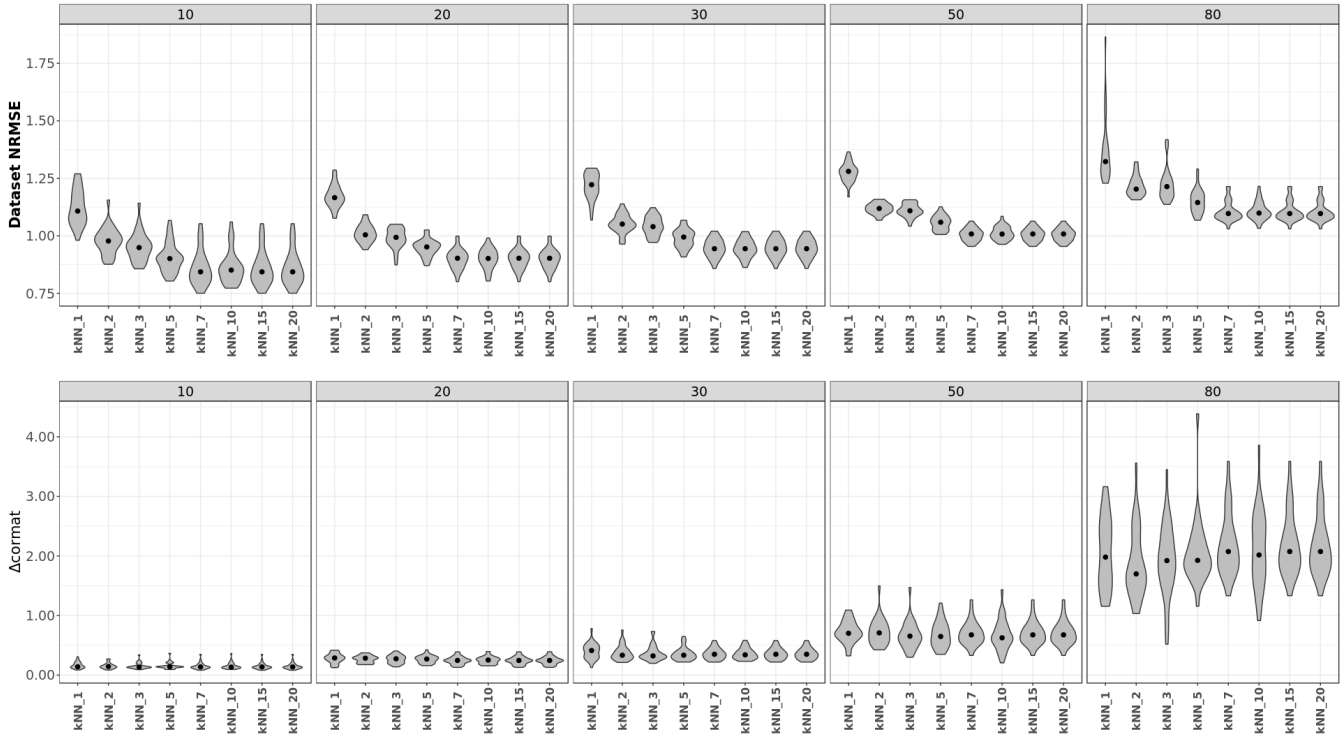

Figure S2. Dataset NRMSE and correlation matrix error for kNN imputations at different missingness levels (10% to 80%) using k=1, 2, 3, 5, 7, 10, 15, 20.

S3. The MICE algorithm:univariate imputation models and settings

In a *k*-variate dataset, MICE first fills all missing values by a random draw from the observed values within each variable. Then, for the first target variable x_1 (i.e. first variable to be imputed), the observed values of x_1 and the corresponding imputed values of the predictors are used to fit the individual imputation model (see below for a description of some univariate imputation methods used in this study). Then, missing values of x_1 are replaced by simulated draws obtained from the posterior predictive distribution of x_1, given the model and the predictors. Next, for the following incomplete variable, x_2, posterior predictive distributions are derived from the previously imputed values of $x_1, x_3 \ldots x_k$. This process is repeated for all variables with missing values and constitutes one iteration or cycle (Azur *et al.*, 2011). A relatively small number of iterations (10-20) is usually sufficient for the posterior distributions to stabilize, avoiding dependence on the imputation order. The whole process is repeated *m* times to obtain *m* imputed datasets.

We tested three univariate imputation models within the mice R package. The first method, here labelled as PRD, is the imputation of target variables as a function of predictors in the dataset using a linear model (the 'norm.predict' method in mice). In the second method, predictive mean matching (PMM), imputations are obtained using values from the complete cases, matching the actual missing datum with respect to some metric (van Buuren, 2012). PMM does not need an explicit model for the distribution of the missing values, it produces realistic imputations because they are drawn from the observed distribution and it does not impute outside the data range (van Buuren, 2012). PMM is also generally robust to non-normality and preserves non-linear relationships amongst the variables, although its performance may decrease at high missingness levels (Morris *et al.*, 2014). The third method, is based on a random forest (RF) algorithm, a recursive partitioning technique which sequentially splits a dataset into the most homogeneous subsamples and creates tree-like structures (Breiman, 2001). Applications of RF in data imputation have shown that they are well-suited to deal with mixed data types, complex interactions and non-linearities within the datasets (Stekhoven & Bühlmann, 2012; van Buuren, 2012). Here, in this comparison of univariate imputation models, we used the implementation of the random forest algorithm in MICE, a described in Doove *et al.*, (2014).

Apart from the choice of univariate models and predictors, already addressed in the main text of this paper, there are several settings in the specification of the imputation model within MICE. These settings are: (i) choosing the form of the imputation model for each variable, (ii) deciding how to impute variables that are functions of other (incomplete) variables, (iii) choosing the imputation order, (iv) setting up the starting imputations and the number of iterations and (v) choosing the number of imputed datasets.

(i) Form of the imputation model

We tested three forms for the imputation model, predictive mean matching, simple linear predictions and random forests (see S3). Despite the good results of random forest-based methods observed elsewhere (Stekhoven & Bühlmann, 2012; Penone *et al.*, 2014), they did not perform better than PMM (Fig. S3-S4). The reduction of stochasticity within MICE, implemented by the PRD method did not result in improved imputations either (Fig. S3-S4).

(ii) Imputation of derived variables and data transformation

In our dataset $B_L{:}A_S$ is a derived variable, calculated as the ratio of leaf biomass to sapwood area. We approached the imputation of $B_L{:}A_S$ using two different variants of MICE: 'passive imputation' (Mice_PAS) and 'just another variable' imputation (Mice_JAV). In Mice_PAS, leaf biomass and sapwood area enter the multivariate dataset, imputations are performed and then $B_L{:}A_S$ is calculated. This approach preserves the consistency between original and derived variables (van Buuren, 2012). In contrast, in the Mice_JAV approach, $B_L{:}A_S$ is calculated first and then imputed. Our results show that the Mice_JAV approach is superior (Fig. S3-S4), as also observed in other studies (Seaman *et al.*, 2012). In Mice_TRN we first log-transformed the trait values prior to imputation, but this did not result in a better performance compared to Mice_JAV (Fig. S3-S4).

(iii) Imputation order

Preliminary tests of the influence of the visiting order in the MICE algorithm did not show substantial differences on imputation performance (data not shown). We therefore used the default setting of imputing from left to right positions in the data frame (in our case, variables were in alphabetical order).

(iv) Starting imputations and number of iterations

The starting imputations were a random draw within each variable and the number of iterations t was set to 20, higher than the mice default value of 5. This increased number of iterations ensures stabilisation of the parameters of the imputation model and minimises the effects of the imputation order (van Buuren & Groothuis-Oudshoorn, 2011; van Buuren, 2012). We monitored the convergence of the MICE algorithm by plotting the mean and variance of imputations per stream and checking that the traces corresponding to the different streams intermingled with each other (van Buuren & Groothuis-Oudshoorn, 2011; van Buuren, 2012).

(v) Number of multiply imputed datasets.

As the purpose of this study is to explore different imputation approaches and imputation is computationally demanding, we set the number of imputed datasets to the default value of $m = 5$. This value is within the classic recommendation of $m = [3,10]$, although in a final application of MICE a value of m set closer to the average percentage of missing data may be desired (van Buuren, 2012). A larger value of $m = 50$ was used to obtain the final, imputed -trait maps (Fig. 8 in the manuscript).

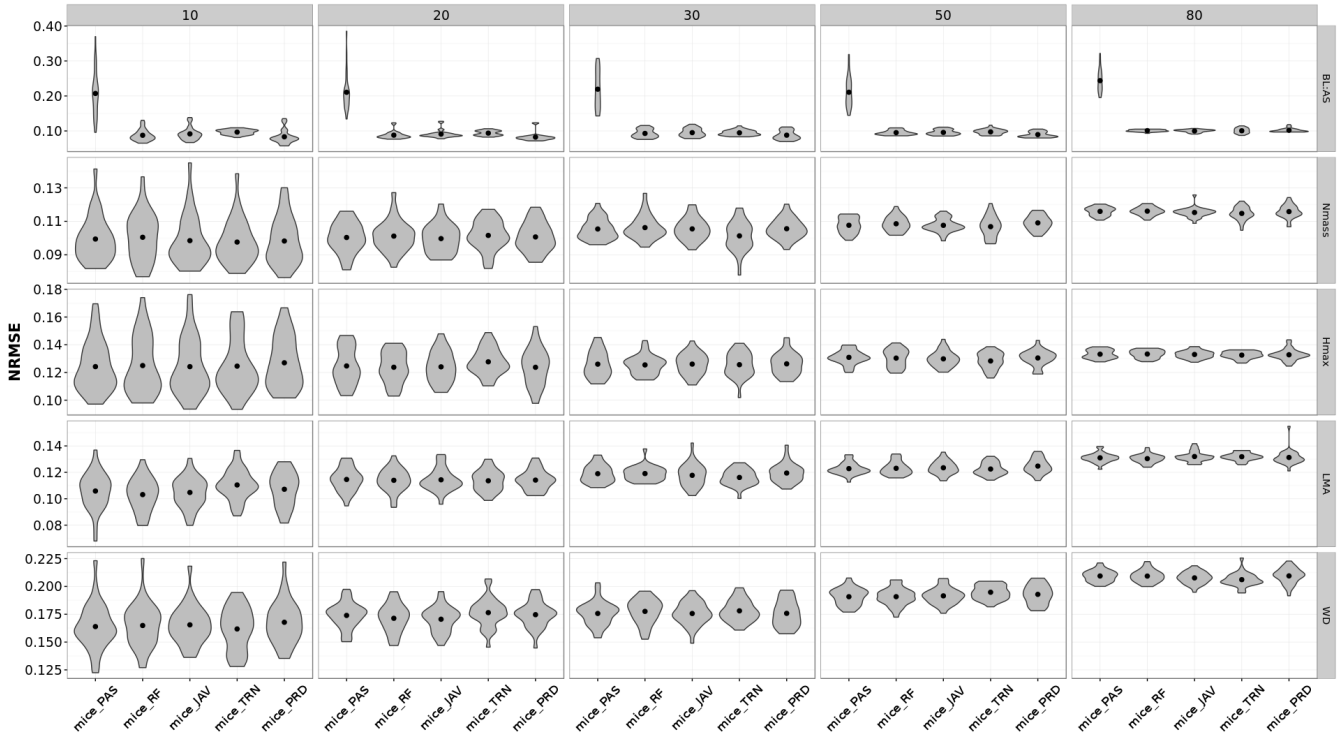

Figure S3. Trait-specific imputation performance (NRMSE) at increasing missingness levels (10% to 80%) using different MICE settings (see text for details). The following approaches used the predictive mean matching method (PMM) as the univariate imputation model: passive imputation of derived variables (mice_PAS), derived variables imputed as 'just another variable' (mice_JAV), imputation using log-transformed variables (mice_TRN). mice_PRD differed from PMM in that mice_PRD used the predicted trait from the sequential, multiple regression models in the MICE framework, without the stochasticity in regression coefficients that is introduced in PMM. mice_RF uses a random-forest algorithm for univariate imputation instead of PMM (see text). The original trait distributions in the IEFC complete data set (Observed) are also shown. Traits: leaf biomass to sapwood area ratio, $B_L{:}A_S$ (t m^{-2}); nitrogen per unit mass, N_{mass} (%mass); maximum tree height, H_{max} (m); leaf mass per area LMA (mg cm^{-2}); wood density, WD, (gm cm^{-3}).

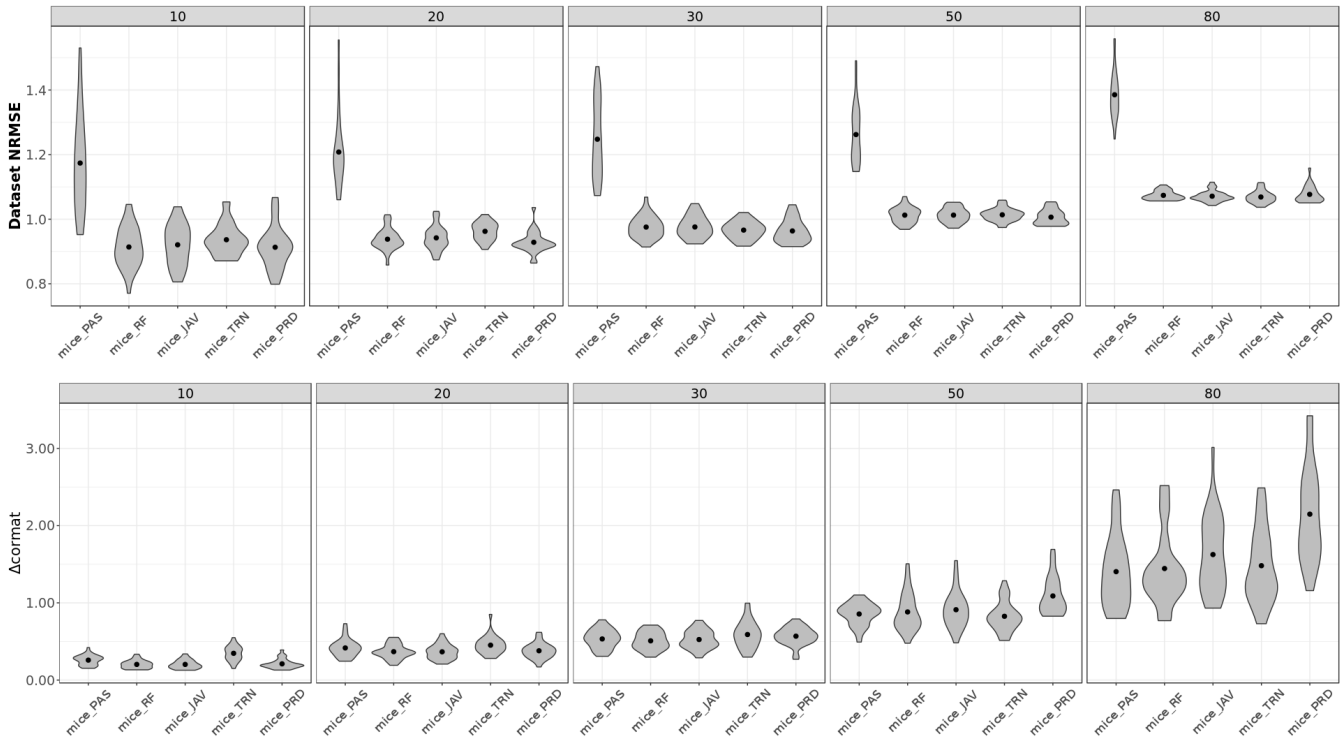

Figure S4. Dataset-specific imputation performance (dataset NRMSE and correlation matrix error) at increasing missingness levels (10% to 80%) using different MICE settings (see text for details). The following approaches used the predictive mean matching method (PMM) as the univariate imputation model: passive imputation of derived variables (Mice_PAS), derived variables imputed as 'just another variable' (Mice_JAV), imputation using log-transformed variables (Mice_TRN). Mice_PRD differed from PMM in that Mice_PRD used the predicted trait from the sequential, multiple regression models in the MICE framework, without the stochasticity in regression coefficients that is introduced in PMM. Mice_RF uses a random-forest algorithm for univariate imputation instead of PMM (see text).

120 **S4. Mean imputations compared to MICE and, Ordinary kriging and kNN imputations using only trait information**

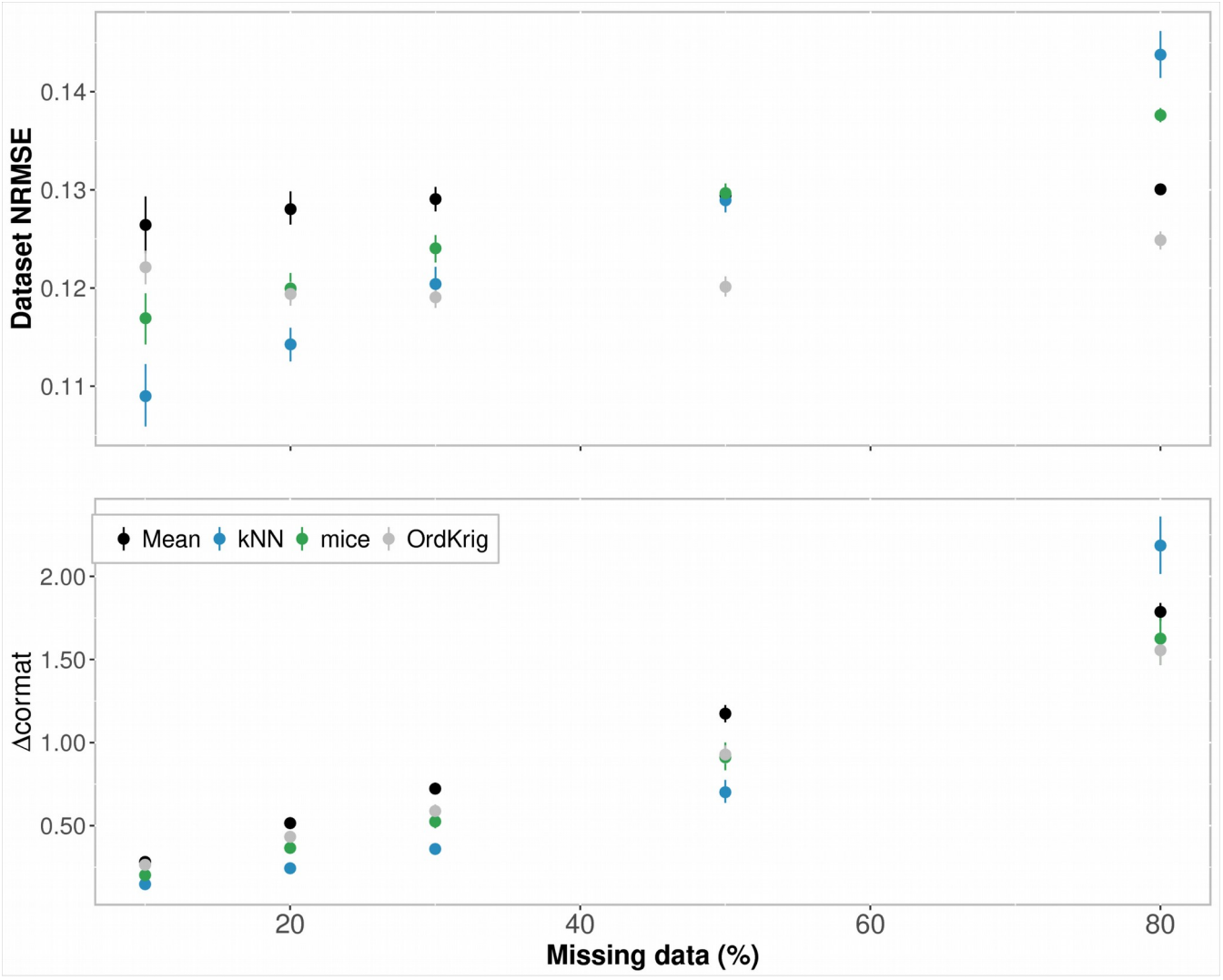

Figure S5. Data set -averaged NRMSE and Δcormat at increasing missingness levels (10% to 80%) for different imputation methods: overall trait mean (Mean), MICE, OrdKrig and kNN, both using only the trait matrix in the predictor set (mice) or in the distance calculation (kNN).

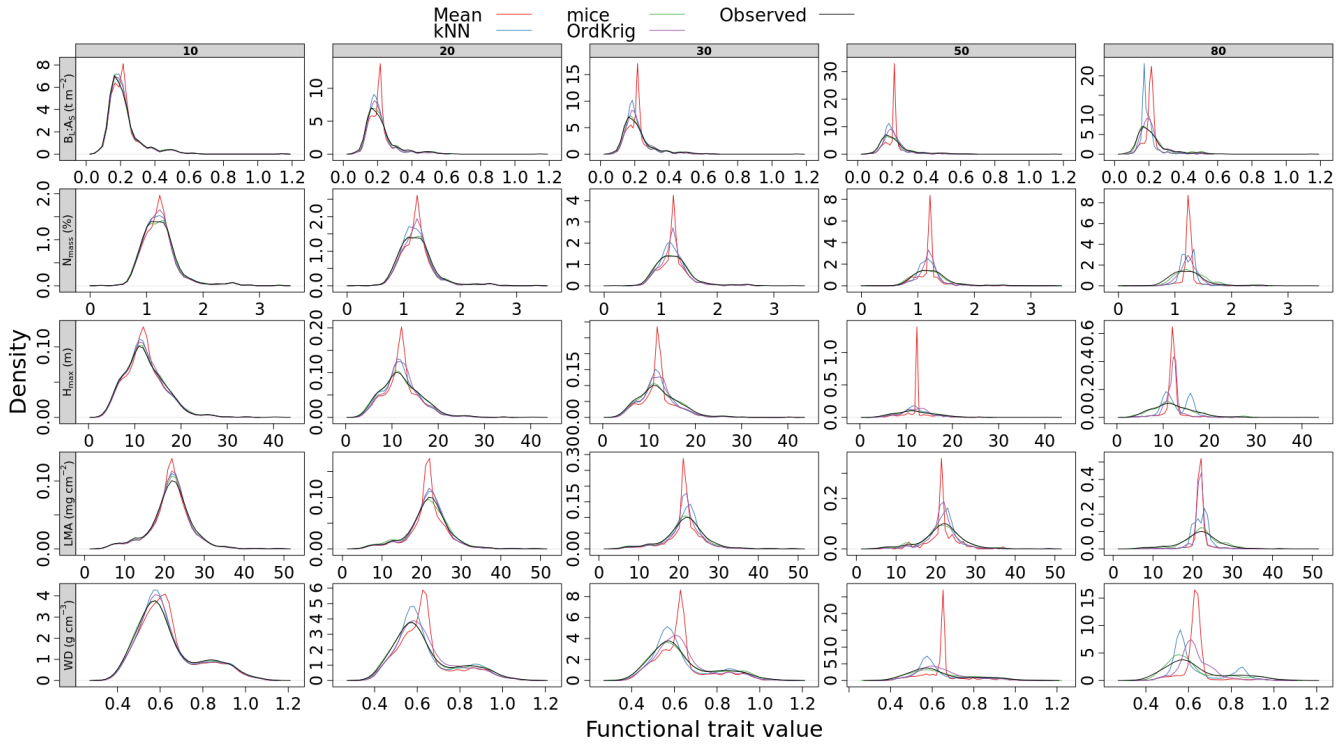

Figure S6. Imputed trait distributions at increasing missingness levels (10% to 80%) for different imputation methods: overall trait mean (Mean), MICE, OrdKrig and kNN, both using only the trait matrix in the predictor set (mice) or in the distance calculation (kNN). The original trait distributions in the IEFC complete data set (Observed) is are also shown. The original trait distributions in the IEFC complete data set (Observed) are also shown. Traits: leaf biomass to sapwood area ratio, $B_L{:}A_S$ (t m^{-2}); nitrogen per unit mass, N_{mass} (%mass); maximum tree height, H_{max} (m); leaf mass per area LMA (mg cm^{-2}); wood density, WD, (gm cm^{-3}).

S5. MICE imputations using different levels of ecological information

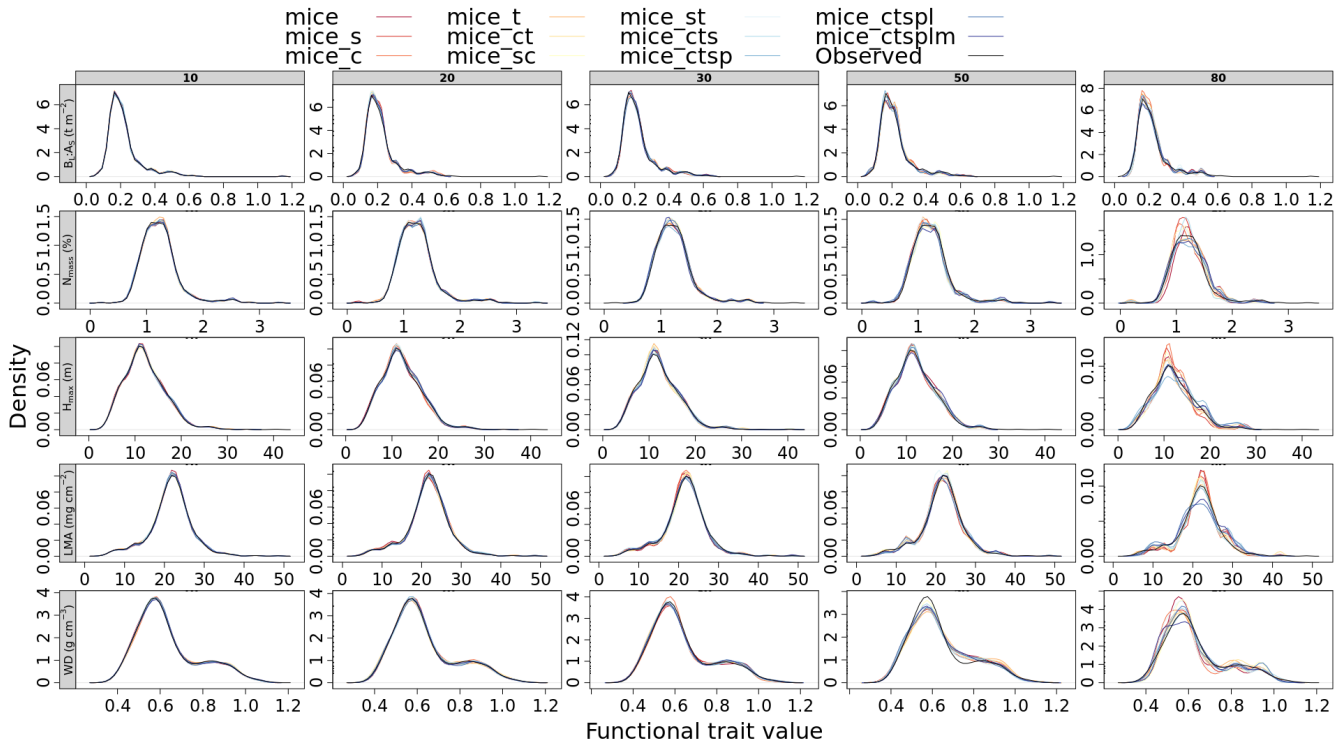

Figure S7. Imputed trait distributions at increasing missingness levels (10% to 80%) for MICE imputations using different combinations of additional predictor sets: species identity ('s'), climate ('c'), forest structure ('t'), spatial variables ('p'), lithology ('l') and sampling month ('m'). See Fig. 1 for an overall view of the experimental design and the Methods section for a detailed description of the variables included in each predictor set. The original trait distributions in the IEFC complete data set (Observed) are also shown. Traits: leaf biomass to sapwood area ratio, $B_L{:}A_S$ (t m^{-2}); nitrogen per unit mass, N_{mass} (%mass); maximum tree height, H_{max} (m); leaf mass per area LMA (mg cm^{-2}); wood density, WD, (gm cm^{-3}).

S6. Species mean imputations compared to MICE , Kriging and kNN imputations using optimum levels of ecological information

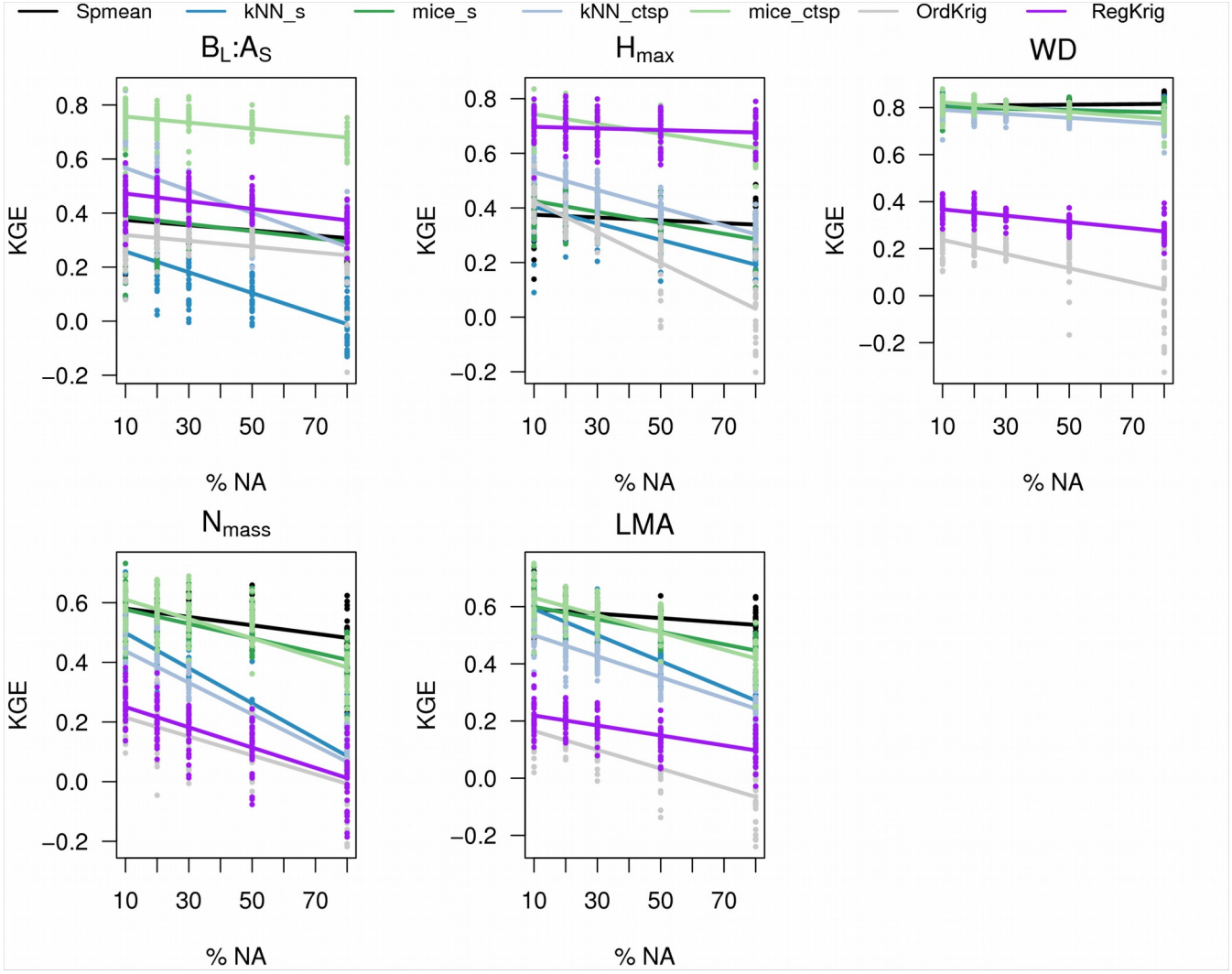

Figure S8. Variation in trait-specific KGE with increasing missingness levels (10% to 80%) and for different imputation methods: species mean (Spmean), miceMICE and kNN with species as predictor (mice_s and kNN_s, respectively) and , MICE and kNN with species, climate, forest structure and spatial variables as predictors (mice_ctsp and kNN_ctsp, respectively), OrdKrig and RegKrig. Lines depict the LME fits. Traits: leaf biomass to sapwood area ratio, $B_L{:}A_S$ (t m^{-2}); nitrogen per unit mass, N_{mass} (%mass); maximum tree height, H_{max} (m); leaf mass per area LMA (mg cm^{-2}); wood density, WD, (gm cm^{-3}).

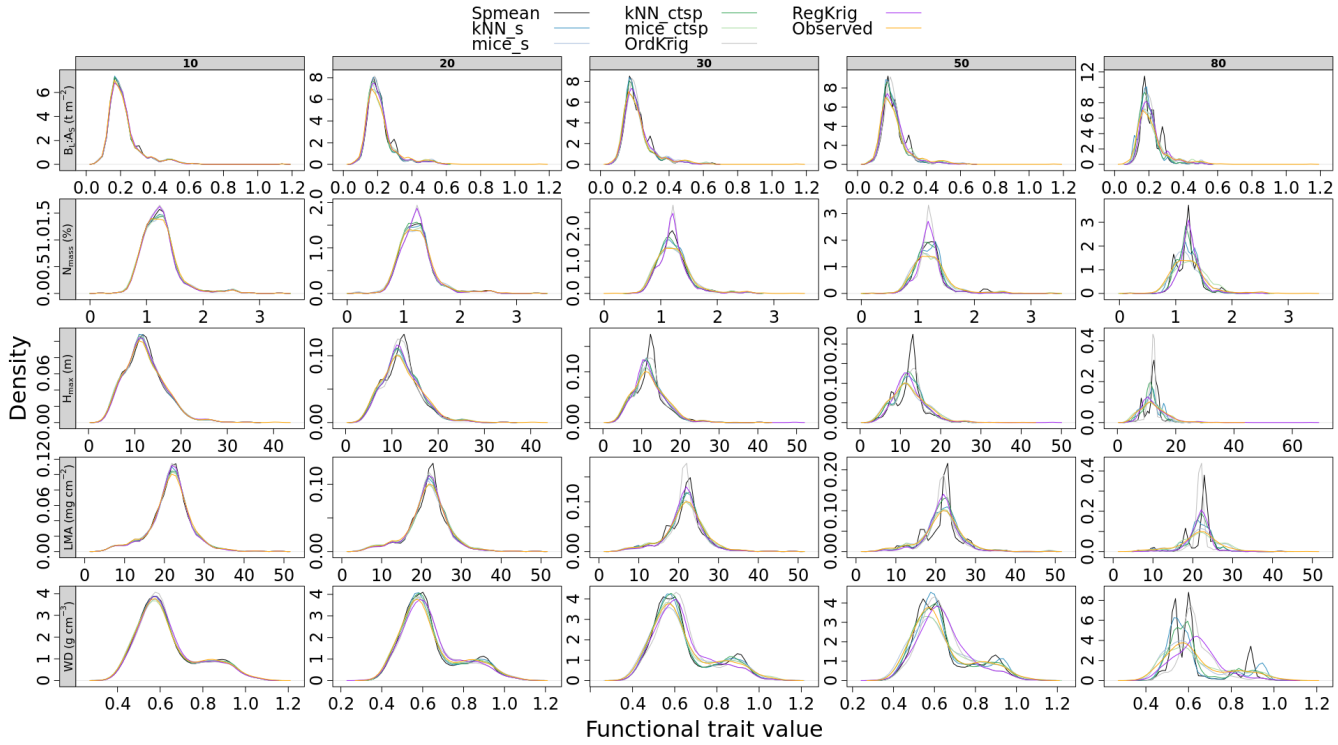

Figure S9. Imputed trait distributions for different imputation methods at increasing missingness levels (10% to 80%): species mean (Spmean), MICE and kNN with species as predictor (mice_s and kNN_s, respectively) and , MICE and kNN with species, climate, forest structure and spatial structure as predictors (mice_ctsp and kNN_ctsp, respectively), OrdKrig and RegKrig. The original trait distributions in the IEFC complete dataset (observed) are also shown. Traits: leaf biomass to sapwood area ratio, $B_L{:}A_S$ (t m^{-2}); nitrogen per unit mass, N_{mass} (%mass); maximum tree height, H_{max} (m); leaf mass per area LMA (mg cm^{-2}); wood density, WD, (gm cm^{-3}).

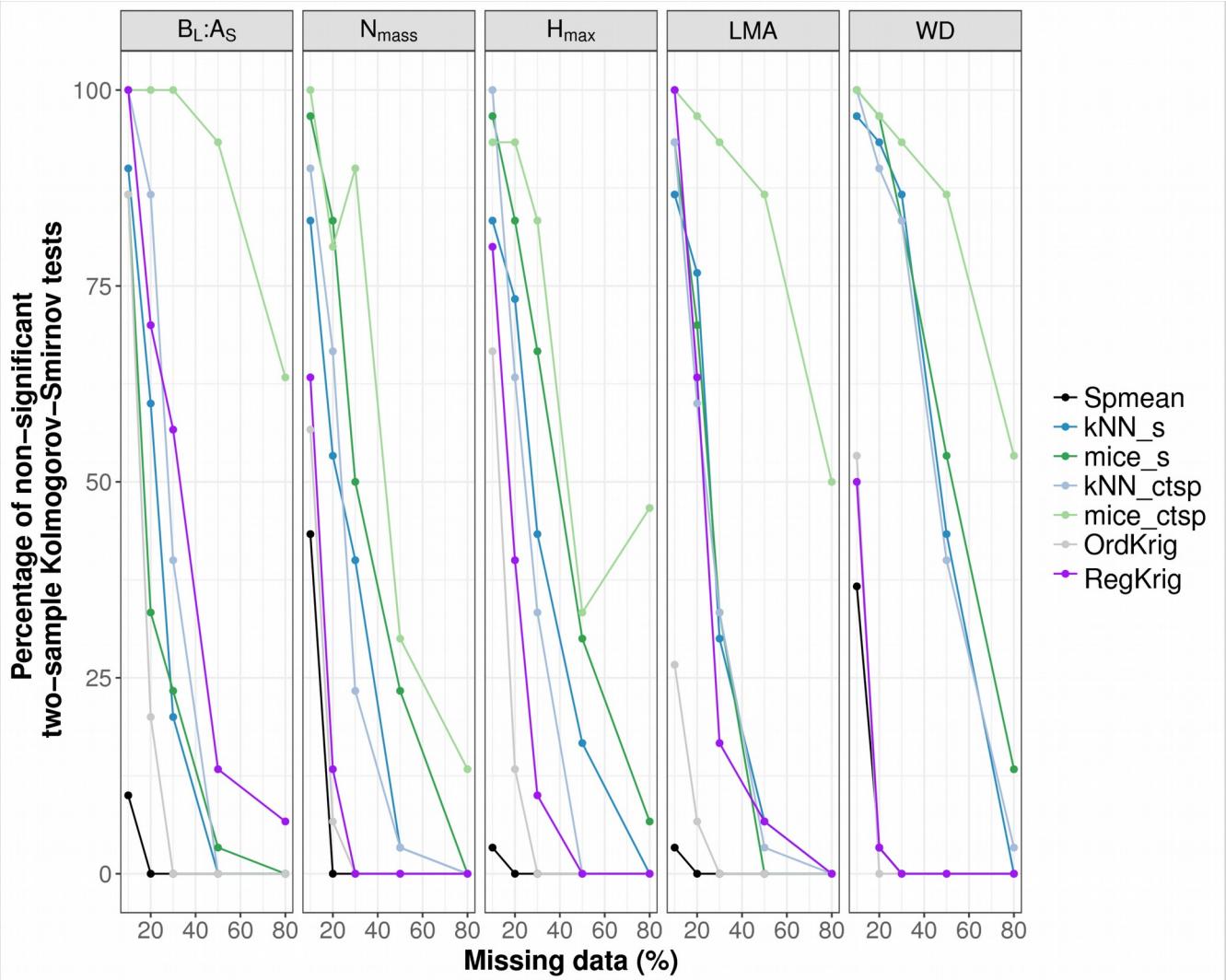

Figure S10. Percentage of non-significant two-sample Kolmogorov-Smirnov tests at increasing missingness levels (10% to 80%): species mean (Spmean), MICE and kNN with species as predictor (mice_s and kNN_s, respectively) and , MICE and kNN with species, climate, forest structure and spatial structure as predictors (mice_ctsp and kNN_ctsp, respectively), OrdKrig and RegKrig. Traits: leaf biomass to sapwood area ratio, B_L:A_S (t m^{-2}); nitrogen per unit mass, N_{mass} (%mass); maximum tree height, H_{max} (m); leaf mass per area LMA (mg cm^{-2}); wood density, WD, (gm cm^{-3}).

135

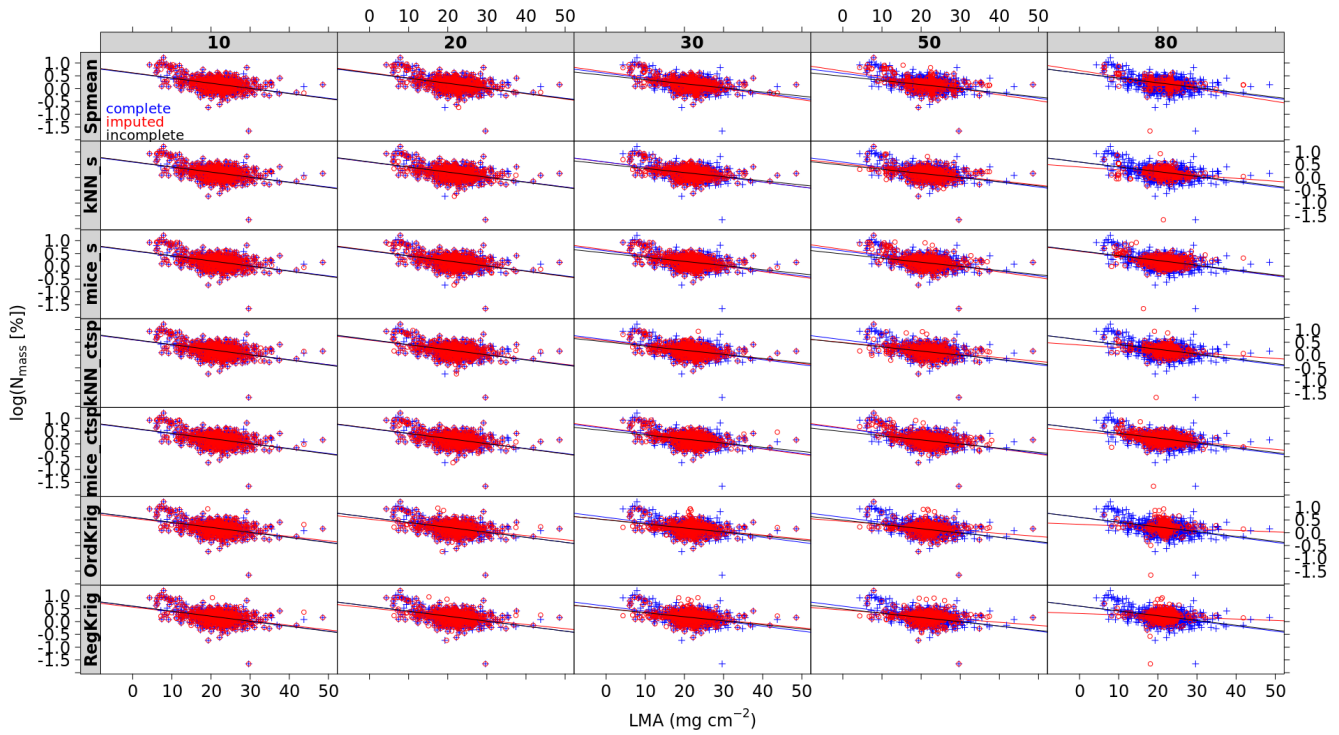

Figure S11. Linear regressions between log-transformed N_{mass} (nitrogen per unit mass, %mass) and LMA (leaf mass per area LMA, mg cm^{-2}) in the IEFC complete trait dataset (blue), and in incomplete IEFC trait datasets at different missingness levels (from 10% to 80%), either imputed (red) or non-imputed (black). Imputation methods are: 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 structure as predictors (mice_ctsp and kNN_ctsp, respectively), OrdKrig and RegKrig.

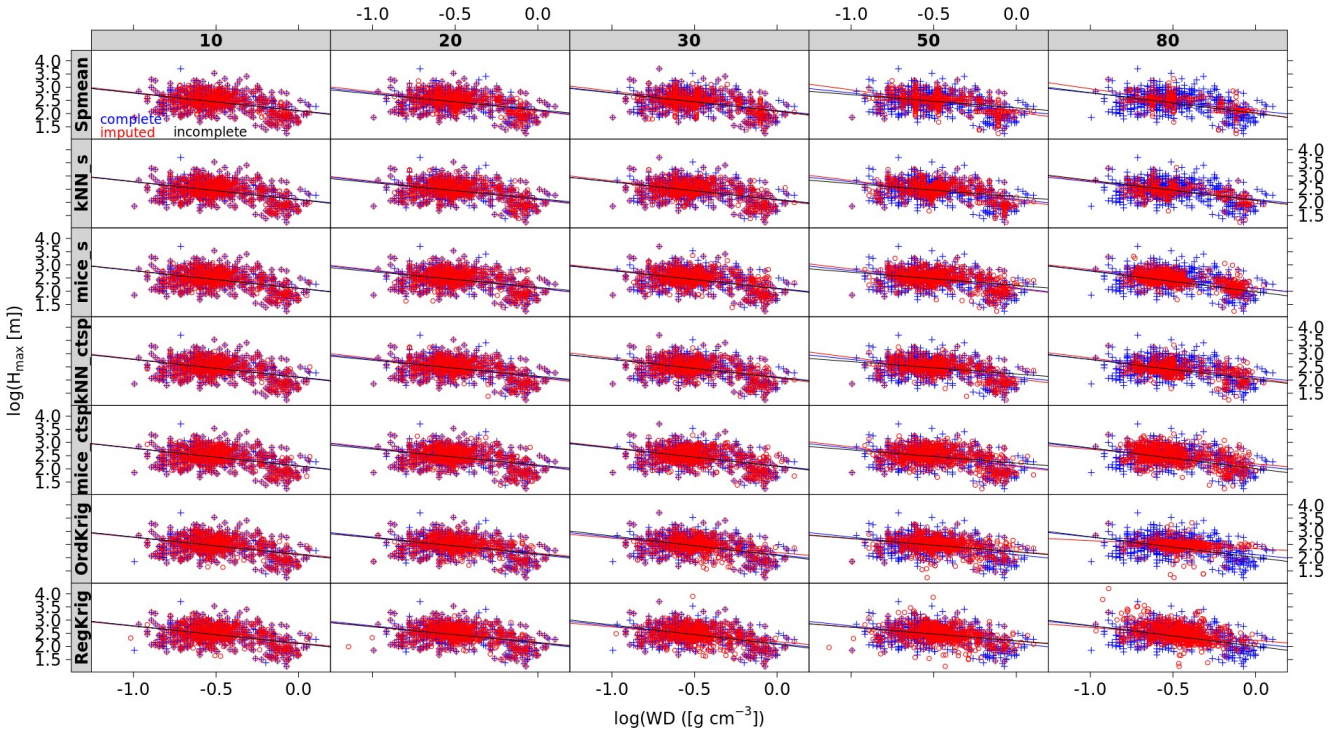

Figure S12. Linear regressions between log-transformed H_{max} (maximum tree height, m) and log-transformed WD (wood density, gm cm^{-3}) in the IEFC complete trait dataset (blue), and in incomplete IEFC trait datasets at different missingness levels (from 10% to 80%), either imputed (red) or non-imputed (black).

Table S1. Tukey pairwise comparisons of the LME coefficients relating Δcormat to increasing missingness levels for selected imputation methods. Letters represent significant differences (P < 0.05) and imputation methods are listed in increasing order of the model coefficient.

	Coefficient	SE	df	Lower CL	Upper CL	
mice_ctsp	1.27E-02	6.27E-04	148	1.15E-02	1.39E-02	a
kNN_ctsp	1.46E-02	6.27E-04	148	1.33E-02	1.58E-02	ab
mice_s	1.61E-02	6.27E-04	148	1.49E-02	1.73E-02	bc
RegKrig	1.68E-02	6.27E-04	148	1.55E-02	1.80E-02	bcd
kNN_s	1.71E-02	6.27E-04	148	1.58E-02	1.83E-02	cd
OrdKrig	1.84E-02	6.27E-04	148	1.71E-02	1.96E-02	d
Spmean	2.20E-02	6.27E-04	148	2.08E-02	2.33E-02	e

S7. Imputing traits for the main forest species in Catalonia

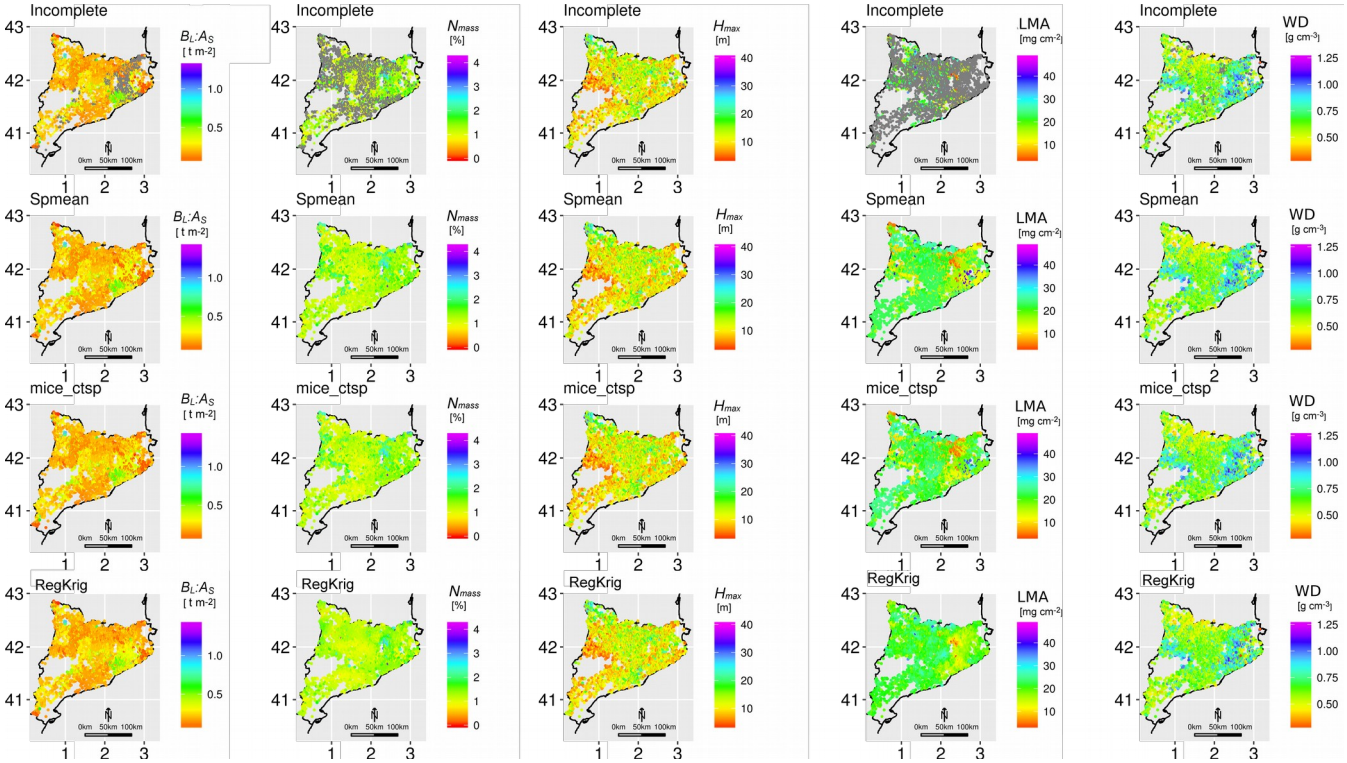

Figure S13. Maps with the distribution of functional traits across the selected plots in the IEFC. The first row shows the incomplete dataset, with missing values in grey. The second row shows species mean imputations ('Spmean'), the third row shows the 'mice_ctsp' imputations (MICE imputation using species identity, climate, forest structure and topography as predictors) and the fourth row shows the universal kriging imputations (RegKrig). Traits: leaf biomass to sapwood area ratio, B_L:A_S (t m^{-2}); nitrogen per unit mass, N_{mass} (%mass); maximum tree height, H_{max} (m); leaf mass per area LMA (mg cm^{-2}); wood density, WD, (gm cm^{-3}).

References

Azur, M. J. et al. 2011. Multiple imputation by chained equations: what is it and how does it work? - Int. J. Methods Psychiatr. Res. 20: 40–49.

Breiman, L. 2001. Random forests. - Machine learning 45: 5–32.

Doove, L. L. et al. 2014. Recursive partitioning for missing data imputation in the presence of interaction effects. - Computational Statistics & Data Analysis 72: 92–104.

Eskelson, B. N. I. et al. 2009. The roles of nearest neighbor methods in imputing missing data in forest inventory and monitoring databases. - Scandinavian Journal of Forest Research 24: 235–246.

Gower, J. C. 1971. A General Coefficient of Similarity and Some of Its Properties. - Biometrics 27: 857–871.

Gracia, C. et al. 2000. Inventari Ecològic i Forestal de Catalunya: Centre de Recerca Ecològica i Aplicacions Forestals. 10 volumes.

Laforest-Lapointe, I. et al. 2014. Intraspecific variability in functional traits matters: case study of Scots pine. - Oecologia 175: 1337–1348.

Lenth, R. V. 2016. Least-Squares Means: The *R* Package **lsmeans**. - Journal of Statistical Software 69: 1–33.

McRoberts, R. E. 2012. Estimating forest attribute parameters for small areas using nearest neighbors techniques. - Forest Ecology and Management 272: 3–12.

McRoberts, R. E. et al. 2002. Stratified estimation of forest area using satellite imagery, inventory data, and the k-Nearest Neighbors technique. - Remote Sensing of Environment 82: 457–468.

Morris, T. P. et al. 2014. Tuning multiple imputation by predictive mean matching and local residual draws. - BMC Med Res Methodol 14: 1–13.

Ninyerola, M. et al. 2000. 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: 1823–1841.

Penone, C. et al. 2014. Imputation of missing data in life-history trait datasets: which approach performs the best? - Methods Ecol Evol 5: 961–970.

Pinheiro, J. et al. 2012. nlme: Linear and Nonlinear Mixed Effects Models.

Seaman, S. R. et al. 2012. Multiple imputation of missing covariates with non-linear effects and interactions: an evaluation of statistical methods. - BMC Medical Research Methodology 12: 46.

Stekhoven, D. J. and Bühlmann, P. 2012. MissForest—non-parametric missing value imputation for mixed-type data. - Bioinformatics 28: 112–118.

Templ, M. et al. 2013. VIM: Visualization and Imputation of Missing Values. R package version 3.0. 3.1.

Troyanskaya, O. et al. 2001. Missing value estimation methods for DNA microarrays. - Bioinformatics 17: 520–525.

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

van Buuren, S. and Groothuis-Oudshoorn, K. 2011. mice: Multivariate Imputation by Chained Equations in R. - Journal of Statistical Software 45: 3.

White, I. R. et al. 2011. Multiple imputation using chained equations: Issues and guidance for practice. - Statist. Med. 30: 377–399.