



When does gap filling of trait data confound taxonomic and functional analyses?



Julia Joswig^{1, 2}, Jens Kattge^{2, 3}, Guido Kraemer⁴, Miguel D. Mahecha^{3, 4}, Nadja Rüger^{3, 4, 5},
Michael E. Schaepman¹, Franziska Schrodtt⁶, Christian Wirth^{2, 3, 4}, Meredith C. Schuman¹

¹ University of Zürich, ² MPI for Biogeochemistry, ³ German Centre for Integrative Biodiversity Research (iDiv), ⁴ University of Leipzig, ⁵ Smithsonian Tropical Research Institute, ⁶ University of Nottingham

Importance of complete plant trait data

Data on plant traits are increasingly used to understand relationships between biodiversity and ecosystem processes. Large trait databases are sparse because they are compiled from many smaller and usually more local databases. This sparsity severely limits the potential for both multivariate and global data analyses

Towards a guideline

1. Describe deviations introduced by gap-filling
 - error
 - clustering (taxonomic or functional groupings)
 - trait correlation patterns.
2. Indicators of data susceptibility to deviations.
 - amount of data
 - the number and particularity of observations within clusters
 - BHPMF confidence product.

Recommendations (preliminary)

Gap-filling:

Balancing the benefit of adding (*envelope*) trait data against its drawbacks.

- ✗ Increases (slightly) the error, skewed residuals
- ✓ Reduces taxonomic clustering bias
- ✓ Stabilizes trait-trait relationships
- ✗ May change trait-trait relationships slightly

Usage of gap filled data:

- ✗ Intraspecific, taxonomic or phylogenetic analyses, but clusters with more individuals and low functional diversity have lower deviations
- ✓ Trait-trait relationship analyses
- ✓ Probably useful for across-species and functional analyses

Deviation by BHPMF

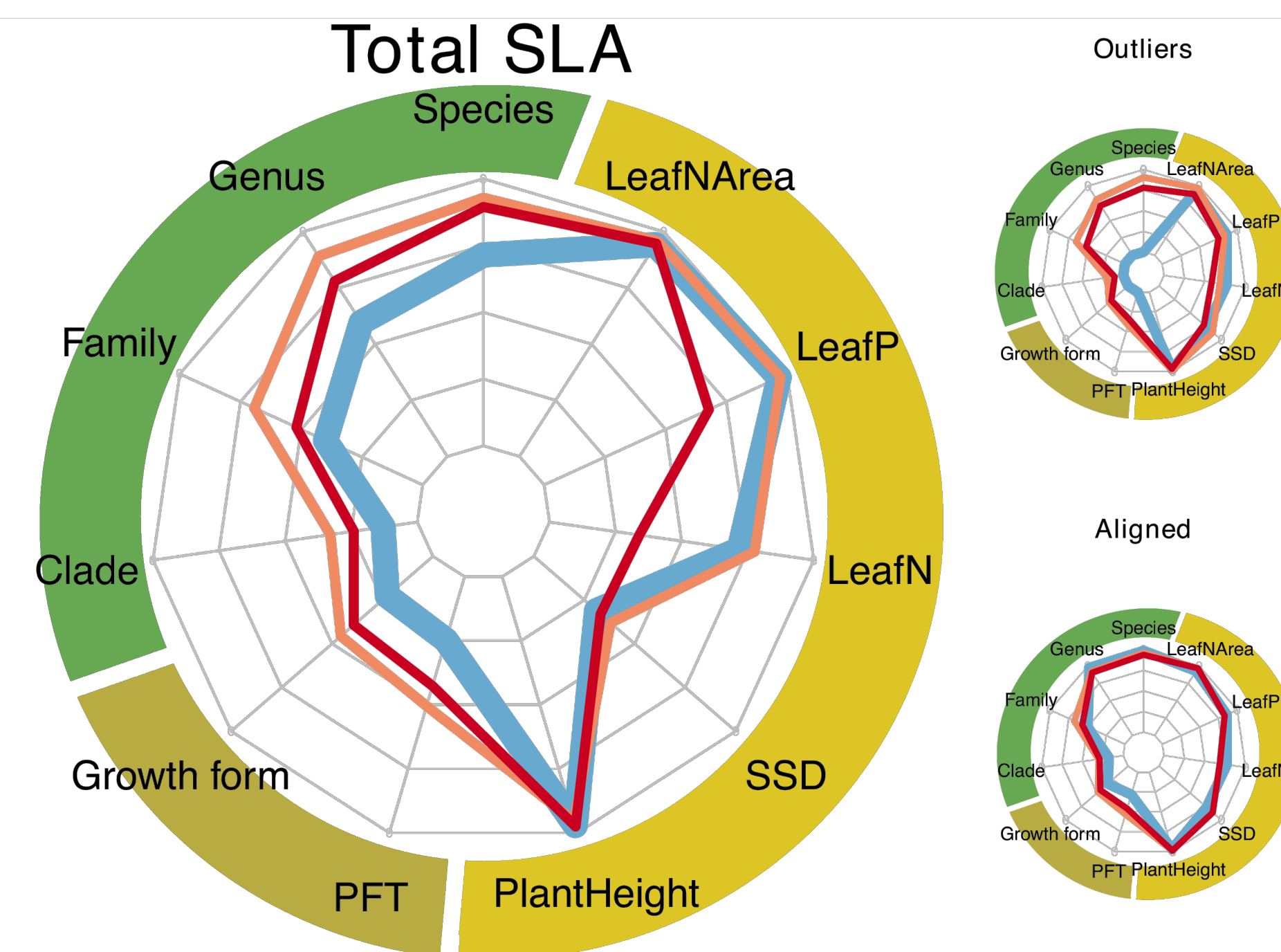


Figure 1: Measures of clustering and trait-trait relationship as median (n=1136). Blue= observed, orange=predicted, red=predicted WITH *envelope*. Outer circle= high clustering or strong trait-trait relationship.

BHPMF - Bayesian Hierarchical Probabilistic Matrix Factorization

- Commonly used gap-filling algorithm
- Method of choice for both TRY [3], and the large and widely used database sPLOT
- Makes use of the species' relationships described by their [taxonomy and their trait-trait relationships](#) [1, 2]



Contact: julia.joswig@geo.uzh.ch

Workflow

Data

- TRY-derived data [3] of 241,653 individuals, 1136 individuals are completely observed for 6 traits (tropical rainfor data). The observed data is our test data set, the rest is termed here the *envelope* data.
- Erase data points up to 80% missingness in the test data (n=3 per % gaps)

BHPMF gap-filling

We gap-filled the test data of various missingness levels (n=3) with and without *envelope* data with the R-package BHPMF.

Analyses

We analysed the gap-filled data on the multi-trait scale to the value level.

1. Error: RMSE, residuals, distribution
2. Clustering: Silhouette index, coefficient of variation, [distance to co-cluster individuals](#) (clusters: [taxonomic and functional groups](#))
3. Trait-trait relationships: Variance explained, Pearson correlation coefficient, [distance to lm-prediction](#)

References

- [1] Fazayeli Fazayeli; et al. Uncertainty quantified matrix completion using Bayesian Hierarchical Matrix factorization. *International Conference on Machine Learning and Applications (ICMLA)*, 2014.
- [2] Franziska Schrodtt; et al. BHPMF - a hierarchical Bayesian approach to gap-filling and trait prediction for macroecology and functional biogeography. *Global Ecology and Biogeography*, 24(12):1510--1521, 12 2015.
- [3] Jens Kattge; et al. TRY plant trait database - enhanced coverage and open access. *Global Change Biology*, 26(1):119--188, 1 2020.