Predicting Missing Values in Biodiversity Datasets Using Phylogenetics and Spatial Mapping

Jay Jain, Quentin Read, Andrew Finley, Phoebe Zarnetske

Department of Forestry, Michigan State University, East Lansing, MI

Introduction

- Large biodiversity datasets with many missing values have become common in ecology.
- Filling in missing values is critical to understanding biodiversity in the age of climate change.
- Models based on both phylogenetic and spatial factors can be used to predict missing trait values.
- Such models use the Monte Carlo method to solve multi-dimensional integrals

Objective

- To demonstrate that a Bayesian method of imputation that combines phylogenetic and spatial methods will outperform existing imputation methods.

Methods

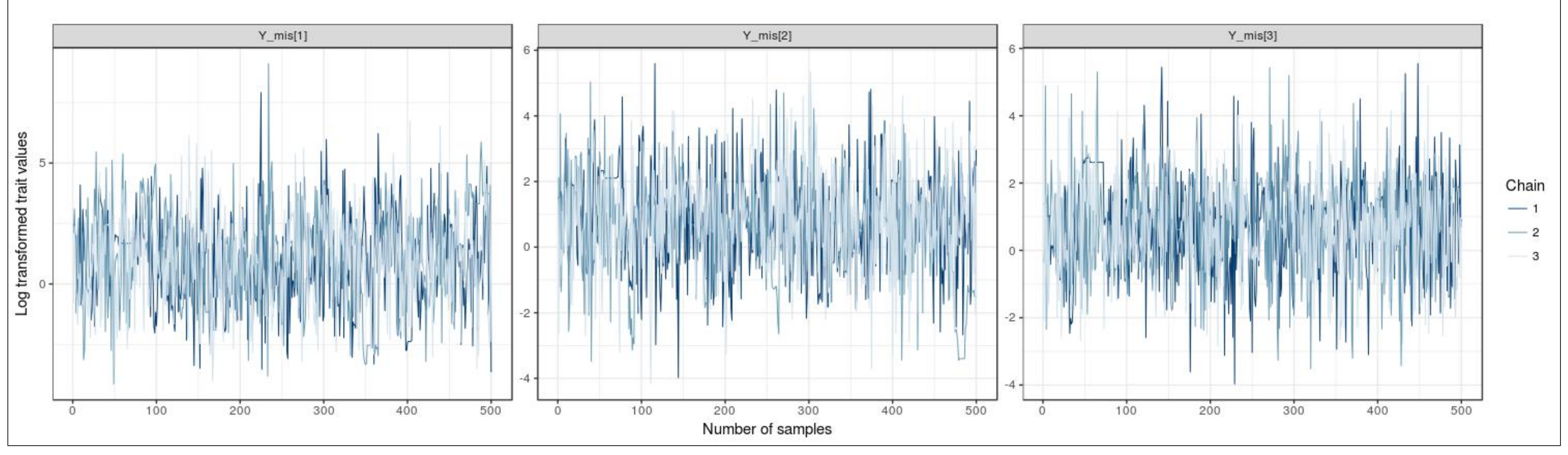
- Data was removed at random from a North American forest tree species dataset.
- Trait values were imputed using packages available in the R programming languages.
- The imputed values were compared to the known values with 95% confidence intervals.
- The combined model was developed using STAN, a programming language for statistical inference written in C++ .

$$y \sim X\beta + Z\alpha + \varepsilon$$

y = imputed traits
Z = identity matrix
X = predictor matrix
α = phylogenetic random effects
β = fixed effects/slope
ε = residuals

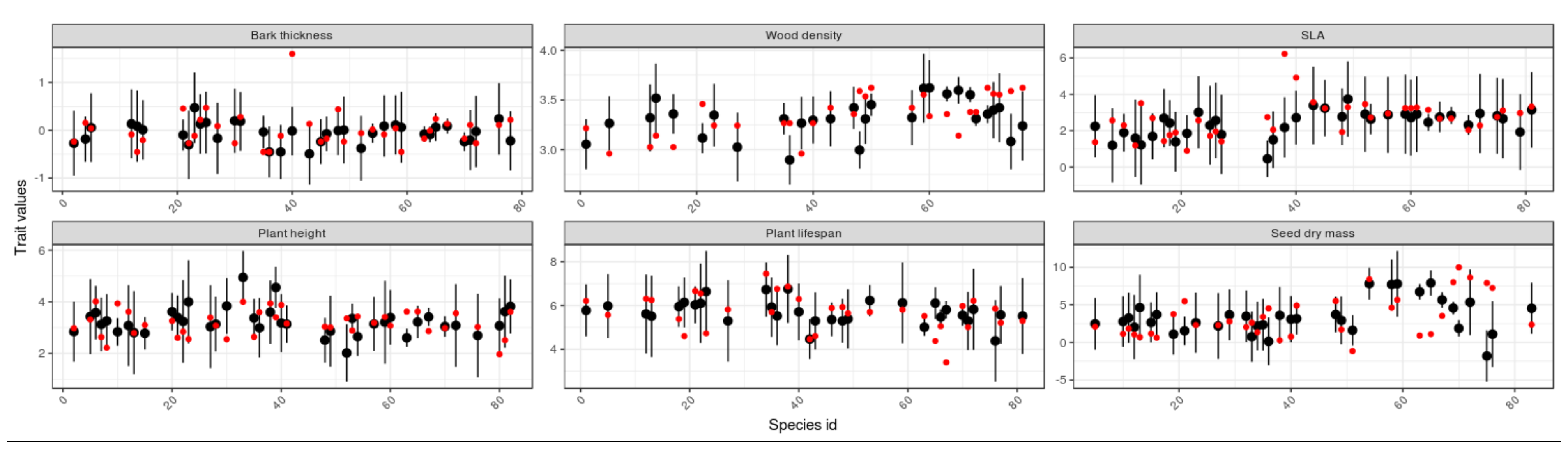
- The Ornstein-Uhlenbeck process was used to model the phylogenetic tree reconstruction.

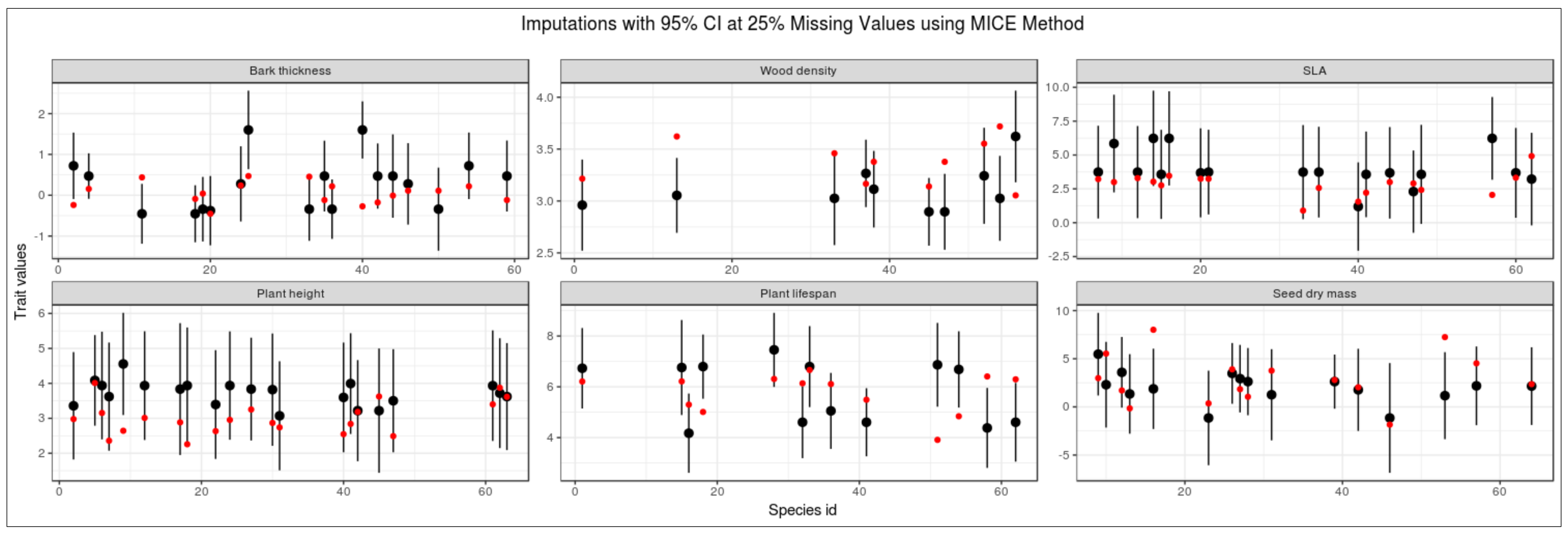
Monte Carlo Method Convergence



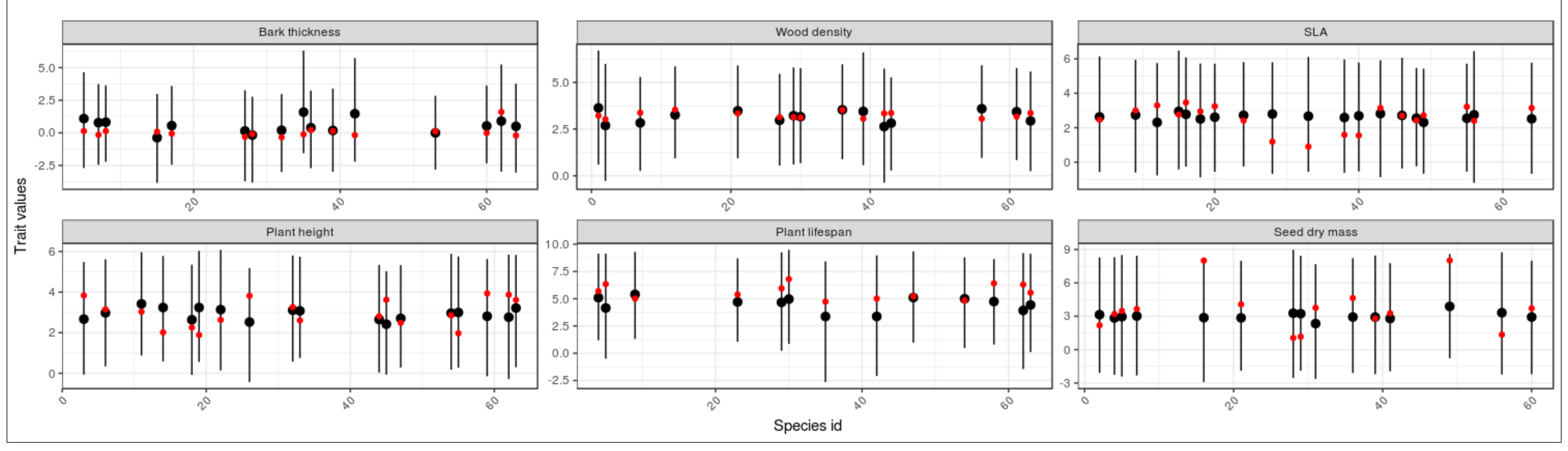
Results

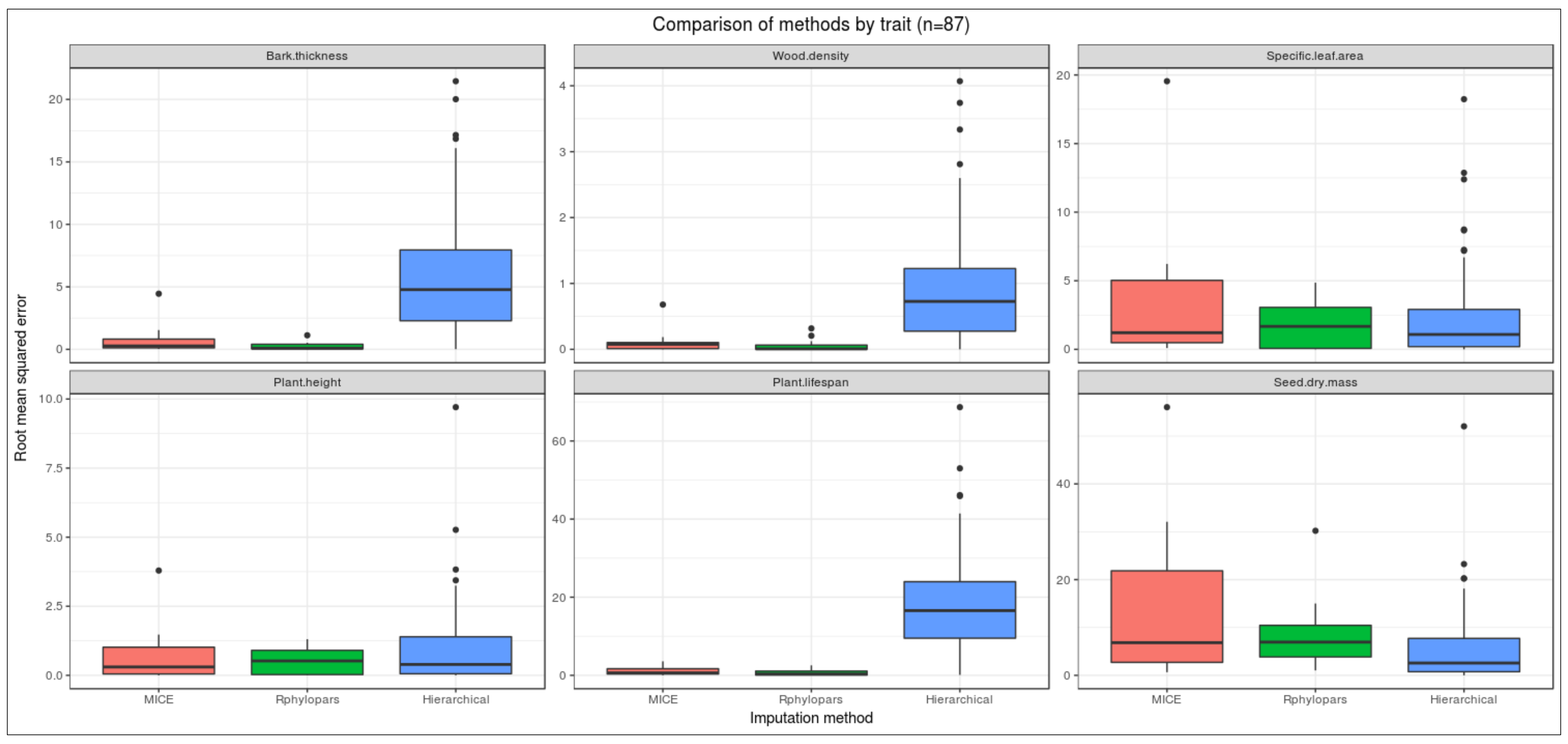
Imputations with 95% CI at 25% Missing Values using Phylogenetic Method





Imputations with 95% CI at 25% Missing Values using Hierarchical Model





Conclusions & Future Work

- The proposed Bayesian model imputed values better for some traits than others compared to phylogenetic and multivariate imputation.
- Continue to test the model at varying rates of missing data while using different phylogenetic and spatial parameters.

Acknowledgements

We thank Jens Stevens for providing the trait data and Nina Lany for assistance in fitting the model.
This material is based upon work supported by the National Science Foundation under grant #1560168.