Geophysical Research Abstracts
Vol. 18, EGU2016-16906, 2016
EGU General Assembly 2016

Assessment of imputation methods using varying ecological information to fill the gaps in a tree functional trait database

Rafael Poyatos (1), Oliver Sus (2), Albert Vilà-Cabrera (1), Jordi Vayreda (1), Llorenç Badiella (3), Maurizio Mencuccini (4,5), Jordi Martínez-Vilalta (1,6)

(1) CREAF, Bellaterra, Spain (r.poyatos@creaf.uab.es), (2) Deutscher Wetterdienst, 63067 Offenbach, Germany., (3) Applied Statistics Service, Autonomous University of Barcelona, (4) ICREA at CREAF, Bellaterra, Spain., (5) School of Geosciences, University of Edinburgh, UK., (6) Universitat Autònoma de Barcelona, Bellaterra, Spain.

Plant functional traits are increasingly being used in ecosystem ecology thanks to the growing availability of large ecological databases. However, these databases usually contain a large fraction of missing data because measuring plant functional traits systematically is labour-intensive and because most databases are compilations of datasets with different sampling designs. As a result, within a given database, there is an inevitable variability in the number of traits available for each data entry and/or the species coverage in a given geographical area. The presence of missing data may severely bias trait-based analyses, such as the quantification of trait covariation or trait-environment relationships and may hamper efforts towards trait-based modelling of ecosystem biogeochemical cycles. Several data imputation (i.e. gap-filling) methods have been recently tested on compiled functional trait databases, but the performance of imputation methods applied to a functional trait database with a regular spatial sampling has not been thoroughly studied. Here, we assess the effects of data imputation on five tree functional traits (leaf biomass to sapwood area ratio, foliar nitrogen, maximum height, specific leaf area and wood density) in the Ecological and Forest Inventory of Catalonia, an extensive spatial database (covering 31900 km2). We tested the performance of species mean imputation, single imputation by the k-nearest neighbors algorithm (kNN) and a multiple imputation method, Multivariate Imputation with Chained Equations (MICE) at different levels of missing data (10%, 30%, 50%, and 80%). We also assessed the changes in imputation performance when additional predictors (species identity, climate, forest structure, spatial structure) were added in kNN and MICE imputations. We evaluated the imputed datasets using a battery of indexes describing departure from the complete dataset in trait distribution, in the mean prediction error, in the correlation matrix and in selected bivariate trait relationships. MICE yielded imputations which better preserved the variability and covariance structure of the data and provided an estimate of between-imputation uncertainty. We found that adding species identity as a predictor in MICE and kNN improved imputation for all traits, but adding climate did not lead to any appreciable improvement. However, forest structure and spatial structure did reduce imputation errors in maximum height and in leaf biomass to sapwood area ratios, respectively. Although species mean imputations showed the lowest error for 3 out the 5 studied traits, dataset-averaged errors were lowest for MICE imputations with all additional predictors, when missing data levels were 50% or lower. Species mean imputations always resulted in larger errors in the correlation matrix and appreciably altered the studied bivariate trait relationships. In conclusion, MICE imputations using species identity, climate, forest structure and spatial structure as predictors emerged as the most suitable method of the ones tested here, but it was also evident that imputation performance deteriorates at high levels of missing data (80%).