Geophysical Research Abstracts
Vol. 15, EGU2013-5839, 2013
EGU General Assembly 2013

Filling gaps in large ecological databases: consequences for the study of global-scale plant functional trait patterns

Franziska Schrodt (1,3), Hanhuai Shan (2), Farideh Fazayeli (2), Anuj Karpatne (2), Jens Kattge (1), Arindam Banerjee (2), Markus Reichstein (1), and Peter Reich (3)

(1) Biogeochemical Model-Data Integration Group, Max Planck Institute for Biogeochemistry, Jena, Germany, (2) Department of Computer Science and Engineering, University of Minnesota, Minneapolis, MN, United States, (3) Department of Forest Resources, University of Minnesota, St Paul, MN, United States

With the advent of remotely sensed data and coordinated efforts to create global databases, the ecological community has progressively become more data-intensive. However, in contrast to other disciplines, statistical ways of handling these large data sets, especially the gaps which are inherent to them, are lacking. Widely used theoretical approaches, for example model averaging based on Akaike's information criterion (AIC), are sensitive to missing values. Yet, the most common way of handling sparse matrices – the deletion of cases with missing data (complete case analysis) – is known to severely reduce statistical power as well as inducing biased parameter estimates. In order to address these issues, we present novel approaches to gap filling in large ecological data sets using matrix factorization techniques.

Factorization based matrix completion was developed in a recommender system context and has since been widely used to impute missing data in fields outside the ecological community. Here, we evaluate the effectiveness of probabilistic matrix factorization techniques for imputing missing data in ecological matrices using two imputation techniques. Hierarchical Probabilistic Matrix Factorization (HPMF) effectively incorporates hierarchical phylogenetic information (phylogenetic group, family, genus, species and individual plant) into the trait imputation. Advanced Hierarchical Probabilistic Matrix Factorization (aHPMF) on the other hand includes climate and soil information into the matrix factorization by regressing the environmental variables against residuals of the HPMF. One unique opportunity opened up by aHPMF is out-of-sample prediction, where traits can be predicted for specific species at locations different to those sampled in the past. This has potentially far-reaching consequences for the study of global-scale plant functional trait patterns.

We test the accuracy and effectiveness of HPMF and aHPMF in filling sparse matrices, using the TRY database of plant functional traits (http://www.try-db.org). TRY is one of the largest global compilations of plant trait databases (750 traits of 1 million plants), encompassing data on morphological, anatomical, biochemical, phenological and physiological features of plants. However, despite of unprecedented coverage, the TRY database is still very sparse, severely limiting joint trait analyses. Plant traits are the key to understanding how plants as primary producers adjust to changes in environmental conditions and in turn influence them. Forming the basis for Dynamic Global Vegetation Models (DGVMs), plant traits are also fundamental in global change studies for predicting future ecosystem changes. It is thus imperative that missing data is imputed in as accurate and precise a way as possible.

In this study, we show the advantages and disadvantages of applying probabilistic matrix factorization techniques in incorporating hierarchical and environmental information for the prediction of missing plant traits as compared to conventional imputation techniques such as the complete case and mean approaches. We will discuss the implications of using gap-filled data for global-scale studies of plant functional trait – environment relationship as opposed to the above-mentioned conventional techniques, using examples of out-of-sample predictions of foliar Nitrogen across several species' ranges and biomes.