

Imputing species-level plot basal area and tree density attributes from remotely sensed data in north-central Idaho

Andrew T. Hudak, Nicholas L. Crookston, and Jeffrey S. Evans
USDA Forest Service Rocky Mountain Research Station
RMRS Forestry Sciences Laboratory, 1221 South Main St., Moscow, ID 83843
Emails: ahudak@fs.fed.us, ncrookston@fs.fed.us, jevans02@fs.fed.us

Abstract: Meaningful relationships between forest structure attributes measured on the ground in an unbiased manner, and remotely sensed data measured comprehensively across the same forested landscape, allow the production of maps of forest attributes of interest. Imputation approaches are useful for mapping structural attributes at the species level while preserving the covariance structure of the data. We compared six alternative approaches for imputing basal area and tree density attributes aggregated at the plot scale and species level (Y variables), from topographic variables and canopy structure metrics derived from discrete return airborne lidar data, as well as Advanced Land Imager (ALI) satellite band data (X variables). The X and Y variables were associated using either Euclidean distance, Mahalanobis distance, Canonical Correlation Analysis (aka Most Similar Neighbor, or MSN), Canonical Correspondence Analysis (aka Gradient Nearest Neighbor, or GNN), or the Random Forest approach of generating multiple permutations of classification trees. This last case was achieved by running the Random Forest algorithm as a supervised classification of each Y variable and then concatenating the resulting classification trees; the number of terminal nodes shared between the X and Y variables served as a proxy for nearest neighbor distances, as calculated by the other imputation methods. To compare and evaluate these alternative approaches, we computed the Pearson correlation between observed and imputed values for the nine most common conifer species sampled. Plot totals of basal area (m^2/ha) and tree density (trees/ha) for all species were also added as variables. It is questionable whether the correlation statistic is most appropriate for comparing imputation approaches. For this reason, we also computed the root mean squared difference (RMSD) between imputed and observed values, and compared imputation methods with this statistic as well.

When considering all species, we found that Canonical Correspondence Analysis (GNN) and Random Forest (RF) produced comparably high correlations between imputed and observed values, followed by Euclidean distance, while Canonical Correlation Analysis (MSN) performed markedly worse, and Mahalanobis distance the worst. Contrary to our expectations, this pattern also was observed with regard to the RMSD statistic, with RF usually producing the highest values. However, other considerations reflected favorably on the RF method. The shape of the distribution of values imputed by RF reproduced the positive skew in the distribution of observed values better than did the other methods, especially GNN which produced Gaussian distributions. Increasing the number of neighbors (k) improved the correlation and RMSD statistics across all methods, as did applying inverse distance weighting to the nearest neighbors rather than simply averaging them. We concluded that GNN and RF were the best imputation methods considered for this problem. The RF results were particularly promising and add to a growing body of research suggesting that the Random Forest approach to predictive modeling in the forestry sciences has considerable merit.