living planet symposium
13–17 May 2019 | Milan

Enhancing NDVI-Based Biomass Models Through Feature Selection and Spatiotemporal Cross-Validation

Thomas P. Higginbottom, Elias Symeonakis

Manchester Metropolitan University, Manchester M15GD, UK

The accurate mapping and quantification of above ground biomass (AGB) is required for a number of applications, including carbon accounting, fire and grazing management, amongst others. Accordingly, relating field measurements of AGB to satellite-derived indicators, most prominently the Normalised Difference Vegetation Index (NDVI) has been a feature of the remote sensing literature for over 30 years. Recently, there has been an increase in the use of machine learning methods and the incorporation of auxiliary environmental variables for spatiotemporal modelling. However, there is increasing evidence that these models may be vulnerable to artefacts of data structure, such as spatial autocorrelation and inappropriate auxiliary variables, which may hinder the development of accurate models. In this study, a robust methodology for the creation of moderate-resolution AGB estimates is presented. We obtained AGB data from an 18-year long dataset comprising 533 sites within the Kruger National Park of South Africa. We then generated a 36 1km-resolution NDVI product by downscaling the GIMMS 3g NDVI using Empirical Orthogonal Teleconnections (EOT) and the MODIS MYD13A2. AGB was then predicted based on a series of NDVI-metrics and auxiliary environmental variables in a Cubist regression model framework. Our analysis consisted of two components: i) a comparison of validation approaches, including a k-fold cross validation (CV) and multiple spatial/temporal CVs; and ii) a variable selection component, incorporating forward feature selections (FFS) on the above validation strategies. Prediction accuracies differed considerably, with the Root Mean Squared Error ranging from 1310 to 1844 kg ha^{-1}, depending on the variables and validation strategy employed. Errors were consistently higher with spatial or temporal validation strategies. Spatial overfitting was prominent in most models, which we attribute to spatial autocorrelation within the predictor variables. Comparatively, the NDVI-biomass relationship was highly variable between years, with unseen years being poorly modelled. This potentially results from changing species composition and moisture content on an annual basis. The FFS was effective at correcting these issues, where possible, by constructing models with appropriate variable combinations. For temporal models, the profile of auxiliary variables was increased leading to a more deterministic prediction approach. This study contributes to the growing literature highlighting the potential pitfalls of machine learning for spatiotemporal predictions, and offers strategies for their detection and mitigation.