Performance evaluation and hyperparameter tuning of statistical and machine-learning models using spatial data

Patrick Schratz
(PhD Student at Department of Geography, GIScience group,
Friedrich-Schiller-Universität Jena)

20.06.2018, 16:00 (s.t.)
Seminarraum, Institut für Statistik

Abstract:
Machine-learning algorithms have gained popularity in recent years in the field of ecological modeling due to their promising results in predictive performance of classification problems. While the application of such algorithms has been highly simplified in the last years due to their well-documented integration in commonly used statistical programming languages such as R, there are several practical challenges in the field of ecological modeling related to unbiased performance estimation, optimization of algorithms using hyperparameter tuning and spatial autocorrelation. We address these issues in the comparison of several widely used machine-learning algorithms such as Boosted Regression Trees (BRT), k-Nearest Neighbor (WKNN), Random Forest (RF) and Support Vector Machine (SVM) to traditional parametric algorithms such as logistic regression (GLM) and semi-parametric ones like generalized additive models (GAM). Different nested cross-validation methods including hyperparameter tuning methods are used to evaluate model performances with the aim to receive bias-reduced performance estimates. As a case study the spatial distribution of forest disease Diplodia sapinea in the Basque Country in Spain is investigated using common environmental variables such as temperature, precipitation, soil or lithology as predictors. Results show that GAM and RF (mean AUROC estimates 0.708 and 0.699) outperform all other methods in predictive accuracy. The effect of hyperparameter tuning saturates at around 50 iterations for this data set. The AUROC differences between the bias-reduced (spatial cross-validation) and overoptimistic (non-spatial cross-validation) performance estimates of the GAM and RF are 0.167 (24%) and 0.213 (30%), respectively. It is recommended to also use spatial partitioning for cross-validation hyperparameter tuning of spatial data.