Geophysical Research Abstracts, Vol. 9, 04071, 2007
SRef-ID: 1607-7962/gra/EGU2007-A-04071

The random forests technique: an application in eco-hydrologic distribution modelling

J. Peters (1), N.E.C. Verhoest (1), B. De Baets (2), R. Samson (3)

(1) Department of Forest and Water Management, Ghent University, Gent, Belgium, (2) Department of Applied Mathematics, Biometrics and Process Control, Ghent University, Gent, Belgium, (3) Department of Applied Biological Sciences, University of Antwerp, Antwerp, Belgium

Introduction: Exploring the distribution of plant species and vegetation types is a central goal in ecohydrology. Numerous studies examined hydrological and hydrogeochemical gradients in relation to plant species or vegetation type distributions in various ecosystems. Most approaches developed for distribution modelling have their roots in quantifying species-environment or vegetation-environment relationships. Distribution models are mostly empirical models relating field observations to environmental predictors based on statistically or theoretically derived responses.
Material and methods: In this study, two statistical techniques are evaluated: (i) the widely used multiple logistic regression technique within the generalized linear modelling (GLM) framework, and (ii) a recently developed machine learning technique called 'random forests'. The latter is an ensemble learning technique which generates many classification trees and aggregates the individual results. The two techniques are used to develop distribution models to predict vegetation type occurrences of eleven vegetation types of the Flemish (Belgium) lowland valley ecosystems based on spatially distributed measurements of environmental conditions. This spatially distributed data set consists of 1705 grid cells covering an area of 47.32 ha.
Results:

(i) The random forest model is tested on overfitting. The generalization error converges when more classification trees are added to the ensemble. The random forest model does not overfit.

(ii) The models are applied to independent data sets using 2-fold cross-validation.

Predicted vegetation types are compared with observations, and the McNemar test indicates a better performance of the random forest model compared with the multiple logistic regression model at the 0.05 significance level.

(iii) The random forest model calculates probabilities of occurrence for the different vegetation types for each grid cell. Inspection of these values for correctly classified grid cells indicates a strong predictive power in the central areas of homogeneous vegetation sites with a decreasing confidence towards the margins of these areas.

Synthesis and applications: Ecohydrological distribution models are useful tools to predict future vegetation changes under different management strategies or changing environmental conditions. Incorporating the random forest technique in those models has the ability to lead to better model performances.