

Modelling the relationship between zonal forest types and climatic variables in Hungary

Bálint Czúcz¹, László Gálhidy², Csaba Mátyás² and Franck Torre³

Introduction

Distribution of zonal forest types are critically influenced by climatic conditions, and climate change will probably affect their future distributions in many countries including South-East European ones.

Our aim is to study the relationship between climatic variables and zonal forest types in Hungary based on their current distribution according to forest inventory data.

We use two national databases as data sources:

- Hungarian Forest Inventory Data (2001, four main zonal forest types)
- Climate surfaces from the Hungarian Meteorological Service (1961-1990, 20 climatic variables).

We used regression tree analysis for the modelling, selecting the most important climatic variables that explain the present-day distribution of zonal forest types.

On this poster we present a simple case study with two of the tree species (*Fagus sylvatica* and *Quercus petraea*), in order to show the power and the weaknesses of this modelling technique, as well as our efforts in improving it.

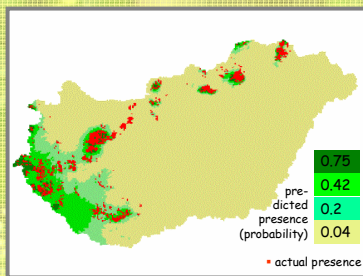
Regression Trees

Classification and Regression Trees are simple models of a nonlinear relationship between a response variable and several predictor variables. They are constructed using recursive binary partitioning over a training data set. In each step one predictor variable is selected, and according to its value the whole data base is split into two child partitions. The successive splits are implemented with the aim of achieving maximal homogeneity in the 'child's.

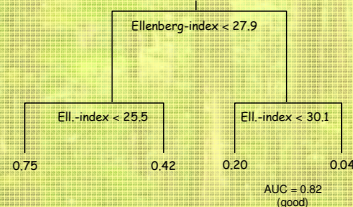
The main advantages of tree-based methods are:

- There can be many possible "predictor" variables during the analysis, collinearities don't matter
- Predictor variables do not have to be normally distributed
- Ability to describe nonlinear relationships
- The resulting models are easy to interpret, even for non experts

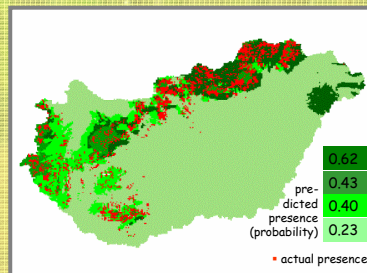
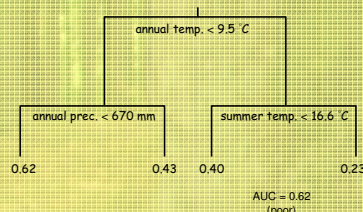
The calculations of this study are performed in the R statistical environment, where we used the package 'rpart' to build the trees. To evaluate the models we used the receiver operating characteristics (ROC) analysis ('ROCR' package) and the area under curve (AUC) index to compare the results.



Predicted map (left) and CART-diagram (below) of the *European beech (Fagus sylvatica)* in Hungary



Predicted map (right) and CART-diagram (below) of the *sessile oak (Quercus petraea)* in Hungary



Simple linear combination (SLC) splits

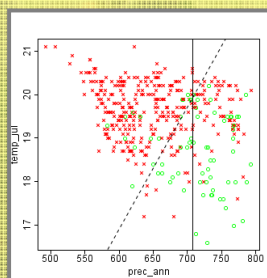
This is a new concept in decision tree theory. Previously there used to be principally two kind of splits in decision trees:

- linear combination (LC) splits (oblique trees):
 - the splitting is always done over a linear combination of all the predictor variables (thus in the space of the predictor variables, the objects are split by an arbitrarily directed hyperplane)

- univariate splits (orthogonal trees):
 - the splitting is always performed after a single splitting variable (thus in the space of the predictor variables, the objects are split by a hyperplane parallel to the axes)

LC splits are known to perform better in prediction, but are still rarely applied, because such models lose one of the most appealing features of the tree-based models: the ease of interpretation. „There is a tradeoff to consider in allowing multivariate tests: simple tests may result in large trees that are difficult to understand, yet multivariate tests may result in small trees with tests that are difficult to understand.” (Utgoff & Brodley, 1990) Though this tradeoff has been recognized quite long ago, until recently there was no good solution.

The idea of SLC splits arises from the recognition that increasing complexity caused by including more terms in a LC split can be handled in a similar way to the complexity caused by including further nodes into the model. The selection of splitting variable is transformed into a sequential procedure, where new terms are entered into the SLC one by one only if the larger SLC performs significantly better than the previous one. (This procedure can be incorporated in a simple and straightforward way into the framework of conditional inference based decision trees, using the same permutation tests as in the variable selection and the stopping rule.)



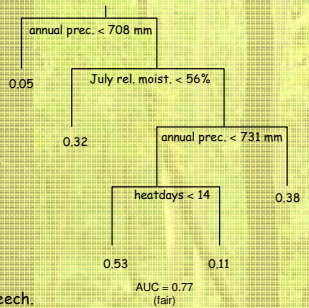
Simple illustration of the difference between univariate and LC splits: this is (a random sample from) the beech database plotted against the two predictors from the Ellenberg-index. The solid line shows the root split of the original tree-model, and the dashed line is the same of the tree with Ellenberg-index included. The difference in split homogeneity can be clearly seen

A thought-provoking issue: the „Ellenberg-effect“

• It is worth to compare the „beach-tree“ on the left with this one →

• Despite its more complex structure, this tree performs worse than the previous one.

• The only difference between the two tree-models is that, in the first model we introduced an additional climatic variable, the so-called Ellenberg index (July temperature/annual precipitation). This index is known to be in strong relationship with the climatic needs of several plant species, including beech.



Lessons from the „Ellenberg-effect“

• The reason for the success of Ellenberg-index is that the real factors conditioning the ecological processes are rarely the ones that we are able to measure.

• Including simple combinations of predictors in the model may significantly improve both simplicity and performance.

• There might be several other combinations of the predictors, that could improve the model even better. (These can be thought of as bioclimatic indices 'not yet invented'...)

(Our) ways of improving tree-based bioclimatic models:

- Applying new, modern modelling algorithms: we intend to use a new decision tree method based on conditional inference (package 'party' in R). Important new features:
 - permutation tests in each step of the model building
 - unbiased splitting variable selection at the nodes
 - a powerful stopping rule with no risk of overfitting
 - one-phase tree growing without pruning
 - capable of coping with multivariate responses
 - allowing of sensible transformations to predictor and response variables

• Introducing *simple linear combination splits* into the model (see right box) → in order to be able to identify 'new bioclimatic indices'

• Building *logistic regression trees* by incorporating the appropriate transformation into the model for the responses

• Building one tree with multivariate response instead of several univariate response trees can also improve our understanding

Main references

- Breiman, L., Friedman, J.H., Olshen, R.A. & Stone, C.J. (1984): Classification and Regression Trees: Chapman and Hall (Wadsworth, Inc.): New York.
- Utgoff, P.E., & Brodley, C.E. (1990): An incremental method for finding multivariate splits for decision trees. *Proceedings of the Seventh International Conference on Machine Learning* (pp. 58-65). Austin, TX: Morgan Kaufmann.
- Hothorn, T., Hornik, K. & Zeileis, A. (2006): Unbiased Recursive Partitioning: A Conditional Inference Framework, *Journal of Computational and Graphical Statistics* (accepted).