CHOOSING THE OPTIMAL SIZE OF MULTIVARIATE REGRESSION TREES (MRT)

M.-H. Ouellette[1], P. Legendre[1]

[1]*Université de Montréal, Québec, Canada*

Email: *marie-helene.ouellette@umontreal.ca*

The Multivariate Regression Tree (MRT) is a divisive hierarchical clustering method that groups multivariate objects in a tree-like manner according to the fit provided by several explanatory variables. Each split of the tree groups the objects in clusters that are homogenous with respect to the explanatory variable selected to model that split. To select the optimal size of a tree, cross-validation (CV) is commonly used. Unfortunately, CV over-estimates the prediction power of the model if the user exploits data from a single source. On top of that, the relative error (R^2) is influenced by the number of explanatory variables in the model and by the sample size: it gets larger as we use additional explanatory variables. As a result of this property, it cannot be used for model selection. We thus define an $R^2{}_a$ (adjusted R^2) that replaces CV for selection of the optimal size of the MRT. With simulations, we show that $R^2{}_a$ is suited to fulfill this function. This new coefficient will not only be useful to select the optimal tree size; it will also be valuable to compare models that have been built from different data sets. This method can be of great interest to biologists. For instance, it can be useful to identify the factors that influence the relative abundance (or richness) of species found at multiple sites. Some species have no pattern; other species, that may form species associations, are related to environmental factors of interest.