Université de Montréal

L'arbre de régression multivariable et les modèles linéaires généralisés revisités : applications à l'étude de la diversité bêta et à l'estimation de la biomasse d'arbres tropicaux

par

Marie-Hélène Ouellette

Département de sciences biologiques

Faculté des Arts et des Sciences

Thèse présentée à la Faculté des études supérieures

En vue de l'obtention du grade de Philosophiae Doctor (Ph.D.) en sciences biologiques

avril, 2011

Université de Montréal

Faculté des études supérieures

Cette thèse intitulée :

L'arbre de régression multivariable et les modèles linéaires généralisés revisités : applications à l'étude de la diversité bêta et à l'estimation de la biomasse d'arbres tropicaux

Présentée par :

Marie-Hélène Ouellette

A été évaluée par un jury composé des personnes suivantes :

Daniel Boisclair, président rapporteur

Pierre Legendre, directeur de recherche

François-Joseph Lapointe, membre du jury

David R. Larsen, examinateur externe

Paul Charbonneau, représentant du doyen de la FES

Résumé

En écologie, dans le cadre par exemple d'études des services fournis par les écosystèmes, les modélisations descriptive, explicative et prédictive ont toutes trois leur place distincte. Certaines situations bien précises requièrent soit l'un soit l'autre de ces types de modélisation ; le bon choix s'impose afin de pouvoir faire du modèle un usage conforme aux objectifs de l'étude.

Dans le cadre de ce travail, nous explorons dans un premier temps le pouvoir explicatif de l'arbre de régression multivariable (ARM). Cette méthode de modélisation est basée sur un algorithme récursif de bipartition et une méthode de rééchantillonage permettant l'élagage du modèle final, qui est un arbre, afin d'obtenir le modèle produisant les meilleures prédictions. Cette analyse asymétrique à deux tableaux permet l'obtention de groupes homogènes d'objets du tableau réponse, les divisions entre les groupes correspondant à des points de coupure des variables du tableau explicatif marquant les changements les plus abrupts de la réponse.

Nous démontrons qu'afin de calculer le pouvoir explicatif de l'ARM, on doit définir un coefficient de détermination ajusté dans lequel les degrés de liberté du modèle sont estimés à l'aide d'un algorithme. Cette estimation du coefficient de détermination de la population est pratiquement non biaisée. Puisque l'ARM sous-tend des prémisses de discontinuité alors que l'analyse canonique de redondance (ACR) modélise des gradients linéaires continus, la comparaison de leur pouvoir explicatif respectif permet entre autres de distinguer quel type de patron la réponse suit en fonction des variables explicatives. La comparaison du pouvoir explicatif

entre l'ACR et l'ARM a été motivée par l'utilisation extensive de l'ACR afin d'étudier la diversité bêta.

Toujours dans une optique explicative, nous définissons une nouvelle procédure appelée l'arbre de régression multivariable en cascade (ARMC) qui permet de construire un modèle tout en imposant un ordre hiérarchique aux hypothèses à l'étude. Cette nouvelle procédure permet d'entreprendre l'étude de l'effet hiérarchisé de deux jeux de variables explicatives, principal et subordonné, puis de calculer leur pouvoir explicatif. L'interprétation du modèle final se fait comme dans une MANOVA hiérarchique. On peut trouver dans les résultats de cette analyse des informations supplémentaires quant aux liens qui existent entre la réponse et les variables explicatives, par exemple des interactions entres les deux jeux explicatifs qui n'étaient pas mises en évidence par l'analyse ARM usuelle.

D'autre part, on étudie le pouvoir prédictif des modèles linéaires généralisés en modélisant la biomasse de différentes espèces d'arbre tropicaux en fonction de certaines de leurs mesures allométriques. Plus particulièrement, nous examinons la capacité des structures d'erreur gaussienne et gamma à fournir les prédictions les plus précises. Nous montrons que pour une espèce en particulier, le pouvoir prédictif d'un modèle faisant usage de la structure d'erreur gamma est supérieur. Cette étude s'insère dans un cadre pratique et se veut un exemple pour les gestionnaires voulant estimer précisément la capture du carbone par des plantations d'arbres tropicaux. Nos conclusions pourraient faire partie intégrante d'un programme de réduction des émissions de carbone par les changements d'utilisation des terres.

Mots clés
Arbre de régression multivariable ; diversité bêta ; estimation de la biomasse d'arbres tropicaux ; modèle linéaire généralisé ; recapture du carbone

Abstract

In ecology, in ecosystem services studies for example, descriptive, explanatory and predictive modelling all have relevance in different situations. Precise circumstances may require one or the other type of modelling; it is important to choose the method properly to insure that the final model fits the study's goal.

In this thesis, we first explore the explanatory power of the multivariate regression tree (MRT). This modelling technique is based on a recursive bipartitionning algorithm. The tree is fully grown by successive bipartitions and then it is pruned by resampling in order to reveal the tree providing the best predictions. This asymmetric analysis of two tables produces homogeneous groups in terms of the response that are constrained by splitting levels in the values of some of the most important explanatory variables.

We show that to calculate the explanatory power of an MRT, an appropriate adjusted coefficient of determination must include an estimation of the degrees of freedom of the MRT model through an algorithm. This estimation of the population coefficient of determination is practically unbiased. Since MRT is based upon discontinuity premises whereas canonical redundancy analysis (RDA) models continuous linear gradients, the comparison of their explanatory powers enables one to distinguish between those two patterns of species distributions along the explanatory variables. The extensive use of RDA for the study of beta diversity motivated the comparison between its explanatory power and that of MRT.

In an explanatory perspective again, we define a new procedure called a cascade of multivariate regression trees (CMRT). This procedure provides the possibility of computing an MRT model where an order is imposed to nested explanatory hypotheses. CMRT provides a framework to study the exclusive effect of a main and a subordinate set of explanatory variables by calculating their explanatory powers. The interpretation of the final model is done as in nested MANOVA. New information may arise from this analysis about the relationship between the response and the explanatory variables, for example interaction effects between the two explanatory data sets that were not evidenced by the usual MRT model.

On the other hand, we study the predictive power of generalized linear models (GLM) to predict individual tropical tree biomass as a function of allometric shape variables. Particularly, we examine the capacity of gaussian and gamma error structures to provide the most precise predictions. We show that for a particular species, gamma error structure is superior in terms of predictive power. This study is part of a practical framework; it is meant to be used as a tool for managers who need to precisely estimate the amount of carbon recaptured by tropical tree plantations. Our conclusions could be integrated within a program of carbon emission reduction by land use changes.

Keywords
Beta diversity ; carbon recapture ; generalized linear models ; multivariate regression tree ; tropical tree biomass estimation

Table des Matières

Résumé ……..…..**iii**

Mots clés *iv*

Abstract ……..…..**vi**

Keywords *vii*

Table des matières ……..…... **viii**

Listes des tableaux ……..……..……....**xi**

Listes des figures ……..…**xiii**

Listes des sigles et abréviations……..……..…**xxv**

Remerciements ……....**xxviii**

Avant-propos……..…..**xxx**

Chapitre 1 : Introduction générale et objectifs de la thèse……..**1**

Chapitre 2 : Revue et analyse de la littérature ……....**8**

Chapitre 3 : An adjusted R^2 statistic for multivariate regression tree analysis……17

Summary *17*

Introduction *18*

Definitions and proofs *21*

Coefficient of determination 22

Degrees of freedom 25

Bias assessment with simulated ecological data *29*

Results *31*

Number of groups or leaves 31

Estimation of the population tree size 34

Assessment of τ for small samples 35

Case studies *35*

Discussion *38*

Ecological implications 38

Adjustment, number of leaves and tuning parameter 40

Comparing $R^2_{MRT(GDF)}$ with differing impurity measures 41

Conclusion *42*

Acknowledgment *43*

Appendix 1 *44*

Appendix 2 *47*

Chapitre 4 : Cascade Multivariate Regression Tree: a novel approach for modelling nested explanatory sets…….…….. 78

Summary *78*

Introduction *80*

CMRT: the procedure *83*

R^2 partition *87*

Software *88*

Case studies *89*

Doubs River fish 89

Oribatid mite 96

Discussion 98

General remarks on the procedure 98

The case studies 101

Hierarchical hypotheses in ecology 103

Extension of the cascade 105

Relating CMRT to nested MANOVA 106

Conclusion *106*

Acknowledgment *107*

Chapitre 5 : Bootstrap assessment of the prediction accuracy of aboveground tree biomass estimation for five native species in a young Panamanian tropical plantation ……..……..……...108

Abstract *108*

Introduction *109*

Material and methods *112*

Comparing GLMs for AGB estimation 112

Sampling 116

Results *121*

Discussion *134*

Choice of modelling techniques 134

Data transformation 135

Comparing regressions : the bootstrap approach 140

Conclusion *141*

Chapitre 6 : Discussion générale, conclusions et perspectives d'étude……..... *142*

Références ……..……. 154

Listes des tableaux

Chapitre 5

Table 5.1 : List of the basic equations used. *H*, *D*, *S* and *AGB* stand for height, diameter, density and above-ground biomass respectively. …….**114**

Table 2: Tables of comparison between Gamma and Gaussian model.

a) Table showing the number of comparisons where Gamma performed better than Gaussian family on the total number of comparisons for all basic equations per data set. ……...**125**

b) Table showing the percentage of comparisons where Gamma performed better than Gaussian family on the total number of comparisons for all basic equations. ……..**125**

c) Table showing the number of comparisons where Gamma performed better than Gaussian family on the total number of comparisons by type of transformation or link. ……..**126**

Table 5.3: List of best predictive models for all species pooled (n=150) chosen according to the 0.632 predictive error estimator ($MSE_{.632}$) given for all basic equations and measures of tree diameter. …….**128**

Table 5.4: List of best predictive models for *Anacardium excelsum* chosen according to the 0.632 predictive error estimator ($MSE_{.632}$) given for all basic equations and measures of tree diameter.…….**129**

Table 5.5: List of best predictive models chosen for *Cedrela odorata* according to the 0.632 predictive error estimator ($MSE_{.632}$) given for all basic equations and measures of tree diameter. ……**130**

Table 5.6: List of best predictive models chosen for *Hura crepitans* according to the 0.632 predictive error estimator ($MSE_{.632}$) given for all basic equations and measures of tree diameter (in other words, the best predictive model for each of Tables 6-20 of the Appendix). The $MSE_{.632}$ value is given, as is the regression technique and the transformation or link of the response when applicable. ……**131**

Table 5.7: List of best predictive models chosen for *Luehea seemanii* according to the 0.632 predictive error estimator ($MSE_{.632}$) given for all basic equations and measures of tree diameter (in other words, the best predictive model for each of Tables 6-20 of the Appendix). The $MSE_{.632}$ value is given, as is the regression technique and the transformation or link of the response when applicable. ……..…**132**

Table 5.8: List of best predictive models chosen for *Tabebuia rosea* according to the .632 predictive error estimator ($MSE_{.632}$) given for all basic equations and measures of tree diameter (in other words, the best predictive model for each of Tables 6-20 of the Appendix). The $MSE_{.632}$ value is given, as is the regression technique and the transformation or link of the response when applicable. ……**133**

Listes des figures

Chapitre 3

Figure 3.1: Boxplot triplets of ρ^2 estimates (see legend) for trees with different numbers of leaves. The triplets are shown in the same order: R^2, $R^2_{(p)}$ and finally $R^2_{(GDF)}$. Simulations were carried out on population 1 with a sample size of 100 ……**33**

Figure A3.2: Boxplot triplets of ρ^2 estimates (see legend) for trees with different number of leaves. All triplets are shown in the same order: R^2, $R^2_{(p)}$ and finally $R^2_{(GDF)}$. Simulations were carried on population 1 with a sample size of 50……..…..**47**

Figure A3.3: Boxplot triplets of ρ^2 estimates (see legend) for trees with different number of leaves. All triplets are shown in the same order: R^2, $R^2_{(p)}$ and finally $R^2_{(GDF)}$. Simulations were carried on population 1 with a sample size of 20……..……. **48**

Figure A3.4: Boxplot triplets of ρ^2 estimates (see legend) for trees with different sample sizes. All triplets are shown in the same order: R^2, $R^2_{(p)}$ and finally $R^2_{(GDF)}$. Simulations were carried on population 1 and trees with 2 leaves were build (underfitted trees)……..**49**

Figure A3.5: Boxplot triplets of ρ^2 estimates (see legend) for trees with different sample sizes. All triplets are shown in the same order: R^2, $R^2_{(p)}$ and finally $R^2_{(GDF)}$. Simulations were carried on population 1 and trees with 3 leaves were build (underfitted trees).……..**50**

Figure A3.6: Boxplot triplets of ρ^2 estimates (see legend) for trees with different sample sizes. All triplets are shown in the same order: R^2, $R^2_{(p)}$ and finally $R^2_{(GDF)}$. Simulations were carried on population 1 and trees with 4 leaves were build (fitted trees).……..……..**51**

Figure A3.7: Boxplot triplets of ρ^2 estimates (see legend) for trees with different sample sizes. All triplets are shown in the same order: R^2, $R^2_{(p)}$ and finally $R^2_{(GDF)}$. Simulations were carried on population 1 and trees with 5 leaves were build (overfitted trees).……..……..…..……**52**

Figure A3.8: Boxplot triplets of ρ^2 estimates (see legend) for trees with different sample sizes. All triplets are shown in the same order: R^2, $R^2_{(p)}$ and finally $R^2_{(GDF)}$. Simulations were carried on population 1 and trees with 6 leaves were build (overfitted trees).……..……..……..……**53**

Figure A3.9: Boxplot triplets of ρ^2 estimates (see legend) for trees with different sample sizes. All triplets are shown in the same order: R^2, $R^2_{(p)}$ and finally $R^2_{(GDF)}$. Simulations were carried on population 1 and trees with 7 leaves were build (overfitted trees).……..……**54**

Figure A3.10: Boxplot triplets of ρ^2 estimates (see legend) for trees with different sample sizes. All triplets are shown in the same order: R^2, $R^2_{(p)}$ and finally $R^2_{(GDF)}$. Simulations were carried on population 1 and trees with 10 leaves were build (overfitted trees).. ……..……**55**

Figure A3.11: Boxplots of ρ^2 estimates (see abscissa) for trees with different τ tuning parameter values in GDF estimates. All triplets are shown in the same order: R^2, $R^2_{(p)}$ and finally $R^2_{(GDF)}$. Simulations were carried on population 1 and trees with 4 leaves were build (fitted trees).……**56**

Figure A3.12: Boxplot triplets of ρ^2 estimates (see legend) for trees with different number of leaves. All triplets are shown in the same order: R^2, $R^2_{(p)}$ and finally $R^2_{(GDF)}$. Simulations were carried on population 2 with a sample size of 100.……...……**57**

Figure A3.13: Boxplot triplets of ρ^2 estimates (see legend) for trees with different number of leaves. All triplets are shown in the same order: R^2, $R^2_{(p)}$ and finally $R^2_{(GDF)}$. Simulations were carried on population 2 with a sample size of 50……..**58**

Figure A3.14: Boxplot triplets of ρ^2 estimates (see legend) for trees with different number of leaves. All triplets are shown in the same order: R^2, $R^2_{(p)}$ and finally $R^2_{(GDF)}$. Simulations were carried on population 2 with a sample size of 20.…… **59**

Figure A3.15: Boxplot triplets of ρ^2 estimates (see legend) for trees with different sample sizes. All triplets are shown in the same order: R^2, $R^2_{(p)}$ and finally $R^2_{(GDF)}$. Simulations were carried on population 2 and trees with 2 leaves were build (underfitted trees).. ……**60**

Figure A3.16: Boxplot triplets of ρ^2 estimates (see legend) for trees with different sample sizes. All triplets are shown in the same order: R^2, $R^2_{(p)}$ and finally $R^2_{(GDF)}$. Simulations were carried on population 2 and trees with 5 leaves were build (underfitted trees).……**61**

Figure A3.17: Boxplot triplets of ρ^2 estimates (see legend) for trees with different sample sizes. All triplets are shown in the same order: R^2, $R^2_{(p)}$ and finally $R^2_{(GDF)}$. Simulations were carried on population 2 and trees with 6 leaves were build (fitted trees)……**62**

Figure A3.18: Boxplot triplets of ρ^2 estimates (see legend) for trees with different sample sizes. All triplets are shown in the same order: R^2, $R^2_{(p)}$ and finally $R^2_{(GDF)}$. Simulations were carried on population 2 and trees with 7 leaves were build (overfitted trees)……...**63**

Figure A3.19: Boxplot triplets of ρ^2 estimates (see legend) for trees with different sample sizes. All triplets are shown in the same order: R^2, $R^2_{(p)}$ and finally $R^2_{(GDF)}$. Simulations were carried on population 2 and trees with 10 leaves were build (overfitted trees). ……**64**

Figure A3.20: Boxplots of ρ^2 estimates (see abscissa) for trees with different τ tuning parameter values in GDF estimates. Simulations were carried on population 2 and trees with 6 leaves were build (fitted trees).……**65**

Figure A3.21: Barplots summarizing the Monte Carlo study (1000 runs) of the v-fold cross-validation 1se rule (500 multiple validations) to pick the population 2 size of tree (6) for sample sizes 20, 50 and 100 and for random, stratified and importance sampling strategies. Frequency of the tree size 5, 6 and 7 are the only ones depicted here.…….……..**66**

Figure A3.22: Barplots summarizing the Monte Carlo study (1000 runs) of the v-fold cross-validation min rule (500 multiple validations) to pick the population 2 size of tree (6) for sample sizes 20, 50 and 100 and for random, stratified and importance sampling strategies. Frequency of the tree size 5, 6 and 7 are the only ones depicted here.…….……..**68**

Figure A3.23: Multivariate regression tree model of the Doubs fish data set with Hellinger transformed response data. This output is provided by the *MRT()* function of MVPARTwrap. The main difference with the regular output of function *mvpart()* is the vertical scale, which is R^2 here. For each leaf, we find the number of objects in the node and the number of the group. The node numbers are in parentheses in the center of the nodes; the variation explained by each split is printed underneath. ……**70**

Figure A3.24: Geographical map of the MRT partition results for the Doubs' fish data set. The group numbers correspond to the numbers given in figure 23……..…**71**

Figure A3.25: RDA analysis triplot illustration (scaling 1, 'wa' scores) of the Doubs' fish data set with forward selection on the raw explanatory variables. The sites are color-coded according to the partition of the MRT with the same colors as figure 24. The numbers corresponds to the order from the source. Species are abbreviated by three capital letters: CHA (Bullhead *Cottus gobio*), TRU (Brown trout *Salmo trutta fario*), VAI (Minnow *Phoxinus phoxinus*), LOC (Stone Loach *Nemacheilus barbatulus*), OMB (Grayling *Thymallus thymallus*), BLA (Souffia or Western Vairon *Telestes soufia agassizi*), HOT (Nase *Chondrostoma nasusi*), TOX (Southwest european nose *Chondostroma toxostoma*), VAN (Common dace *Leuciscus leuciscus*), CHE (Chub *Leuciscus cephalus cephalus*), BAR (Common barbel *Barbus barbus*), SPI (Spirlin *Spirlinus bipunctatus*), GOU (Gudgeon *Gobio gobio*), BRO (Northern pike *Esox lucius*), PER (European perch *Perca fluviatilis*), BOU (European Bitterling *Rhodeus amarus*), PSO (Pumpkinseed sunfish *Lepomis gibbosus*), ROT (Rotfedern *Scardinius erythrophtalmus*), CAR (Common carp *Cyprinus carpio*), TAN (Tench *Tinca tinca*), BCO (Common bream *Abramis brama*), PCH (Black bullhead *Ictalurus melas*), GRE (Ruff *Acerina cernua*), GAR (Roach *Rutilus rutilus*), BBO (Silver bream *Blicca bjoerkna*), ABL (Bleak *Alburnus alburnus*), ANG (European eel *Anguilla anguilla*). Moreover, the explanatory variables selected by forward selection represented in this triplot

are distance to the source (das), biological oxygen demand (dbo), slope (pen), altitude (alt) and finally dissolved oxygen (oxy)..……..……..…..... **72**

Figure A3.26: Bleak abundances from the Doubs Hellinger transformed fish data set as a function of specific explanatory variables (distance to the source (das), oxygen content (oxy) and biological demand for oxygen (dbo)). The sites are color-coded according to the partition of the MRT with the same colors as figure 24. The numbers corresponds to the order from the source……..**74**

Figure A3.27: RDA analysis triplot illustration (scaling 1, 'wa' scores) of the spider data with original explanatory variables chosen by forward selection. Five explanatory variables were selected by the forward selection procedure, which were water content % of dry weight (Water.content), the cover by herb layer in % (Herb.cover), reflection of soil surface at cloudless sky × 100 (soil reflection), lux (AEG Lux-meter measure) at cloudless sky × 1000 (Illuminance.cloudless.sky) and finally cover by fallen leaves and twigs in % (Leaves.twigs). The species names are abbreviated as follows (- indicates no common name found): Alop.acce (- *Alcopecosa accentuata*), Alop.cune (- *Alopecosa cuneata*), Alop.fabr (Great fox-spider *Alopecosa fabrilis*), Arct.lute (- *Arctosa lutetiana*), Arct.peri (- *Aulonia perita*), Aulo.albi (- *Aulonia albimana*), Pard.lugu (- *Pardosa lugubris*), Pard.mont (Pin-stripe wolf-spider *Pardosa monticola*), Pard.nigr (- *Pardosa nigriceps*), Pard.pull (Common wolf spider *Pardosa pullata*), Troc.terr (Ground wolf-spider *Trochosa terricola*) and finally Zora.spin (- *Zora spinimana*). ……..……..……...**75**

Figure A3.28: Multivariate regression tree model of the spider data set with Hellinger transformed response. This output is provided by the MRT function of MVPARTwrap. At each leaf we find the number of objects in the node and the number of the group. The node numbers are in parentheses in the center of each node; the variation explained by each split is printed underneath. Five explanatory variables were selected by the forward selection procedure, which were water content % of dry weight (Water.content or Wtr.content), cover by fallen leaves and twigs in % (Leaves.twigs) and finally cover by *Corynephorus canescens* in % (Gray clubawn grass) which is noted here as Corynephorus.……..**76**

Figure A3.29: RDA analysis triplot illustration (scaling 1, 'wa' scores) of the spider data with polynomial of environment variables. Five explanatory variables were selected by the forward selection procedure, which were water content % of dry weight polynomial of degree two (Water.content2), cover by fallen leaves and twigs in % (Leaves.twigs), the cover by herb layer in % (polynomial degree one and two respectively Herb.cover and Herb.cover2), lux (AEG Lux-meter measure) at cloudless sky × 1000 (Illuminance.cloudless.sky), hummus content in % of dry weight (Humus) and finally polynomial of degree two of percentage of bare sand (Bare.sand2). The species Latin and common names (when available) are listed in the Figure 27 description…….……**71**

Chapitre 4

Box 4.1: Terminology review for MRT and CMRT analyses. In this diagram, we have four drops (four trees): one in wave 1 and three in wave 2……..…...**84**

Figure 4.1: (a) Diagram of the CMRT procedure along with (b) a general R^2 diagram. In (b) we depict the variation explained by the whole cascade in a rectangle whose size (left + right portions) represents the total variation in the response data (100%). The shaded area on the left represents the variation of the response data explained by the first wave (main analysis). The shaded area or areas (there may be more than one) on the right represent the variation explained by the subordinate drops of the second wave. For each shaded rectangle in the white area on the right, its width represents the proportion of the relative error (RE) of the first wave while its height represents the R^2 of the subsequent response explained by the subordinate drop. The white area is the variation that remains unexplained at the end of the waves…….....**85**

Figure 4.2: Original MRT analysis of the Doubs River fish data. For each node, its identification number in parentheses, e.g. (1), corresponds to the one found in the `summary.MRT` function of the `MVPARTwrap`. Under the number is found the percentage of explained variation. For each leaf, the number in parentheses, e.g. (#3), is the one found in the `summary.MRT` function of the `MVPARTwrap` package; the number of objects in the leaf is also shown, e.g. n = 4.……...……**91**

Figure 4.3: CMRT analysis results for the Doubs River data. Each drop is on the left; on the right we find the corresponding geographical map of the groups. The number (#) and size (n) of each leaf are shown. The number and percentage of explained variation are given for each node. Three explanatory variables appear in this figure: mean minimum discharge (deb), ammonium concentration (amm) and dissolved oxygen (oxy)……..…….**92**

Figure 4.4: Output of the ***CasMRTR2()*** function for the Doubs River fish data. The global R^2 is 55.6%, the portion of the global R^2 explained by the subordinate drop 3 is 14.36%, and only that one has any extra variation to be explained. The drop number corresponds to the number of the leaf in the tree of the first drop (Figure 4.3). The VA percentage (41.24%) is the variation explained by the main explanatory variable, which happens to be the 'mean discharge' variables…….……. **93**

Figure 4.5: MRT analysis for the oribatid mite data. Details: see legends of Figs. 4.2 and 4.3.……...…….**97**

Figure 4.6: Summary of the CMRT analysis results for the oribatid mite data with the explanatory variable shrub as the primary (main) effect. Details: see legend of Figs. 4.1 and 4.4. The explanatory variables used to split the objects were the shrub states (none, few, many; the variable is noted 'Shrubs'), the substrate density (dry matter) in g/dm^3 noted 'SubsDens', and finally the water content in g/dm^3 noted 'WatrCont'……..…….**99**

Figure 4.7: Output of the ***CasMRTR2()*** function for the oribatid mite data. The global R^2 is 36.07%; the portion of the global R^2 explained by subordinate drops 2 and 3 together is 19.74. The VA percentage (16.32%) is the

proportion of the response variation explained by the main explanatory variable, which happens to be shrub presence or absence.……...……..…**100**

Chapitre 5

Box 5.1 Representation of the different estimators of the prediction error organized along an optimism axis. This is not an exhaustive list. Let y_i be the response of the i^{th} object, $\hat{y}_i^{(p)}$ its predicted value, N the size of the full data set, N_{bs} the size of the bootstrap sample, P the number of bootstrap runs, and the subscript *(test)* the designation of an object that was not in the training set (thus not used to compute the model). MSE_{emp} is the most optimistic estimate because the sum of squares and the model are both calculated on the full data set. It is followed by MSE_{bs}, a bootstrap estimate where P models are computed and the mean prediction error is calculated over those P bootstrap samples. It is known to have a large downward bias (overoptimistic). MSE_{bs2} is known to have a smaller bias (as the bias is estimated by bootstrap); it is also called the ordinary bootstrap estimate. MSE_0 is the most pessimistic estimate because only the objects not used to compute the model are used to estimate the prediction error. $MSE_{.632}$ is a compromise between MSE_0 and MSE_{emp}; it is generally a good choice (e.g. Davison & Hinkley 1997)……..…..**117**

Figure 5.1 : Scatterplot of residuals as a function of fitted values for a link model (Chave1, BA, power link for Ae). In this case, the Gamma family better

grasps the residual structure, and has a smaller $MSE_{.632}$ value than the Gaussian family for this data set for power link…..……..……**123**

Figure 5.2 : Scatter plot of residuals as a function of fitted values for a transformation model (Chave1, BA, power transformation for Ae). In this case the Gaussian family better grasps the residual structure of the transformed data.……....**124**

Figure 5.3: Scatter plot of residuals as a function of fitted values for the best Cm model (Brown basic equation with BD tree diameter measure and log transformation)……...**136**

Figure 5.4: Assessment of residuals using the Ae data set, with BA as a tree diameter measure and Chave 1 basic equation. Here we show, from left to right, and top to bottom, the residuals as a function of the fitted values of models with increasing values of $MSE_{.632}$ (a < b < … < h). The models were (a) Gaussian with power transformation, (b) Gamma with power transformation, (c) Gaussian with *ln* transformation, (d) Gamma with *ln* link, (e) Gaussian with *ln* link, (f) Gamma with *ln* transformation, (g) Gamma with power link, and (h) Gaussian with power link. We observe that heteroscedasticity seems to be getting larger as we go down the figure for most models, but it is a subjective assessment. This strengthens the argument that further assessment of the predictive accuracy of the models by means of AIC or bootstrapping is necessary.……..…….**138**

Listes des sigles et abréviations

% : percent
°C : degree Celcius
'1se' rule : 1 standard error rule
ABL : bleak, *Alburnus alburnus*
ACR : analyse de redondance canonique
Ae : *Anacardium excelsum*
AGB : above ground biomass
AIC : Akaike information criterion
alt : altitude
amm : ammonium
ARM : arbre de régression multivariable
ARMC : arbres de régression multivariable en cascade
BA : basal area
BD : basal diameter
CART : classification and regression tree analysis
CCA : correspondance canonical analysis
CMRT : cascade multivariate regression tree
Ca : *Cordia alliodora*
Co : *Cedrela odorata*
CVRE : cross-validation relative error (also notre CV Error)
$\Delta \mathbf{T}_{nxm}$: matrix of standard normal deviates
D : diameter
DBH : diameter at breast height
DBHall : diameter at breast height of all stems
das : distance to the source
dbo : biological oxygen demand
dm : decimetre
deb : mean discharge
df : degrees of freedom
dur : hardness
e_i : residual value of object *i*
F: statistique F
GDF : generalized degrees of freedom
GLM : generalized linear models
GR^2 : general R^2 definition
g : group membership or number of leaves of an MRT
H : hat matrix of height of tree
h : hour
Hc : *Hura crepitans*
h_{ij} : element *i*,*j* of matrix H
i : line index
IWLS : iterative weighted least-squares
IndVal : indicator value

j : column index
k : k^{th} of Monte carlo runs
kg : kilogram
K : total number of Monte carlo runs
L : litre
ln : Neperian logarithm
LM : linear model
Ls : *Luehea seemanii*
MANOVA : analyse de variance multivariée / multivariate analysis of variance
MCO : Moyenne des carrés ordinaires
MEPC : moyenne des erreurs prédictives mises au carré
'min' rule : minimum CVRE rule
MLG : modèle linéaire généralisé
mg : milligram
MRT : multivariate regression tree
MS : mean square error
MSE : mean square error
$\mathbf{MSE_{.623}}$: weighted average of MSE_0 and MSE_{emp}
$\mathbf{MSE_0}$: as the mean sum of squared error computed only on the objects that are not members of the training set
$\mathbf{MSE_{bs}}$: estimated prediction error
$\mathbf{MSE_{emp}}$: empirical risk
MSPE : mean square prediction error
m : number of response variable
m : meter
μ : mean
N : north
n : number of objects
$N(\mu,\sigma^2)$: normal distribution of mean μ and variance σ^2
OLS : ordinary least squares
oxy : dissolved oxygen content
p : number of independent variables
P : partition or subsample
ρ^2 : population coefficient of determination
$\rho^2_{\mathbf{MRT}}$: population coefficient of determination of the MRT model
R^2 : coefficient de détermination / coefficient of determination
$R^2_{\mathbf{a}}$: coefficient de détermination ajusté / adjusted coefficient of détermination
$R^2_{\mathbf{MRT}}$: coefficient de détermination pour l'ARM / coefficient of détermination for the MRT analysis
$R^2_{\mathbf{Y|X}}$: coefficient de détermination pour l'ACR / coefficient of détermination for the RDA
$R^2_{\mathbf{MRT(GDF)}}$: coefficient de détermination ajusté pour l'ARM via GDF/ adjusted coefficient of détermination for the MRT with GDF
$R^2_{\mathbf{MRT(a)}}$: coefficient de détermination ajusté pour l'ARM via le nombre de noeuds/ adjusted coefficient of détermination for the MRT with number of nodes

$\boldsymbol{R^*(T)}$: true mean squared error of a tree T
$\boldsymbol{R(T)}$: resubstitution estimate of true mean squared error of a tree T
$\boldsymbol{R(T)^{cv}}$: cross-validation estimate of true mean squared error of a tree T
RE : relative error (also noted Error)
$\boldsymbol{RE^*(T)}$: true relative mean squared error of a tree T
RDA : canonical redundancy analysis
RSS : residual sum of squares
RT : univariate regression trees
$\boldsymbol{r}$: simple correlation
σ : population standard deviation
σ^2 : population variance
S : subordinate explanatory variable or density
$\boldsymbol{s}$: standard deviation (also noted SE)
s : second
SubsDens : substrate density
τ : tuner value
$\boldsymbol{T}$: a tree (in the clustering sense)
Topo : micro topography
Tr : *Tabebuia rosea*
V,$\boldsymbol{v}$: number of subsets
W : covariate matrix or west
WatrCont : water content
$\boldsymbol{x_i}$: explanatory variable in column i of X
X : explanatory table or position in Cartesian plot
Y : response table or position in Cartesian plot
$\hat{Y}$: fitted response table
$\boldsymbol{y_{ij}}$: an observation

Remerciements

En premier lieu je dois remercier mon directeur de thèse, Pierre Legendre, qui m'a soutenue tout au long de ce travail, sans montrer le moindre doute en mes capacités de remplir mon mandat, malgré toutes les embûches rencontrées et le temps qui s'écoulait. Il a été présent avec ses conseils judicieux du début à la fin. Ce fut un honneur d'être ton étudiante Pierre, et j'espère avoir été à la hauteur de tes attentes :o)

J'aimerais également remercier Daniel Borcard, chercheur sénior du laboratoire, pour ses petits conseils émis sur le bord d'une table, mais qui étaient tellement importants et pertinents. J'aimerais également le remercier pour m'avoir choisie comme chef-démo pendant toutes ces années, ce qui m'a permis de développer mes capacités d'enseignante au niveau universitaire. C'était tellement agréable que c'est difficile d'appeler ça du travail.

Sur le plan personnel, d'une importance au delà de ce que les gens pourraient bien imaginer, mes remerciements à tous les gens qui de près ou de loin m'ont montré leur support, soit en me proposant de partager leur maison, en écoutant mes inquiétudes, ou en me changeant les idées : merci à Michel, Liette, Catherine, Stéphanie, Gary, Marie-Line, Alex, Guillaume, Cindy, Sam, Françoise, Kristy, Isabelle, mon père, ma mère, mon frère, Murielle, Brad et toute ma famille. L'ordre n'a aucune importance. Vous êtes des anges.

Finalement, le dernier et bien loin d'être le moindre, j'aimerais remercier mon partenaire de vie qui m'a tellement manquée pendant ces dernières années, qui a su m'aimer suffisamment pour me laisser réaliser mon rêve d'enfance, qui m'a donné un support incroyable malgré la distance physique qui nous a séparés. *Mike, it takes a*

good man to be true to a long distance relationship, but it takes a great one to encourage it. You we're with me 24 hours a day in spirit, I could feel it. Thank you for your patience, your respect, your courage, your love, your understanding, thank you for believing in me, in us.

Avant-propos

Le travail présenté dans cette thèse comprend six chapitres. Le premier est une introduction générale ; il est suivi d'un chapitre de revue de la littérature et de trois chapitres comtenant chacun un manuscrit qui a été soumis pour publication. L'ouvrage se termine par un chapitre de discussion et une conclusion générale.

Le premier chapitre décrit la problématique générale ainsi que les objectifs de chacun des manuscrits inclus dans la thèse. Le deuxième chapitre, que j'ai rédigé seule, présente une revue de la littérature reliée aux différents sujets abordés dans la thèse. Les manuscrits du corps de la thèse furent tous rédigés par moi-même et révisés par mon directeur de thèse, Pierre Legendre. J'ai par la suite effectué les corrections proposées.

Le premier manuscrit (chapitre 3) traite de la définition et de l'étude du biais statistique d'un coefficient de détermination ajusté pour l'arbre de régression multivariable (ARM). Nous insisterons sur le fait que ce coefficient témoigne du pouvoir explicatif du modèle et qu'il permet entre autres de comparer ce pouvoir à celui d'autres méthodes de modélisation multivariable comme l'analyse canonique de redondance (ACR) qui est très populaire en écologie.

Dans le deuxième manuscrit (chapitre 4), nous décrivons une procédure que nous appelons l'arbre de régression multivariable en cascade (ARMC) et qui consiste en l'utilisation dans un ordre hiérarchique prédéterminé de deux tableaux explicatifs dans une analyse ARM. Cette nouvelle démarche permet de forcer l'ordre dans lequel les deux tableaux explicatifs sont considérés. Il est ultimement possible d'en tirer des

informations supplémentaires sur les causes de la variation des variables réponse, par exemple des espèces, par rapport aux résultats de l'ARM habituel.

Pour le troisième manuscrit (chapitre 5) que j'ai rédigé avec plusieurs co-auteurs, ma contribution consiste en la mise en oeuvre des modèles linéaires généralisés (MLG) pour l'élaboration d'équations allométriques afin d'estimer la biomasse aérienne d'arbre tropicaux. Bien connue des modélisateurs statistiques, cette méthode semble méconnue des praticiens du domaine de la foresterie. Les données m'ont été fournies par mes co-auteurs qui les ont également récoltées. Plus précisément, Diana M. T. Sharpe, Benjamin Wadham-Gagnon et un étudiant Panaméen, Jose Luis Bonilla, ont mesuré et récolté les arbres, puis ils les ont pesés après les avoir débités. J'ai rédigé l'article que mes co-auteurs ont ensuite lu et commenté, puis j'ai effectué les corrections nécessaires.

Le chapitre 6, que j'ai rédigé, présente une discussion ainsi que la conclusion générale de la thèse. Y sont également présentées les deux bibliothèques de fonctions R que j'ai rédigées afin de réaliser les calculs et les simulations rapportés dans la thèse.

Chapitre 1

Introduction générale et objectifs de la thèse

Pour que les problèmes environnementaux deviennent des sujets de préoccupation publique, les écologistes s'intéressent de plus en plus à l'approche économique de l'environnement en insistant sur les services nécessaires et indispensables fournis par les écosystèmes (e.g. de Groot et al. 2002, Kremen & Ostfeld 2005). Comme l'ont fait Boyd & Banzhaf (2007) ainsi que Fisher et al. (2009), nous définissons un 'service écologique' comme un aspect ou une fonction d'un écosystème dont les humains tirent profit. Certains de ces services sont vitaux pour les populations humaines, comme la disponibilité de l'eau potable et le recyclage des déchets (de Groot et al. 2002). L'évaluation et l'étude des fonctions, biens et services fournis par les écosystèmes comporte plusieurs aspects (Kremen 2005) qui requièrent à un moment donné ou à un autre un processus de modélisation. Les modèles construits sont établis afin d'évaluer les conditions pour lesquelles les services rendus sont optimaux. Par exemple, la récolte étant un service de provision, l'évaluation de la performance d'une nouvelle pratique comme l'agriculture de conservation requiert un processus de modélisation bien spécial. Ce type d'agriculture se caractérise par un travail minimal du sol, une association et rotation culturale, puis une couverture permanente du terrain (e.g. Hobbs 2007, Knowler & Bradshaw 2007). Ainsi, on optimise le rendement en profitant d'un sol

qui a des propriétés physiques, biologiques et chimiques améliorées. En réalité, cette pratique est une manière plus efficace de cultiver en profitant des propriétés de l'écosystème naturel. Elle vise les récoltes à petites échelles, comme ce qu'on trouve dans les pays en voie de développement. Des méthodes de modélisation comme l'ANOVA peuvent être utilisées afin de comparer la performance de différents types de culture (e.g. Sommer, Wall et al. 2007). Les services de régulation nécessitent également un type de modélisation. Par exemple, les écosystèmes forestiers régulent le débit de l'eau dans le bassin versant: des caractéristiques comme l'interception de l'eau par la canopée, l'absorption dans la litière et la conservation de l'eau par le sol peuvent être responsables de cette régulation (liste non exhaustive), et des modèles peuvent être développés afin d'évaluer la capacité et les bénéfices de chacune de ces caractéristiques (e.g. Guo, Xiao et al. 2000). L'importance des débits a des retombées économiques importantes pour l'exploitation de l'hydroélectricité par exemple.

En modélisation statistique, les *variables réponse*, appelées aussi dépendantes ou à expliquer, sont les variables dont on veut expliquer la variation à l'aide de variables explicatives, appelées aussi indépendantes. Dans tout domaine d'application valorisant la modélisation en tant qu'outil d'évaluation de la relation entre deux tableaux, un choix s'impose au niveau du type de modélisation à employer : descriptif, explicatif ou prédictif ? Chacun de ces types peut être pertinent à un moment ou à un autre, et il différera selon les objectifs précis de l'étude. Les modèles descriptifs s'imposent lorsque le seul but est de résumer la structure de la réponse en fonction de variables (hypothétiquement) explicatives d'une manière évocatrice et condensée, en évoquant l'association entre la réponse et les variables explicatives et non un lien de cause à effet (lien de causalité) entre deux phénomènes mesurés (e.g.

Shmueli 2010). C'est avec les modèles explicatifs que l'on cherche précisément à évaluer des hypothèses de causalité.

Nous nous intéresserons dans cette thèse à l'utilisation d'un type particulier de modèle multivariable qui permet d'étudier le lien entre la composition en espèces aux sites et un ensemble de variables explicatives. L'étude des facteurs explicatifs de la variation spécifique entre les sites d'une région géographique donnée correspond à l'étude des déterminants de la diversité bêta (Whittaker 1960, 1972, Legendre et al. 2005). La diversité bêta est la variation de la composition spécifique dans une région géographique donnée ; c'est une mesure du taux de changement de la composition en espèces à travers l'espace géographique. L'approfondissement de nos connaissances concernant les conditions favorables ou défavorables aux espèces et leur application à des fin de conservation contribue au maintien de cette biodiversité (e.g. Angeler et al. 2008, Hodgson et al. 2009, Hodgson et al. 2011). Cette même diversité est à la base de plusieurs autres fonctions écosystémiques (de Groot et al. 2002, Balvanera et al. 2006) puisque la composition du biota et les conditions qui la maintiennent caractérisent l'écosystème qui fournit les services (e.g. Marrs et al. 2007). Ainsi dans le cadre de ce travail, nous traitons de l'identification des facteurs hypothétiquement causaux sous-jacents au patron de distribution des espèces. Les facteurs hypothétiquement causaux sont les facteurs à l'étude qui, selon une théorie ou une hypothèse préétablie, pourraient avoir un lien de cause à effet avec la variable réponse étudiée. Un des principaux défis de l'écologie des communautés est de reconnaître, parmi une multitude de facteurs, ceux qui structurent principalement les communautés étudiées (MacArthur 1972, Burnham & Anderson 2002, Guisan & Thuiller 2005). L'étude des causes hypothétiques sous-jacentes aux patrons de

distribution des espèces apparaissant dans le paysage contribue à l'approfondissement de notre compréhension des processus modelant la composition spécifique observée; l'étude des relations entre les communautés et leur environnement est une pratique courante pour rencontrer cet objectif (voir par exemple Legendre & Fortin 1989, Jackson & Harvey 1993, Diniz-Filho & Bini 1996, Rodriguez & Lewis 1997, Jenkins & Buikema 1998, Boyce & McDonald 1999, Jackson et al. 2001, Peres-Neto et al. 2006).

On peut relever dans la littérature plusieurs types mathématiques et statistiques de modélisation utilisés à ces fins (e.g. Guisan & Zimmermann 2000, Peres-Neto et al. 2006, Legendre & Legendre 1998). Au cours des dernières années, on remarque un intérêt grandissant pour la modélisation de tableaux de composition spécifique en fonction d'un tableau explicatif *via* l'arbre de régression multivariable (ARM). Il est maintenant nécessaire de pouvoir comparer son pouvoir explicatif à celui de l'analyse canonique de redondance (ACR) qui est une méthode de modélisation multivariable très répandue en écologie depuis le milieu des années 1980 (Birks et al. 1998, Peres-Neto et al. 2006). Nous traitons de cette problématique dans le 3^{e} chapitre de cette thèse en définissant un coefficient de détermination ajusté (R^2_{GDF}) pour l'ARM, ce qui permet de comparer directement le pouvoir explicatif des deux méthodes. Ce faisant, nous espérons pouvoir identifier des jeux de données pour lesquels la structure est mieux captée par l'ARM que par l'ACR, ce qui permet d'obtenir de l'information distincte concernant le patron de répartition géographique des espèces. En effet, l'ARM groupe des sites qui minimisent les variations intragroupes d'abondance des espèces en formant un arbre binaire dont les divisions correspondent à des points de coupure des variables explicatives (Segal 1992, De'ath

2002, Larsen & Speckman 2004). Cette méthode diffère fondamentalement de l'ACR qui modélise plutôt le tableau réponse en fonction de combinaisons linéaires des variables explicatives. D'un point de vue pratique, le modèle produit par l'ARM met en évidence les seuils des variables explicatives auxquels répondent le plus fortement des groupes de sites (objets) caractérisés par leur homogénéité en composition spécifique. Dans le cadre particulier de l'ARM, il n'y a aucune restriction imposée par la distribution statistique des espèces, la nature des variables explicatives et leur lien (qui peut être linéaire ou non) avec le tableau réponse.

Toujours dans une optique explicative, nous développons dans le 4^{e} chapitre un moyen de hiérarchiser l'utilisation de deux tableaux explicatifs dans le cadre d'une analyse ARM. Nous appelons cette nouvelle méthode l'ARM en cascade pour laquelle nous utiliserons le sigle ARMC (CMRT en anglais). Il n'est pas rare en écologie qu'un chercheur ait à traiter plusieurs tableaux explicatifs à la fois et qu'il cherche à établir quels sont leurs effets combinés ou isolés sur le tableau réponse (Legendre & Legendre 1998). En analyse écologique jusqu'à maintenant, l'ACR partielle a été utilisée pour partitionner la variation d'un tableau réponse en fonction de deux ou plusieurs tableaux explicatifs (Borcard et al. 1992, Peres-Neto et al. 2006). Cette forme d'analyse permet d'estimer la contribution unique de chaque tableau à l'explication de la variation de la réponse en considérant que leurs effets sont linéaires et additifs. L'ARMC se distingue de l'ACR partielle puisqu'en forçant l'ARM à considérer un des tableaux explicatifs comme le principal déterminant de la composition spécifique et à traiter le deuxième comme un effet subordonné, l'influence du tableau subordonné change en fonction des groupes produits par l'ARM du tableau principal. L'effet des deux tableaux est alors additif mais pas

linéaire. On obtient donc un modèle qui s'interprète comme une MANOVA hiérarchique. Ceci nous permet d'étudier les hypothèses de causalité dans l'ordre qui nous convient sans avoir à suivre l'ordre d'incorporation des variables choisi par l'ARM sur des bases statistiques. Vraisemblablement, puisqu'on impose un ordre prédéfini aux hypothèses, on peut extraire de ce nouveau modèle des informations différentes de celles fournies par l'ARM simple, en particulier en ce qui concerne l'interaction entre les deux tableaux explicatifs dans l'explication du tableau réponse. On peut par la suite identifier les espèces indicatrices qui peuvent caractériser les groupes de sites ainsi délimités.

Il existe d'autres types de services écosystémiques, entre autre les services de régulation. On retrouve dans cette catégorie l'étude des flux de carbone dont une des composantes est sa séquestration en vue de réguler les émissions d'origine anthropique. La déforestation et la dégradation des forêts augmentent considérablement ces émissions (Houghton 1999, Houghton & Hackler 2001, Brown 2002). Même s'il apparaît impossible de stopper ces pratiques, surtout dans les pays en voie de développement, nous espérons pouvoir tamponner une partie des émissions en instaurant des systèmes de crédit de carbone (Harmon 2001). L'objectif est de particulièrement encourager les propriétaires de terres à privilégier des conditions qui permettent d'établir des plantations et de favoriser la repousse tout en maximisant la recapture du carbone. Il est primordial pour l'instauration de tels programmes de pouvoir mesurer avec précision la quantité de carbone (par l'estimation de la biomasse) émis par les changements d'utilisation des terres et celle qui est reprise *via* les plantations et la croissance des arbres (Kraenzel 2003, Losi 2003, Pelletier et al. 2010). Au chapitre 5, nous utilisons les modèles linéaires

généralisés (MLG) dans le cadre d'une modélisation prédictive qui a pour but d'estimer la biomasse aérienne de six espèces d'arbres tropicaux fréquemment utilisées dans les plantations au Panama. Le développement de ces modèles vise ultimement l'estimation de la biomasse aérienne d'arbres qui ne font pas partie du jeu de données utilisé pour les construire. On s'insère donc dans un cadre prédictif. La modélisation prédictive concerne les études pour lesquelles l'objectif principal est la prédiction de la réponse de nouveaux objets non inclus dans le jeu de données d'apprentissage (celui utilisé pour calculer les paramètres du modèle) pour lesquels on a mesuré les valeurs des variables explicatives. Les variables explicatives dans ce cas sont des caractéristiques allométriques (diamètre du tronc, hauteur, etc.). Nous cherchons à utiliser les meilleures pratiques de modélisation, en particulier l'utilisation d'une structure d'erreur différente de l'erreur gaussienne (*via* les MLG) et la comparaison des modèles sur des bases prédictives (AIC, estimation de l'erreur de prédiction) et non explicatives (comparaison des R^2_a comme on le fait souvent). Nous visons à obtenir des modèles de biomasse d'arbres produisant les prédictions les plus précises que possible.

Chapitre 2

Revue et analyse de la littérature

Les écologues s'entendent pour dire qu'une grande partie des services fournis par les écosystèmes ont un rapport étroit avec la biodiversité (de Groot et al. 2002, Balvanera et al. 2006, Palumbi et al. 2009). Le terme *biodiversité* recouvre plusieurs types de diversité de nature biologique, entre autres la diversité organismique ou spécifique, génétique, fonctionnelle et phylogénique pour en nommer quelques-unes (Loreau 2010). On peut conceptuellement organiser l'ensemble de ces diversités (et d'autres) dans un diagramme à deux axes indépendants. Le premier axe décrit le niveau organisationnel de la diversité étudiée, qui part de la molécule, passe par l'organisme et la population pour se rendre jusqu'à l'écosystème. Le deuxième décrit la composante de la diversité étudiée, par exemple la richesse (le nombre de différent élément du niveau organisationnel choisit), l'équitabilité (l'arrangement compositionnel du niveau organisationnel choisit) et la disparité (différence entre les éléments du niveau organisationnel). Cette représentation révèle une multitude de combinaisons de niveaux organisationnels et de composantes possibles. Elles ne sont pas toutes nécessairement étudiées dans la littérature. L'une d'elles, la diversité bêta qui est la diversité, ou la variation, de la composition spécifique entre les sites dans une région géographique donnée (Whittaker 1960, 1972, Legendre et al. 2005), est l'objet d'un grand nombre d'études. Il est primordial de pouvoir reconnaître les

facteurs clefs qui expliquent cette variation pour la maintenir ou la restaurer afin de préserver ou de rétablir les services écosystémiques qu'elle fournit (Cherwin et al. 2009, McGovern et al. 2011). La modélisation explicative entre la composition spécifique et des variables hypothétiquement explicatives est donc de mise afin d'identifier les facteurs responsables de ces patrons de distribution. Dans la littérature écologique, on fait abondamment usage de l'analyse canonique de redondance (ACR, Rao 1964) afin d'étudier le lien entre la diversité bêta représentée par le tableau de composition spécifique et les variables explicatives choisies (e.g. Birks, Peglar, & Austin 1996, Bojsen & Jacobsen 2003, Chust, Chave et al 2006, Legendre et al. 2005, Peres-Neto et al. 2006, Legendre 2008, Urban, Skelly et al 2006). Cette méthode d'ordination est une extension des modèles de régression linéaire multiple à un tableau réponse multivariable, ce qui permet d'utiliser la composition spécifique des sites comme réponse de l'analyse au lieu de se limiter à une seule espèce, à la richesse spécifique ou à la somme des individus de toutes les espèces observées aux différents sites. Le tableau réponse doit être linéairement relié aux variables explicatives pour être convenablement expliqué par l'ACR. Legendre et al. (2005) ont lancé la discussion en montrant qu'il est possible d'étudier de multiples hypothèses de processus pouvant être à l'origine de la diversité bêta en traitant plusieurs tableaux explicatifs par des ACR partielles. Ce constat permet entre autres d'étudier et de soupeser le pouvoir explicatif de variables représentant la variabilité spatiale et environnementale, puis ultimement de partitionner de manière exclusive et commune la variation expliquée par ces deux tableaux (partitionnement de la variation : Borcard et al. 1992, Borcard & Legendre 1994). Ainsi, on peut décrire l'importance des patrons spatiaux et du contrôle environnemental sur la variation de

la composition spécifique, puis tester leur signification. Peres-Neto et al. (2006) renchérissent en discutant le problème de contraction des fractions de la variation estimées par les modèles d'ACR partielle. Ces fractions sont assujetties aux mêmes problèmes que le coefficient de détermination R^2 : les pourcentages de variation expliquée estimés sont plus grands pour un échantillon que ceux de la population associée et donc un ajustement (une réduction) s'impose, en tenant compte du nombre d'objets et du nombre de variables explicatives utilisées. Les auteurs parviennent à définir cet ajustement et à comparer les fractions estimées par différents modèles d'ACR. Legendre (2008) a montré l'utilité de ces ajustements dans un cadre d'analyse de la diversité bêta. Depuis, plusieurs applications de ces principes ont été publiées (e.g. Laliberté et al. 2009, Legendre et al. 2009).

En parallèle à l'ACR, Segal (1992), De'ath (2002) et Larsen & Speckman (2004) ont développé l'arbre de régression multivariable, une généralisation multivariable des modèles de *Classification and Regression Tree analysis* (CART, Breiman et al. 1984) dans un cadre écologique avec des promesses de pouvoir prédictif. Comme l'ACR, cette méthode de modélisation est asymétrique. Elle est basée sur un algorithme de bipartition récursif : les objets sont divisés en deux groupes un grand nombre de fois jusqu'à ce qu'on obtienne un grand arbre. Les divisions sont choisies de manière à ce que la composition des sites soit la plus homogène possible tout en étant fonction de points de coupure dans les valeurs des variables explicatives. Finalement, un algorithme de rééchantillonnage, la validation croisée, est utilisé pour l'élaguer et obtenir l'arbre qui donne les meilleures prédictions en minimisant l'erreur attendue sur la prédiction d'une nouvelle observation.

Plus précisément, la validation croisée à *v*-recouvrements est un processus de ré-échantillonnage qui fait partie intégrante du processus de construction de l'arbre. La première étape de la validation croisée débute avant même la première bipartition. La première étape consiste à diviser l'ensemble des objets du nœud racine en *v* groupes les plus égaux possibles. Chacun de ces groupes, un à la fois, est retiré de l'ensemble et un arbre est construit à partir des objets restants. On nomme les groupes d'objets retirés les groupes tests, et les objets restants, les groupes d'apprentissage. Chaque groupe test est associé à un groupe d'apprentissage. Une fois les *v* arbres construits, on utilise les objets des différents groupes tests pour calculer ce que l'on appelle l'erreur de la validation croisée relative (CVRE), qui est le résultat de la division de la dispersion autour des prédictions par la dispersion totale de la réponse. Cette erreur relative de validation croisée peut varier de 0, pour un très bon modèle prévisionnel, à des valeurs près de 1 pour un modèle ayant un faible potentiel prévisionnel. On calcule cette valeur pour toutes les tailles d'arbres. La taille d'arbre optimale choisie est celle qui procure la plus petite erreur de validation croisée relative.

Depuis ces premières publications, on trouve dans la littérature écologique un nombre impressionnant d'applications de l'ARM, mais pas toutes dans un cadre prédictif (par exemple Work et al. 2004, Koivula & Vermeulen 2005, Claudet et al. 2006). L'arbre de régression multivariable ne présente pas en général de mesure de variation expliquée, et donc il est très difficile d'évaluer si ce modèle explique bien le phénomène étudié. Par contre, l'arbre de régression multivariable est intéressant même pour une modélisation explicative puisqu'il ne présente aucune supposition de relation linéaire entre le tableau réponse et explicatif comme en ACR (De'ath 2002,

Larsen & Speckman 2004). De surcroît, parce que la relation entre la réponse et les variables explicatives a la forme d'un arbre, celui-ci se décompose en un ensemble de règles d'interprétation simples qui, parce qu'elles sont binaires, sont faciles à utiliser dans un contexte de gestion des ressources. Par exemple, on peut identifier une composition spécifique particulière qui répond à certains seuils des variables explicatives. Notons par contre que l'étude de plus d'un tableau explicatif n'existait pas encore en analyse ARM, ce qui fait que la partition de la variation expliquée n'a pu être réalisée jusqu'ici entre deux tableaux explicatifs en ARM.

Dans le cadre explicatif, on trouve de nombreuses applications écologiques auxquelles l'arbre de régression multivariable convient parfaitement. Par exemple, on remarque un intérêt pour ce type de méthode lors de la modélisation de l'habitat des communautés en synécologie (écologie des communautés). Le raisonnement est issu de la théorie de la niche qui a été révisée par Pulliam (2000), les auteurs principaux à l'origine de cette théorie étant Grinnell (1917), Elton (1927) et Hutchinson (1957). On part du principe qu'il existe des relations explicatives entre l'occurrence des espèces et certains facteurs de leur environnement. L'association entre les espèces et leur habitat est une conséquence de l'interaction entre les stratégies d'histoire de vie des espèces et le filtrage de l'habitat. En effet, on soutient que la structure et la dynamique de l'habitat physique sont le cadre dans lequel les communautés sont organisées (« habitat templet theory » : Southwood 1977, 1988 ; Townsend & Hildrew 1994). L'abondance et la distribution des espèces sont influencées par plusieurs facteurs qui peuvent être issus de l'habitat lui-même ou de la présence des autres espèces (Schlosser 1982, Brown 1984, Moyle & Vondracek 1985, Taylor et al. 1993), alors que l'influence de chacun des facteurs peut varier selon l'échelle à

laquelle on l'observe ou le mesure (Wiens et al. 1986, Wiens et al. 1987, Menge & Olson 1990, Hanski 1991, Legendre & Legendre 1998, Jackson et al. 2001). On s'attend alors à ce que des changements de composition spécifiques soient observés lorsque l'habitat se dégrade. Pour le chercheur qui désire identifier différents types de composition spécifique homogènes répondant aux caractéristiques de l'habitat, une méthode comme l'ARM produit directement une partition des sites en groupes ayant des compositions spécifiques les plus homogènes possible, répondant à des seuils des variables de l'habitat. Après avoir obtenu les groupes, le chercheur peut identifier les espèces indicatrices de ces groupes à l'aide de méthodes comme IndVal (Dufrêne & Legendre 1997, De Càceres et al. 2010).

Dans un autre ordre d'idée, d'autres services écosystémiques concernent la régulation, par exemple la régulation du flux d'éléments importants comme le carbone. L'estimation de ces flux a beaucoup d'importance dans la lutte contre les changements climatiques. Les forêts jouent un rôle important dans la séquestration du carbone: elles accumulent une bonne partie du carbone terrestre et ont la capacité de le conserver à long terme (Shvidenko et al. 2005). Elles peuvent donc jouer le rôle de source ou de puits de carbone, dépendant de leur utilisation. On attribue à la déforestation dans les tropiques un des plus grands impacts sur le cycle du carbone, par comparaison aux autres changements d'utilisation des terres (Shvidenko et al. 2005). Pour contrer cet effet, on propose la reforestation ou le boisement. Des systèmes de crédit ont été et sont en voie d'être instaurés afin d'encourager les propriétaires de terres à s'engager dans des pratiques de reforestation ou de boisement (Harmon 2001). Ces systèmes nécessitent l'estimation de la biomasse fixée dans les arbres. Il est, dans ce cadre, important de pouvoir estimer le plus précisément possible

la biomasse sans avoir recours à des processus destructifs. On se sert de modèles allométriques basés sur des mesures non destructives pour l'arbre (mesure de la taille du tronc, de la hauteur, etc.) et on utilise ces mesures dans une modélisation prédictive plutôt que explicative. Comme rapporté dans le chapitre 1 de cette thèse, les modèles sont construits afin d'estimer la biomasse aérienne d'arbres qui ne sont pas inclus dans le jeu de données utilisé pour la construction des modèles.

La plupart des publications sur ce type de modélisation utilisent des statistiques explicatives comparant les modèles linéaires, *via* le R^2 par exemple. Les premiers modèles allométriques étaient en effet comparés en utilisant des R^2 (e.g. Overman et al. 1994, Ketterings et al. 2001, Brown 2002, Losi et al. 2003, Wang 2006), ce qui est une pratique courante mais peu efficace pour identifier les modèles qui produisent les meilleures prédictions. Dans le cadre de cette thèse, nous avons plutôt recours à une approche prédictive. La comparaison de modèles allométriques *via* les AIC (e.g. Henry et al. 2010, Chave et al. 2005) est un exemple de l'approche statistique prédictive.

Les auteurs cités ci-dessus se limitent tous au modèle linéaire ou au modèle linéaire général (Kim & Timm 2007) pour estimer les paramètres des équations. Dans le cadre de cette thèse, nous utiliserons plutôt les modèles linéaires généralisés (MLG, McCulloch 2000) qui sont plus flexibles que la régression linéaire par moindres carrés ordinaires. Cette flexibilité vient de l'utilisation d'une fonction lien qui permet d'optimiser la variable réponse dans l'espace d'origine et de la possibilité d'utiliser une structure d'erreur différente de celle gaussienne.

Afin de mieux comprendre la différence entre l'utilisation d'une fonction lien et d'une transformation, il importe de présenter les modèles sous forme d'équation

d'espérance de la réponse dans les deux cas. Voici l'exemple d'une transformation logarithme (ln) et de la fonction lien correspondante.

Transformation		***Fonction lien***	
$E(\ln(y)) = a + b\ln(x)$	éq. 1	$\ln(E(y)) = a + b\ln(x)$	éq. 4
$\ln(y) = a + b\ln(x) + \varepsilon$	éq. 2	$E(y) = e^{a+b\ln(x)}$	éq. 5
$y = e^{a}\, e^{b\ln(x)}\, e^{\varepsilon}$	éq. 3	$y = e^{a}\, e^{b\ln(x)} + \varepsilon$	éq. 6

En procédant à une transformation de la réponse y au préalable, c'est-à-dire en modélisation directement la variable aléatoire ln(y) comme dans la colonne « Transformation » du tableau ci-dessus, on cherche à caractériser la distribution de cette variable aléatoire ln(y). On suppose dans un modèle linéaire que ln(y) est fonction de ln(x) et que cette fonction s'exprime selon l'équation d'une droite $a + b\ln(x)$. L'espérance de ln(y) est donc égale à $a + b\ln(x)$ (éq. 1), ce qui veut dire que ln(y) est égal à $a + b\ln(x)$ plus une erreur aléatoire ε d'une distribution connue (éq. 2). En élevant chacun des membres de l'équation 2 à l'exponentielle, on obtient que $y = e^{a}\, e^{b\ln(x)}\, e^{\varepsilon}$ (éq. 3). On remarquera tout de suite que l'erreur n'est plus additive. Dans le cadre d'une fonction lien (éqs 4 à 6), on cherche à modéliser directement la variable aléatoire y. On peut supposer dans un cadre de modèle linéaire que y est fonction de ln(x) (éq. 4) et que cette fonction s'exprime selon l'équation $e^{a+b\ln(x)}$. L'espérance de y est donc égale à $e^{a+b\ln(x)}$ (éq. 5), ce qui veut dire que y est égale à $e^{a}\, e^{b\ln(x)}$ plus une erreur aléatoire ε d'une distribution connue (éq. 6). Ici l'erreur demeure additive au reste du modèle.

Par soucis de comparaison, nous construisons les modèles avec un lien et avec une transformation. Pour comparer le pouvoir prédictif de modèles dans lesquels nous avons fait différentes transformations de la réponse, nous utiliserons comme statistique la moyenne des erreurs prédictives mises au carré (MEPC ou MSPE en anglais, e.g. Efron & Tibshirani 1997).

Chapitre 3

An adjusted R^2 statistic for multivariate regression tree analysis

Ce chapitre a été soumis pour publication dans une revue internationale : Methods in Ecology and Evolution.

Marie-Hélène Ouellette[*] and Pierre Legendre

Département de sciences biologiques, Université de Montréal, C.P. 6128, Succursale Centre-Ville, Montréal, Qc H3C 3J7, Canada.

[*]Correspondence author.

SUMMARY

1. Multivariate regression tree analysis (MRT) has grown to considerable importance on the ecological statistical modelling scene. Proper means of comparing its explanatory power with the widely used canonical redundancy analysis (RDA) is imperative. MRT is an asymmetric two-matrix analysis based on binary recursive partitioning, the final model chosen by minimization of the cross-validation relative error (*CVRE*). In contrast, RDA is a multivariate generalization of multiple linear regression; its fit is assessed by the adjusted coefficient of determination (R_a^2). These two performance measures are not directly comparable.

2. In this paper, we define an R_a^2 statistic for MRT analysis. We show that R^2_{MRT} is the relative error minus 1 in the least-squares case. Seeking proper adjustment, we compare by Monte Carlo simulations two different ways of estimating the number of degrees of freedom (*df*) used in Ezekiel's R_a^2 formula: the number of leaves (groups) and a generalized *df* estimation algorithm (GDF) that involves bootstrapping.

3. Results of the Monte Carlo study show that for fitted sample trees and all sample sizes, the ρ^2 estimate based on *df* computed by the GDF algorithm is less biased. The resulting unbiased estimate of R_a^2 for MRT analysis is called $R^2_{MRT(GDF)}$. Two illustrative examples are presented.

4. The $R^2_{MRT(GDF)}$ statistic provides an unbiased estimate of the percentage of variation of the response explained by explanatory variables and a mean of comparing its explanatory power to that of RDA, which is widely used to study beta diversity.

KEY-WORDS: degrees of freedom, explanatory power, Mutlivariate regression tree (MRT), species composition drivers

INTRODUCTION

Species are distributed in landscapes where they form patterns that are driven by different causal variables. The assessment of the relationships between species distributions and a set of explanatory variables potentially driving these patterns is an essential component of ecological studies (see for example Legendre & Fortin 1989, Jackson & Harvey 1993, Diniz-Filho & Bini 1996, Boyce & McDonald 1999, Jackson et al. 2001, Peres-Neto et al. 2006, Legendre & Legendre 2012). Different modelling methods are required to perceive different types of patterns. For instance, multivariate regression tree analysis (MRT, De'ath 2002, Larsen & Speckman 2004)

is ideal to identify thresholds of the explanatory variables that explain the most drastic changes occurring in species composition. MRT is a recursive partitioning algorithm that splits objects (e.g. sampling sites) into homogenous groups according to the response, with the splits constrained by explanatory variables. The tree is grown by splitting the data a large number of times, then it is subsequently pruned (reduction of the number of groups) via a resampling method called *v*-fold cross-validation (Breiman et al. 1984) to obtain the best predictive tree size. MRT is an asymmetric method of multivariate analysis, meaning that there is a response data table **Y** and an explanatory table **X** – for example species abundances and explanatory environmental variables. In the general case considered in this paper, both **Y** and **X** are multivariate although they can be univariate in particular cases. For logical reasons, in explanatory modelling, these two matrices are not interchangeable because they play different roles in the analysis. MRT has gained broad popularity within the ecologist community since its first publications (Segal 1992, De'ath 2002, Larsen & Speckman 2004). For instance we find applications in the fields of microbial ecology (Auguet et al. 2010), paleolimnology (Davidson et al. 2010) and forest ecology (Chen et al. 2010), to name only a few.

Comparing the explanatory power of a model that has an underlying hypothesis of linear relationships like RDA, which brings out linear relationships between the response and the gradients, to a method modelling abrupt changes like MRT, which is capable of detecting thresholds where abrupt changes in community composition occur, has valuable ecological implications. For data sets where the latter is chosen as the best descriptive model, it suggests that species form crisper (in the clustering sense) rather than continuously changing assemblages along the

ecological gradients. If the sampling design was irregular along the gradient, the discontinuity could be sampling-induced and may suggest a different sampling strategy. If not, this could incite the search for refined rules of management regulation that could lead to policy making (Huggett 2005, Sonderegger et al. 2009) and could be very efficient if used carefully (Huggett 2005). This could also provide novel insight on the problem under study. For habitat studies, abrupt changes in assemblages could mean strong specificity in species, environmental physical barriers, other dispersal limitations or strong anthropogenic perturbation. In time or space, abrupt changes in assemblages could point at strong successive sharp drivers at the time of the shifts (strong anthropogenic perturbation for example) compared to smoother transitions. Indicator species analysis (Dufrêne & Legendre 1997, De Càceres et al. 2009) and discriminant species (species that contribute the most to the explained variation of the response in the MRT) are statistical tools available to identify species that strongly respond to these phenomena.

Even if the interest in modelling species distributions has moved towards prediction (Guisan & Thuiller 2005, Elith & Leathwick 2009), explanatory assessment for ecological understanding is still a vital aspect of current research (Peres-Neto et al. 2006, Elith & Leathwick 2009). Explanatory modelling aims at providing clues about the causes of the patterns exhibited. Even though MRT analysis is fundamentally a predictive statistical method since it uses holdout data (v-fold cross-validation) within its computing procedure, its explanatory power can also be of interest for the purpose of studying causality.

The coefficient of determination (R^2) and its adjusted form (R^2_a) are widely used by ecologists in the canonical analysis framework (Peres-Neto et al. 2006) as

estimates of the percentage of variation of the response explained by the model. It is common knowledge that R^2 is a biased estimate: it tends to be larger than the population value ρ^2. This is why an adjustment must be performed on this estimate, noted R^2_a, by using appropriate degrees of freedom (*df*). Canonical analysis is very popular amongst ecologist, and the growing interest in MRT raises the need to compare the results of those two analyses. Canonical analysis bases its optimization on minimizing the sum of squares of the residuals of a linear model, and the tests of statistical significance are based on the coefficient of determination. Accordingly, we seek to define an R^2 and an R^2_a statistics for MRT analysis. The latter will require a sound definition of the degrees of freedom (*df*) of an MRT model.

DEFINITIONS AND PROOFS

Cross-validation provides an excellent mean of pruning a multivariate regression tree (Breiman et al. 1984) and choosing the best predictive model, so R^2 and R^2_a are not needed for that purpose. More precisely, MRT analysis is often implemented with *v*-fold cross-validation as a pruning procedure. This resampling method starts before the first bipartition, by splitting all objects into *v* test subsets. A tree is then build with each learning set obtained by removing one of the test subsets from the whole set; *v* trees are thus obtained. For each tree size, we calculate the cross-validation relative error:

$$CVRE = \frac{\sum_{k=1}^{10}\sum_{i=1}^{n_i}\sum_{j=1}^{m}\left(y_{ij(k)} - \hat{y}_{j(k)}\right)^2}{\sum_{i=1}^{N}\sum_{j=1}^{m}\left(y_{ij} - \bar{y}_j\right)^2} \quad \text{(eq. 1)}$$

where $y_{ij(k)}$ is an observation of the test set *k*, $\hat{y}_{j(k)}$ is the predicted value for this observation in tree *k* computed from the corresponding learning set, n_k is the number of observations in test set *k*, and *m* is the number of variables in response matrix **Y**.

In reality, comparison of the explanatory power of canonical analysis and MRT is a problem met by ecologists who are using both methods; it is not necessarily a general concern. In this section we sought to properly define an R^2, along with *df*, for an MRT model and also an R^2_a for the least-squares based MRT analysis.

COEFFICIENT OF DETERMINATION

The coefficient of determination (R^2) represents the proportion of variation of the response variable(s) explained by a model; it is a commonly used measure of explanatory power in linear modelling. In fact R^2 implies the comparison between two models: one is the full model, the one for which we wish to assess the enhancement over the reduced model, the latter referred to as the 'no relationship' model. We can formulate the following general R^2 (GR^2) definition (Anderson-Sprecher 1994):

$$GR^2 = 1 - \frac{RSS(full)}{RSS(reduced)} \text{ (eq. 2)}$$

RSS (*reduced*) refers to the residual sum of squares of the reduced model: as stated above, this model is the one with no relationship. In linear regression with intercept, this model would reduce to the model with intercept only (all slopes equal to 0). Anderson-Sprecher (1994) argues that the R^2 definition can also be applied to non-linear regression, of which MRT is a special case, as long as it is based on least squares and there is an intuitive reduced model nested in the full, ensuring that this general equation is readily interpretable. In the MRT analysis framework, we can

examine the improvement of the sum of squared error of the model computed over the no-splits model (*reduced*), which is the total sum of squares of the response matrix, and define an R^2_{MRT} as

$$R^2_{MRT} = 1 - \frac{RSS(full)}{RSS(reduced)} = 1 - \frac{\sum_{i=1}^{n}\sum_{j=1}^{m}\left(y_{ij(g)} - \bar{y}_{j(g)}\right)^2}{\sum_{i=1}^{n}\sum_{j=1}^{m}\left(y_{ij} - \bar{y}_{j}\right)^2} \quad \text{(eq. 3)}$$

where n in the number of objects (sites), m is the number of response variables (species), y_{ij} is an observation of the $\mathbf{Y}_{i\times j}$ response matrix, $\bar{y}_j$ is the mean of the response j and finally (g) designates membership to group g. The numerator of eq. 3 is the sum of the within-group sums of squares.

We can draw the same conclusions from the regression tree models theory (Breiman et al. 1984) extended to the multivariate case. Performance measures of these models are based on estimating the true mean squared error $R^*(T)$ of a particular regression tree T. One of these estimates is the well-known deviance of T also called the *resubstitution estimate*, noted $R(T)$, which is in fact equal to $RSS(full)/n$. It is generally considered to be an explanatory measure of performance because the data set used to compute the model is the same as the one used to calculate its error. This is one of the reasons why it systematically gives an over-optimistic estimate of the risk of the model over the population (De'ath 2002).

Another estimate can be obtained via resampling by v-fold cross-validation. This estimate is obtained by dividing the whole data set into V subsets (we use V = 10 here) each containing as much as possible the same number of objects. For each of the V subsets, a model is built using all objects except those that pertain to the subset

under consideration (this new set is called the learning set), and each model with its corresponding test set is used to calculate an estimate of the mean squared error. We will note this estimate $R(T)^{cv}$:

$$R(T)^{CV} = \frac{1}{n}\sum_{v=1}^{10}\sum_{i=1}^{n_i}\sum_{j=1}^{m}\left(y_{ij(v)} - \hat{y}_{j(v)}\right)^2 \text{ (eq. 4)}$$

This performance measure is the basis for a predictive approach as the intersection of the learning set used to compute the model and the test set used to estimate the error is the empty set: they have no objects in common.

This being said, the mean squared error (MSE) estimated by one or the other means stated above does depend on the response's scale thus it is useful to normalize these performance measures by dividing them by the sum of squares around the mean, which is the total variation of the response, divided by n. In general we call this value the relative mean squared error (RE^*):

$$RE^*(T) = R^*(T) \Big/ R^*(\mu) \text{ (eq. 5)}$$

When we use $R(T)$ as an estimate for $R^*(T)$, the estimate is called the relative error:

$$RE = \frac{\sum_{i=1}^{N}\sum_{j=1}^{m}\left(y_{ij(g)} - \hat{y}_{j(g)}\right)^2}{\sum_{i=1}^{N}\sum_{j=1}^{m}\left(y_{ij} - \bar{y}_j\right)^2} \text{ (eq. 6)},$$

and when we use $R(T)^{cv}$, it is called the cross-validation relative error, noted *CVRE* (see eq. 1).

In the light of the previous definition of R^2_{MRT}, we find that:

$$RE = \frac{R(T)}{R(\mu)} = 1 - R^2_{MRT} \text{ (eq. 7),}$$

thus the coefficient of determination for the final MRT is 1 minus the relative error of the tree, a value provided by most software.

Breiman et al. (1984) had originally stated that in general the mean squared error "is not a variance and it does not make sense to refer to (1 – relative error) as the proportion of variance explained. Neither is the relative error equal to the square of the sample correlation between the response and the predicted values." Yet, if we dummy-code the partition of the model and use this in an RDA as the matrix of explanatory variables with the response as is, the $R^2_{Y|X}$ of the RDA will be exactly R^2_{MRT} (demonstration in Appendix 1). The resulting $R^2_{Y|X}$ (Peres-Neto et al. 2006) is a weighted mean of the R^2 of individual models computed on each variable $\mathbf{y}_j$, j=1,…,m as a function of the dummy-coded partition with weights proportional to the species variances divided by the total variance.

DEGREES OF FREEDOM

Now that a proper R^2_{MRT} has been defined, we need to be able to calculate the number of degrees of freedom (*df*) for an MRT model in order to compute an adjusted R^2_{MRT} (Anderson-Sprecher 1994):

$$R^2_{MRT} = 1 - \frac{MS(full)}{MS(reduced)} = 1 - \frac{RSS(full)/df(full)}{RSS(reduced)/df(reduced)} \text{ (eq. 8)}$$

which is equivalent to Ezekiel's (1930) formulation of R_a^2 (correction for shrinkage) It would be a mistake to directly use the adjustment of the R^2 given by the RDA analysis for our purpose because the *df* for the RDA and the MRT analysis are not equivalent, as we explain in this section. For linear models, we define the *df* of a

model as the number of parameters estimated, p, because those two quantities are equal in this case. In linear models, the orthogonal projection matrix (eq. 9), also called the **H** or *hat* matrix, is defined as follows:

$$\hat{\mathbf{Y}} = \mathbf{HY} \text{ where } \mathbf{H} = \mathbf{H}\left[\mathbf{X}'\mathbf{X}\right]\mathbf{X}' \text{ (eq. 9)}$$

$$0 \leq h_{ij} \leq 1 \text{ (eq. 10)}$$

$$\sum_{i=1}^{n} h_{ii} = p \text{ (eq. 11)}$$

The **H** matrix, whose order is (n x n), contains diagonal values h_{ii} that estimate the influence (or weight) of observation i in determining the fitted value i. In linear models, it can be shown that the number of parameters (rank of $\boldsymbol{X}$) and the trace of the **H** matrix are equal (Neter et al. 1996). It can also be shown that it is the geometry, thus the linearity of the model that confers this equality: when we estimate the slopes and ordinates in OLS regression, the following has to hold:

$$\sum_{i=1}^{n} e_i = 0 \text{ (eq. 12)}$$

$$\sum_{i=1}^{n} x_i e_i = 0, \ \forall x \in X \text{ (eq. 13)}$$

where e_i is the residual associated to object i and x_i is the i^{th} explanatory variable in the **X** matrix. The cost in *df* for fitting this model, thus estimate the parameters, is p: the model is constrained to lie in a space of dimension p described by eqs. 12 and 13, with the additional constraint that $p \leq (n–1)$. Even if p is equal to the dimension of the constraining space in linear modelling, it is not necessarily the case for all modelling procedures, especially for complex statistical procedures requiring minimum assumptions. They generally convey a complex underlying model for which the

geometry is complicated, thus the degrees of freedom are not directly calculable as in linear regression (Ye 1998).

Considering the need to define a general procedure for obtaining *df*, Ye (1998) defined the concept of generalized *df* (GDF). The basis of this definition is the interpretation we can make of the diagonal elements of **H**: it is the number of values in the calculation of a statistic that can vary freely without violating any given restrictions (Eisenhauer 2008). In modelling, the limiting condition is the fitted model (fitted values), and the control of observation *i* on its corresponding fitted value is the portion of this observation that cannot vary in order to respect that rule (Walker 1940). The larger h_{ii} is, the larger influence this observation has on the shape of the model.

The calculation depends on the modelling procedure and the underlying true model and is defined by the author as "the sum of the sensitivities of each fitted value to perturbations in the corresponding observed value" (Ye 1998). For modelling procedures that do not include a **H** matrix estimation, the influence of each observation on the final model can be calculated by perturbing randomly (adding random noise) each observation, computing the model, and relating the fitted values to the perturbations. The stronger the relationship between the fitted values and the perturbations, the stronger the influence of the observation is on the fitted values.

Following Ye's (1998) proposal for classification and regression tree (CART) models, we use an algorithm to estimate *df* for a particular MRT model (algorithm 1 reported in Ye's paper) to ultimately estimate the value of *MS(full)*. The algorithm is Monte Carlo based and is defined as follows. Let **Y** be the response matrix, **X** the explanatory matrix, and $\Delta\mathbf{T}_{nxm}$ a matrix of standard normal deviates with columns

multiplied by $\tau \times s(\mathbf{Y}_j)$ where $s(\mathbf{Y}_j)$ is the standard deviation of the j^{th} column. If $\tau = 1$, $\Delta\mathbf{T}_{nxm}$ is a matrix of normal deviates with the j^{th} column's standard deviation equal to $s(\mathbf{Y}_j)$ and mean equal to 0. The multiplicative constant τ is a tuner value: as it gets larger, the perturbation of the original data gets larger. Box 1 contains a summary of the procedure in code script format.

Box 1: Summary of Ye's algorithm 1 adapted to MRT analysis.

Repeat ***k=1,…, K*** **times :**

{

- *Generate* ΔT_{nxm}. In each of K successive runs, store the first column of ΔT_{nxm} in ΔT_k, $k = 1,\ldots, \ldots K$.
- *Evaluate (get fitted values) of* $Y + \Delta T_{nxm} \sim X$ *using MRT analysis.* Store the first column of the resulting matrix of fitted values in $\hat{Y}_k$.

}

The resulting matrices ΔT_k and $\hat{Y}_k$ are of order $n \times K$.

For all $i=1,\ldots,n$ rows, where i designates the rows of and ΔT_k, *we compute the model* $\hat{Y}_k \sim h_i \Delta T_{ik} + b_i$.

Sum all values h_i to obtain the GFD estimate.

In an MRT model, the fitted value of observation *i* is the multivariate centroid of all observations that are placed by the MRT procedure in the same group (leaf) as *i*.

The same result (number of GDF) would be obtained for any response variable (all columns of ΔT_{nxm} and the fitted values matrix); calculation for all variables of **Y** is thus futile.

According to the simulation results in Ye (1998), the error variance (*RSS*(*full*)/*df*(*full*) or *MSE*) can be estimated without bias using the GDF calculation. By simulation (n = 100, σ^2 = 0.25), he showed that for a constant number of groups (19), the *MSE* estimate of a fitted CART model with GDF calculation of *df* was practically unbiased (equal to the simulated 0.25) contrary to the *MSE* with *df* calculated as the number of nodes (or splits) of the tree. According to these simulations, a tree with 19 nodes (this value was chosen randomly as an example of an overfitted tree and the simulated number of groups was 5) can have up to an estimated 79 *df* (calculated with Ye's algorithm number 1). For fitted or larger trees, the MSE estimate remained unbiased.

BIAS ASSESSMENT WITH SIMULATED ECOLOGICAL DATA

We carried out simulations to assess the behaviour and effectiveness (bias) of two adjustments for R^2_{MRT} based on Ezekiel's formulation of R^2_a (Ezekiel 1930). The adjustments differ in the calculation of *df*. It is either (1) the number of nodes of the model (adjustment noted $R^2_{MRT(p)}$) or (2) the estimate provided by Ye's (1998) algorithm (adjusted value noted $R^2_{MRT(GDF)}$). By comparing these methods for computing the *df* of the tree, we will determine if they are equivalent; if they are not, we will find which one provides the best adjustment for the R^2_{MRT}. We assess the bias of these adjustments by Monte Carlo simulations (100 runs), using a procedure similar to those of Kromrey & Hines (1995) and Peres-Neto et al. (2006). From populations of size 10 000 (see Appendix 2 for full description), samples of different sizes (20, 50, 100) are drawn at random with replacement, and the R^2_{MRT}, $R^2_{MRT(p)}$ and $R^2_{MRT(GDF)}$ statistics are estimated from those samples. For $R^2_{MRT(GDF)}$, the computation

is performed using the `R2aMRT` function found in the MVPARTWRAP package soon available on the R-Forge site, with K = 1000 runs, which confers to the $R^2_{MRT(GDF)}$ estimate a variance similar to the R^2_{MRT} and $R^2_{MRT(p)}$; the response species data were Hellinger transformed (Legendre & Gallagher 2001). The Hellinger transformation takes the square root of the profiles of relative species abundances, $y'_{ij} = y_{ij}/y_{i+}$, where y_{ij} is the abundance of species j at site i and y_{i+} is the total number of individuals at site i. This transformation is appropriate before analysing frequency data in linear models.

Contrary to Ye's simulations, ours were carried out using realistic species abundance data and for sample sizes under or equal to 100 to better represent the field's most common data features. In total, 2 statistical populations were simulated.

The first population was structured by 4 gradients, 3 of which had 3 species linearly associated with them (with regression coefficients of 1, 7, and 0.5 for the three species respectively) whereas the fourth gradient had 3 random species. The ρ^2_{MRT} of the MRT model (4 leaves) computed for the Hellinger-transformed response data was 24.50%; the RDA model provided an ρ^2 of 25.89% explained variation.

The second population was simulated with 5 guilds of 3 species and no linear gradient, providing a ρ^2_{MRT} of 60.61% with 6 leaves, and an RDA ρ^2 of 9.83%.

The large populations (10 000 objects) were generated by the SIMSSDR function of the RSIMSSDCOMPAS R package, which allows the simulation of deterministic environment and species composition data, related linearly or in a niche manner. This package will soon be available on R-Forge (http://r-forge.r-project.org/), an online platform on which one can make available to other users R

packages that are under development. For species data generated under a niche model, asymmetric physiological responses and other types of restrictions and simulation options are available but are not described here. All gradients were simulated on a square 100×100 grid and consisted of a diagonal gradient coming from top to bottom and from left to right, to which we added standard normal deviates $N(0,1)$. This means that all four gradients in population 1 we're very similar.

In a first Monte Carlo study, the number of leaves was fixed for the computed trees, with underfitted (lower number of leaves than the population), fitted (same number of leaves as the population) and overfitted trees (more leaves than the simulated populations) to outline the effect of the chosen number of leaves. In these simulations, cross-validation was not used to select the size of the trees. A tuning value τ (defined in section 'Degrees of freedom') of 0.5 was used following Ye's suggestion that in his experience, values between 0.5σ and σ give similar results for GDF estimation. In a second Monte Carlo study, τ values were varied to confirm Ye's statement that results do not vary in terms of τ. Both studies were conducted using both statistical populations.

RESULTS

NUMBER OF GROUPS OR LEAVES

Results of the Monte Carlo study on the estimated R^2 and its two adjusted values were compiled for both populations. All results are depicted in boxplots comparisons to illustrate the range of the estimates, with whiskers extending to 1.5 times the interquartile range. Points represent values out of these limits. Three figures for each population (Figs. A3.2-A3.3 in Appendix 2; Fig. 3.1 presented here for

population 1 and Figs. A3.12-A3.15 for population 2 in Appendix 2) depict the estimated values as a function of the number of leaves for a fixed sample size (20, 50 or 100). Other figures (Figs. A3.4-A3.10 for population 1 and Figs. A3.16-A3.22 for population 2 in Appendix 2) show results of similar simulations as a function of sample size for a fixed number of leaves. The first important observation is that the trios of estimates of ρ^2 are always ordered in the same manner in both populations for all sizes of trees and all sizes of samples: $R^2_{MRT} > R^2_{MRT(p)} > R^2_{MRT(GDF)}$. For a fixed sample size, when the trees are underfitted (the number of leaves is smaller than the population size, four for population 1, and six for population 2), it is problematic to assess which estimate is less biased as the trio of values can be spread on both sides of the population ρ^2 value line (Figs. A3.4-A3.5 and Figs. A3.16-A3.17 respectively in Appendix 2). On the other hand, if we focus on the fitted or overfitted models (larger than the population size, Figs. A3.6-A3.10 and Figs. A3.18-A3.22 respectively in Appendix 2), the trios of sample estimates are on or above the population value, and we see that $R^2_{MRT(GDF)}$ is always less biased than the other members of each trio. A visible bias still remains in $R^2_{MRT(GDF)}$ for overfitted models; this bias seems stronger for population 2 and larger for small sample sizes in both populations, although it is less biased than the other two estimates. According to the Monte Carlo simulations presented as a function of tree sizes (see size 4 in Figs. A3.1-A3.3 in Appendix 2 and size 6 in Figs. A3.12-A3.15 in Appendix 2), bias is much smaller for the correctly fitted sample trees. In this setting, bias seems almost null for all sample sizes of population 1, and the same applies to populations 2 for sample sizes equal to or greater than 50.

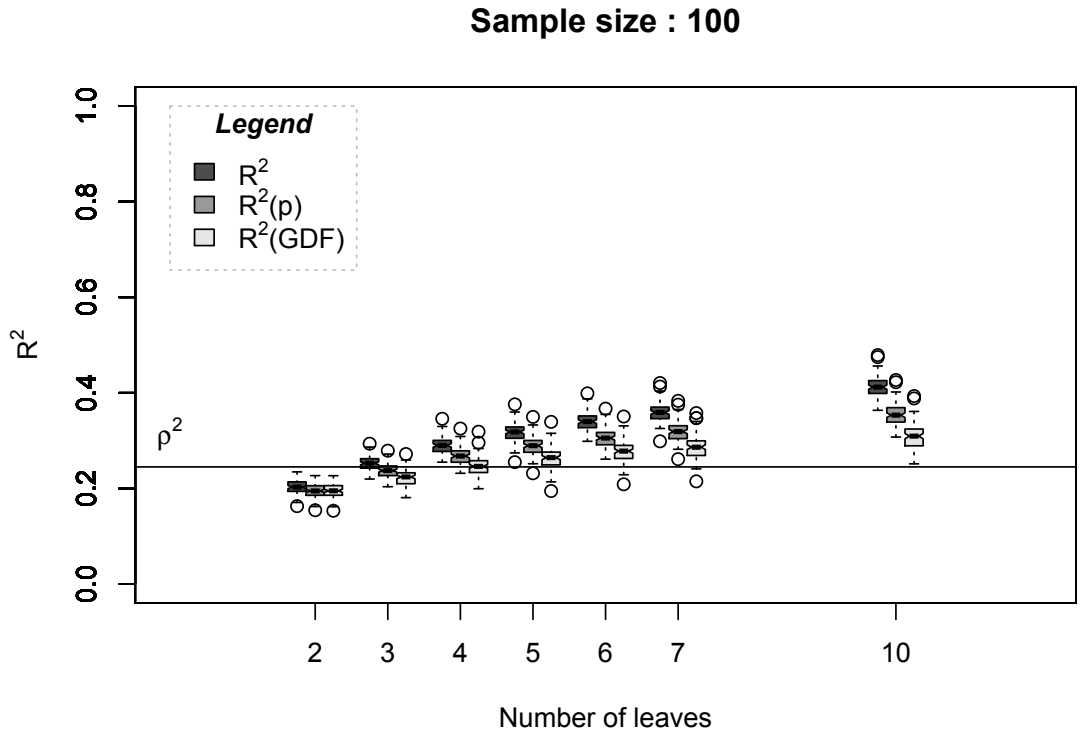


Figure 3.1: Boxplot triplets of ρ^2 estimates (see legend) for trees with different numbers of leaves. The triplets are shown in the same order: R^2, $R^2_{(p)}$ and finally $R^2_{(GDF)}$. Simulations were carried out on population 1 with a sample size of 100.

Overall, for the correctly fitted trees, $R^2_{MRT(GDF)}$ is a better estimator of ρ^2 than R^2_{MRT} because its mean sample estimate is much closer to the true value.

ESTIMATION OF THE POPULATION TREE SIZE

In the light of the previous results, picking the right size of tree is the key to minimizing the bias of $R^2_{MRT(GDF)}$. For real data sets, we do not know what the true number of leaves is, thus it has to be estimated. Two rules based on *CVRE* of a model can be used to pick the size of a sample tree: the minimum *CVRE* rule ('min' rule) and the minimum *CVRE* minus 1 standard error rule ('1se' rule) which produces a more parsimonious solution (Breiman et al. 1984, De'ath 2002, Legendre & Legendre 2012). The choice of rule, or the use of any other rule, is left at the user's discretion in practice.

Here we present Monte Carlo simulations regarding the ability of these two rules to pick the proper size of sample trees. For population 2 that is of size 6, 1000 Monte Carlo runs were carried out with samples of sizes 20, 50 and 100 picked with replacement from the populations. The sample trees were grown following both rules, and the resulting sample tree size was noted. For both populations, three sampling designs, namely random, random stratified and importance sampling (sampling weighted by the size of the leaves, Gentle 2003) among leaves were used to weight the importance of prior knowledge about the structure of the data. Results are reported in Figs. A3.21 (Appendix 2) for the '1se' rule and A3.22 (Appendix 2) for the 'min' rule, both for population 2. For the random sampling design, the 'min' rule found the true size of the tree more often. For stratified sampling, both rules showed higher percentages of size six trees than for random sampling, implying that the use

of prior knowledge on the data structure in sampling could improve the odds of selecting the proper size of tree. Oddly enough, the importance sampling performed more poorly than the stratified sampling for all sample sizes in the '1se' rule study. We observe the same trend for the min rule, except for the size 50 samples. We conclude that the min rule is the best one to identify the population size (i.e. the correct number of leaves) of the tree, and that prior knowledge, if available, should be used to sample equally all leaves.

ASSESSMENT OF τ FOR SMALL SAMPLES

As stated earlier, we used 0.5 as the τ tuning parameter in the previously reported simulations. In the hope of lowering the bias of $R^2_{MRT(GDF)}$ and to confirm Ye's statement, we assessed the effect of using a range of τ values [0.3, 0.5, 0.7, 1, 2, 5, 10] for small samples in both populations. We illustrate the results in Figs. A3.11 and A3.20 in Appendix 2 for populations 1 and 2 respectively. Overall there was no observable change in recovery of the correct ρ^2.

CASE STUDIES

We illustrate the use and interest of R^2_a in MRT analysis using two sets of real data.

The first data set is a subset (30 sites) of the Doubs River fish assemblage data of Verneaux (1973). The Doubs River is in the Jura mountains near the France-Switzerland border. This subset is available in the R package ADE4. It is also distributed with the electronic material provided as companion to the book of Borcard et al. (2011) at http://www.bio.umontreal.ca/numecolR/. The latter was used to do the following analyses. Three data tables describe the 30 sites: fish species composition,

explanatory variables describing the water quality and river morphology, and spatial coordinates of the sites. Fish species composition is considered an ecological indicator of the different water bodies along the river. The community composition data were Hellinger-transformed prior to the analyses. We first analyzed the data using MRT analysis (original explanatory variables) and RDA (after forward selection of the explanatory variables). For MRT, the 'min' criterion in cross-validation identified 8 groups as the best partition. The $R^2_{MRT(GDF)}$ statistic was 66.24% for the MRT model and 57.97% for the RDA model (see Fig. A3.23 for the MRT tree, Fig. A3.24 for the geographical map of the partition, and Fig. A3.25 for the RDA triplot results, Appendix 2). The MRT model identified the distance to the source and the biological oxygen demand as the two most discriminating explanatory variables. RDA identified the same explanatory variables along with other variables that were in strong negative correlation with them: altitude and dissolved oxygen. Calculating the difference between the unexplained variation of each species by both models, we were able to identify that the bleak *Alburnus alburnus* (noted ABL) was a species for which the difference was greater. In the RDA, this species was positively correlated with the distance to the source (das) and to the biological oxygen demand (dbo) and negatively with dissolved oxygen. Actually, ABL does not have a linear relationship with these explanatory variables (Fig. A3.26, Appendix 2). The bleak first appears at site 17 and is present in all sites farther from the source with a constant relative abundance except for a jump in abundances at sites 23-25 which are heavily eutrophized (high concentrations of phosphorus, nitrate and ammonium, high biological oxygen demand, low dissolved oxygen concentration, Fig. A3.26,

Appendix 2). This form of relationship cannot be accounted for in the RDA triplot because it is not linear, and no transformation would remediate this problem. In this example, MRT performed descriptively better than RDA because of this special step relationship, which is due to anthropogenic pollution from known agricultural runoff at site 23 (Borcard et al. 2011).

The second data set is from Aart & Smeek-Enserink (1975) where hunting spider and environmental condition data collected at 28 sites in a sand dune area of the Netherlands have strong linear relationships; in addition, some scatter plots of the spider abundances as functions of the explanatory variables, shown by Aart & Smeek-Enserink (1975), showed polynominal relationships. The first RDA model was computed with the raw explanatory variables; it provided an R^2_a of 71.18% (triplot in Fig. A3.27, Appendix 2); the MRT analysis produced an $R^2_{MRT(GDF)}$ of 79.88% (tree in Fig. A3.28, Appendix 2). Since Aart & Smeek-Enserink (1975) had shown that some species had polynomial relationships with environmental variables, we raised them to power 2 and added them to the equation, used forward selection (Blanchet et al. 2008) to select meaningful explanatory variables, and obtained with RDA an R^2_a of 80.62 % (triplot in Fig. A3.29, Appendix 2). These results illustrate that when the relationships between the response and explanatory variables are linear (or have been linearized), MRT can still stand out in its selection of explanatory variables: the first split generated by the presence or absence of the grass *Corynephorus canescens* (Poaceae) explains almost half of the variation accounted for by the multivariate regression tree whereas it does not even appear among the variables selected for the RDA. Despite of that difference, the two analyses have

similar values of adjusted R^2. One can easily see why when examining the scatter plots of the spider species as a function of *Corynephorus*: it is nearly impossible to linearize the relationship with the response (not shown here). Some spider species, in particular *Alopecosa cuneata*, *Arctosa lutetiana*, *Aulonia albimana*, *Pardosa lugubris*, *Pardosa nigriceps* and *Zora spinimana*, are completely absent when *Corynephorus* is present. A tree split represents this type of relationship much better than a linear trend.

DISCUSSION

ECOLOGICAL IMPLICATIONS

Seeking proper means of comparing MRT and RDA models finds its origin in the univariate analysis of ecological data. Regression trees (RT) have shown their usefulness by proving to be more powerful in some cases in terms of prediction than the linear modelling techniques that are usually preferred; see for example Rejwan et al. (1999) and Vayssiéres et al. (2000). The multivariate version of RT, MRT, identifies explanatory thresholds delineating the largest changes in community composition whereas linear modelling is limited to account for linear relationships. Changes in the response variables corresponding to differences between groups may be smooth or abrupt. A model describing abrupt changes in species composition may be especially useful in defining suitable habitat conditions for species assemblages and in discovering ecological thresholds that arise from environmental pressure related to global climate change, as well as species losses due to anthropogenic and other types of disturbances. Natural resource managers, who must provide simple and applicable rules, will especially welcome the simplicity of interpretation of an MRT

model. Applied ecologists need to be able to recognize when the site classification rules depicted in a tree better describe the distribution pattern of the species than a continuous linear relationship. By properly weighting the explanatory powers of MRT and RDA models, we can compare their results and establish which type of pattern the species assemblage follows, and in return provide insight on the proper management actions that may be required.

In our case studies, we confirmed that for data sets that are known to be linearly related, RDA outperformed MRT, whereas for response data (species) that were related to the environment in the form of step functions, the MRT analysis was better suited. In general, when sampling is carried out over a sufficiently large range of an environmental variable, the relationship between species abundances and that environmental variable is unimodal with an optimum located away from the extremes, as seen for the hunting spider data (Aart & Smeek-Enserink 1975). Some specific conditions may, in some instances, favour the appearance of step function relationships between species and their environment, like strong anthropogenic disturbances.

We suggest that RDA and MRT remain complementary in their use even when one model shows stronger explanatory power. Most assessments will include species linearly related to the environment and others with the relationships in the form of step functions; thus a full description of the pattern of species distribution may require both analyses. Comparison between RDA and MRT unexplained species variation can be used to identify species non-linearly related to the explanatory variables when the MRT analysis performs descriptively better. Also, when MRT outperforms RDA in terms of R^2_a, users should attempt to transform the explanatory

variables in such a way as to linearize the relationships: if the exercise is successful, the species are then related linearly (or in a polynomial way) to the environmental variables. In such a case, RDA should be preferred.

ADJUSTMENT, NUMBER OF LEAVES AND TUNING PARAMETER

To define an adjusted form of R^2_{MRT}, we implemented the estimation of residual variance of a model given by the generalized *df* (GDF) estimation of Ye (1998). We can interpret GDF as the cost of the modelling process, so that under suitable conditions, an unbiased estimate of the error variance can be obtained. There are several differences between GDF and the traditional *df*. The number of parameters estimated in the modelling process no longer corresponds to the number of *df* estimated by GDF, which may not even be an integer. GDF estimation depends on both the modelling procedure and the underlying true model.

There is an apparent contradiction between Ye's simulations results and ours. In Ye's simulation results, the estimated residual variance of the model did not decrease with the number of nodes. In our simulations, $R^2_{MRT(GDF)}$ increased with the number of nodes in the case of overfitted trees. The explanation lies in the nature of Ye's simulated data: they contained really crisp clusters that were perfectly associated with the explanatory variables and the within-cluster variance was very small (0.25). As a consequence, his clusters had low variance and, when split into two, the means of the new clusters were about the same as the original cluster. This lead to a very similar sum of squares, thus about the same residual variance estimate. When clusters are not crisp, like in our case, the means of the two new clusters tend to differ from the original cluster, minimizing the sum of squares and thus the

variance; this leads to a smaller residual variance and thus a larger $R^2_{MRT(GDF)}$. In the light of this difference, it becomes compellingly important to properly fit the tree prior to the calculation of $R^2_{MRT(GDF)}$ if we wish to accurately estimate ρ^2. Our Monte Carlo simulations showed that cross-validation and decision through the 'min' rule were best suited for this purpose. This procedure should become the way of choosing the number of leaves when assessment of the explanatory value of the model is the goal.

The stronger bias of $R^2_{MRT(GDF)}$ observed for population 2 in sample size 20 may be due to the population structure. It was generated with 6 groups of very different sizes, the smallest with 110 objects and the largest with 3189 objects, contrarily to population 1 that had group sizes ranging from 2265 to 2812. Monte Carlo simulations with random sampling on population 1 were thus more likely to correctly represent each cluster. With this consideration, we cautiously state that sample size does not matter, and $R^2_{MRT(GDF)}$ remains a practically unbiased estimator of ρ^2 for fitted sample trees.

COMPARING $R^2_{MRT(GDF)}$ WITH DIFFERING IMPURITY MEASURES

The *impurity measure* of an MRT is the value that should be minimized when a split is performed. The usual impurity measure is the within-group sum of squares over leaves (OLS). The use of different impurity measures in the computation of MRT has been suggested in the literature (De'ath 2002), thus it is important to extend our definition of R^2_{MRT} to these other measures. For example, one can use the sum of squares around the median instead of the mean. In this case, an analogous parallel measure may be defined around the chosen measure of variation, relating the

variation of the full model over the reduced model into a proportionate reduction in this particular variation (Anderson-Sprecher 1994). Note the importance that both the full *and* reduced model be based on the same variation measure.

Caution must be taken when comparison between R^2 is required because the R^2 of a model based on sum of squares, like RDA, is not comparable to the R^2 of a model computed using a different measure of variation even if the exact same data set is used (Anderson-Sprecher 1994). When comparison between RDA and MRT is the objective of the study, the R^2_{MRT} should be calculated using a sum of squares even if the modelling procedure was not based on that variation measure. In that same line of idea, the response data must be transformed in the same way prior to the analyses. For example, if a log or Hellinger transformation was used before RDA modelling, the same should be done before MRT modelling because a non-linear transformation of the data changes the estimated values of the within-group sums of squares.

We also stress that the descriptive R^2_a model comparison should not be employed for predictive purposes, when prediction of new observations is the main objective of the study. For this, an information-theory based comparison (Burnham & Anderson 2002), for example AIC or cross-validation, should be used or developed if not available. This was not the aim of the present study.

CONCLUSION

Ye's (1998) GDF estimation of the degrees of freedom associated with a particular model was shown to be the best route towards an unbiased coefficient of determination for MRT analysis for a fitted sample tree. This measure of the adjusted variation explained by an MRT is readily comparable to the RDA adjusted coefficient

of determination as long as it is least-square based and the response data have been pre-transformed in the same way. With this new coefficient, it is possible to compare the explanatory power of a linear model of relationship (RDA) between multivariate response data and a set of explanatory variables of choice, to a model like MRT that favors discontinuity. In practice, the comparison between the descriptive power of RDA and MRT may provide insights on the shape of the species distributions along the explanatory variables. It can also suggest that transformations of the explanatory variables are needed to linearize the relationships, or by contrast confirm that the relationships cannot be linearized.

ACKNOWLEDGMENT

This research was supported by Natural Sciences and Engineering Research Council of Canada (NSERC) grant no. 7738 to P. Legendre. We wish to thank Daniel Borcard, Steven Walker and Guillaume Guénard for comments on manuscript drafts.

APPENDIX 1

COROLARY 1: *Let P be the partition of k groups of a tree T obtained via an MRT analysis of X on Y. Then* $R^2_{Y|P} = 1\text{-}RE(T)$.

PROOF: We first get the final MRT model as usual, using cross-validation as a pruning method. From this model, we extract the partition (i.e. the modeling result) and use it as the explanatory variables of an RDA. This can be done using the function MRT from the MVPARTwrap library (see help file for details).

Let *n* be the number of objects (sites), $S = \{s_1, s_2, \ldots, s_n\}$ be the set of objects, *Y* be the response matrix (sites (*n*) x species (*m*)), *X* be the explanatory matrix (sites (*n*) x variables (*p*)) and finally P be the partition of *k* groups obtained via an MRT analysis of *X* on *Y* ($Y \sim X$), thus $P = \{P_1, P_2, .. P_k\}$, $P_i \in S$, $P_i \neq \phi$, $\cup P_i = S$, $P_i \cap P_j = \phi$ if $i \neq j$ is a partition of the sites that minimizes the intra-group sum of squares of the *Y* matrix, while respecting the order of some of the vector of matrix *X* (given by the bipartition of the MRT analysis).

We recode the partition in a proper form, to be used into an RDA analysis. So take $P = \{P_1, P_2, .. P_k\}$, and recode in a dummy variable matrix, to have a column representing each k groups. So let *G* be a matrix of size *n* x *k* be defined as $g_{ij} = 1$ if $s_i \in P_j$, and 0 otherwise, $\forall$ i=1, …, n, $\forall$ j=1,…,k. This new matrix *G* is a matrix that represents the partition of the $Y \underset{MRT}{\sim} X$ model in dummy variables.

We now have all we need to proceed with the RDA analysis.

We get the following formulation to the coefficient of determination in the MRT context:

$$R^2_{Y|P} = \frac{trace\left(\hat{Y}'\hat{Y}\right)}{trace(Y'_{cent}Y_{cent})} = 1 - \frac{trace\left[\left(Y_{cent} - \hat{Y}\right)'\left(Y_{cent} - \hat{Y}\right)\right]}{trace\left(Y'_{cent}Y_{cent}\right)}$$

where $\hat{Y} = G(G'G)^{-1}G'Y_{cent}$ represents the matrix of predicted values. This is identical to calculating predicted values for individual multiple regressions of each column of Y on G; $Y_{cent} = \left(I - P\right)Y$ is centered by column means, I is the identity matrix, and P is a square matrix with all 1/n elements. In ecological applications the species are solely centered, not standardized, and $R^2_{Y|P}$ is a weighted mean of the R^2 of individual models with weights proportional to the species variances divided by the total variance (Peres-Neto et al. 2006). In this particular setting, $G'G$ is a square *k* x *k* is a diagonal matrix with the number of objects in each group on the diagonal, thus $(G'G)^{-1}$ is a square *k* x *k* diagonal matrix with 1 on the number of objects in each group on the diagonal, and finally F= $G(G'G)^{-1}G'$ is a square *n* x *n* symetric matrix with $f_{ij} = f_{ji} = 1/n_k$ if $(s_i \cup s_j) \cap P_k \neq \phi$ for at least and only a *j*, and 0 otherwise, $\forall$ i,j=1, …, n. By multiplying $G(G'G)^{-1}G'Y_{cent}$ we get the predicted values, which correspond exactly to the mean abundance per species for each node: in other words, this is exactly the predicted values given by the MRT model.

Now $Y_{cent} - \hat{Y}$ is equivalent to the difference around the group's mean calculated per species on the Y matrix, thus $\left(Y_{cent} - \hat{Y}\right)'\left(Y_{cent} - \hat{Y}\right)$ is the diagonal matrix with the sum of squares of the Y values around the group means for each species on the diagonal, i.e. the residual sum of squares of the MRT model :

$$\left(Y_{cent}-\hat{Y}\right)'\left(Y_{cent}-\hat{Y}\right)=\begin{bmatrix} SSE_1 & & & \\ & SSE_2 & & \\ & & SSE_{\dots} & \\ & & & SSE_p \end{bmatrix}$$

where SSE_i is the sum of squares for each species around the mean of the group.

By taking the trace of this matrix, we get exactly the resubstitution estimate for R*(t) defined earlier, just like by taking the trace of $Y'_{cent}Y_{cent}$ we get $R^*(\mu)$, thus $R^2_{Y|P}$ is equivalent to $1-RE(T)=1-R(T)/R(\mu)$. We thus conclude that 1-R(t) is the proportion of variation explained by the final partition, the MRT model.

APPENDIX 2

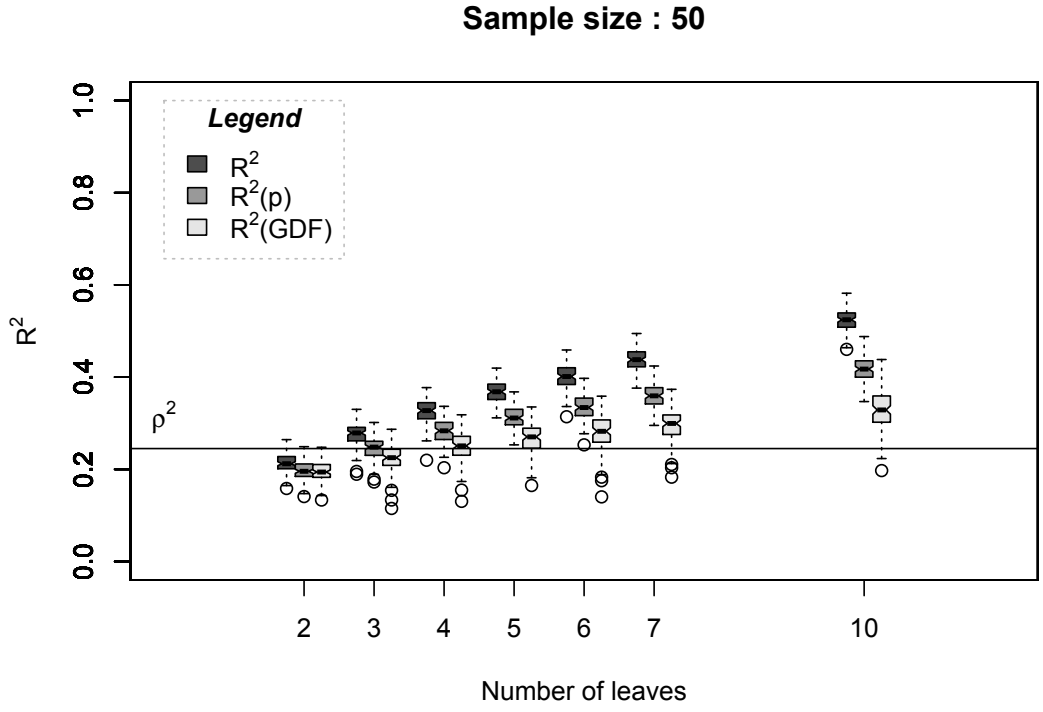


Figure A3.2: Boxplot triplets of ρ^2 estimates (see legend) for trees with different number of leaves. All triplets are shown in the same order: R^2, $R^2_{(p)}$ and finally $R^2_{(GDF)}$. Simulations were carried on population 1 with a sample size of 50.

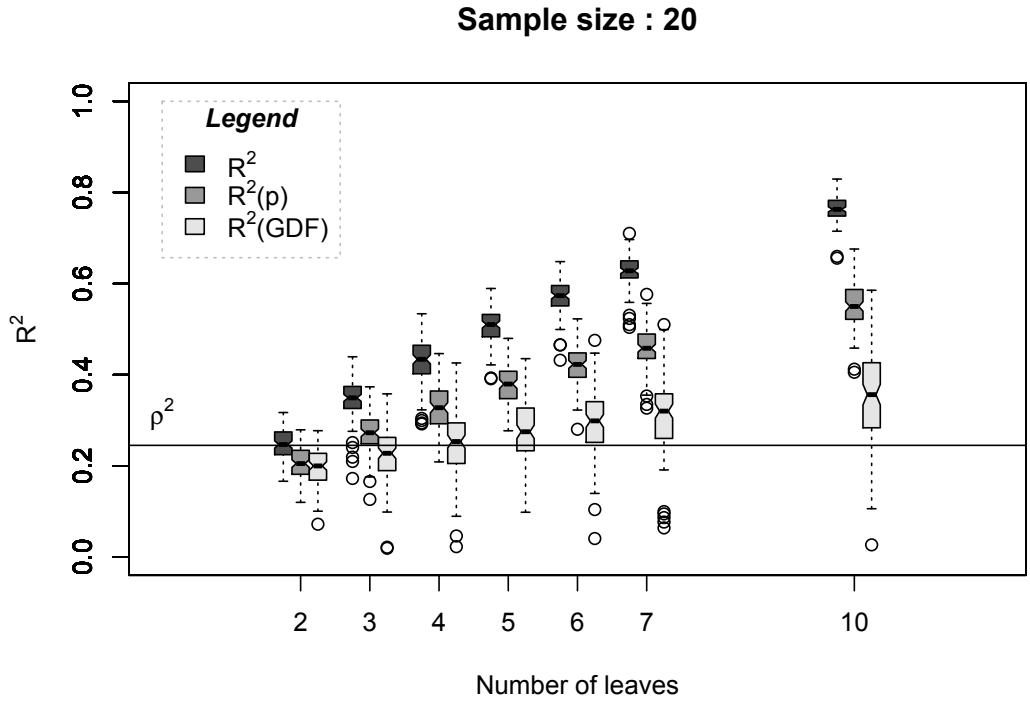


Figure A3.3: Boxplot triplets of ρ^2 estimates (see legend) for trees with different number of leaves. All triplets are shown in the same order: R^2, $R^2_{(p)}$ and finally $R^2_{(GDF)}$. Simulations were carried on population 1 with a sample size of 20.

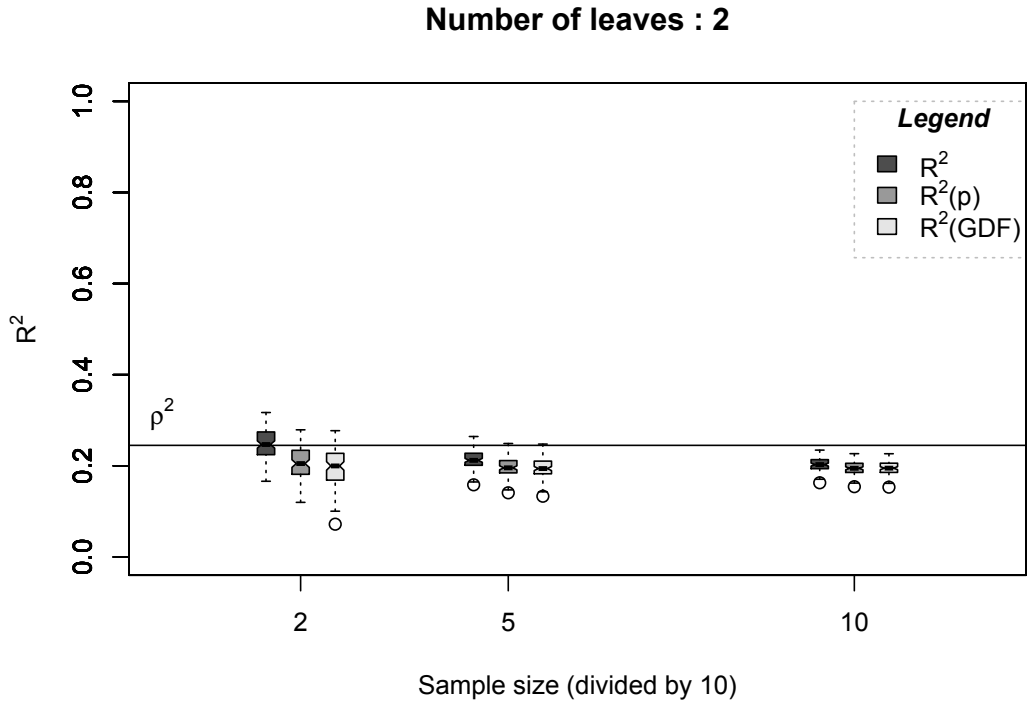


Figure A3.4: Boxplot triplets of ρ^2 estimates (see legend) for trees with different sample sizes. All triplets are shown in the same order: R^2, $R^2_{(p)}$ and finally $R^2_{(GDF)}$. Simulations were carried on population 1 and trees with 2 leaves were build (underfitted trees).

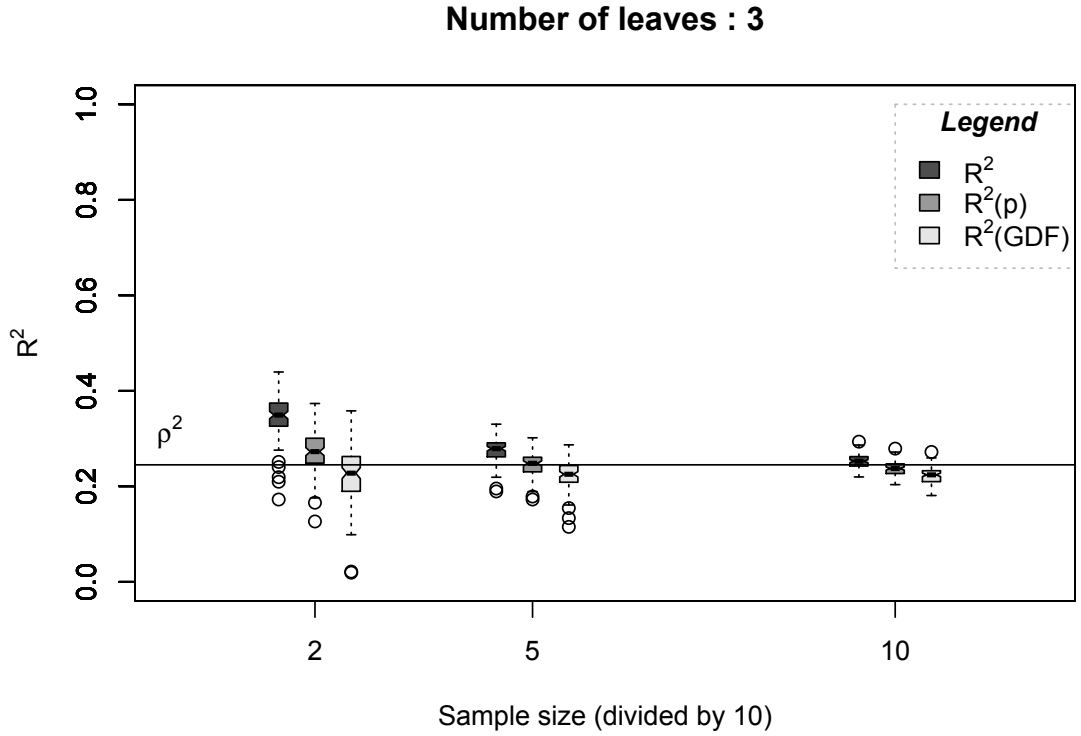


Figure A3.5: Boxplot triplets of ρ^2 estimates (see legend) for trees with different sample sizes. All triplets are shown in the same order: R^2, $R^2_{(p)}$ and finally $R^2_{(GDF)}$. Simulations were carried on population 1 and trees with 3 leaves were build (underfitted trees).

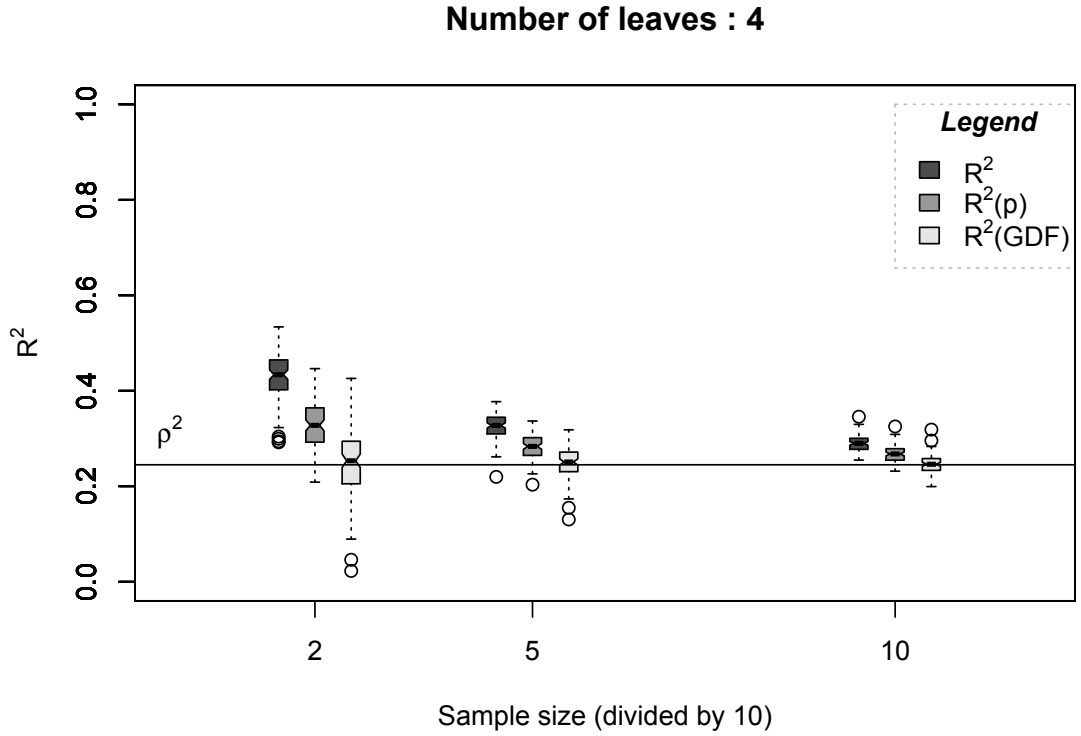


Figure A3.6: Boxplot triplets of ρ^2 estimates (see legend) for trees with different sample sizes. All triplets are shown in the same order: R^2, $R^2_{(p)}$ and finally $R^2_{(GDF)}$. Simulations were carried on population 1 and trees with 4 leaves were build (fitted trees).

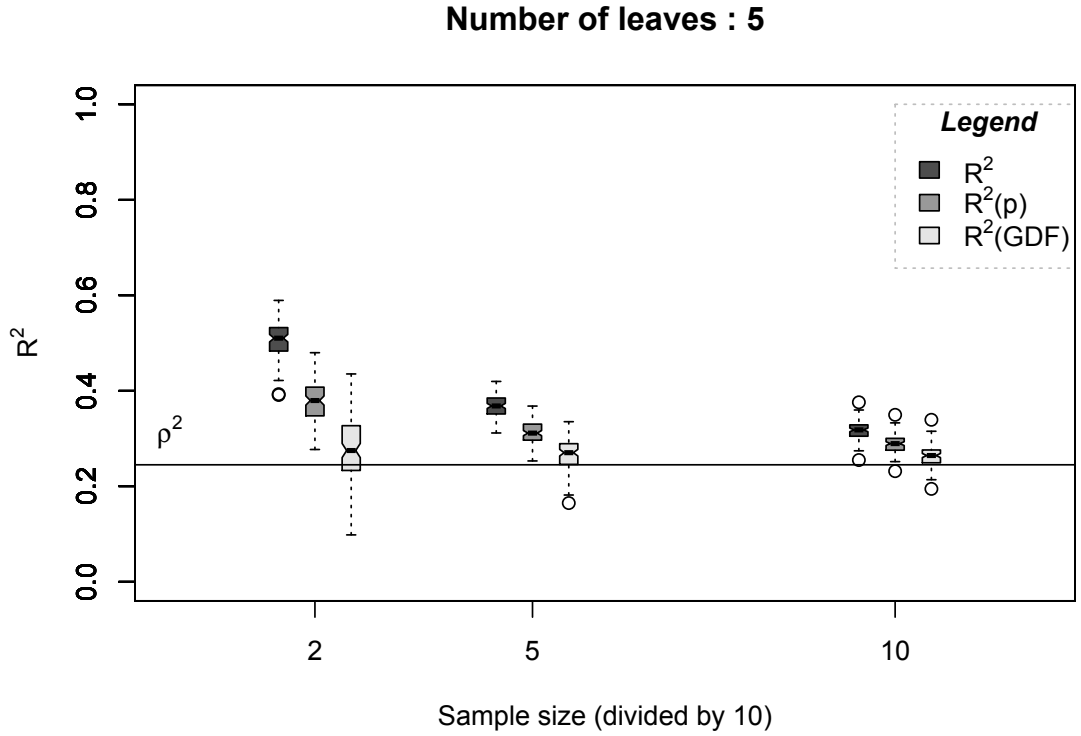


Figure A3.7: Boxplot triplets of ρ^2 estimates (see legend) for trees with different sample sizes. All triplets are shown in the same order: R^2, $R^2_{(p)}$ and finally $R^2_{(GDF)}$. Simulations were carried on population 1 and trees with 5 leaves were build (overfitted trees).

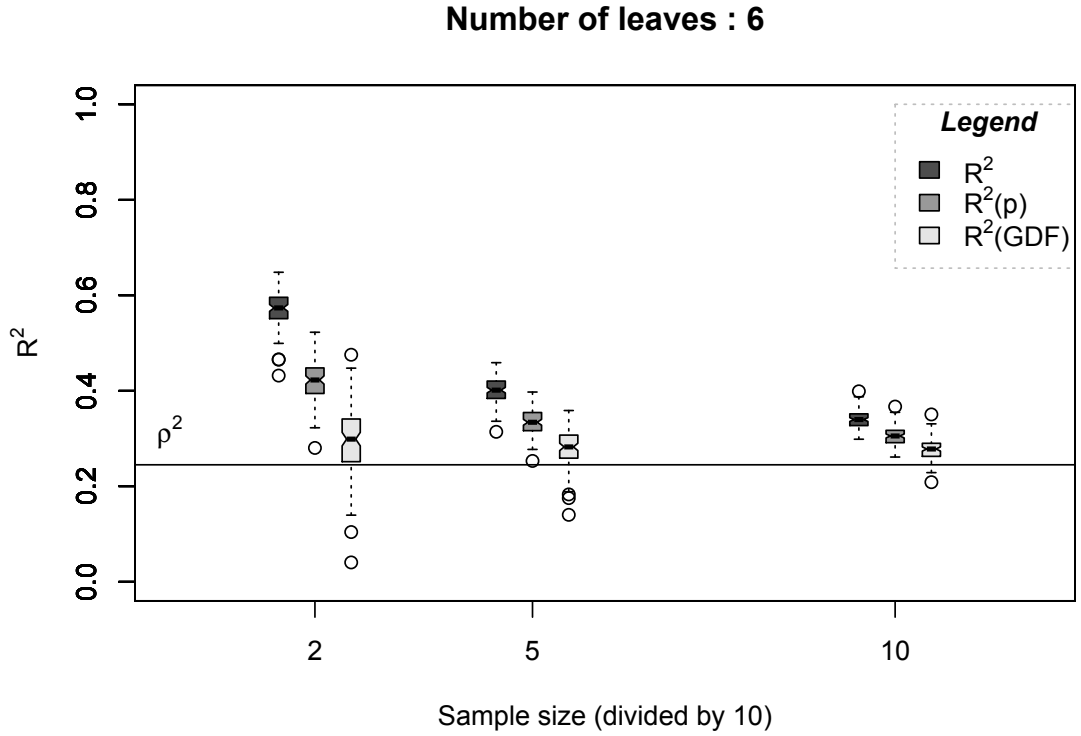


Figure A3.8: Boxplot triplets of ρ^2 estimates (see legend) for trees with different sample sizes. All triplets are shown in the same order: R^2, $R^2_{(p)}$ and finally $R^2_{(GDF)}$. Simulations were carried on population 1 and trees with 6 leaves were build (overfitted trees).

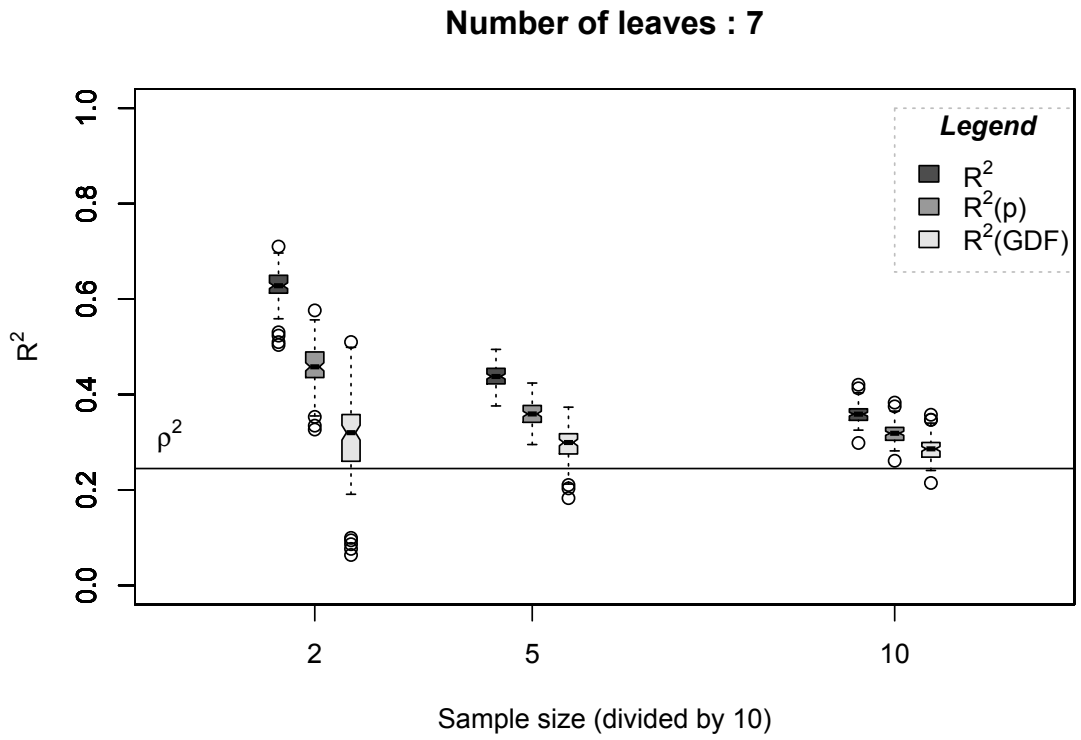


Figure A3.9: Boxplot triplets of ρ^2 estimates (see legend) for trees with different sample sizes. All triplets are shown in the same order: R^2, $R^2_{(p)}$ and finally $R^2_{(GDF)}$. Simulations were carried on population 1 and trees with 7 leaves were build (overfitted trees).

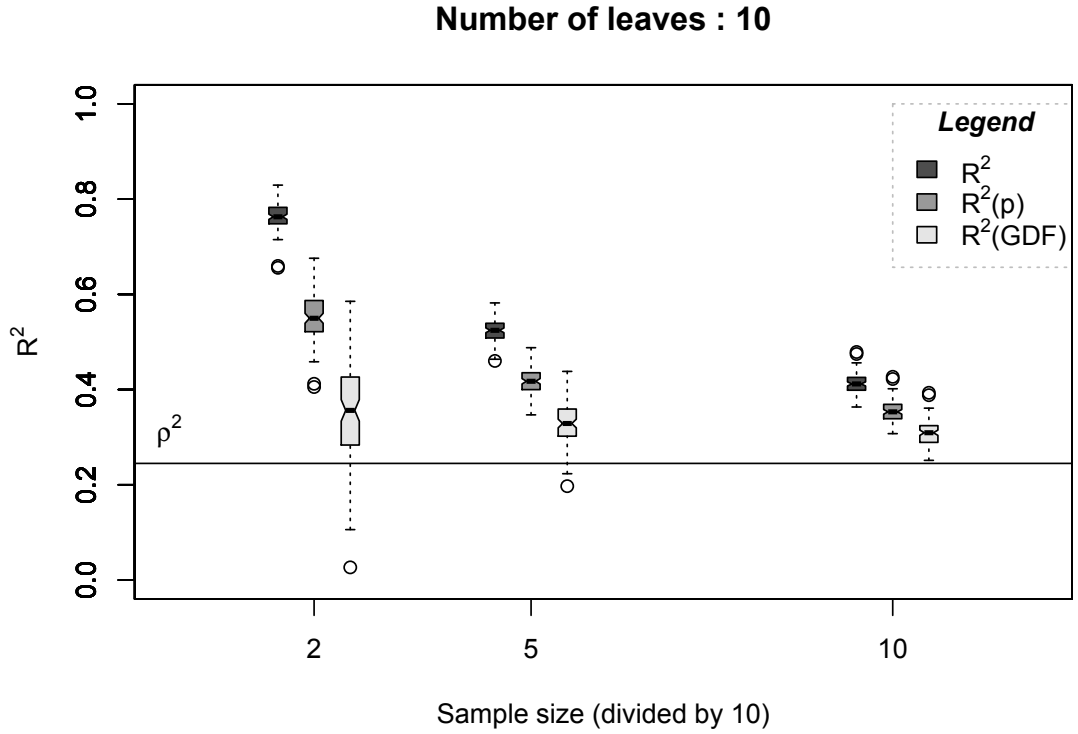


Figure A3.10: Boxplot triplets of ρ^2 estimates (see legend) for trees with different sample sizes. All triplets are shown in the same order: R^2, $R^2_{(p)}$ and finally $R^2_{(GDF)}$. Simulations were carried on population 1 and trees with 10 leaves were build (overfitted trees).

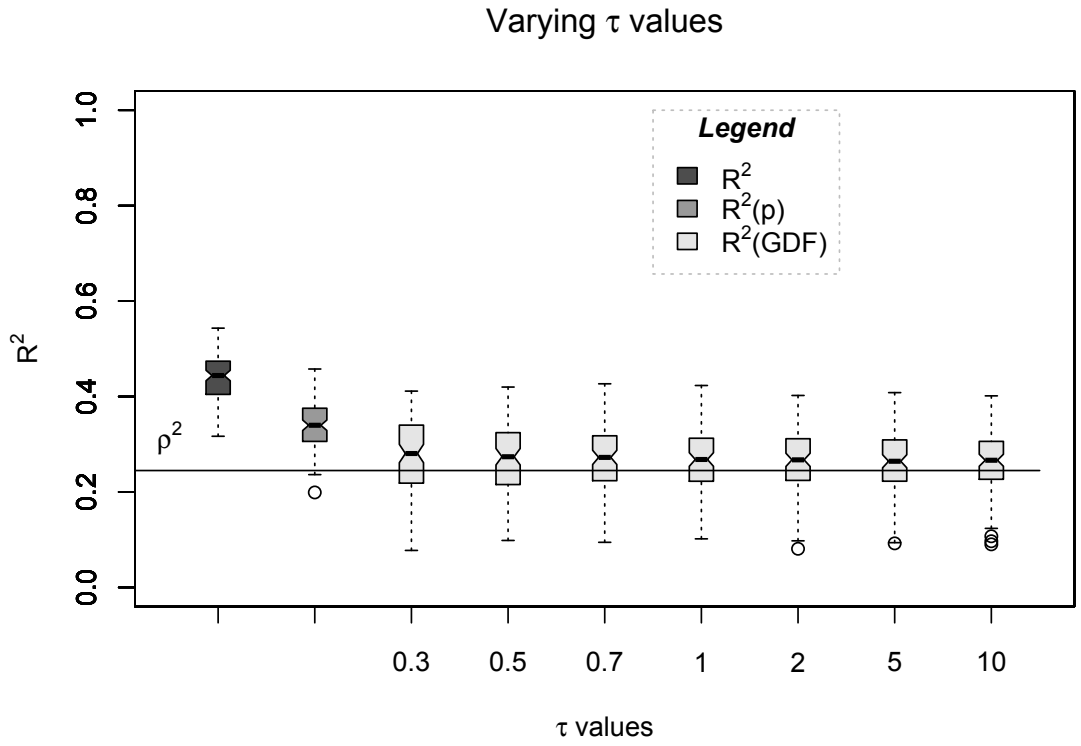


Figure A3.11: Boxplots of ρ^2 estimates (see abscissa) for trees with different τ tuning parameter values in GDF estimates. All triplets are shown in the same order: R^2, $R^2_{(p)}$ and finally $R^2_{(GDF)}$. Simulations were carried on population 1 and trees with 4 leaves were build (fitted trees).

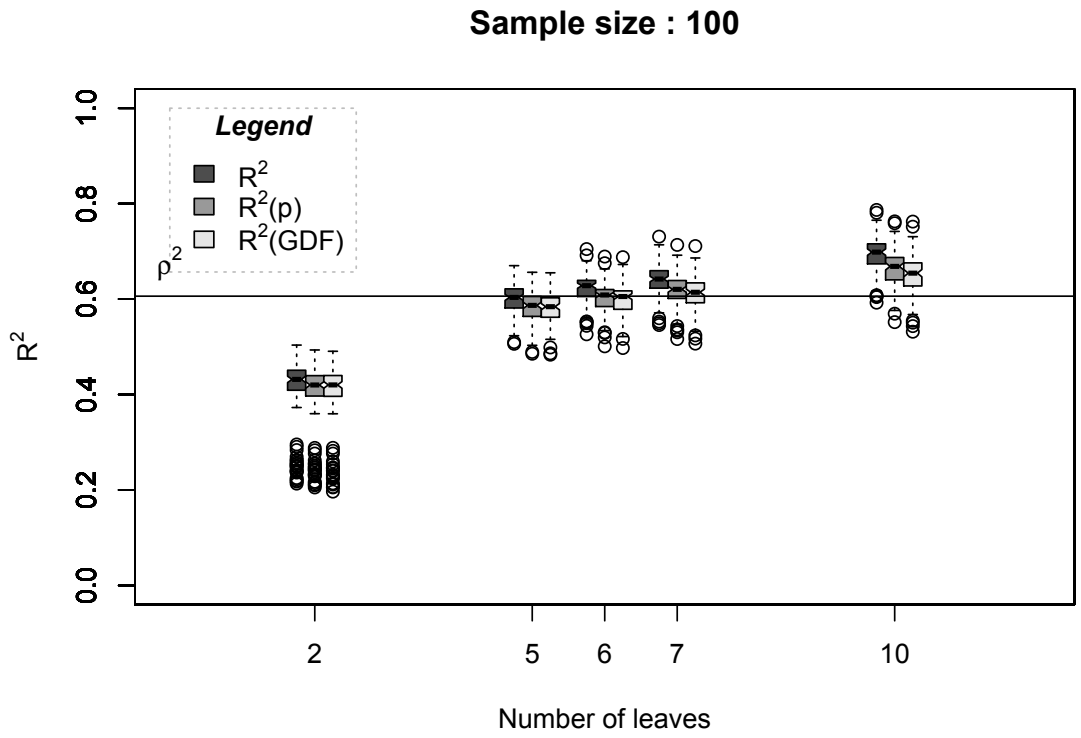


Figure A3.12: Boxplot triplets of ρ^2 estimates (see legend) for trees with different number of leaves. All triplets are shown in the same order: R^2, $R^2_{(p)}$ and finally $R^2_{(GDF)}$. Simulations were carried on population 2 with a sample size of 100.

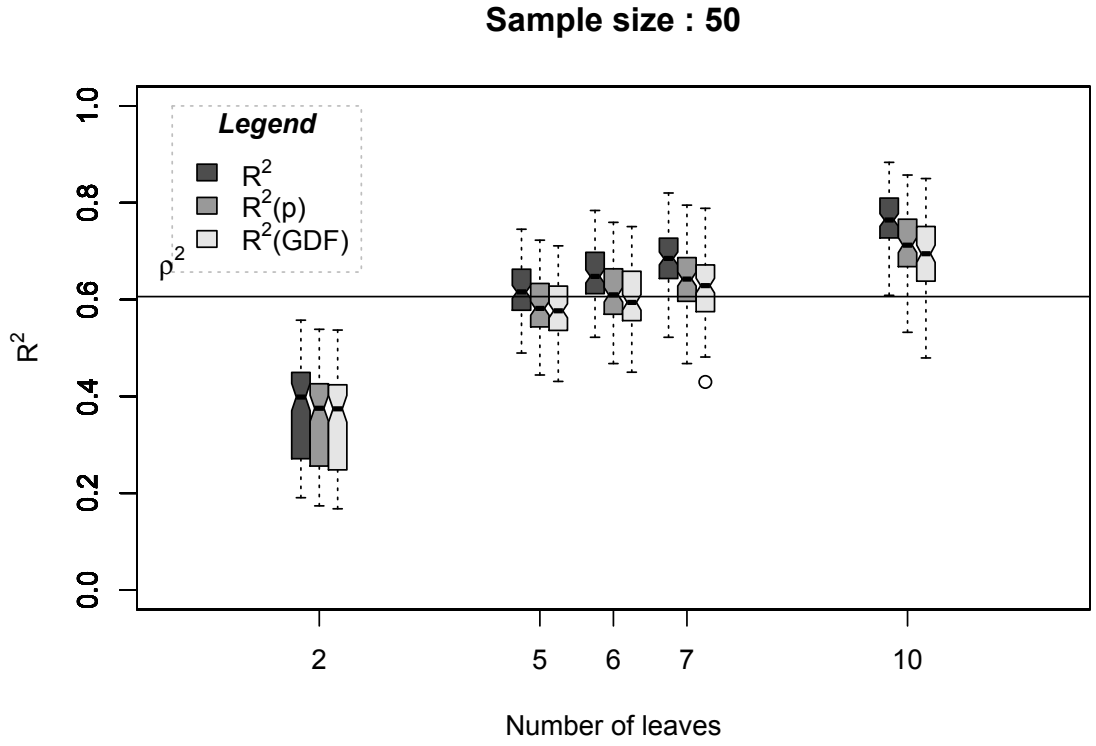


Figure A3.13: Boxplot triplets of ρ^2 estimates (see legend) for trees with different number of leaves. All triplets are shown in the same order: R^2, $R^2_{(p)}$ and finally $R^2_{(GDF)}$. Simulations were carried on population 2 with a sample size of 50.

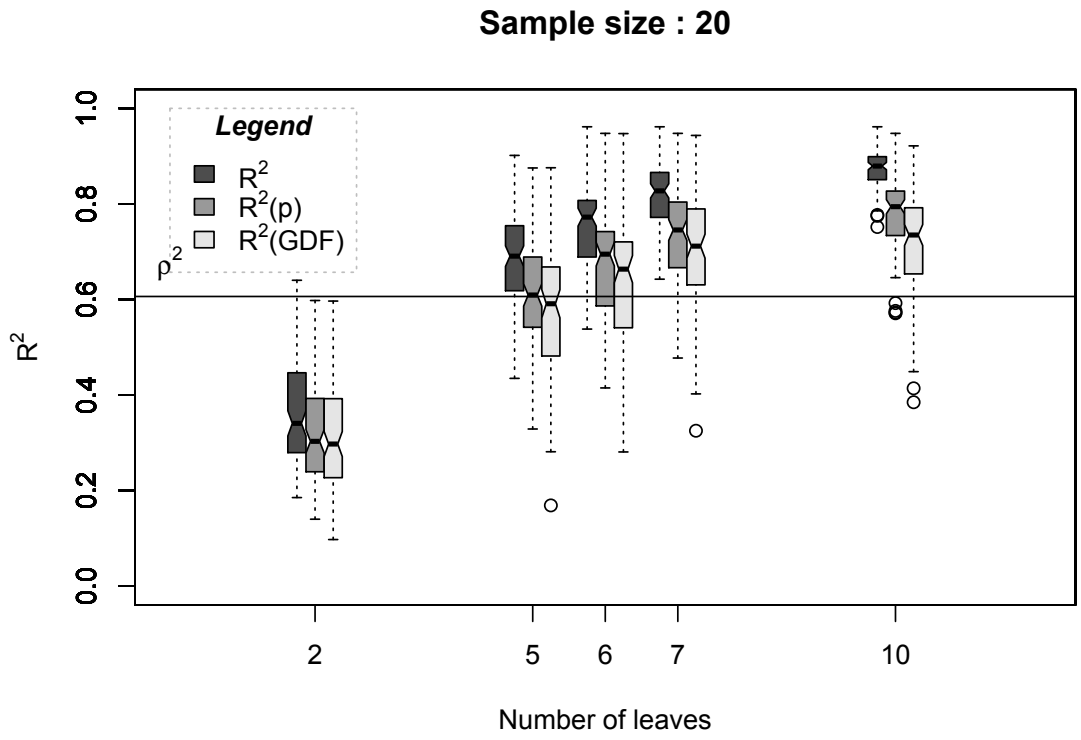


Figure A3.14: Boxplot triplets of ρ^2 estimates (see legend) for trees with different number of leaves. All triplets are shown in the same order: R^2, $R^2_{(p)}$ and finally $R^2_{(GDF)}$. Simulations were carried on population 2 with a sample size of 20.

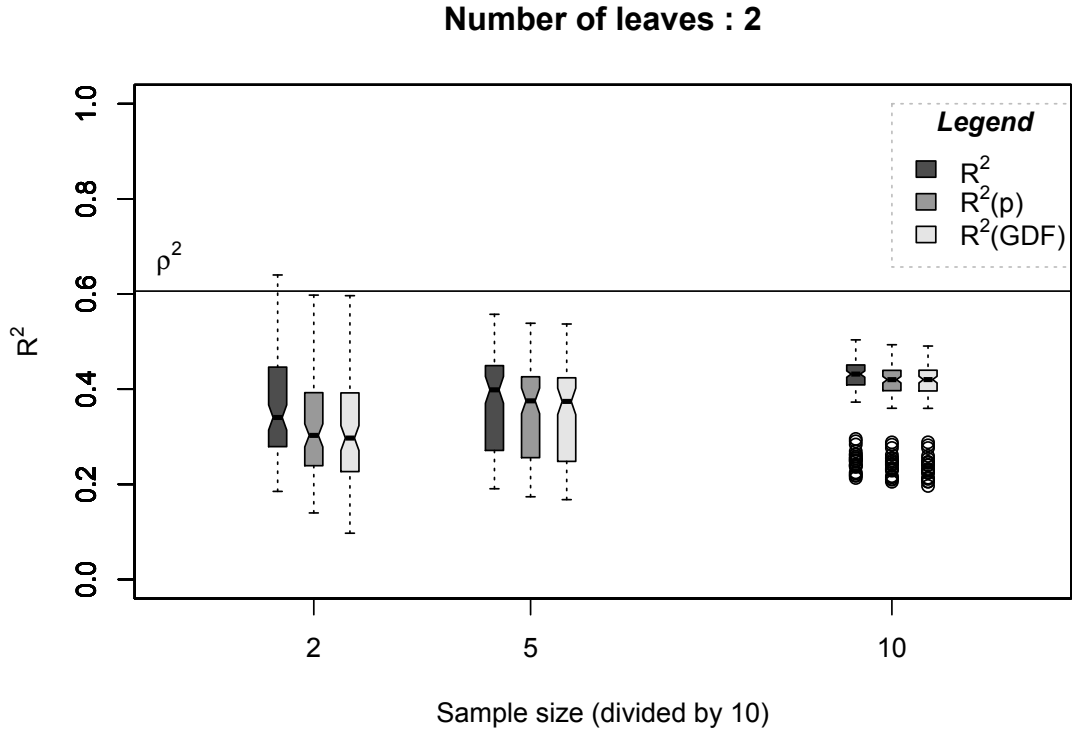


Figure A3.15: Boxplot triplets of ρ^2 estimates (see legend) for trees with different sample sizes. All triplets are shown in the same order: R^2, $R^2_{(p)}$ and finally $R^2_{(GDF)}$. Simulations were carried on population 2 and trees with 2 leaves were build (underfitted trees).

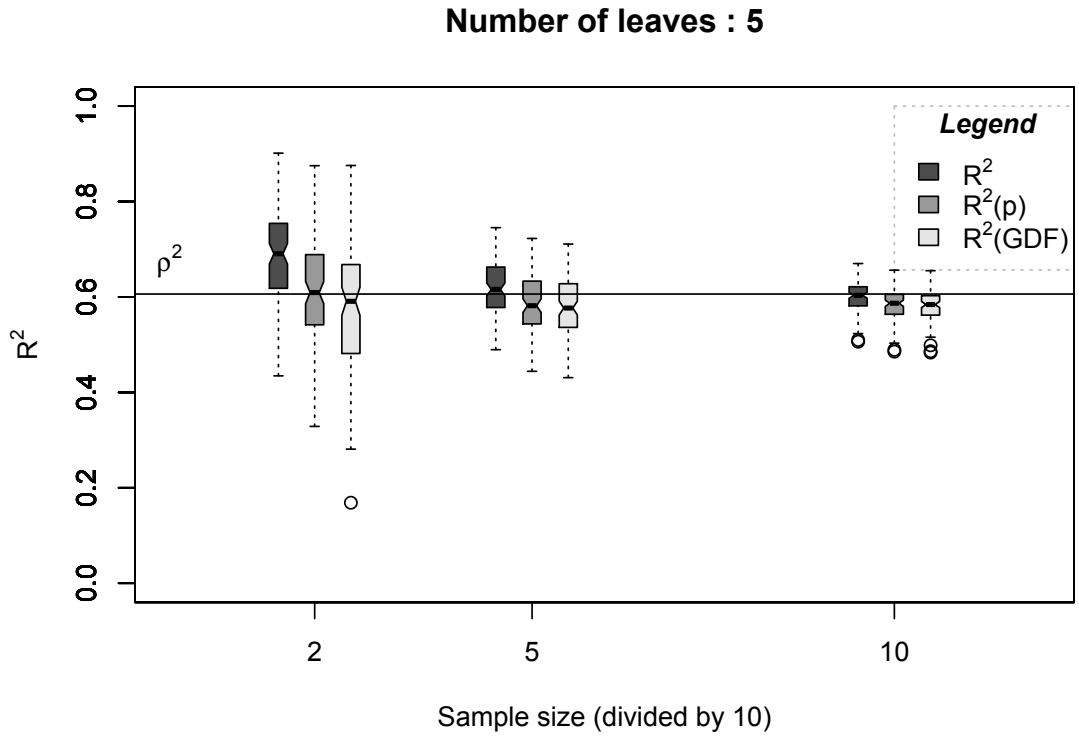


Figure A3.16: Boxplot triplets of ρ^2 estimates (see legend) for trees with different sample sizes. All triplets are shown in the same order: R^2, $R^2_{(p)}$ and finally $R^2_{(GDF)}$. Simulations were carried on population 2 and trees with 5 leaves were build (underfitted trees).

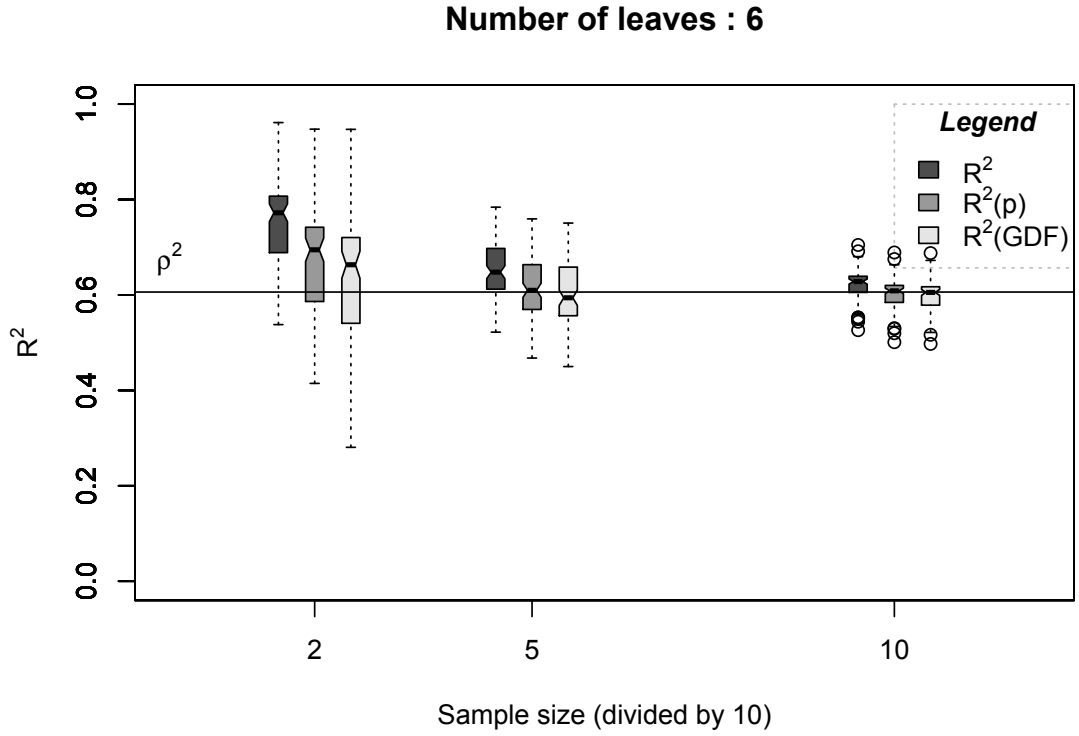


Figure A3.17: Boxplot triplets of ρ^2 estimates (see legend) for trees with different sample sizes. All triplets are shown in the same order: R^2, $R^2_{(p)}$ and finally $R^2_{(GDF)}$. Simulations were carried on population 2 and trees with 6 leaves were build (fitted trees).

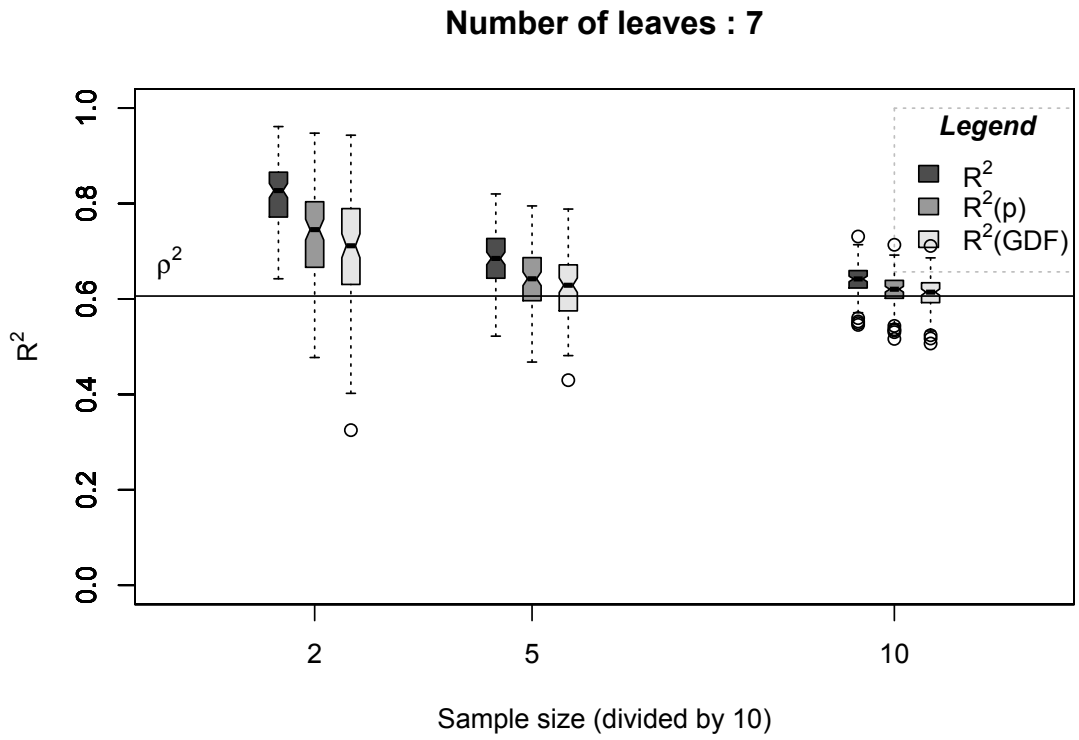


Figure A3.18: Boxplot triplets of ρ^2 estimates (see legend) for trees with different sample sizes. All triplets are shown in the same order: R^2, $R^2_{(p)}$ and finally $R^2_{(GDF)}$. Simulations were carried on population 2 and trees with 7 leaves were build (overfitted trees).

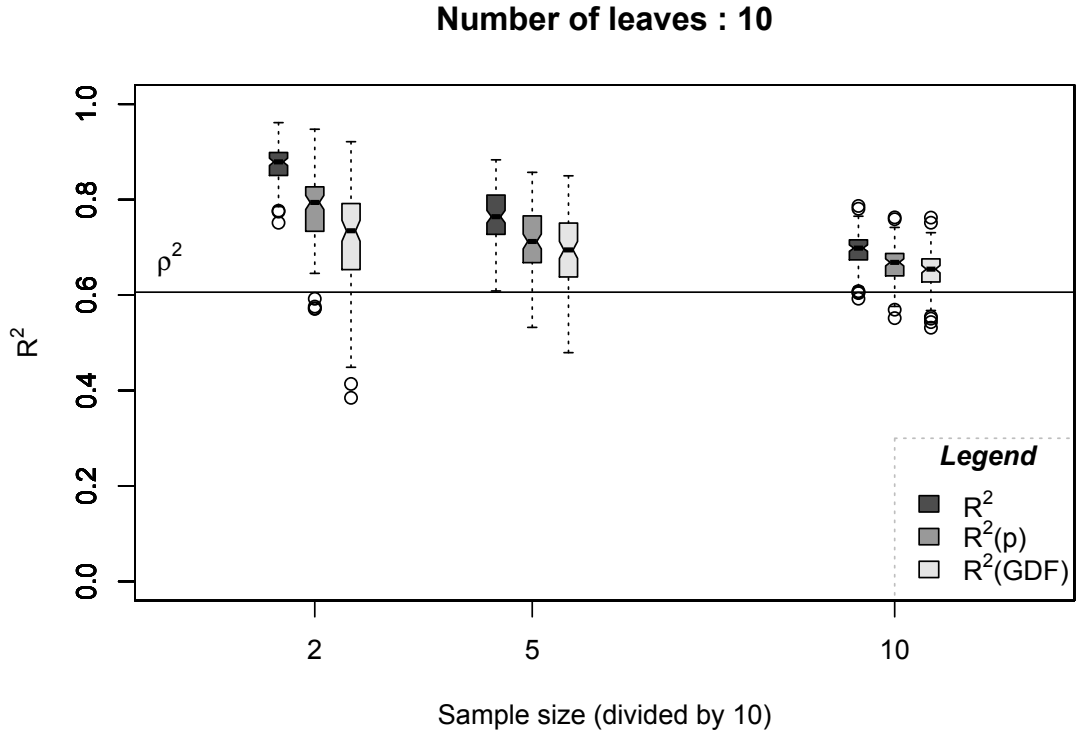


Figure A3.19: Boxplot triplets of ρ^2 estimates (see legend) for trees with different sample sizes. All triplets are shown in the same order: R^2, $R^2_{(p)}$ and finally $R^2_{(GDF)}$. Simulations were carried on population 2 and trees with 10 leaves were build (overfitted trees).

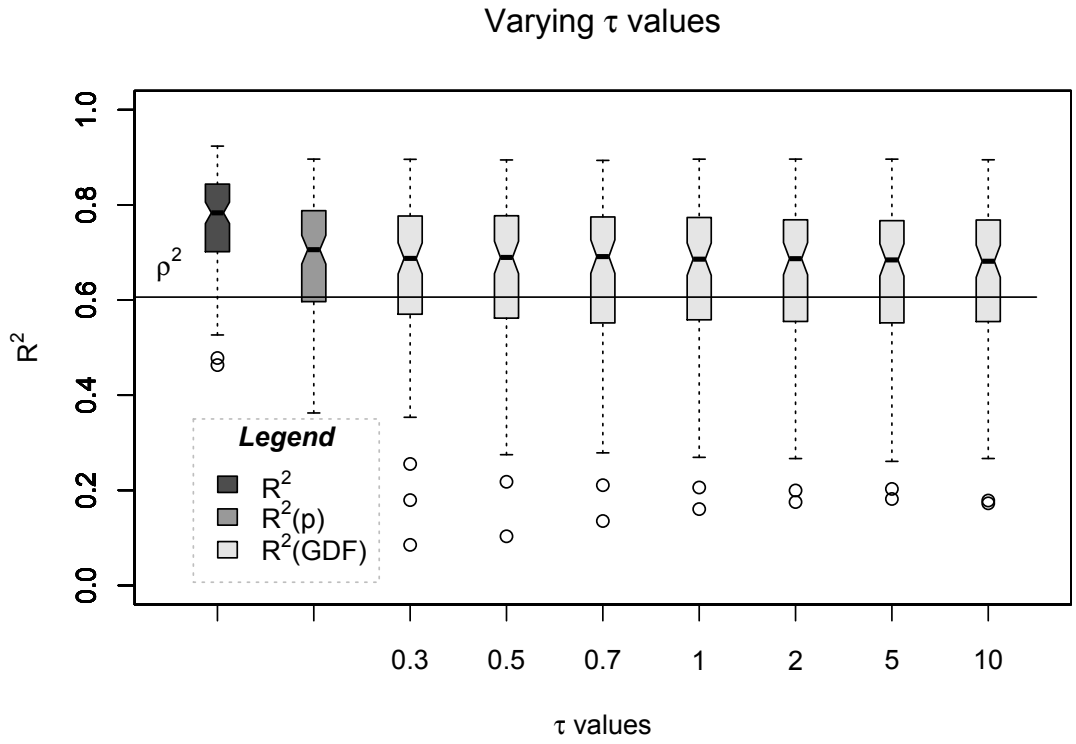


Figure A3.20: Boxplots of ρ^2 estimates (see abscissa) for trees with different τ tuning parameter values in GDF estimates. Simulations were carried on population 2 and trees with 6 leaves were build (fitted trees).

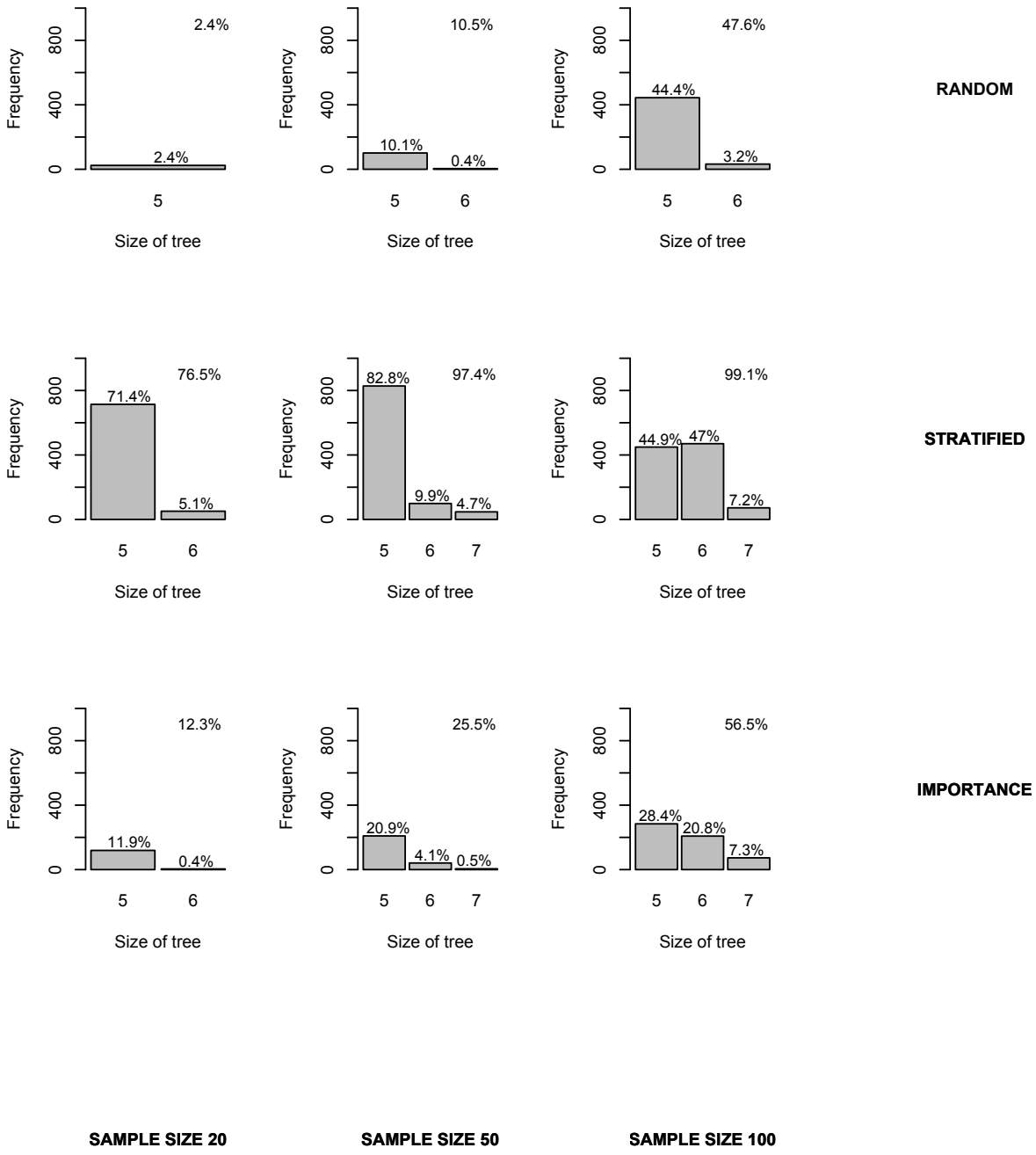


Figure A3.21: Barplots summarizing the Monte Carlo study (1000 runs) of the v-fold cross-validation 1se rule (500 multiple validations) to pick the population 2 size of tree (6) for sample sizes 20, 50 and 100 and for random, stratified and importance

sampling strategies. Frequency of the tree size 5, 6 and 7 are the only ones depicted here.

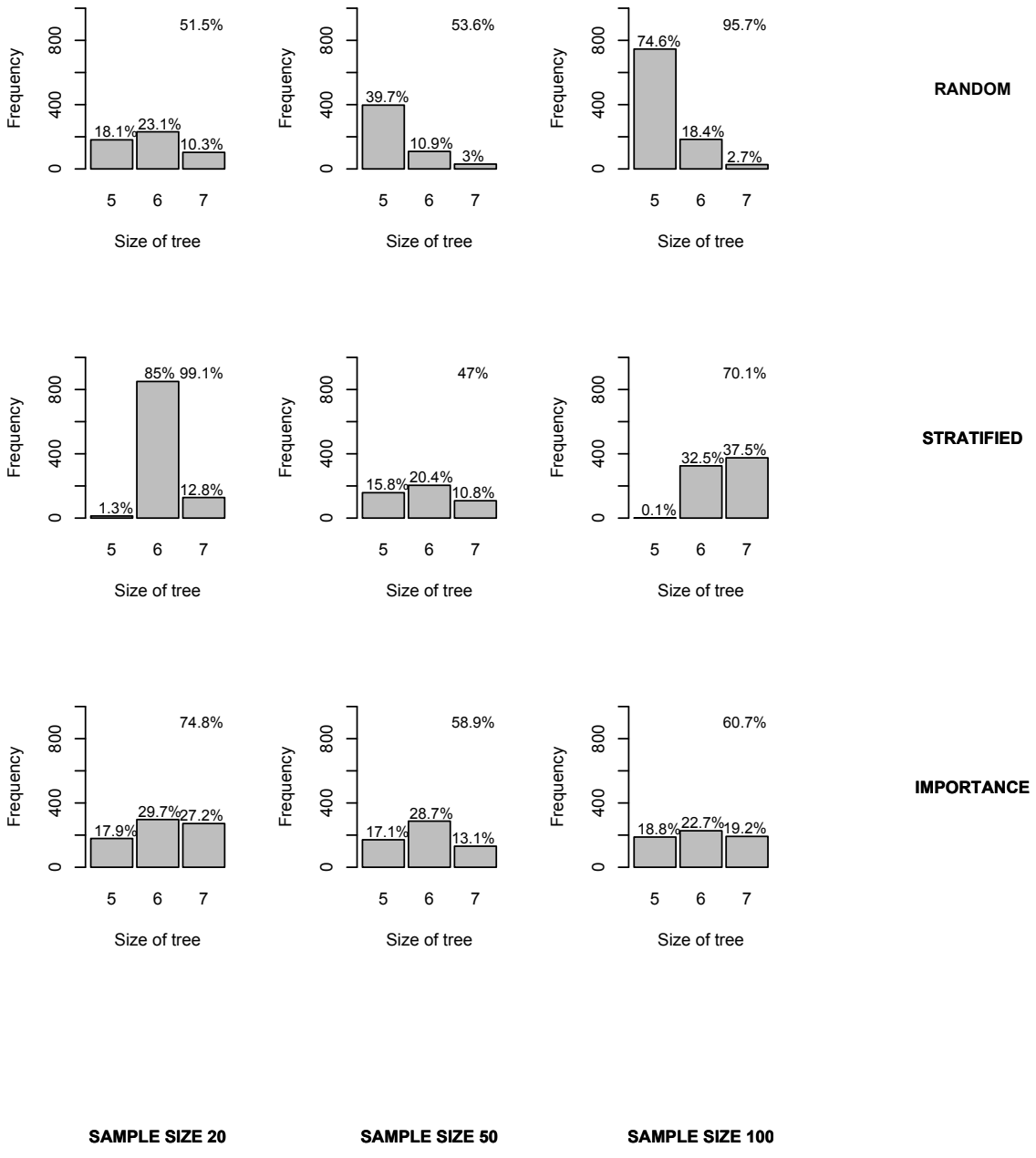


Figure A3.22: Barplots summarizing the Monte Carlo study (1000 runs) of the v-fold cross-validation min rule (500 multiple validations) to pick the population 2 size of tree (6) for sample sizes 20, 50 and 100 and for random, stratified and importance

sampling strategies. Frequency of the tree size 5, 6 and 7 are the only ones depicted here.

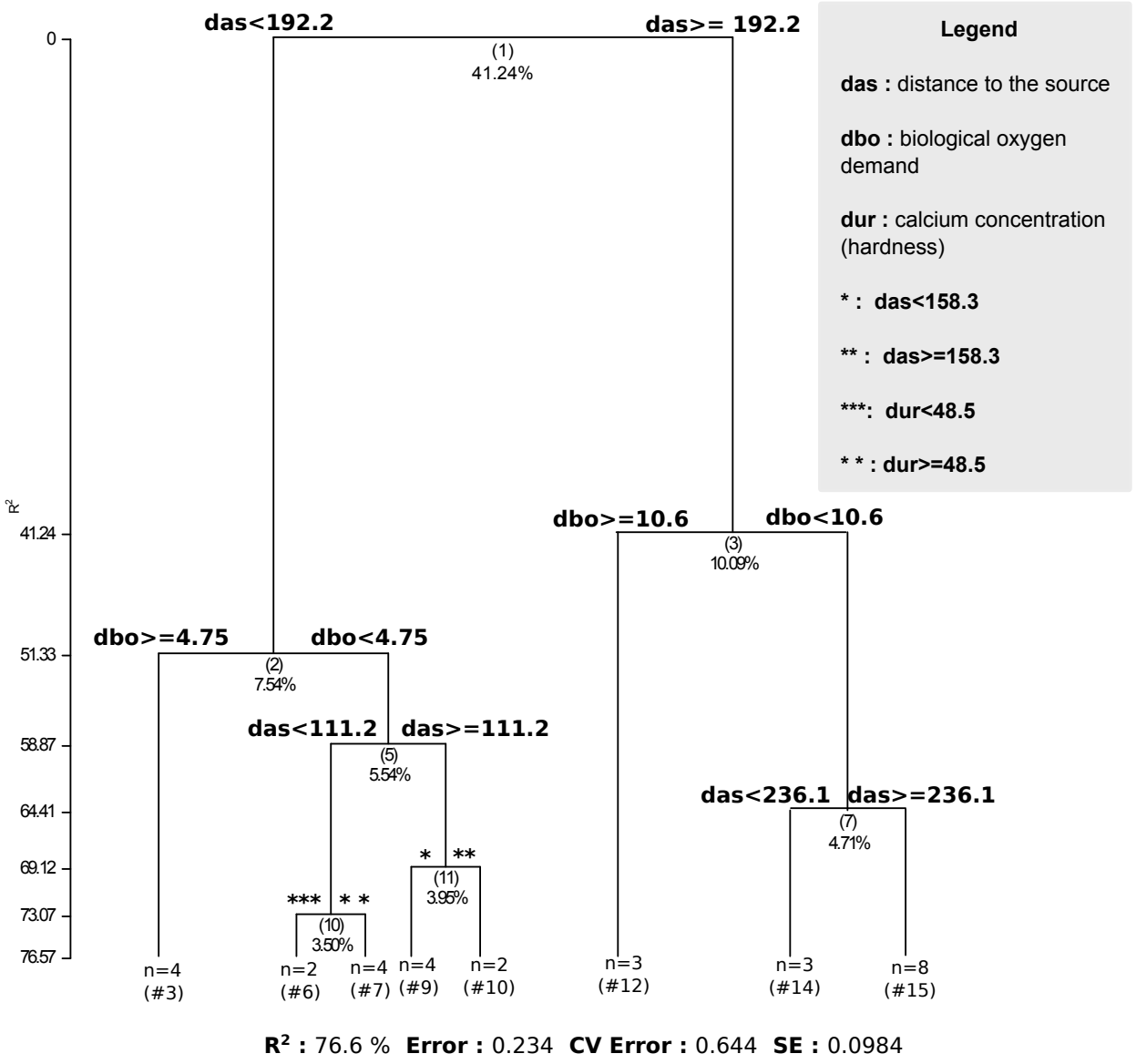


Figure A3.23: Multivariate regression tree model of the Doubs fish data set with Hellinger transformed response data. This output is provided by the ***MRT()*** function of MVPARTwrap. The main difference with the regular output of function ***mvpart()*** is the vertical scale, which is R^2 here. For each leaf, we find the number of objects in the node and the number of the group. The node numbers are in parentheses in the center of the nodes; the variation explained by each split is printed underneath.

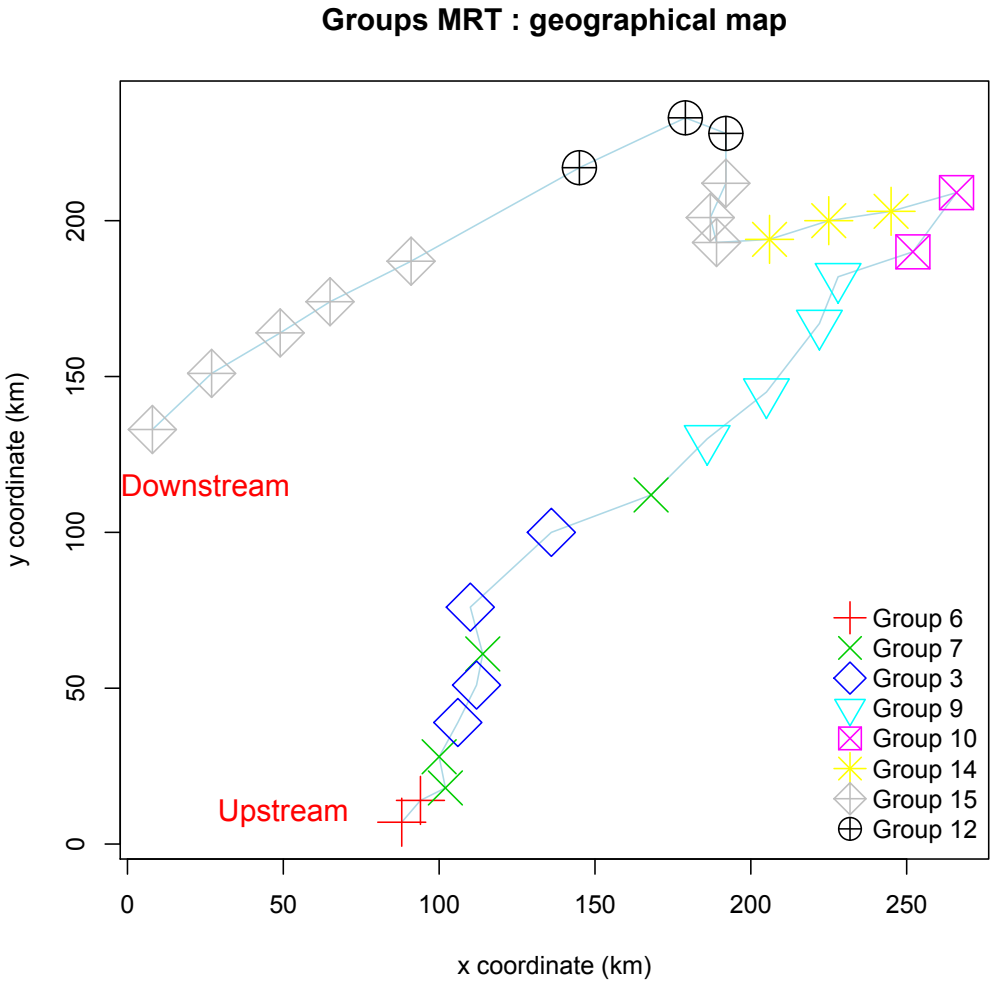


Figure A3.24: Geographical map of the MRT partition results for the Doubs' fish data set. The group numbers correspond to the numbers given in figure 23.

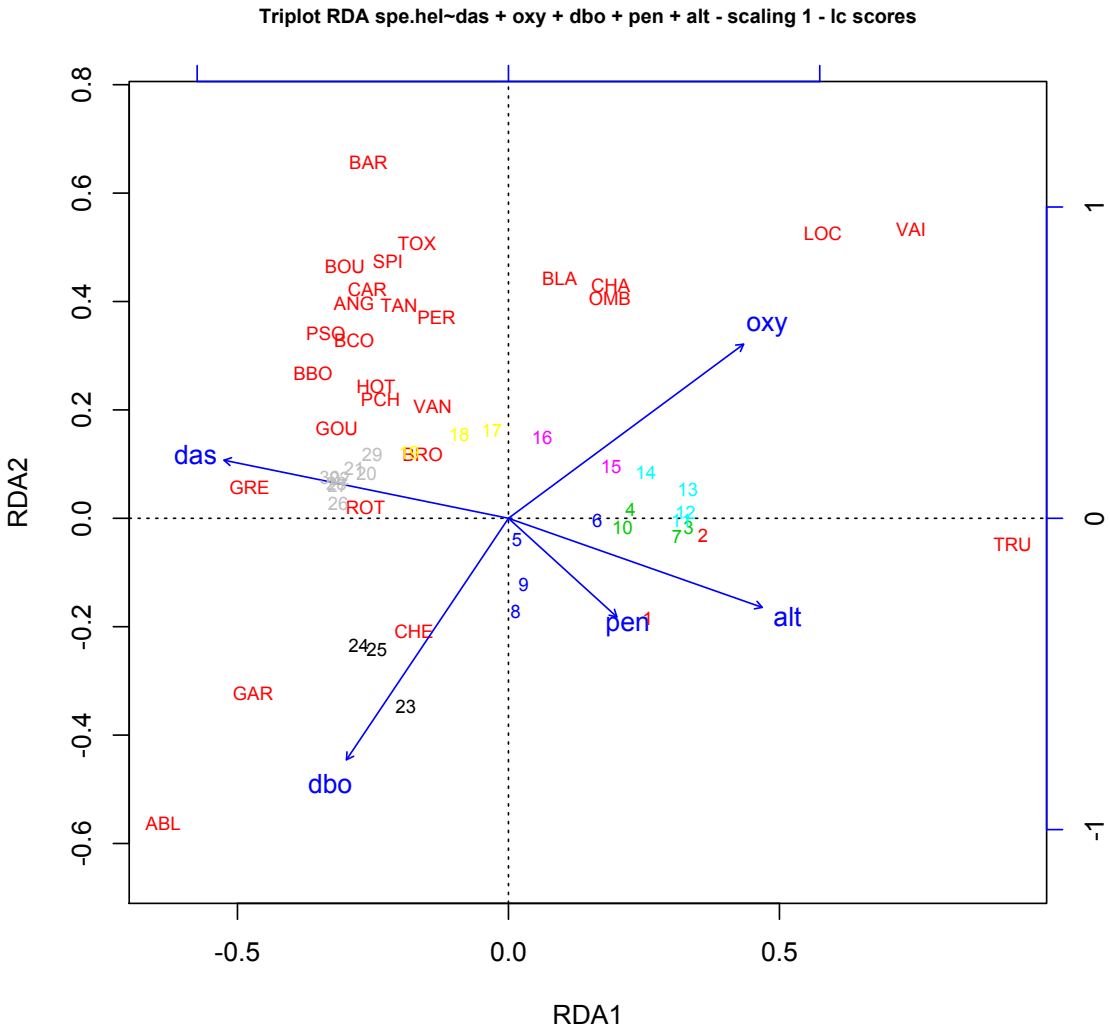


Figure A3.25: RDA analysis triplot illustration (scaling 1, 'wa' scores) of the Doubs' fish data set with forward selection on the raw explanatory variables. The sites are color-coded according to the partition of the MRT with the same colors as figure 24. The numbers corresponds to the order from the source. Species are abbreviated by three capital letters: CHA (Bullhead *Cottus gobio*), TRU (Brown trout *Salmo trutta fario*), VAI (Minnow *Phoxinus phoxinus*), LOC (Stone Loach *Nemacheilus barbatulus*), OMB (Grayling *Thymallus thymallus*), BLA (Souffia or Western Vairon *Telestes soufia agassizi*), HOT (Nase *Chondrostoma nasusi*), TOX (Southwest european nose *Chondostroma toxostoma*), VAN (Common dace *Leuciscus leuciscus*), CHE (Chub *Leuciscus cephalus cephalus*), BAR (Common barbel *Barbus barbus*), SPI (Spirlin *Spirlinus bipunctatus*), GOU (Gudgeon *Gobio gobio*), BRO (Northern pike *Esox lucius*), PER (European perch *Perca fluviatilis*), BOU (European Bitterling *Rhodeus amarus*), PSO (Pumpkinseed sunfish *Lepomis gibbosus*), ROT (Rotfedern *Scardinius erythrophtalmus*), CAR (Common carp *Cyprinus carpio*), TAN (Tench *Tinca tinca*), BCO (Common bream *Abramis brama*), PCH (Black bullhead *Ictalurus*

melas), GRE (Ruff *Acerina cernua*), GAR (Roach *Rutilus rutilus*), BBO (Silver bream *Blicca bjoerkna*), ABL (Bleak *Alburnus alburnus*), ANG (European eel *Anguilla anguilla*). Moreover, the explanatory variables selected by forward selection represented in this triplot are distance to the source (das), biological oxygen demand (dbo), slope (pen), altitude (alt) and finally dissolved oxygen (oxy).

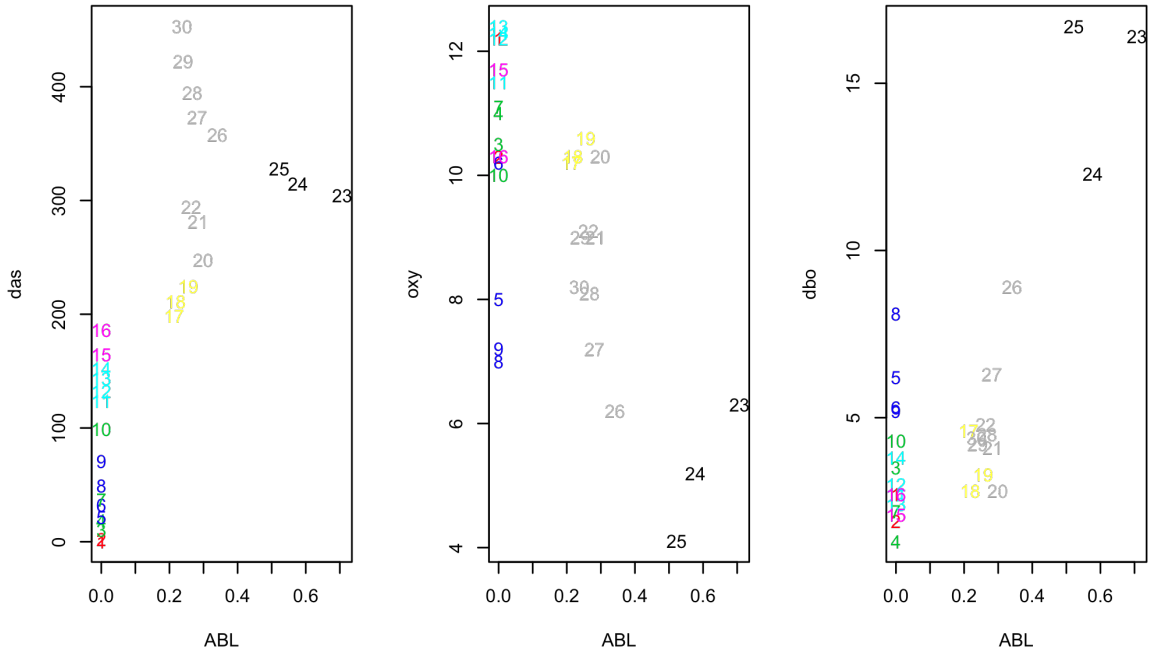


Figure A3.26: Bleak abundances from the Doubs Hellinger transformed fish data set as a function of specific explanatory variables (distance to the source (das), oxygen content (oxy) and biological demand for oxygen (dbo)). The sites are color-coded according to the partition of the MRT with the same colors as figure 24. The numbers corresponds to the order from the source.

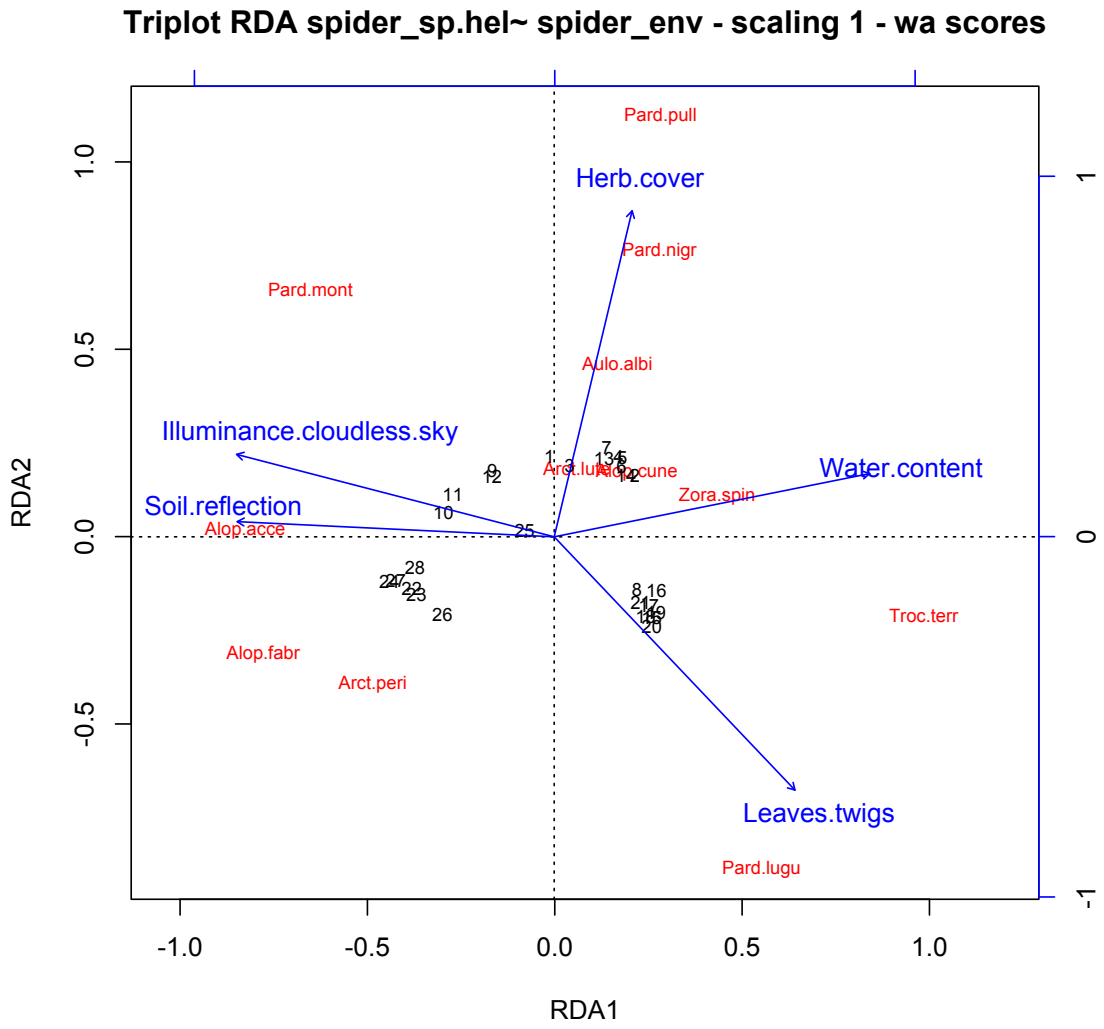


Figure A3.27: RDA analysis triplot illustration (scaling 1, 'wa' scores) of the spider data with original explanatory variables chosen by forward selection. Five explanatory variables were selected by the forward selection procedure, which were water content % of dry weight (Water.content), the cover by herb layer in % (Herb.cover), reflection of soil surface at cloudless sky × 100 (soil reflection), lux (AEG Lux-meter measure) at cloudless sky × 1000 (Illuminance.cloudless.sky) and finally cover by fallen leaves and twigs in % (Leaves.twigs). The species names are abbreviated as follows (- indicates no common name found): Alop.acce (- *Alcopecosa accentuata*), Alop.cune (- *Alopecosa cuneata*), Alop.fabr (Great fox-spider *Alopecosa fabrilis*), Arct.lute (- *Arctosa lutetiana*), Arct.peri (- *Aulonia perita*), Aulo.albi (- *Aulonia albimana*), Pard.lugu (- *Pardosa lugubris*), Pard.mont (Pin-stripe wolf-spider *Pardosa monticola*), Pard.nigr (- *Pardosa nigriceps*), Pard.pull (Common wolf spider *Pardosa pullata*), Troc.terr (Ground wolf-spider *Trochosa terricola*) and finally Zora.spin (- *Zora spinimana*).

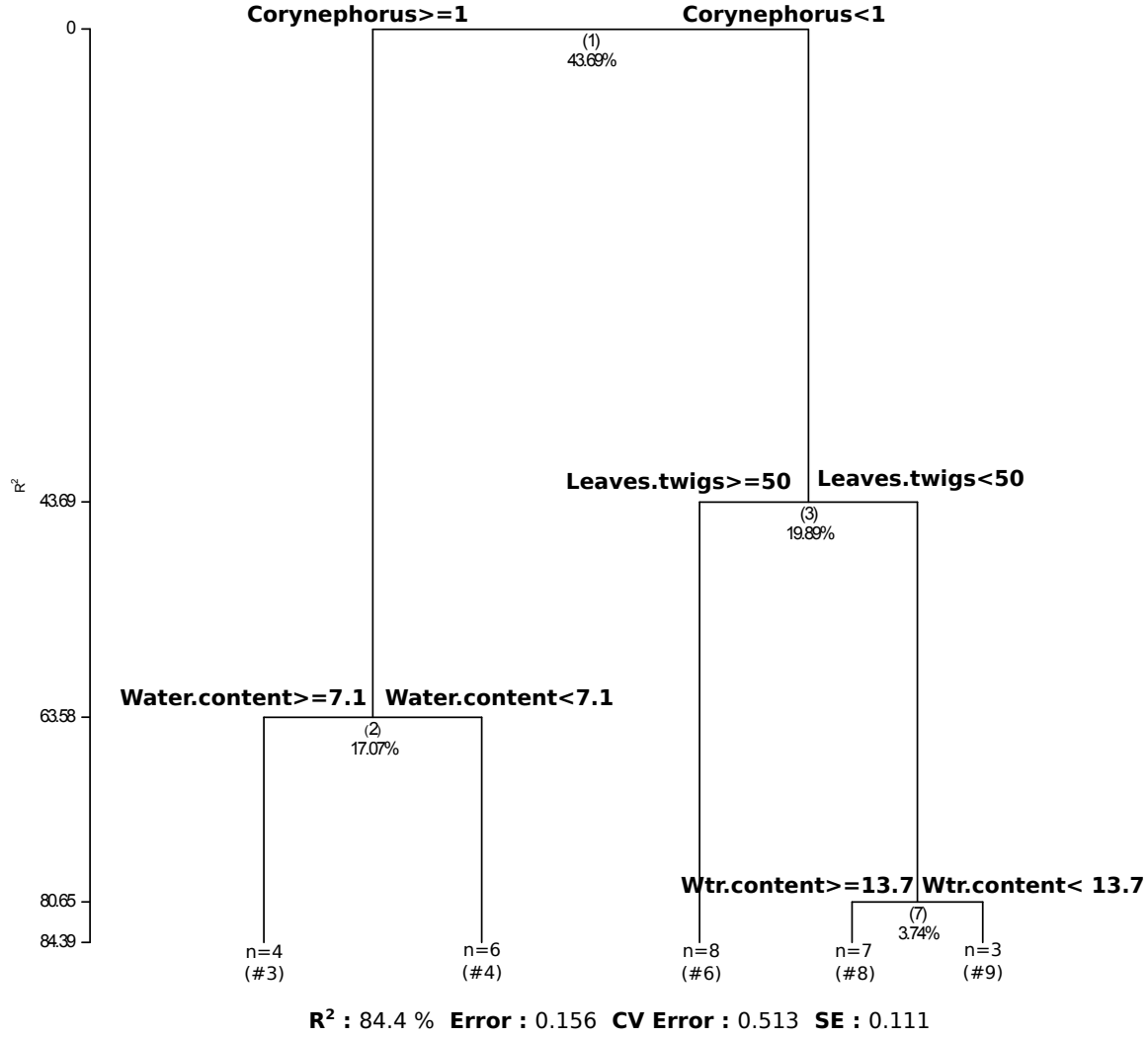


Figure A3.28: Multivariate regression tree model of the spider data set with Hellinger transformed response. This output is provided by the MRT function of MVPARTwrap. At each leaf we find the number of objects in the node and the number of the group. The node numbers are in parentheses in the center of each node; the variation explained by each split is printed underneath. Five explanatory variables were selected by the forward selection procedure, which were water content % of dry weight (Water.content or Wtr.content), cover by fallen leaves and twigs in % (Leaves.twigs) and finally cover by *Corynephorus canescens* in % (Gray clubawn grass) which is noted here as Corynephorus.

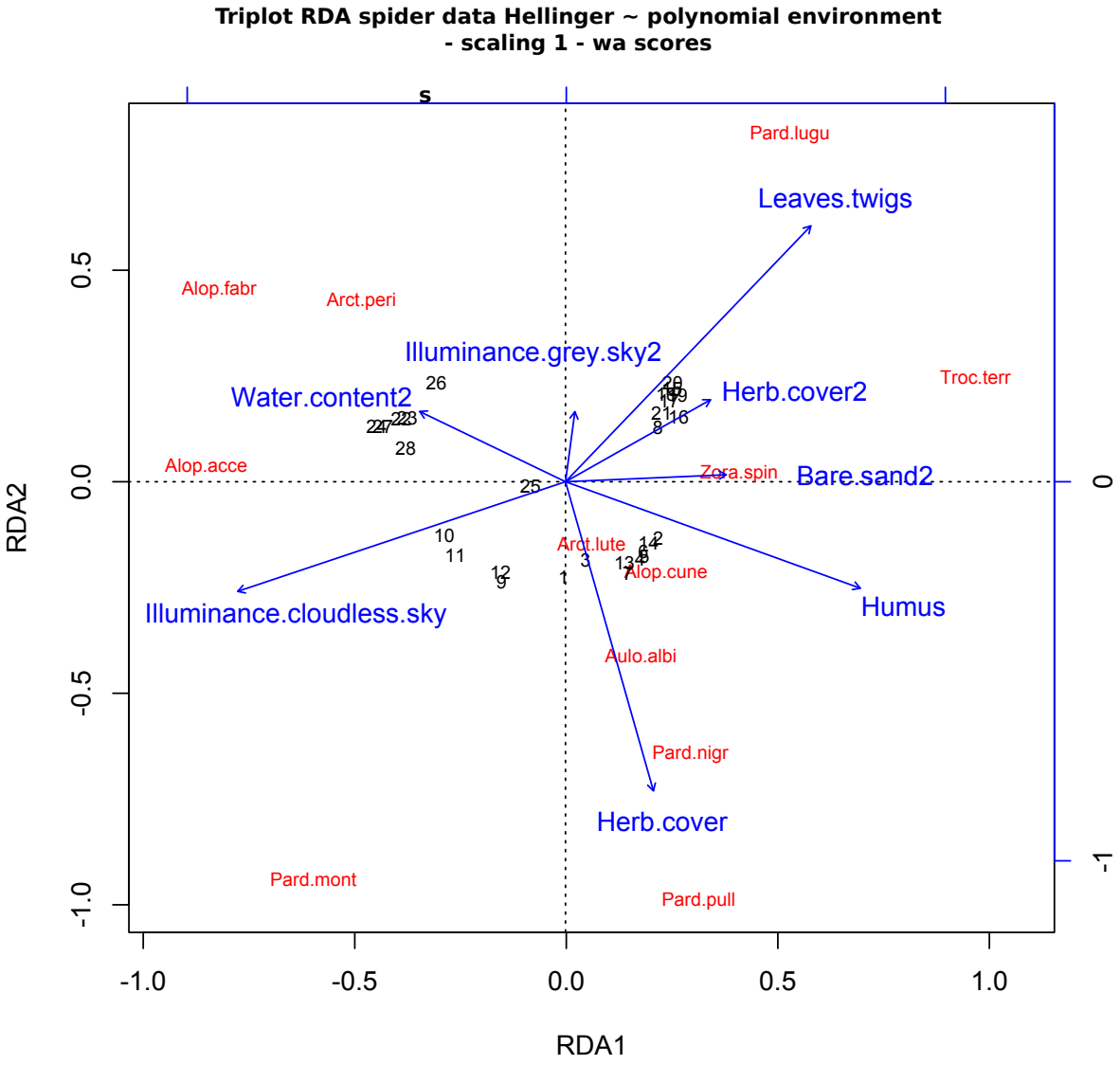


Figure A3.29: RDA analysis triplot illustration (scaling 1, 'wa' scores) of the spider data with polynomial of environment variables. Five explanatory variables were selected by the forward selection procedure, which were water content % of dry weight polynomial of degree two (Water.content2), cover by fallen leaves and twigs in % (Leaves.twigs), the cover by herb layer in % (polynomial degree one and two respectively Herb.cover and Herb.cover2), lux (AEG Lux-meter measure) at cloudless sky × 1000 (Illuminance.cloudless.sky), hummus content in % of dry weight (Humus) and finally polynomial of degree two of percentage of bare sand (Bare.sand2). The species Latin and common names (when available) are listed in the Figure 27 description.

Chapitre 4

Cascade Multivariate Regression Tree: a novel approach for modelling nested explanatory sets

Ce chapitre a été soumis pour publication dans une revue internationale : Methods in Ecology and Evolution.

Marie-Hélène Ouellette* Pierre Legendre and Daniel Borcard

Département de sciences biologiques, Université de Montréal, C.P. 6128, Succursale Centre-Ville, Montréal, Qc H3C 3J7, Canada.

*Correspondence author.

SUMMARY

1. Ecological data analysis frequently calls for the assessment of the relationship between species composition and a set of explanatory variables of interest. The assessment may have to be pursued while taking into account the influence of another set of explanatory variables. The hypothetical nature and structure of the influence of an explanatory set on the effect of a distinct explanatory set guides the proper choice of modelling methodology for a combined explanatory assessment. For example, to model the effect of an explanatory set on a response while controlling for (or in the presence of) another explanatory set, if their effects are thought to be additive and linear, partial linear (regression or canonical) analysis is adequate. These assumptions do not always fit the circumstances, however — for example when we wish to

explore explanatory data organized in a nested manner (in terms of scale for example). To study the influence of a set of explanatory variables of interest as it changes as a function of another explanatory set, a different method is required for proper explanatory assessment.

2. Here we describe a framework where the relationship between the response data and a main set of explanatory variables is not linear. It may, for example, be hypothesized to be in the form of abrupt changes in the response following thresholds of the explanatory variables, or any other non-linearizable relationship. The influence of a second set of explanatory variables is determined a posteriori, after the influence of the main explanatory set has been recognized. This is useful when one of the sets is thought to have a main effect and the second set's influence changes as a function of the first.

3. To pursue this type of assessment, we use a *cascade of multivariate regression trees* (CMRT). We ultimately decompose the total dispersion of a response matrix between two explanatory data sets in a hierarchical manner. By handling each leaf (group) resulting from the main MRT analysis as separate independent data sets in following analyses, we can separate the explanatory power of the first partition from those of the subordinate partitions computed using a second explanatory set. A preliminary biological hypothesis will guide the choice of which set of explanatory variables should be used to compute the main partition. The method could be extended to more than two explanatory data sets whose effects on the response data are hierarchical.

4. CMRT allows for the first time users to impose a nested structure to their causal hypotheses in multivariate regression tree analysis.

5. To illustrate this new procedure, we used the well-known Doubs fish and oribatid mite data sets, which are readily available in R.

6. R functions are provided in an R package (MVPARTwrap). Hence the new method of analysis can easily be applied by users.

KEYWORDS: *cascade; multivariate regression tree; ecological community*; *nested explanatory assessment*

INTRODUCTION

Modelling field data in ecology often translates into the study of the effect of more than one set of explanatory variables on a response data set (Legendre & Legendre 1998). Species assemblages, in particular, can respond to a great number of environmental factors, and a lot of these may play an important explanatory role, but their effects on the response are not necessarily independent from one another.

The most common methodologies used to assess the influence of multiple explanatory data sets in ecology are linear regression modelling and ANOVA, as well as their multivariate extensions: canonical analysis (RDA and CCA) and MANOVA (Legendre & Anderson 1999; Anderson 2001a; McArdle & Anderson 2001). In the linear modelling framework, where we want to model a response as a function of two sets of explanatory variables, we use partial linear regression in the univariate case, and partial canonical analysis in the multivariate case (partial redundancy analysis, partial RDA: Davies & Tso 1982; partial canonical correspondence analysis, partial CCA: ter Braak 1988). The effect of two or several explanatory data sets on response data can be untangled by variation partitioning (Borcard, Legendre & Drapeau 1992; Borcard & Legendre 1994; Anderson & Cribble 1998; Peres-Neto et al. 2006). The

effects of both explanatory sets are then hypothesized to be additive over the data set. Partial RDA and partial CCA both allow a constrained ordination of the response on the explanatory variables to be computed while controlling for the linear effect of a matrix of covariables **W**. In the MANOVA case, the effect of two (or more) factors is assessed, and interaction may be tested if replicates are available.

In this paper we use available statistical tools in a new combination to show how to tackle ecological data assessment in the event where the relationship between a main explanatory data set and the response is non-linear. An extreme example is when strong discontinuities in species composition exist along particular variables of a main explanatory data set. In non-linear situations, thresholds better describe the relationship between the two data sets than linear models. Subsequently, the variation of each leaf (a group at the end of the tree) depicted by the discontinuities is to be independently explained by other explanatory variables of interest in a (possibly) different manner. Thus we study the effect of both explanatory sets simultaneously by keeping in mind that the effect of one set might change as a function of the other. Multivariate regression tree analysis (MRT) is the perfect tool to undertake such a task, and we call the global procedure by the name *Cascade multivariate regression tree analysis* (CMRT).

MRT analysis has stimulated growing interest in several ecological fields during the past few years. For instance we find applications of MRT in microbial ecology (Auguet, Barberan & Casamayor 2010), limnology (Davidson et al. 2010), forestry (Chen et al. 2010), reefs studies (DeVantier et al. 2006), entomology (Koivula & Vermeulen 2005), ornithology (Ouellette et al. 2005), arachnology (Pinzón & Spence 2010) and wetland studies (Sheaves, Abrantes & Johnston 2007).

This method, introduced in the ecological literature by De'ath (2002) and Larsen & Speckman (2004), is a recursive binary partitioning algorithm that splits objects into homogenous groups in the response matrix with the groups constrained by the explanatory variables. MRT is particularly useful to detect abrupt changes in community composition along an environmental gradient, since thresholds in the explanatory variables are used to delimit the leaves in the resulting tree. In the procedure, the data set is split a large number of times to form the tree, then a pruning procedure is applied to reduce the large tree and obtain the best predictive tree size. Pruning is achieved by a resampling method called *v*-fold cross-validation (Breiman et al. 1984). First, all objects are split into *v* test subsets. Then, *v* trees are built from the *v* learning sets constructed by removing the *v* test set one at a time from the whole set of objects. All trees are fully grown, and subsequently for each tree size, cross-validation relative error is calculated as follows:

$$CVRE = \frac{\sum_{k=1}^{10}\sum_{i=1}^{n_i}\sum_{j=1}^{m}\left(y_{ij(k)} - \hat{y}_{j(k)}\right)^2}{\sum_{i=1}^{N}\sum_{j=1}^{m}\left(y_{ij} - \bar{y}_j\right)^2}$$

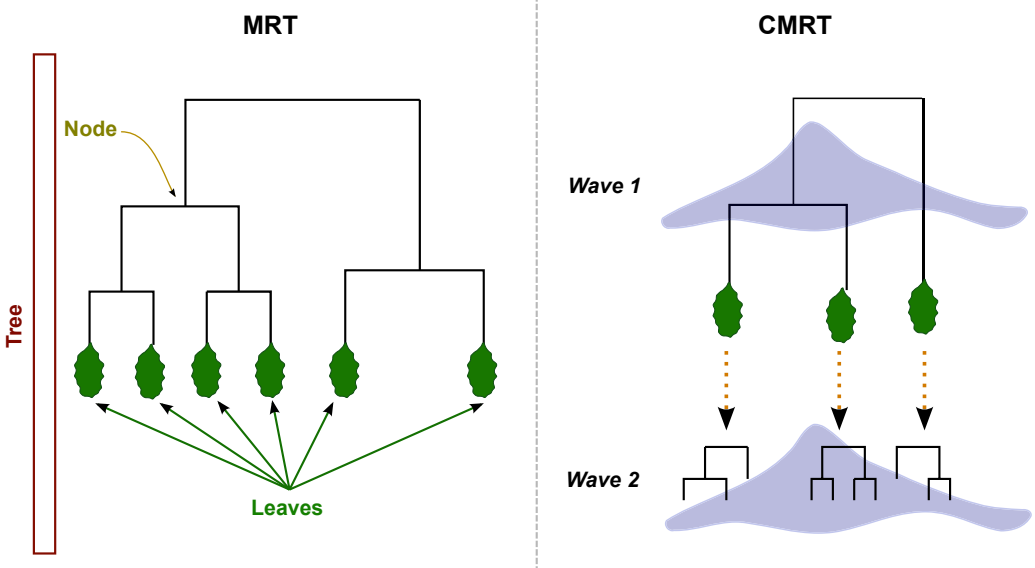
where $y_{ij(k)}$ is one observation of the test set k, $\hat{y}_{j(k)}$ is the predicted value of this observation in the k tree computed from the corresponding learning set, n_k is the number of observations in the test set k, and m is the number of variables in the response matrix **Y**. If the response data contain species abundances, the predicted response is a particular species composition, each of them corresponding to a leaf.

CMRT is a procedure that focuses on modelling the response data in the form of assemblages constrained by two sets of explanatory variables that are taken into

account in an order that reflects their hypothesized nested influence. The explanatory variables may be of any mathematical type since quantitative and qualitative explanatory variables can be used by MRT analysis, which achieves the partitioning. Moreover, because it is based on MRT analysis, this new procedure does not require that the relationships between the response and explanatory variables be linear, normally distributed, and homoscedastic. It can also deal with missing values. These features make CMRT a valuable modelling technique for ecological data, where stringent statistical assumptions are seldom met.

CMRT: THE PROCEDURE

Because CMRT is a new procedure, we first provide the necessary associated terminology. We use the word *wave* to describe each level of the nested structure imposed by the user, and the word *drop* for each response data set analysed at each level for which a tree is produced; see Box 4.1 for a review of the terminology and Figure 4.1 for a diagram of the general structure. The number of waves is the number of explanatory data sets in the user's nested structure. Before launching the procedure, it is essential to identify which of the explanatory sets will have the main effect, and which will have the subordinate effect. This decision should not be taken lightly since it strongly influences the inferences that can be drawn from the resulting model; see Discussion.



Leaf: group of objects found at the end of the tree.

Node: split of objects in two groups.

Tree: set of nodes and leaves, build by MRT algorithm.

Drop: a tree found in the CMRT gobal model. In this diagram, we have four drops. The group of objects used to build the drops are provided by the leaves of the previous wave.

Wave: a set of drops that were build from the same explanatory variables. In this diagram, we have two waves.

Subsequent drops: drops other than drop 1, subsequent from wave 1.

CMRT : A set of waves.

Box 4.1: Terminology review for MRT and CMRT analyses. There are four drops four trees in this diagram: one in wave 1 and three in wave 2.

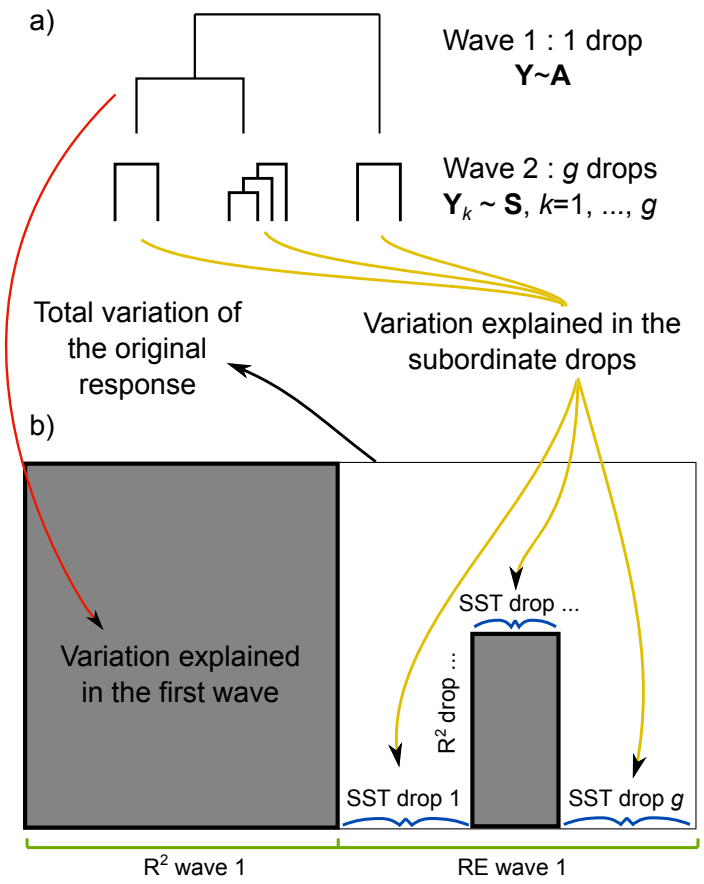


Figure 4.1: (a) Diagram of the CMRT procedure along with (b) a general R^2 diagram. In (b) we depict the variation explained by the whole cascade in a rectangle whose area (left + right portions) represents the total variation in the response data (100%). The shaded area on the left represents the variation of the response data explained by the first wave (main analysis). The shaded area or areas (there may be more than one) on the right represent the variation explained by the subordinate drops of the second wave. For each shaded rectangle in the white area on the right, its width represents the proportion of the relative error (RE, unexplained variation) of the first wave while its height represents the R^2 of the subsequent response explained by the subordinate drop. The white area is the variation that remains unexplained at the end of the waves.

Let **Y** be the response matrix whereas **A** and **S** are respectively the main and subordinate explanatory tables. Several criteria may be used to decide which are the main and subordinate explanatory sets. The criterion may be scale: large, medium and small scales, or else landscape and microhabitat scales. The hierarchy could also be based on the nature of the explanatory data sets, for example: morphometry of the river (main) and land use impact (subordinate). See the *Hierarchical hypotheses in ecology* subsection of the Discussion for more examples. In the procedure, an MRT model is first computed with **Y** as the response and **A** as the explanatory table. Cross-validation is carried out to prune the tree: this is the first wave of the cascade. The first wave thus consists of analyzing a single drop through an MRT model. The explanatory set is hypothesized to vary as a function of **A** (main effect), identifying the groups of sites with the most homogeneous species composition along the studied gradients at large scale.

It is important for this first wave of analysis to set the complexity parameter high enough to identify only the largest variation in species composition. The complexity parameter of an MRT model is the minimum contribution to the R^2 of the tree for a split to be considered. The value of the complexity parameter selected for the first drop will shape the partition produced by this first wave by limiting the number of splits, and it is left at the user's discretion: a split will not be performed unless it explains at least the chosen R^2 value.

Let $\boldsymbol{g}$ be the number of leaves resulting from the first wave. In a second step, the response variation in each leaf, noted $\mathbf{Y}_{\boldsymbol{k}}$, $k = 1, \ldots, g$, is modelled independently with the **S** explanatory table to form the subordinate drops. For these drops, the complexity parameter may be reduced to the usual value (the default value is 0.01 in

the ***mvpart()*** R function; it is passed from the ***rpart.control()*** R function; both functions are found in the MVPART package) as the second wave is intended to model finer variation in species composition.

The combined model, called the cascade, is exactly that: a cascade of models, depicting in a nested manner the explanatory power of two sets of explanatory variables (Figure 4.1). The distinctiveness of CMRT analysis lies in its ability to force the order in which two or more explanatory data sets are used in MRT analyses. Two general conclusions may emerge from a cascade: either the explanatory variables and splits are the same for all leaves identified in the first wave, which means that the subordinate effect is the same over all subordinate data sets, or they are not. Therefore the sequence of subordinate drops may be assessed to identify splits and explanatory variable differences between drops for a subjective interaction investigation analogous to a test of interaction in MANOVA.

R^2 PARTITION

A coefficient of determination (R^2) is obtained for the global analysis; it is depicted in the diagram corresponding to wave 1 (Figure 4.1b). The R^2 of a single MRT tree (or a drop) is 1 minus the relative error defined by De'ath (2002). Thus a single coefficient of determination (R^2) can be computed for each drop. The subordinate drops are computed from the unexplained variation of the first drop; to be able to sum the explained variations of the subordinate drops to the main drop R^2, we need a common denominator, which is the total variation of the response. To do so, we weight the subordinate R^2 by the proportion of variation of the response data not explained by the first drop. The explained variation of the first and second waves can

now be added and the global R^2 can be partitioned between the main drop and the subordinate drops. It is the independence among the subordinate drops that makes the global R^2 calculation admissible.

In the diagram provided by the ***CasMRTR2()*** function of the MVPARTWRAP package, the surface area of the outer rectangle containing all other rectangles represents the total variation of the response data, and should be thought of as a rectangle of unit area. Each drop has a shaded box that represents the portion of variation of the original response that it explains. The box for the drop of the first wave is at the far left. Its width represents the R^2 of the first drop and its height represents 1, so its surface area is proportional to the explained variation by the first drop, or the first wave; see Box 4.1 for an illustration of this equivalence. To the right of this box are shown the boxes for the subordinate drops (the drops of wave 2 in the example). The widths of these boxes are proportional to the unexplained variation of the response table in the corresponding leaves of the first drop, so their sum is equal to the relative error of the first drop. In turn, this means that the length of the bottom side of the large rectangle is one. The heights of the rectangles represent the R^2 of the subordinate drops. The surface of these rectangles is thus proportional to the explained variation of the original response data in each subordinate drop.

SOFTWARE

The `MVPARTwrap` R package is available on R-Forge at the address http://r-forge.r-project.org/projects/mvpartwrap/.

CASE STUDIES

We illustrate the CMRT procedure by using two data sets that have been studied with different types of analyses by Borcard, Gillet & Legendre (2011) and are readily available in R (R Development Core Team 2010). For both case studies, a complexity parameter of 0.10 was used for the first wave, and the usual 0.01 value was used for the second wave. Also, both community response matrices were Hellinger transformed prior to the analysis (Legendre & Gallagher 2001).

DOUBS RIVER FISH

The Doubs River fish data were collected by Verneaux (1973) who considered the fish species composition to be an ecological indicator of the different water bodies along the Doubs River in the Jura Mountains, near the France-Switzerland border. The data set presented here is a subset of the original data in Verneaux' thesis, merely 35 sites, described by three data tables: the fish species composition, explanatory variables describing the water quality and river morphology, and finally the spatial coordinates of the sites. It is provided as electronic material with the book of Borcard, Gillet & Legendre (2011). In the original MRT analysis (Figure 4.2), the distance to the source provides the first split; actually, this split identifies two zones that had been identified by Verneaux as the Salmonid region (upstream) and the Cyprinid region (downstream). To illustrate the distinctiveness of the CMRT procedure, we use the morphological variables 'mean discharge' and 'slope' as the main explanatory set and the physical and chemical variables (calcium concentration (hardness), pH, phosphate, nitrate, ammonium, dissolved oxygen and biological

oxygen demand) as the subordinate explanatory set, in order to seek new insights about the ecology of this river.

The resulting cascade is shown in Figure 4.3. In the first drop, the sites are split by a mean discharge of 23.65 m^3/s. On the left is the Cyprinid region of Verneaux (1973) (group 3) whereas the Salmonid region (group 2) is found in the right-hand branch of the tree. Indicator species analysis (Dufrêne & Legendre 1997) with Holm correction for multiple testing shows that the Salmonid region is characterized by the brown trout (*Salmo trutta fario,* a Salmonid) and the common minnow (*Phoxinus phoxinus*, a Cyprinid) as indicator species. The Cyprinid region has the bleak (*Alburnus alburnus*), the common nase (*Chondrostoma nasus*), the ruff (*Acerina cernua*), the pumpkinseed sunfish (*Lepomis gibbosus*), the European bitterling (*Rhodeus amarus*), the European eel (*Anguilla anguilla*), the roach (*Rutilus rutilus*), the spirlin (*Spirlinus bipunctatus*), the common carp (*Cyprinus carpio*), the white bream (*Blicca bjoerkna*), the common barbell (*Barbus barbus*), the common bream (*Abramis brama*), the rudd (*Scardinius erythrophthalmus*) and the south-west European nase (*Chondrostoma toxostoma*) as indicator species.

Within each zone identified by the first drop, the water quality variables are used in the subordinate analyses to identify and explain finer differences in species composition. No further splits were found in the Salmonid region (v-fold cross-validation pointed to one group). It was not the case for the Cyprinid region, which showed three species assemblages responding to two explanatory variables: ammonium concentration and dissolved oxygen; see Figure 4.3 for a map of the sites along the river and the cascade of analyses, and Figure 4.4 for a summary of the explained variation.

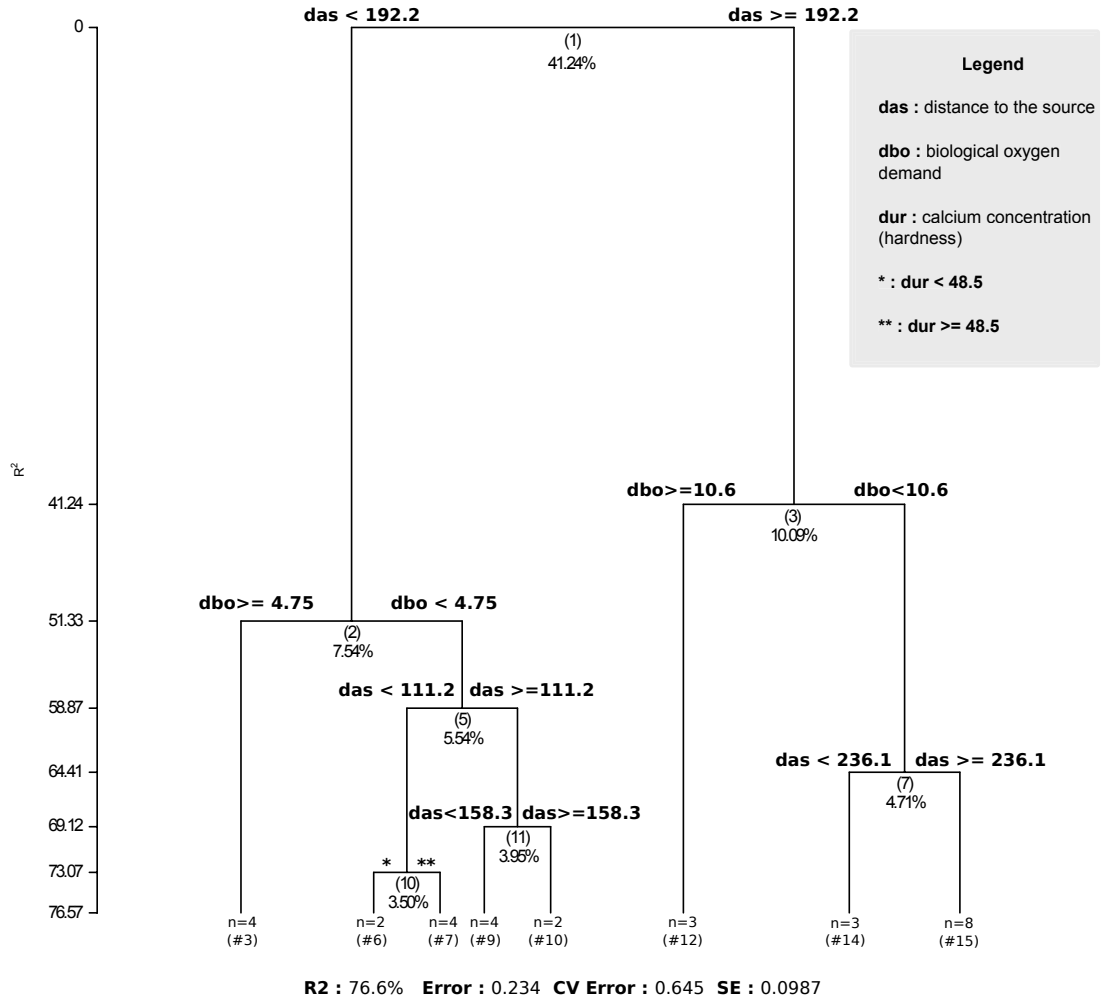


Figure 4.2: Original MRT analysis of the Doubs River fish data. For each node, its identification number in parentheses, e.g. (1), corresponds to the one found in the `summary.MRT` function of the `MVPARTwrap`. Under the number is found the percentage of explained variation. For each leaf, the number in parentheses, e.g. (#3), is the one found in the `summary.MRT` function of the `MVPARTwrap` package; the number of objects in the leaf is also shown, e.g. n = 4.

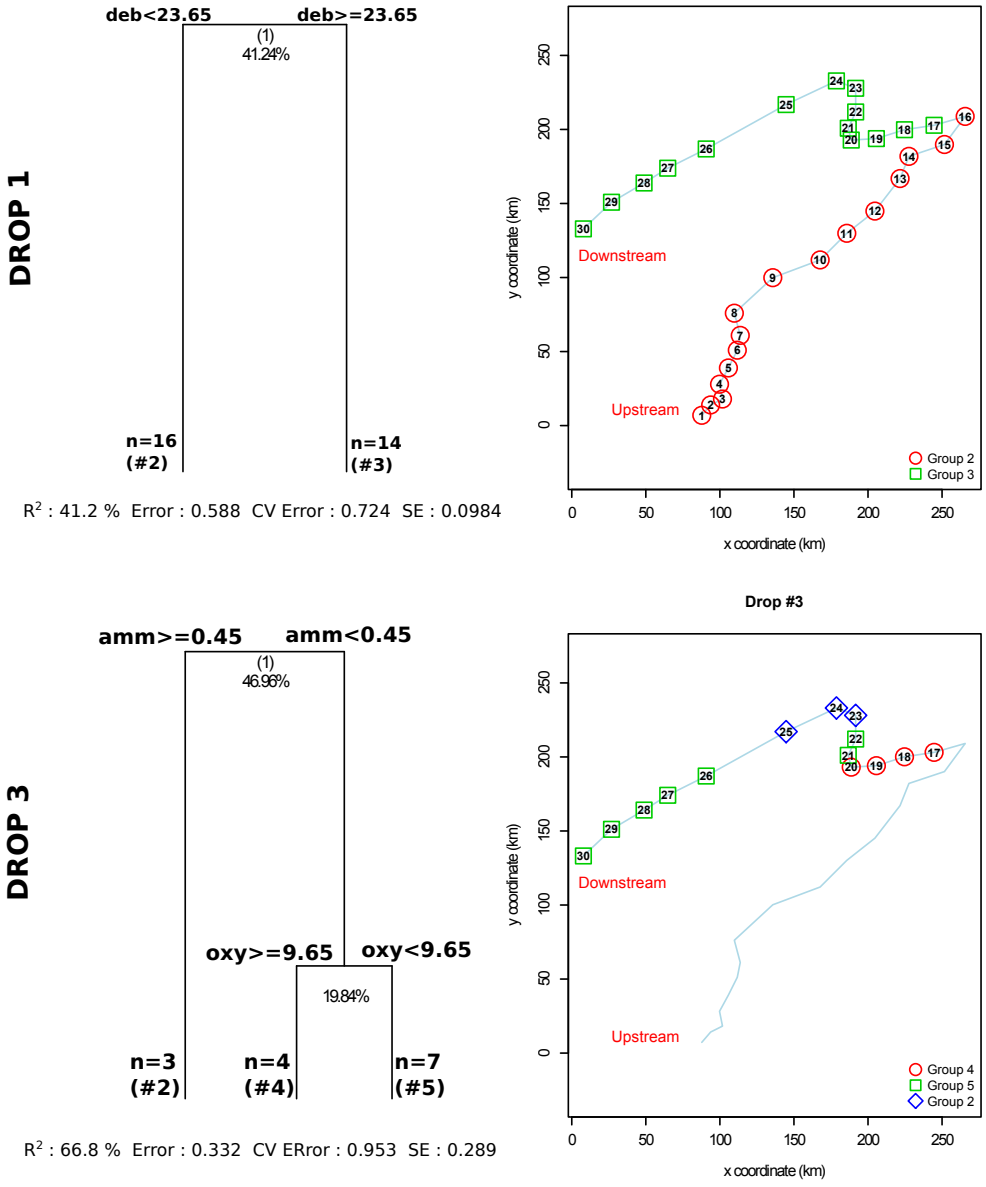


Figure 4.3: CMRT analysis results for the Doubs River data. Each drop is on the left; on the right we find the corresponding geographical map of the groups. The number (#) and size (n) of each leaf are shown. The number and percentage of explained variation are given for each node. Three explanatory variables appear in this figure: mean minimum discharge (deb), ammonium concentration (amm) and dissolved oxygen (oxy).

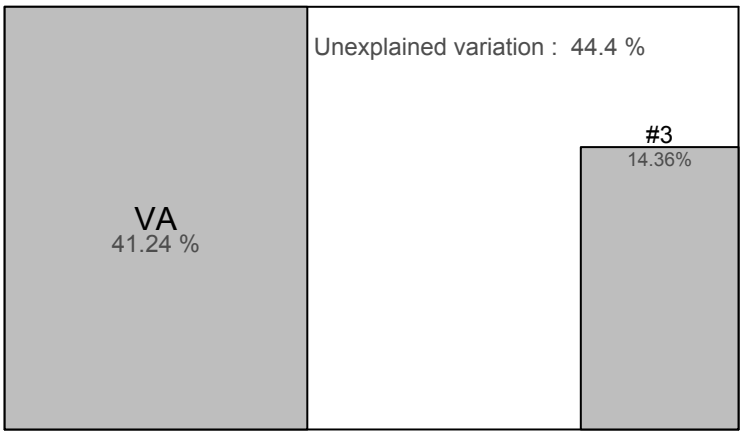


Figure 4.4: Output of the ***CasMRTR2()*** function for the Doubs River fish data. The global R^2 is 55.6%, the portion of the global R^2 explained by the subordinate drop 3 is 14.36%, and only that one has any extra variation to be explained. The drop number corresponds to the number of the leaf in the tree of the first drop (Figure 4.3). The VA percentage (41.24%) is the variation explained by the main explanatory variable, which happens to be the 'mean discharge' variables.

Three groups were depicted in the tree of drop 3. Group 2 of that tree contains sites 23-25, characterized by large concentrations of ammonium (>=0.45 mg/L) and, by correlation, by large concentrations of phosphorus (r = 0.9695) and high biological oxygen demand (r = 0.8858); these two variables, which would produce the same split, are not shown in the tree. The bleak *Alburnus alburnus*, the chub *Leuciscus cephalus cephalus*, and the roach *Rutilus rutilus* are the indicator species of this group (sites 23-25). The bleak is present at sites 21-30 but particularly successful at the highly eutrophized sites 23-25. This species feeds on zooplankton near the surface (Horppila & Kairesalo 1992) which is, for this species, an important habitat for feeding (de Nie 1987) and to lay eggs (Pihu 1996). Thus the indicator value of this species corresponds to the presence of macrophytes, which are in turn associated with high nutrient concentrations (Carr & Chambers 1998). The same applies to the roach for which macrophytes are also an important feeding habitat. As shown by Borcard, Gillet & Legendre (2011, Fig 2.5), this group is found in a zone where there is a significant drop in species richness and where we are more likely to find perturbation-tolerant species.

Group 4, which includes sites 17-20, is also part of drop 3. It is characterized by high levels of dissolved oxygen (>=9.65 mg/L) and small concentrations of ammonium (< 0.45 mg/L). The indicator species in this case are the stone loach (*Nemacheilus barbatulus*, Kottelat & Freyhof 2007), the western vairone (*Telestes soufia agassizi*, Kottelat & Freyhof 2007), the common minnow (*Phoxinus phoxinus,* DORIS 30/7/2010), the southwest European nase (*Chondrostoma toxostoma*, Chappaz, Brun & G. 1989), the spirlin (*Spirlinus bipunctatus*, (Kottelat & Freyhof 2007)) and the common dace *Leuciscus leuciscus* (DORIS 25/2/2010). All these

species have a common preference for intermediate to high oxygen levels (see associated references).

Lastly, from drop 3 we get group 5, which is characterized by low dissolved oxygen levels (< 9.65 mg/L) and small concentrations of ammonium (< 0.45 mg/L). Low dissolved oxygen levels are found in stagnant turbid waters linked to muddy bed, to which all the following species are indicators. First, the European eel (*Anguilla anguilla*) is found near river mouths; this species migrates to the sea for reproduction, and prefers to live close to the bottom in mud or crevasses (Deelder 1984). The bream (*Abramis brama*) prefers slow-flowing waters (Kottelat & Freyhof 2007) and the catfish (*Ictalurus melas*) is found in slow current, pools, and backwaters (Page & Burr 1991), just like the northern pike (*Esox lucius*) (Crossman 1996); *Acerina cernua* (or *Gymnocephalus cernua*) is favoured by eutrophic conditions (Kottelat & Freyhof 2007). The carp (*Cyprinus carpio)* prefers warm, deep, slow to still waters (Kottelat & Freyhof 2007), the silver bream (*Blicca bjoerkna)* still waters (Kottelat & Freyhof 2007), and the pumpkinseed (*Lepomis gibbosus*) vegetated pools (Page & Burr 1991).

In summary, it was not possible to find further splits in the Salmonid region using the physical and chemical explanatory variables. For the Cyprinid region, however, the ammonium and dissolved oxygen variables delimited first a polluted region, sites 23-25. Then, among the less polluted sites, two groups were discriminated by the oxygen level, which is a proxy for less agitated waters, which in turn is a proxy for the type of river bed. Our understanding of the fish communities along the Doubs River was enhanced by CMRT analysis that allowed us to impose a nested structure to our species-environment causal hypotheses.

ORIBATID MITE

The second data set consists of three data tables (species composition of oribatid mites, micro-environmental variables, and spatial coordinates) extracted from 70 peat moss cores collected by Borcard & Legendre (1994) in a small area in the peat blanket surrounding Lac Geai (Québec, Canada), going from the edge of the forest to the open water of this bog lake. The sampling area is only 2.5 m x 10 m in size; the small size of these arthropods calls for small sampling units and extent. In the "non-nested" analysis run with all variables, water content (g/dm^3) was selected for the first split of the MRT (Figure 4.5). Since oribatids are not aquatic, in this extremely wet environment some oribatids will prefers more or less water which confers this explanatory variable a direct effect. The water content also has an indirect effect on the biota by structuring the vegetation for example. Other substrate and micro-environmental variables are available as explanatory variables, in particular the density of the substrate (g/dm^3), type of substrate (7 unordered classes), shrub density (none, few, many) and microtopography (blanket-hummock). This data set is available in the VEGAN R package as well as in the electronic material provided with the book of Borcard, Gillet & Legendre (2011).

In the CMRT analysis, we use the variable 'shrub' as the main effect because shrub density provides a particular microclimate and microsubstrate modification for the mites: it increases shade and tops the original substrate (sphagnum moss) with additional woody matter. The first drop of the cascade divides the sites in two groups separating the sites with no shrubs, with indicator morphospecies *Trimalaconothrus sp.*, *Tectocepheus cf. vietsi* and *Ceratozetidae sp3*, from the sites with a few or many

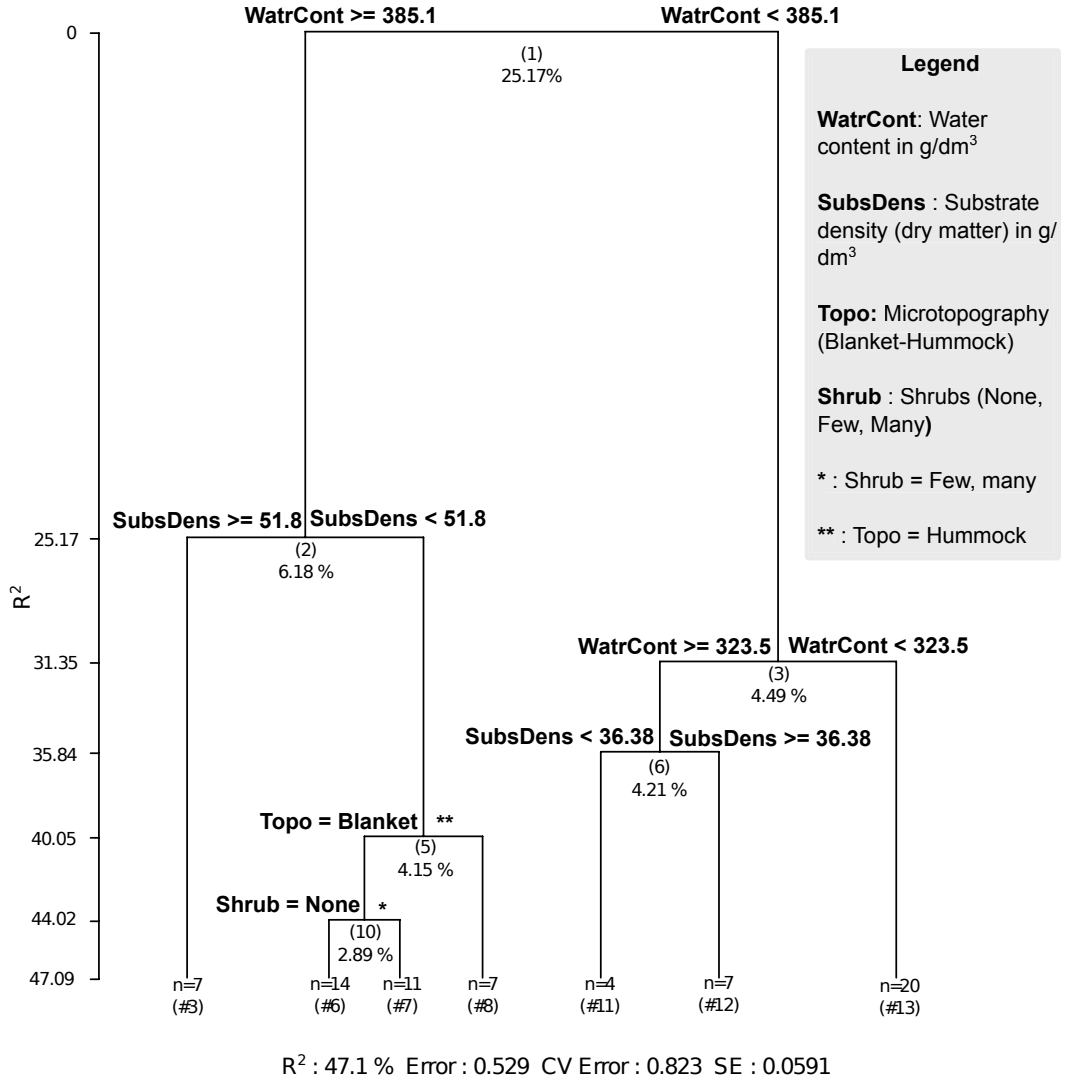


Figure 4.5: MRT analysis for the oribatid mite data. Details: see legends of Figs. 4.2 and 4.3.

shrubs (indicator morphospecies *Tectocepheus velatus*, *Malaconothrus cf. egregius, Oppiella nova, Fuscozetes setosus, Hypochthoniella sp1 & sp2,* and *Galumnidae).*

In subordinate drops 2 and 3 (Figure 4.6), different explanatory variables were used to split each subset of sites in two: for the sites without shrubs, substrate density is the splitting explanatory variable and the splitting point is 50.36 g/dm^3, and for sites with shrubs, water content at 385.1 g/dm^3 is the delimiter. For the sites without shrubs, we have only one indicator morphospecies per group: for low substrate density we have *Oppiella nova* and for high substrate density *Trhypochthonius cf. tectorum*. For the sites with shrubs and high water content, the indicator morphospecies are *Nanhermannia coronata*, *Limnozetes rugosus* and *Limnozetes cf. ciliatus*, whereas for low water content we have *Tectocepheus velatus*, *Fuscozetes setosus*, *Hypochthoniella sp. 2* and *Rhysotritia ardua*. After forcing the shrub variable at the top of the model, the R^2 of the first drop is low (16.3, see Figure 4.7) and the CVRE is high (0.94). Yet, we are still able to extract new insight from the cascade, not available in the global MRT: where there is no shrub, substrate density has stronger control over the species composition, whereas where shrubs are present, water content is the most discriminating explanatory variable.

DISCUSSION

GENERAL REMARKS ON THE PROCEDURE

CMRT offers the opportunity to address ecological hypotheses in a preferential order, allowing one to override the original explanatory order of the variables presented in MRT analysis to explore specific avenues by testing the influence of precise variables on the response data. The peculiarity of the CMRT procedure resides in the

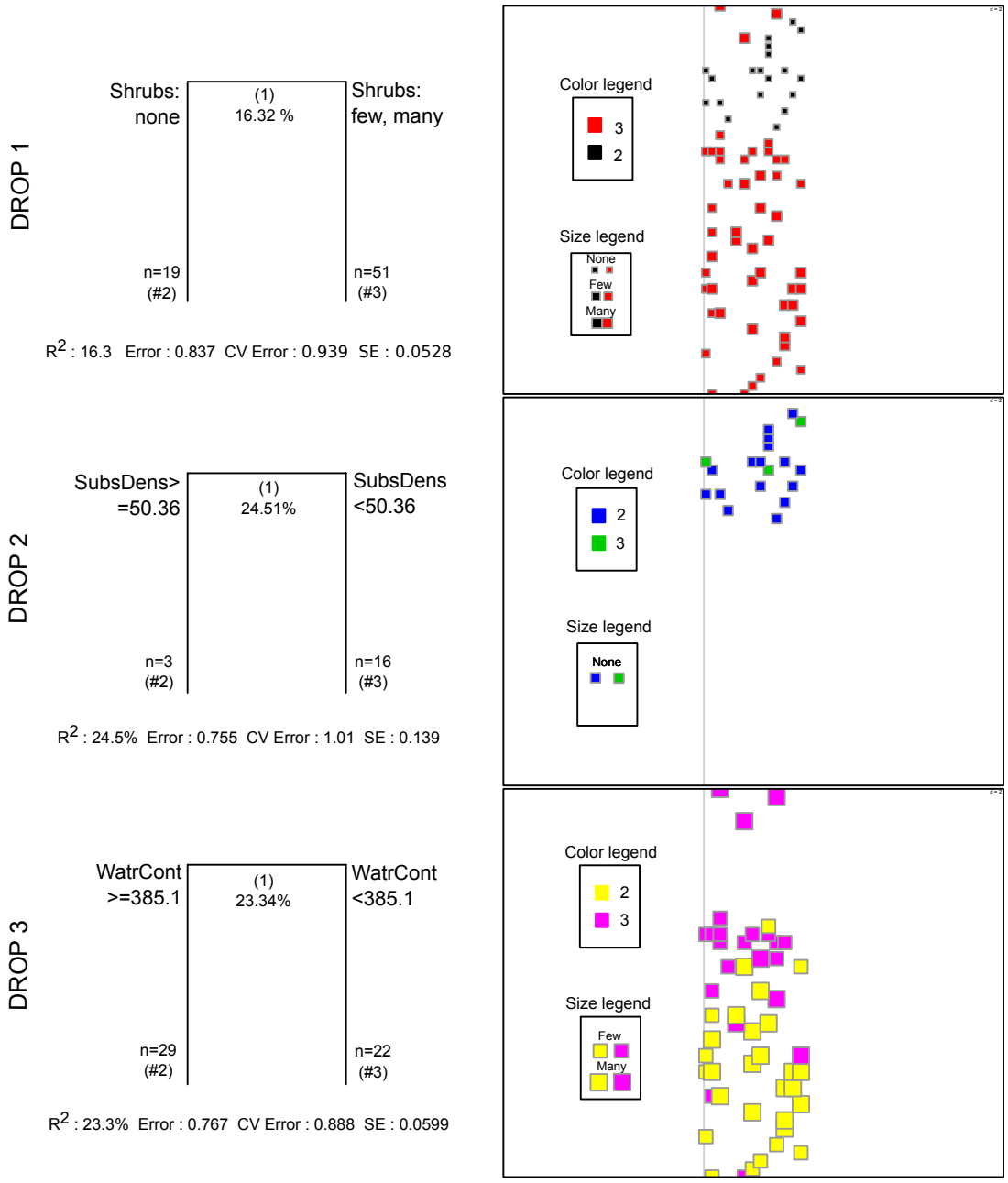


Figure 4.6: Summary of the CMRT analysis results for the oribatid mite data with the explanatory variable shrub as the primary (main) effect. Details: see legend of Figs. 4.1 and 4.4. The explanatory variables used to split the objects were the shrub states (none, few, many; the variable is noted 'Shrubs'), the substrate density (dry matter) in g/dm^3 noted 'SubsDens', and finally the water content in g/dm^3 noted 'WatrCont'.

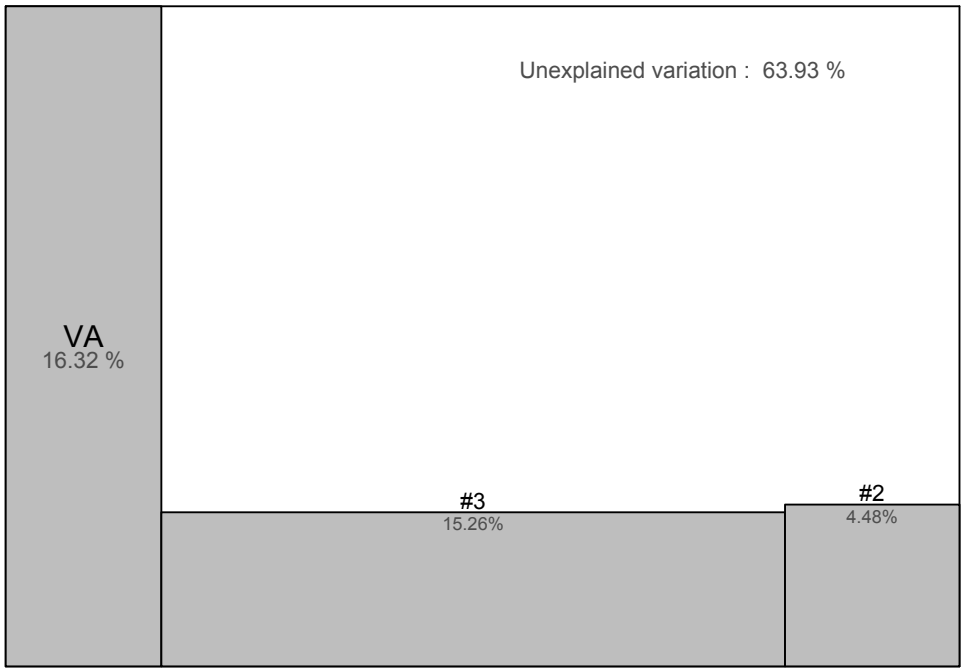


Figure 4.7: Output of the ***CasMRTR2()*** function for the oribatid mite data. The global R^2 is 36.07%; the portion of the global R^2 explained by subordinate drops 2 and 3 together is 19.74. The VA percentage (16.32%) is the proportion of the response variation explained by the main explanatory variable, which happens to be shrub presence or absence.

possibility to pre-select the explanatory set of variables that will be used to compute the first few bipartitions. Ultimately, the cascade provides new insights on the data structure that would not have been available in simple MRT analysis. In order to exploit the CMRT procedure to its fullest potential, the selected explanatory variables for the first wave should be different from the first bipartition of the simple MRT; if it was the same, it would depict the same pattern. It is more interesting to consider some other hypothesis in the CMRT procedure. Actually, if we chose to use in CMRT the original first explanatory variables identified by the simple MRT model, not only would the resulting CMRT model be the same as the MRT result, but it would have a smaller number of leaves because the independent cross-validations conducted in the drops would have reduced power.

In the linear procedures — partial linear regression and canonical analysis (RDA) — where we include the use of covariables, the use of residuals is necessary to partial out the variation explained by one of the explanatory sets (Legendre, Oksanen & ter Braak 2011, Legendre & Legendre 2012). Here, as each leaf of the first wave is treated and modelled separately by the subordinate set of explanatory variables, there is no need to use residuals. Actually, if we used the residuals of the first wave for the subordinate analyses, we would obtain exactly the same cascade structure and R^2 as with the original data; thus this practice is useless.

THE CASE STUDIES

In the Doubs River CMRT, we forced the order in which the explanatory variables were used in the cascade: river morphology was used in the main analysis and the physical and chemical variables of the water in the subordinate analyses. There were some undeniable similarities between the original simple MRT and the

CMRT model. First, the minimum average debit level of 23.65 m^3/s was used for the first split in the CMRT, delimiting exactly the Salmonid and Cyprinid regions; the distance to the source of 192.2 km had produced the same split in the simple MRT analysis. It is not surprising that we found this similarity in the results: by choosing the morphology of the river as the main driver, we obtained the same structure for the first split as the simple MRT analysis, even if we did not include the distance to the source. The strong Kendall correlation (which is well suited for the MRT setting) between these variables (τ(mean discharge, distance to the source) = 0.9540) made it impossible for any other structure to emerge in the first split. Despite this redundancy, we ultimately obtained different results by forcing the subordinate explanatory variables to be of a physical or chemical nature. The Salmonid region is not further split even if in the simple MRT, biological oxygen demand was the next explanatory variable. We have to attribute this outcome to the lower power of individual cross-validations conducted on drops in the CMRT analysis. In this case, the physical and chemical explanatory variables had no predictive power in the Salmonid region. Another difference between simple MRT and CMRT is found in the node delimited by the oxygen level in CMRT: site 20 moves from the lower group (in MRT) to the upper group (in CMRT). This modification of the original configuration only costs 0.44% of explanation power. By examining the data more closely, we find that these sites (17-20) are the only ones with high oxygen levels (10.2-10.6 mg/L) combined with high to moderate ammonium concentrations (0.15-0.30 mg/L). This configuration is a better representation of the subordinate effect of the physical and chemical variables.

In the oribatid mite case, the CMRT and MRT models differed substantially because of the imposed order of the variables in the CMRT analysis. Even if shrub density was not as important as water content in terms of explanation power, separating the data into two groups on the shrub basis had interesting consequences in the subordinate splits. The environmental constraints seemed different in areas with and without shrubs.

NESTED HYPOTHESES IN ECOLOGY

CMRT allows for the first time users to impose a nested structure to their causal hypotheses in multivariate regression tree analysis. Several ecological studies include a natural hierarchical explanatory configuration. For instance, a land use impact study of communities (e.g. fish, phytoplankton, zooplankton) could include explanatory variables about the lake or river morphometry as the main driver along with land use impact variables as the subordinate effect. With the CMRT procedure, inherently, the assessments can be conducted while considering that for each of the groups identified by the morphometry explanatory data, the subordinate effect of land use impact can be studied and identified.

In the analysis of time series, one can use the time sequence as the basis for a primary segmentation (wave 1 analysis) of the data in CMRT, followed by secondary analysis of each segment using environmental variables. The same could be done for a spatial transect. The Doubs River data, which form a spatial series along the course of the river, could be analyzed in that way. Segmentation of the river by MRT, which corresponds to wave 1 of this type of analysis, is shown as an example in Section 4.11.5 of Borcard, Gillet & Legendre (2011). For surveys conducted on a two-

dimensional geographic map, the primary segmentation could be done by spatially-constrained clustering.

Another possible application of CMRT is for space-time surveys. Legendre, De Cáceres & Borcard (2010) showed how one could test the space-time interaction in this type of survey for univariate or multivariate response data. (1) If the interaction is not significant, fairly homogeneous space-time blocks of observations can be identified by wave 1 analysis in CMRT, followed by secondary (wave 2) separate analysis of each block using environmental variables. (2) If a significant interaction between space and time is identified, it indicates that the spatial distribution of the response data, e.g. species, has changed through time (or, *mutatis mutandis*, that the species composition has changed through time at the different sampling sites). In that case, the surveys conducted at different times should be studied separately by CMRT: the observations of each time should be segmented through space in wave 1 analysis, followed by wave 2 analysis of each segment using environmental variables; and *mutatis mutandis* for the time analysis of each separate site.

In some applications, the nested structure may be more or less obvious. For space-time studies, time or space can be used as the main set of explanatory variables. (1) Let us explore a hypothetical situation where tree community composition has been collected in a forest (space) over time, and the study includes the evolution of the distribution of a potentially invasive species. In this case, space will be used as the primary factor. By doing so we isolate geographically contiguous sites that are the most similar at all times. Subsequently, each of these contiguous groups of sites with similar assemblages through space may respond differently in time to disturbances:

for example a drought could boost the invasive ability of a species. The secondary analysis will be done with the explanatory temporal variable. (2) Let us now suppose that our main interest is to study the effect of an unusually long drought. In this case we would use time as the main factor to first focus on the evolution of the species composition through time, pointing perhaps at main extinction events due to this drought. Subsequently, we would study each assemblage identified along the time line and see how they behave in space, or with respect to environmental factors that may condition the structure of the community through space: we may observe a large jump (positive or negative) in the number of invaded sites, that number evolving through time.

EXTENSIONS OF THE CASCADE

The procedure described in this paper was solely based on MRT. It is possible to pursue a cascade analysis using other methods. For example, the first drop may come from a partition either constructed with another method or simply known by previous knowledge of the data. A linear model, if the assumptions of such a procedure are met, may also be used to model the subordinate drops. Thus a mixture of modelling procedures may be used in the framework. The explained variation still holds because the subordinate analyses are independently pursued in each drop and the calculation of an R^2 in each of the independent analyses is properly defined. Moreover, more than two waves could in theory be used. This would require that the data set be large in order to have some variation left to be explained in the third wave of the analysis.

RELATING CMRT TO NESTED MANOVA

The CMRT procedure has some fundamental resemblance to nested MANOVA but users should be aware of important theoretical differences. An important difference is that in CMRT, the structure results from splits of the explanatory variables that best explain the response through an MRT analysis. This means that the usual calculation of degrees of freedom, which are necessary to compute an F statistic and carry out the statistical tests that are computed in MANOVA to test the significance of the 'main factor', the 'subordinate factor' and their interaction (Legendre & Anderson 1999; Anderson 2001b; McArdle & Anderson 2001), is not directly applicable (Ouellette & Legendre 2011). For that reason, these tests are not implemented in ***CascadeMRT()*** R function. However, it is possible to subjectively infer from the cascade if the effect of the subordinate explanatory set on the response data changed as a function of the main set, by examining if the subordinate explanatory variables chosen or their splitting values changed as a function of the main partition.

CONCLUSION

The CMRT procedure is a framework where nested ecological hypotheses are precisely admissible. To do that, users must choose in which order two (or more) explanatory sets are considered in an MRT structure. It is also possible to partition the explained variation (R^2) among the sets and ultimately obtain a coefficient of determination for the complete cascade of MRT analyses. The final CMRT model may be subjectively assessed for interaction between the explanatory sets, to evaluate

if the effect of the subordinate set changed as a function of the group membership produced by the first wave of analysis.

ACKNOWLEDGEMENTS

This study was supported by Natural Sciences and Engineering Research Council of Canada (NSERC) grant no. 7738 to P. Legendre. We wish to thank Daniel Borcard and Steven C. Walker for helpful comments and suggestions that helped in improving the manuscript.

Chapitre 5

Bootstrap assessment of the prediction accuracy of aboveground tree biomass estimation for five native species in a young Panamanian tropical plantation

Ce chapitre a été soumis pour publication dans une revue internationale : Environmental and Ecological statistics.

Marie-Hélène Ouellette[1], Diana M. T. Sharpe[2], Benjamin Wadham-Gagnon[2] and Catherine Potvin[2,3]

1-Département de sciences biologiques
Université de Montréal
C.P. 6128, succursale Centre-ville
Montréal, Qc, H3C 3J7
Office : 514 343 6111 (1233)
Fax : (514) 343-2293

2-Department of Biology
McGill University
1205 Dr Penfield
Montréal, QC, H3A 1B1

3-Smithsonian Tropical Research Institute
Balboa, Panama
Republica de Panama

Abstract

As the urgency of limiting our greenhouse gas emissions rises, the exact magnitude of carbon recapture in regenerating forests is attracting much international interest. International methodological guidelines propose allometric equations relating tree biomass to various traits measured non destructively as an essential tool to estimate forest carbon stocks. Using bootstrapping, we assess the predictive accuracy of generalized linear models (GLM) and some of its different features (different error

distributions, transformations/links): our goal is to establish if the best predictive features are the same over different data sets and basic equations. We thus compare models built based on different well-documented allometric equations and different non-destructive measures of tree traits in the hope of improving non-destructive estimates of above-ground biomass of five tropical tree species increasingly used in reforestation trials in Panama. Bootstrapping was established as the best means of estimating predictive error and determining the best predictive model.

Keywords : carbon stocks; climate change; generalized linear model ; land use change impacts; tropical forests.

INTRODUCTION

Mitigation strategies for climate change consider plantation establishment and forest regrowth as options to reduce emissions from the land use sector (Harmon 2001; Lal 2008; Canadell and Raupach 2008). Under both the Kyoto Protocol and the voluntary carbon market, carbon credits from reforestation may be issued to landowners managing plantations according to a stringent set of conditions (Harmon 2001). Meanwhile, the international community is engaged in negotiating approaches to stimulate actions to reduce emissions from deforestation and forest degradation, to implement sustainable forest management and to increase forest carbon stock and forest conservation (REDD+) (Potvin and Bovarnick 2008). To consider land-use change activities as mitigation action, precise estimates of carbon stocks in hardwood plantations and native forests, and hence reliable means of quantifying biomass (Kraenzel 2003; Losi 2003; Pelletier et al. 2010) are necessary.

The main objective of our study was to establish the best statistical practices for estimating aboveground tree biomass (AGB) based on field measurements through allometric equations by means of the GLM methodology with specific error distributions. To do so we compared different features of the GLM approach and created different candidate equations by varying the tree traits used to estimate biomass and modifying the basic allometric equation to ensure the generality of our conclusions. Note that other means of estimating AGB are proposed in the literature, for example remote sensing (see Patenaude, Milne & Dawson 2005 and Lu 2006 for a review) and GIS-based modeling (based on ancillary data, e.g. Brown, Iverson, Prasad et al. 1993, or a combination of data sources, e.g. Freeman & Moisen 2007), but these are not assessed in this paper.

Generalized linear models (GLM, see for example McCullagh and Nelder 1989), constitute a general framework, which encompasses the least squares linear modeling (LM) when the error distribution of the response is chosen as Gaussian and the observations are independent with constant variance. It allows, if necessary, the variance of the measurements to be non-constant by defining it as a function of the mean. This can be useful when the response is known to have heteroscedastic error. By specifying a statistical family that encompasses the proper error structure, this heteroscedasticity can be taken into account in the modeling procedure. This allows one to avoid problems like large variance in the estimates, which produce a less efficient model and thus less accurate forecasts (Wang and Jain 2003). GLM is based on the concept of maximizing the log-likelihood with the use of an algorithm called IWLS (iterative weighted least-squares). Note that linear models rely on least-squares fitting and are the prevalent technique used to estimate the parameters of allometric

equations (see for example Chave et al. 2005; Bond-Lamberty et al. 2002; Fournier et al. 2003; Van et al. 2000). In our study, GLM was used with either the Gaussian (constant variance) or Gamma (non-constant variance) family for the distribution of errors, various links (natural logarithm, power 0.25) and transformations (natural logarithm with and without bias correction (Sprugel 1983), and power 0.25). In the context of predicting AGB, we hypothesized that a main advantage of GLM is that it enables users to optimize the model parameters in the original response space by transforming the mean by using 'link' when the response must be transformed for linearity. The link is an invertible function *g* that links the expected value $E(\mu_i)$ to the linear predictor βX_i in the following manner: $g(E(\mu_i)) = \beta X_i$. The response is fit by maximum likelihood in the transformed scale, but the expected variance is calculated on the original scale (Myers et al., 2002). This combines the explanatory variables additively as in LM, and leaves no interpretation problem at the response original scale and the regression coefficients level. A second advantage of the GLM procedure is that the distribution function of the residuals can follow, among others, the Gaussian, Binomial or Poisson distributions. GLM can therefore accommodate continuous, binary or discrete response variables. As AGB is a continuous variable, we compared Gaussian and Gamma distributions, both suited for this mathematical type. To our knowledge, parameters of allometric equations for AGB have never been estimated by GLM before. Some authors have reported the use of a procedure they abbreviated 'GLM', but they were referring to the *general* linear model and not the *generalized* linear model (Senn 2003). That both procedures are abbreviated by the same letters unfortunately leads to confusion (Senn 2003). A drawback of using LM

to estimate biomass is that the response data is usually log-transformed prior to the analysis (Brown et al. 1989; Overman et al. 1994; Ketterings et al. 2001; Chave et al. 2005). Therefore, the explanatory variables are multiplicative so the regression coefficients are not easily interpreted and the predictions are in logarithmic scale.

The basic equations used in our modeling were those described by Chave et al. (2005), Overman et al. (1994), Ketterings et al. (2001) and Brown et al. (1989). Measures of tree diameter were basal area (BA), tree basal diameter (BD) and the sum of diameters at breast height of all the stems in multi-stem individuals (DBHall). In addition to these various model equations, we also sought to compare general models (in which all five species were pooled) to species-specific models. Bootstrapping was used to compare the predictive accuracy of all candidate models. We tested these AGB models using empirical data from a tropical plantation in Sardinilla, Panama (Potvin and Gotelli 2008). The Sardinilla plantation was designed to compare ecosystem functions among reforested plots containing different numbers of species; at the time of our study, it contained more than 3,500 young trees. This large data set consisting of five and six year old trees is relevant to plantations throughout the tropics that are established as carbon sinks, either through the voluntary carbon market or the formal carbon market. Indeed, the first verification of carbon storage often takes place five years after the establishment of a plantation.

MATERIAL AND METHODS

COMPARING GLMS FOR AGB ESTIMATION

In order to obtain the best predictive accuracy for AGB estimation, we sought to establish which combination of options of the GLM procedure are preferable for

bootstrapping. The features of the GLM analysis assessed here are the type of error distribution (Gaussian and Gamma), the type of link, and transformation followed by counter transformation correction for bias when applicable. In terms of transformation or link, we used the natural logarithm and power (0.25) transformations to linearize the relationship between the explanatory variables and the response. These transformations were applied to the response and the explanatory variables, as shown in the equations in Table 5.1.

From a practical perspective, we sought generality in our assessment, and thus evaluated the predictive accuracy of the different features in the context of general models, either with all species pooled, or with each species modeled separately. We also included in our assessment five basic equations (based on Chave et al. (2005), Ketterings et al. (2001), Overman et al. (1994) and Brown et al. (1989) (Table 5.1), and three different measures of tree diameter (BA, BD and DBHall). Our goal was to establish if with different basic equations and diameter measures, the same combination of options of the GLM procedure produced the best predictions for individual species and for all species pooled.

Turning to GLM instead of strictly LM may seem like a natural choice as this procedure is readily appealing, notably because when a link is used the predictions are made in the original space.

All GLM models were fitted using a procedure based on maximizing the log-likelihood function used by the ***glm()*** function of the STATS library in the R language. For power links, we used the function `tweedie()` from the TWEEDIE R library. As mentioned above, we used bootstrapping (Efron and Tibshirani 1993) as a resampling method to assess the predictive accuracy of the models by estimating the expected

Table 5.1 : List of the basic equations used. *H*, *D*, *S* and *AGB* stand for height, diameter, density and above-ground biomass respectively.

$\underline{Chave\ 1}$: $\ln(AGB) = b_0 + b_1\ln(H) + b_2\ln(D)$
$\underline{Chave\ 2}$: $\ln(AGB) = b_0 + b_1\ln(D) + b_2(\ln(D))^2 + b_3(\ln(D))^3)$
$\underline{Kettering}$: $\ln(AGB) = b_0 + b_1\ln(D^2H)$
$\underline{Brown}$: $\ln(AGB) = b_0 + b_1\ln(D^2HS)$
$\underline{Overman}$: $\ln(AGB) = b_0 + b_1\ln(D^2S)$

error made on a prediction: the prediction error. In the simplest bootstrap estimation process, *P* subsamples of the data are repeatedly analyzed, each subsample consisting of a random sample drawn with replacement from the full data set. For each of the subsamples, the model parameters were estimated using the technique of choice, and we applied each fitted model to the original data to obtain *P* estimates of the prediction error, noted $MSE^{(p)}$, where $MSE^{(p)} = \sum_{i=1}^{N} \frac{\left(y_i - \hat{y}_i^{(p)}\right)^2}{N}$. The overall estimated prediction error (MSE_{bs}) is the average of those *P* estimates. Ordinary bootstrap gives an estimate of the prediction error with low variability, but with possible large downward bias, particularly in highly overfitted situations (Efron 1983).

An alternative is to estimate the *bias* (or *optimism*) of the empirical risk (MSE_{emp}, also called apparent error rate in the help file of the ***bootpred()*** function of the R language) and to add it to the empirical risk. The MSE_{emp} is the sum of squared error between the response and the predicted values calculated on the full original data set. The bias is estimated by bootstrapping: for each subsample, we calculated the difference between the MSE_{bs} and MSE_{emp}. The final predictor error estimate is $MSE_{bs2} = MSE_{emp} + bias$. There is a more sophisticated method to estimate the prediction error based on bootstrap that considers the following set back: when using bootstrapping to estimate MSE_{bs}, the whole process is based on sampling *with* replacement, thus some objects belonging to the training set use to build the model (sample with replacement) are also in the test set (original data set). It can be shown that the percentage of objects belonging to both sets tends toward 63.2%. So one can define MSE_0 as the mean sum of squared error computed only on the objects that are not members of the training set. The MSE_0 estimate is known to be pessimistic, so a

reliable estimator, noted $MSE_{.632}$ is the weighted average of MSE_0 and MSE_{emp}, calculated by multiplying them by 0.632 and 0.368 respectively. With the function ***bootpred()*** from the BOOTSTRAP R library, we calculated the bootstrap estimate of prediction error with 500 resamplings as suggested by Efron and Tibshirani (1993), i.e., the apparent error rate, which is defined by the mean of all errors given by the model, the bootstrap estimate of *optimism*, and the 0.632 bootstrap estimate of prediction error. The final prediction error estimates of the models were taken to be the $MSE_{.632}$ values. This estimator is reported to be the less biaised in the literature. See for example Davison & Hinkley 1997 or Mevik & Cederkvist 2004. See Box 5.1 for a review.

SAMPLING

The plantation is based in Sardinilla (9°19'30"N, 79°38'00"W), a small village in the region of Buena Vista, Panama. Six native tree species were selected for planting: *Anacardium excelsum* (Bert. & Balb. Ex Kunth) Skeels (Ae), *Cedrela odorata* L. (Co), *Cordia alliodora* (Ruiz & Pavon) Oken (Ca), *Hura crepitans* L. (Hc), *Luehea seemanii* Triana & Planch (Ls), and *Tabebuia rosea* (Bertol.) DC. (Tr). The species Tr, Co and Ca are amongst the most important native timber species in the region, while Ae and Hc have important local uses. They show contrasting architectures, with the last four species being generally monopodial while the first two produce many stems. Co grows tall with a very small crown while Hc has a very large basal diameter at 10 cm from the ground (BD) compared with its diameter at breast height (DBH). The plantation consists of 24 plots of approximately the same

$$MSE_{emp} = \frac{\sum_{i=1}^{N}\left(y_i - \hat{y}_i^{(p)}\right)^2}{N}$$

empirical risk or apperant error rate

$$MSE_{bs} = \frac{\sum_{k=1}^{P}\left(\frac{\sum_{i=1}^{N_{bs}}\left(y_i - \hat{y}_i^{(p)}\right)^2}{N_{bs}}\right)}{P}$$

$$MSE_{bs2} = MSE_{emp} + bias \qquad bias = \frac{\sum_{k=1}^{P}\left(MSE_{bs} - MSE_{emp}\right)}{P}$$

$$MSE_{.632} = 0.632\,MSE_0 + 0.368\,MSE_{emp}$$

$$MSE_0 = \frac{\sum_{k=1}^{P}\left(\frac{\sum_{i=1}^{N_{p(test)}}\left(y_{i(test)} - \hat{y}_{i(test)}^{(p)}\right)^2}{N_{p(test)}}\right)}{P}$$

MSE_{χ}: estimate of the expected error made on a prediction

Box 5.1 Representation of the different estimators of the prediction error organized along an optimism axis. This is not an exhaustive list. Let y_i be the response of the i^{th} object, $\hat{y}_i^{(p)}$ its predicted value, N the size of the full data set, N_{bs} the size of the bootstrap sample, P the number of bootstrap runs, and the subscript *(test)* the

designation of an object that was not in the training set (thus not used to compute the model). MSE_{emp} is the most optimistic estimate because the sum of squares and the model are both calculated on the full data set. It is followed by MSE_{bs}, a bootstrap estimate where P models are computed and the mean prediction error is calculated over those P bootstrap samples. It is known to have a large downward bias (overoptimistic). MSE_{bs2} is known to have a smaller bias (as the bias is estimated by bootstrap); it is also called the ordinary bootstrap estimate. MSE_0 is the most pessimistic estimate because only the objects not used to compute the model are used to estimate the prediction error. $MSE_{.632}$ is a compromise between MSE_0 and MSE_{emp}; it is generally a good choice (e.g. Davison & Hinkley 1997).

size (45 x 45 m). 12 plots (two for each species) are monocultures, six plots contain different combinations of three tree species, and six plots contain all tree species. Undergrowth was cleared annually to eliminate competing vegetation and to facilitate work within the plantation. The plots were randomly positioned in order to reduce bias in the results caused by potential differences in soil conditions. Plots are square-shaped and hold 225 trees each, planted at 3-m spacing. Trees were planted in 2001 and were six years old at the time of harvest for this study. Although the plantation was established with six species, very high mortality rates limited the number of Ca trees found in the plantation after six years (Potvin and Gotelli 2008); as a consequence, this species was excluded from our analysis.

We harvested 10 individual trees per species per diversity treatment (monoculture, three-species combination, and six-species combination), chosen to be representative of the size range in the species-treatment group of interest. Therefore, within each species-treatment group, we ranked all individuals by height (H) and divided them into three equal size classes (small, medium and large). Of the 10 individuals to be sampled from each species-treatment group, three were chosen randomly from each of the small, medium and large size class and one chosen randomly from the entire data set. Within size classes, individual trees were selected from an Excel spreadsheet using a random number generator (www.random.org) and their X-Y positions within a plot were noted. For monoculture plots, we ensured that the selected trees would not come from the same subplot, unless all subplots had been filled. Secondly, within a size class, all three trees could not come from the same plot. For the three-species plots, we randomly selected one tree within each plot from each of three size classes, with the restriction that no two trees could come from the same

subplot. For the six-species plots and for each species, the 10 individuals to be measured were chosen from across the six plots containing the six-species combinations. We randomly determined which plots would contain one and two samples. Within each plot, we randomly selected one or two trees from the appropriate size class, with the restriction that no two trees could come from the same subplot.

Prior to the harvest, individual tree height (H), basal area at 10 cm from the ground (BD), and diameter at breast height (DBH) (1.3 m) for each stem were measured. BA was calculated for each stem and summed to obtained tree BA. Tree H was measured using a Vertex (Vertex III, Haglof Sweden AB). Each tree to be cut was marked ahead of time with spray paint and identified with a metal tag bearing its location and species code. Trees were cut at the base, as close to the ground as possible, using either a handsaw or chainsaw, depending on the trunk diameter. Large trees were lowered with ropes to avoid damaging other trees. If necessary, branches were removed prior to cutting the tree to avoid hitting neighbouring trees. For trees with multiple stems, we considered the stem with the largest DBH to be the primary trunk. Using a 20 kg capacity scale, we weighed all trunks and branches separately. We then took a sub-sample of two branches from each of the three locations (low, medium, high) on the primary trunk. We weighed these branches, removed all leaves, and then reweighed them to determine the mean fresh mass of leaves. We next took small wood samples from the following parts of each tree: bottom and top of the primary trunk, low, middle and high branches. We weighed each trunk segment using a Salter-AND EK 12 kg scale to determine its fresh mass, and then stored each

segment in a separate labelled paper bag, to be dried in a drying oven (48 h at 60°C) and reweighed to determine its dry mass.

From the data collected, we were able to convert fresh biomass into dry biomass, taking into consideration the fact that water concentrations differ between branches and trunks. Dry to fresh biomass ratios were calculated for trunks and branches separately, based on the wood segments taken from these tissues. Multiplying these ratios with the fresh biomass obtained in the field gave us the dry biomass of trunks and branches. For trunks, we used a dry to fresh biomass ratio that was calculated by averaging the ratios found from the top and bottom trunk segments. Dry biomass of each structural component were then summed to calculate the total aboveground dry biomass of the tree (AGB).

To determine the predictive power of the allometric models, we used traits measured in 2007 on 3,556 six-year old trees growing in the same plantation. Individual tree H, BD and DBH for each stem were measured, as in 2006, in January and February 2007.

RESULTS

In total, for the general model with all species pooled and the species-specific models (six data sets in total), we computed 900 models: 450 GLM-Gaussian and 450 GLM-Gamma, using the five allometric equations, the three independent measures of tree diameter, and five transformations or link functions.

According to bootstrapping, and across both the pooled and species-specific data sets, GLM-Gamma models performed better than GLM-Gaussian models in 42.6% of the 450 comparisons. GLM-Gamma models had a higher predictive

accuracy than GLM-Gaussian in 52% of the cases for *Anacardium excelsum* and only in 31% of the cases for *Luehea seemanii* (Table 5.2a). In the case of tree diameter measures, 53% of the best BD models were GLM-Gamma in contrast to only 37% for the BA models (Table 5.2a). The allometric equation for which GLM-Gamma models had the best predictive accuracy percentage was Brown (51%) while GLM-Gamma models only performed better in 30% of the cases for Overman equation (Table 5.2b). It becomes compelling when we assess the difference between models with transformations and links (Table 5.2c): for power and log transformations, 18.9% and 10.0% of the comparisons favoured Gamma models, while for power and log links we obtain 52.2% and 77.8% of the cases. Power and log transformations of the response made the residuals more homoscedastic, which in turn penalized the Gamma family that has a non-constant variance as a function of the mean ($V(\mu)=\phi\mu^2$ to be precise). This family favours variance of Y that gets larger as Y gets larger, which is what we observe for the raw data along with a link (Figure 5.1) and much less in the transformed data (Figure 5.2). For the few cases where this did not apply (where either Gamma performed better for a transformation model or worse for a link model), we believe that we are in a grey zone where the heteroscedasticity of the residuals is not strong enough for a Gamma distribution, or not homoscedastic enough for the Gaussian model to perform better than its rival for the respective cases.

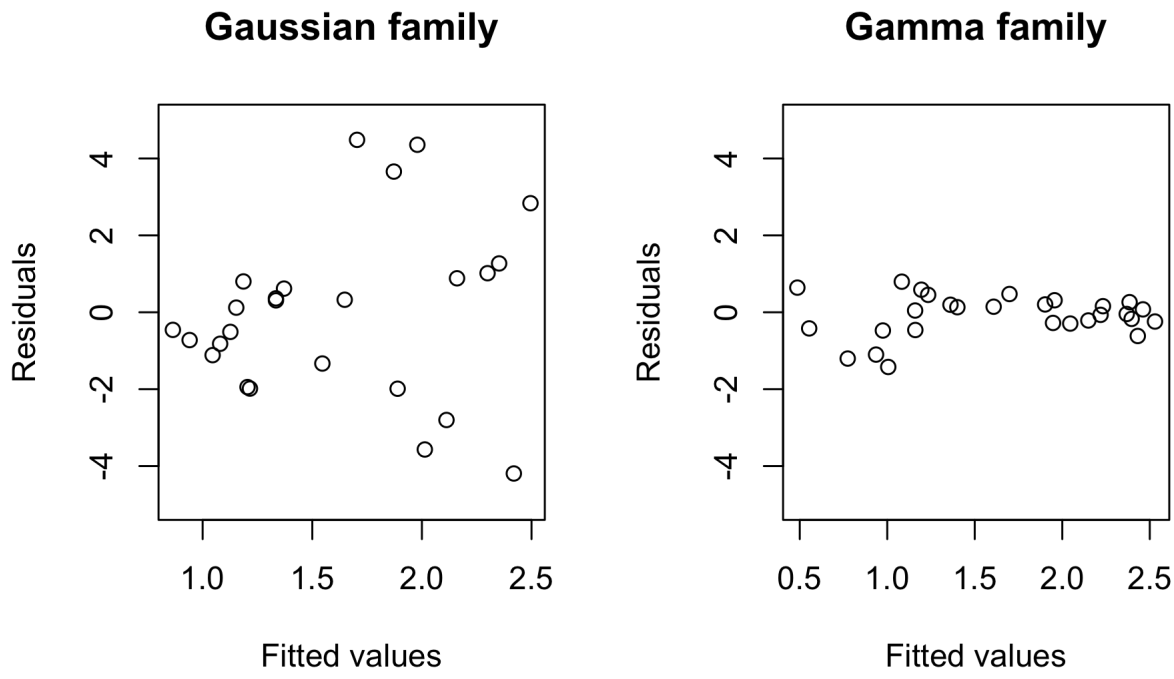


Figure 5.1 : Scatterplot of residuals as a function of fitted values for a link model (Chave1, BA, power link for Ae). In this case, the Gamma family better grasps the residual structure, and has a smaller $MSE_{.632}$ value than the Gaussian family for this data set for power link.

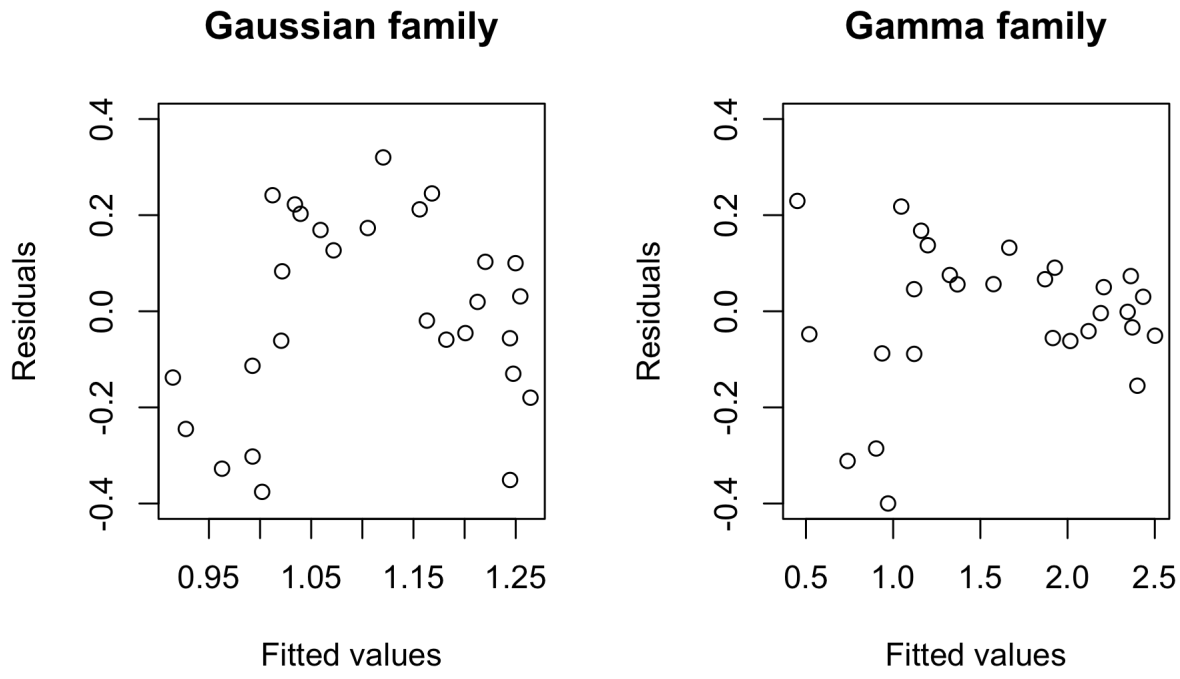


Figure 5.2 : Scatter plot of residuals as a function of fitted values for a transformation model (Chave1, BA, power transformation for Ae). In this case the Gaussian family better grasps the residual structure of the transformed data.

Table 5.2: Tables of comparison between Gamma and Gaussian model.

a) Table showing the number of comparisons where Gamma performed better than Gaussian family on the total number of comparisons for all basic equations per data set.

Basic equation	***Measure of tree diameter***	***General model***	***Ae***	***Cm***	***Hc***	***Tr***	***Ls***	***Total***	***%***
Chave 1	BD	3/5	2/5	2/5	3/5	1/5	1/5	12	
Chave 2	BD	3/5	3/5	3/5	2/5	4/5	3/5	18	
Kettering	BD	4/5	5/5	3/5	3/5	2/5	0/5	17	**53.33**
Brown	BD	4/5	5/5	4/5	2/5	2/5	0/5	17	
Overman	BD	1/5	2/5	3/5	2/5	0/5	2/5	10	
Chave 1	BA	1/5	3/5	1/5	3/5	0/5	3/5	11	
Chave 2	BA	2/5	3/5	3/5	3/5	2/5	2/5	15	
Kettering	BA	3/5	2/5	2/5	2/5	3/5	2/5	14	**37.33**
Brown	BA	3/5	2/5	2/5	1/5	3/5	2/5	13	
Overman	BA	1/5	2/5	0/5	2/5	2/5	2/5	9	
Chave 1	DBH	1/5	3/5	0/5	2/5	2/5	2/5	10	
Chave 2	DBH	0/5	1/5	3/5	1/5	3/5	2/5	10	
Kettering	DBH	3/5	2/5	2/5	3/5	2/5	0/5	12	**40.00**
Brown	DBH	4/5	2/5	2/5	3/5	3/5	2/5	16	
Overman	DBH	0/5	2/5	3/5	2/5	1/5	0/5	8	
TOTAL		33	39	33	34	30	23	**/75**	
%		**44**	**52**	**44**	**45.33**	**40**	**30.67**		

b) Table showing the percentage of comparisons where Gamma performed better than Gaussian family on the total number of comparisons for all basic equations.

Basic equation	***%***
Chave 1	36.66
Chave 2	47.78
Kettering	47.78
Brown	51.11
Overman	30.00

c) Table showing the number of comparisons where Gamma performed better than Gaussian family on the total number of comparisons by type of transformation or link.

Transformation or link	**General model**	***Ae***	***Cm***	***Hc***	***T***	***L***	**%**
Power trans	0/15	4/15	3/15	3/15	4/15	4/15	18.9
Power link	7/15	7/15	9/15	11/15	5/15	8/15	52.2
Log trans	3/15	3/15	1/15	1/15	1/15	0/15	10.0
Log link	14/15	14/15	11/15	14/15	9/15	8/15	77.8
Log trans + bias corr.	10/15	11/15	8/15	5/15	0/15	7/15	45.6

The model giving the most accurate predictions overall for the all-species data set ($MSE_{.632}$ = 138.08) was the one with parameters estimated by GLM with Gaussian family and power transformations while using the Chave 1 basic equation with BD as a tree diameter measure (Table 5.3). Species-specific modeling, however, shows the benefit of a diversity of modeling approaches. GLM-Gaussian modeling with power transformations was preferred for *Anacardium excelsum*, along with Chave 2 equation and BA measure ($MSE_{.632}$ = 138.08), for *Hura crepitans,* along with Chave 1 equation and DBH measure ($MSE_{.632}$ = 355.09), for *Luehea seemanii*, along with Chave 1 and BA measure ($MSE_{.632}$ = 49.57) and for *Tabebuia rosea,* along with Kettering equation and BA measure ($MSE_{.632}$ = 5.59), respectively (Tables 5.4, 5.6, 5.7, 5.8). However *Cedrela odorata* relied on GLM-Gamma with a log transformation and bias correction as the best modeling means with Brown equation and BD for diameter measure ($MSE_{.632}$ = 50.46) (Table 5.5).

Graphical assessment of the linear relationship between the different measures of tree diameter and AGB indicates without ambiguity that the relationship was non-linear and that log transformation was necessary. The bootstrap results for the model with all species pooled and the species-specific models showed that none of the best models had a link of any kind; thus, better predictive results were given by transformations (Tables 5.3-5.8). The log counter-transformation correction for bias was retained amongst the best models only three times, twice for *Cedrela odorata* and once for *Tabebuia rosea* (Tables 5.4 and 5.8 respectively). In all other cases, standard log or power transformations gave the best results.

Table 5.3: List of best predictive models for all species pooled (n=150) chosen according to the 0.632 predictive error estimator ($MSE_{.632}$) given for all basic equations and measures of tree diameter.

Tree diameter measure	***Basic equation***	***LM, GLM, family***	***Transformation or link***	$MSE_{.632}$
BD	Chave 1	GLM Gaussian	Power transformation	138.08
	Chave 2	GLM Gaussian	Power transformation	155.27
	Kettering	GLM Gaussian	Power transformation	163.96
	Brown	GLM Gaussian	Power transformation	156.15
	Overman	GLM Gaussian	Power transformation	166.02
BA	Chave 1	GLM Gaussian	Power transformation	230.92
	Chave 2	GLM Gaussian	Power transformation	245.43
	Kettering	GLM Gaussian	Log transformation	216.22
	Brown	GLM Gaussian	Log transformation	224.31
	Overman	GLM Gaussian	Log transformation	249.85
DBH	Chave 1	GLM Gaussian	Power transformation	174.42
	Chave 2	GLM Gaussian	Power transformation	247.43
	Kettering	GLM Gaussian	Log transformation	195.33
	Brown	GLM Gaussian	Log transformation	236.68
	Overman	GLM Gaussian	Power transformation	269.06

Table 5.4: List of best predictive models for *Anacardium excelsum* chosen according to the 0.632 predictive error estimator ($MSE_{.632}$) given for all basic equations and measures of tree diameter.

Tree diameter measure	***Basic equation***	***LM, GLM, family***	***Transformation or link***	$MSE_{.632}$
BD	Chave 1	GLM Gaussian	Power transformation	23.93
	Chave 2	GLM Gaussian	Power transformation	32.78
	Kettering	GLM Gamma	Power transformation	40.44
	Brown	GLM Gamma	Power transformation	40.54
	Overman	GLM Gamma	Power transformation	37.98
BA	Chave 1	GLM Gaussian	Power transformation	24.52
	Chave 2	GLM Gaussian	Power transformation	19.83
	Kettering	GLM Gaussian	Power transformation	41.25
	Brown	GLM Gaussian	Power transformation	41.72
	Overman	GLM Gaussian	Power transformation	60.48
DBH	Chave 1	GLM Gamma	Power transformation	26.91
	Chave 2	GLM Gaussian	Power transformation	23.41
	Kettering	GLM Gaussian	Power transformation	80.68
	Brown	GLM Gaussian	Power transformation	81.72
	Overman	GLM Gaussian	Power transformation	129.82

Table 5.5: List of best predictive models chosen for *Cedrela odorata* according to the 0.632 predictive error estimator ($MSE_{.632}$) given for all basic equations and measures of tree diameter.

Tree diameter measure	***Basic equation***	***LM, GLM, family***	***Transformation or link***	$MSE_{.632}$
BD	Chave 1	GLM Gaussian	Power transformation	133.96
	Chave 2	GLM Gamma	Power transformation	213.07
	Kettering	GLM Gaussian	Log transformation	163.91
	Brown	GLM Gamma	Log transformation and bias correction	50.46
	Overman	GLM Gaussian	Log transformation	164.46
BA	Chave 1	GLM Gaussian	Log transformation and bias correction	118.49
	Chave 2	GLM Gaussian	Log transformation	170.01
	Kettering	GLM Gaussian	Power transformation	84.49
	Brown	GLM Gaussian	Power transformation	84.70
	Overman	GLM Gaussian	Power transformation	91.14
DBH	Chave 1	GLM Gaussian	Power transformation	143.89
	Chave 2	GLM Gaussian	Power transformation	209.19
	Kettering	GLM Gaussian	Log transformation	191.03
	Brown	GLM Gaussian	Log transformation	185.60
	Overman	GLM Gaussian	Log transformation	291.87

Table 5.6: List of best predictive models chosen for *Hura crepitans* according to the 0.632 predictive error estimator ($MSE_{.632}$) given for all basic equations and measures of tree diameter (in other words, the best predictive model for each of Tables A6-20 of the Appendix). The $MSE_{.632}$ value is given, as is the regression technique and the transformation or link of the response when applicable.

Tree diameter measure	***Basic equation***	***LM, GLM, family***	***Transformation or link***	$MSE_{.632}$
BD	Chave 1	GLM Gamma	Power transformation	897.54
	Chave 2	GLM Gaussian	Log transformation	1144.59
	Kettering	GLM Gaussian	Log transformation	1012.35
	Brown	GLM Gaussian	Log transformation	1070.76
	Overman	GLM Gaussian	Log transformation	1000.61
BA	Chave 1	GLM Gamma	Power transformation	547.06
	Chave 2	GLM Gamma	Power transformation	581.98
	Kettering	GLM Gaussian	Power transformation	610.99
	Brown	GLM Gaussian	Power transformation	619.41
	Overman	GLM Gaussian	Power transformation	516.56
DBH	Chave 1	GLM Gaussian	Power transformation	355.09
	Chave 2	GLM Gaussian	Power transformation	609.23
	Kettering	GLM Gaussian	Power transformation	560.78
	Brown	GLM Gaussian	Power transformation	570.74
	Overman	GLM Gaussian	Power transformation	625.28

Table 5.7: List of best predictive models chosen for *Luehea seemanii* according to the 0.632 predictive error estimator ($MSE_{.632}$) given for all basic equations and measures of tree diameter (in other words, the best predictive model for each of Tables A6-20 of the Appendix). The $MSE_{.632}$ value is given, as is the regression technique and the transformation or link of the response when applicable.

Tree diameter measure	***Basic equation***	***LM, GLM, family***	***Transformation or link***	$MSE_{.632}$
BD	Chave 1	GLM Gaussian	Power transformation	52.63
	Chave 2	GLM Gaussian	Log transformation	68.21
	Kettering	GLM Gaussian	Log transformation	49.84
	Brown	GLM Gaussian	Log transformation	49.57
	Overman	GLM Gaussian	Log transformation	56.16
BA	Chave 1	GLM Gaussian	Power transformation	42.41
	Chave 2	GLM Gaussian	Power transformation	127.03
	Kettering	GLM Gaussian	Log transformation	63.78
	Brown	GLM Gamma	Power transformation	59.47
	Overman	GLM Gamma	Power transformation	91.01
DBH	Chave 1	GLM Gaussian	Log transformation	61.84
	Chave 2	GLM Gaussian	Power transformation	129.55
	Kettering	GLM Gaussian	Power transformation	101.00
	Brown	GLM Gaussian	Power transformation	100.08
	Overman	GLM Gaussian	Power transformation	107.86

Table 5.8: List of best predictive models chosen for *Tabebuia rosea* according to the .632 predictive error estimator ($MSE_{.632}$) given for all basic equations and measures of tree diameter (in other words, the best predictive model for each of Tables A6-20 of the Appendix). The $MSE_{.632}$ value is given, as is the regression technique and the transformation or link of the response when applicable.

Tree diameter measure	***Basic equation***	***LM, GLM, family***	***Transformation or link***	$MSE_{.632}$
BD	Chave 1	GLM Gaussian	Power transformation	8.82
	Chave 2	GLM Gaussian	Power transformation	25.47
	Kettering	GLM Gaussian	Power transformation	14.54
	Brown	GLM Gaussian	Power transformation	14.69
	Overman	GLM Gaussian	Log transformation and bias correction	28.28
BA	Chave 1	GLM Gamma	Power transformation	5.65
	Chave 2	GLM Gamma	Power transformation	33.01
	Kettering	GLM Gaussian	Power transformation	5.59
	Brown	GLM Gaussian	Power transformation	5.66
	Overman	GLM Gaussian	Power transformation	8.21
DBH	Chave 1	GLM Gaussian	Power transformation	8.17
	Chave 2	GLM Gaussian	Power transformation	32.02
	Kettering	GLM Gaussian	Power transformation	11.15
	Brown	GLM Gamma	Power transformation	11.12
	Overman	GLM Gaussian	Power transformation	19.97

DISCUSSION

CHOICE OF MODELING TECHNIQUES

In this paper, we sought to assess if GLM can effectively increase the predictive accuracy of regression models of AGB by (1) using a different error family than the Gaussian, namely the Gamma family, and (2) using a link instead of a transformation. The Gaussian distribution is characterized by its continuity, symmetry and constant variance. It is appropriate when the response variable is continuous, whether positive or negative, and normally distributed. A Gamma distribution is also appropriate for continuous data, but has the additional restriction that the values have to be equal to or greater than 0. The parameters of this distribution are more flexible than the Gaussian, allowing non symmetry, thus a long tail. Even if Gaussian models are thought to be fairly robust to a slight asymmetry (Neter et al. 1996, p. 30), we hypothesized that using a family of error that allows for asymmetry could provide higher prediction accuracy.

Overall, GLM-Gamma modeling performed well and is clearly an option that should be taken into account. One of the motivations for engaging in the current comparison of regression techniques was the inability of developing an allometric equation with a good predictive power for *Cedrela odorata* using LM modeling (Potvin et al., 2011). In an earlier paper, we reported adjusted R^2 ranging 0.9693-0.8646 for species-specific allometric regression in Sardinilla with the exception of *Cedrela odorata* for which the adjusted R^2 was only 0.7685. With coefficients of variation for tree height, BD, BA, DBH ranging between 44% and 69%, our data did not provide any indication that *Cedrela odorata* was more variable than the other

species; consequently, the predictive accuracy of LM would be lowest (Potvin et al., 2011). Interestingly, *Cedrela odorata* is the only species in the current study for which the overall best equation came from GLM-Gamma modeling. Figure 5.3 shows the scatter plot of the fitted values in terms of the residuals.

Unfortunately, it is practically impossible to choose the family and error structure before computing a model. Residual assessment is the key to finding clues about which family (variance or error structure definition) should be examined, but most of the time the only way to find out for sure is to try all possibilities. When the response is always in the same scale (use of link included), AIC (Akaike 1973, 1974) can be used to compare models. If not, estimation of the predictive error by bootstrap is the best (if not the only) way of comparing means (Burnham and Anderson, 2002).

DATA TRANSFORMATION

As mentioned above, a characteristic of the GLM technique is to allow the use of a link function instead of a transformation. The relationship between AGB and the indirect measures of biomass such as DBH are extensively reported to be log-log, and a correction for the bias of counter transforming the response has been reported to be unavoidable when modeling AGB with LM. Such corrections are used by most modellers (e.g., Chave 2005, Wang 2006 and Van 2000), even if they seem to overestimate the bias in some cases (Hepp 1982; Madgwick 1975). In theory, if GLM is used as a modeling technique and a link function is applied as a substitute to the transformation, there is no need for such a correction since the response is predicted in the original scale.

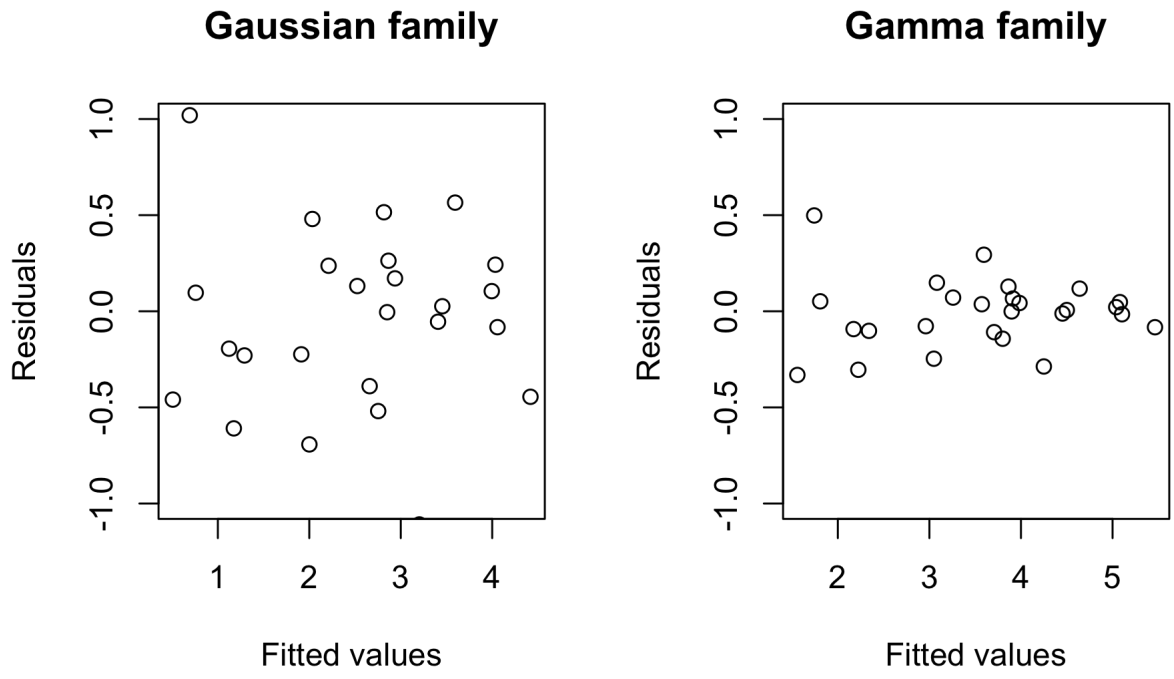


Figure 5.3: Scatter plot of residuals as a function of fitted values for the best Cm model (Brown basic equation with BD tree diameter measure and log transformation).

However, contrary to our expectations, no link was a match to the models with transformed responses in terms of predictive accuracy. In general, the choice between link or transformation is not immediately clear and depends on the specific situation. The nature of the resulting residuals can be used as a guide to the choice of model, with the aim of producing the most homoscedastic residuals. If the variance structure was not properly chosen, a trend will appear in the plot. Note that in the GLM setting, the residual assessment is made on the deviance residuals, which are defined as $d_i = y_i \ln\left(\frac{y_i}{\hat{\mu}_i}\right)$ $d_{i,r} = sign(y_i - \hat{\mu}_i)\sqrt{d_i}$ for the Gamma models. For the Gaussian model, they are equivalent to the usual residuals. The sum of the squares of the d_i values is the deviance of the model, and their assessment is the same as regular residuals in the multiple regression setting (Myers et al. 2002). In our data, if we examine the residual plots (see example in Figure 5.4), link plots showed strong heteroscedasticity for many data sets. While this might not always be the case, in our data it was the heteroscedasticity of the residuals that made the link less appropriate for prediction. In other words, the scale that gives the most homoscedastic residuals should be the one used for optimization of the parameters. We suggest that if we had a way to properly quantify the heteroscedasticity of residuals no matter their nature (to our knowledge, it is not the case, see for example Zaman 2000, Godfrey et al. 2006, Godfrey 2008, and Machado and Silva 2000), this value could be strongly correlated to the $MSE_{.632}$ bootstrap estimates, and could be used instead of bootstrapping to identify the best predictive model.

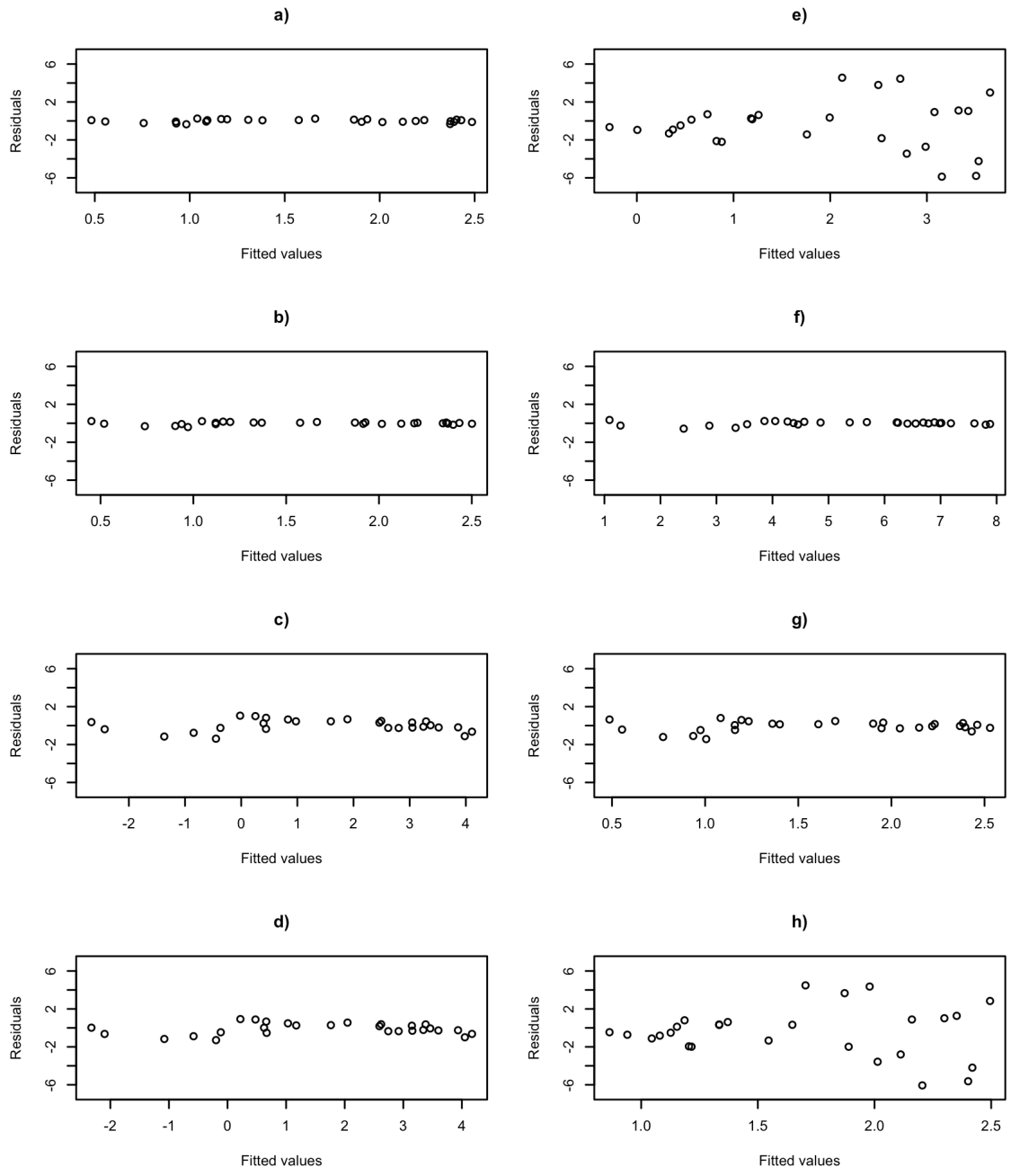


Figure 5.4: Assessment of residuals using the Ae data set, with BA as a tree diameter measure and Chave 1 basic equation. Here we show, from left to right, and top to bottom, the residuals as a function of the fitted values of models with increasing values of $MSE_{.632}$ (a < b < … < h). The models were (a) Gaussian with power transformation, (b) Gamma with power transformation, (c) Gaussian with *ln*

transformation, (d) Gamma with *ln* link, (e) Gaussian with *ln* link, (f) Gamma with *ln* transformation, (g) Gamma with power link, and (h) Gaussian with power link. We observe that heteroscedasticity seems to be getting larger as we go down the figure for most models, but it is a subjective assessment. This strengthens the argument that further assessment of the predictive accuracy of the models by means of AIC or bootstrapping is necessary.

In four of the six data sets, power transformations outperformed the logarithm transformations, and considerably more so when a counter transformation correction for bias was used. This suggests that better predictive accuracy can be achieved with a power transformation to linearize the relationship between the response and the predictors, and that practitioners should consider this alternative transformation (and others) along with the more common natural logarithm transformation. The few appearances of the counter transformation correction for bias in natural logarithm models in the summarizing tables suggest that this correction often gives worse results than straight natural logarithm transformation and should be used with extreme caution.

COMPARING REGRESSIONS: THE BOOTSTRAP APPROACH

In our modeling approach we favour the use of bootstrapping rather than AIC for model selection because the prediction error given by bootstrapping is comparable between all models as long as the predictions are made on the same scale. This is not the case for AIC, as AIC-based coefficients are not necessarily comparable across models if, for example, the data were transformed differently, or different data sets were used in each model. Since we explicitly used different transformations or links in our analysis, bootstrapping was the best method for estimating the prediction error. The coefficient of determination, R^2, and it's adjusted version, R^2_a, could have been used for descriptive comparison purposes. The coefficient of determination can be used as a measure of the explained proportion of variation, i.e., of goodness of fit, but not for predictive accuracy, as stated in Burnham and Anderson (2002, pp. 37 and 95) and shown by McQuarrie and Tsai (1998).

CONCLUSION

This study allows us to propose guidelines for modeling AGB. Although we were not able to identify the exact conditions under which GLM-Gamma performed better then GLM-Gaussian, this novel modeling method nevertheless provides a promising options in situations where traditional LM regression fails to offer the expected predictive power. Secondly, the predictive accuracy of models can be improved by considering transformations other than the typical natural logarithm transformation. For example, using power transformations along with a Gaussian distribution of errors improved the prediction accuracy of the models presented. Log counter transformation correction for bias, which seems to be systematically employed in the literature, should be avoided, or at least assessed, as it offered the worst predictions in most of our models. The flexibility of bootstrapping makes it the best method for estimating prediction error and ultimately determining the optimal allometric equation.

Chapitre 6

Discussion générale, conclusions et perspectives d'étude

L'étude des processus écologiques sous-jacents à la diversité bêta ou à la recapture de carbone par les plantations d'arbres tropicaux nécessite une modélisation explicative ou prédictive, selon le cas. On définit ici la modélisation comme un processus (*sensu* Shmueli 2010) qui compte plus de composantes que le simple ajustement d'un modèle : un aspect important de la modélisation est le choix des moyens utilisés pour évaluer la performance du ou des modèles construits. Par ailleurs, chaque étude de modélisation est caractérisée par un processus qui lui est propre. Le format du lien entre la réponse et les variables explicatives ainsi que les conditions d'application comme la structure des erreurs permettent de choisir la méthodologie. Le type de modélisation que nous choisissons nous indique quel indice de performance nous pouvons ou devons utiliser pour comparer les modèles (Shmueli 2010). C'est dans le type d'incertitude que les deux types de modélisation (explicative et prédictive) diffèrent (Helmer & Rescher 1959) : dans une modélisation prédictive l'incertitude se situe au niveau de l'exactitude d'une prédiction pour un nouvel objet. L'exactitude d'une prédiction se réfère à la différence entre la valeur réelle d'une nouvelle observation et celle prédite par le modèle: on désire que cette différence soit la plus petite possible. En revanche, l'incertitude d'un modèle

explicatif réside dans le degré de précision de l'explication ou encore dans la force du lien entre la réponse et le tableau explicatif. On cherche donc à minimiser la somme des résidus ou alors maximiser la valeur du R^2.

CADRE PREDICTIF ETUDIE : L'ESTIMATION DE LA BIOMASSE D'ARBRES TROPICAUX

Dans un cadre prédictif, la comparaison entre modèles devrait se faire à l'aide d'une mesure appropriée à cet usage. Par exemple, on peut se baser sur l'AIC (Akaike 1974) ou une de ses variantes (voir Burnham & Anderson 2002 pour une revue) ou encore sur l'estimation de l'erreur d'une prédiction à l'aide d'une méthode de rééchantillonnage comme le bootstrap (Efron 1983, Efron and Tibshirani 1993, Efron and Tibshirani 1997) ou la validation croisée (Burnham & Anderson 2002). Sur ces bases, on s'assure d'identifier, parmi tous les modèles considérés, celui qui produit les prédictions les plus justes pour de nouvelles observations n'ayant pas servi à estimer les paramètres du modèle. Malheureusement, il y a parfois de la confusion dans la littérature entre les pouvoirs prédictif et explicatif ou alors on comprend mal quelles mesures conviennent à quel cadre. Certains chercheurs utilisent à tort la comparaison entre des R^2 dans un cadre prédictif (Shmueli 2010). On rencontre en particulier cette confusion dans des travaux où on modélise la biomasse d'arbres à l'aide d'équations allométriques. En effet, plusieurs publications présentent une comparaison de nature explicative au lieu de prédictive entre les modèles allométriques (e.g. Losi et al. 2003, Wang 2006). Dans cette thèse, nous avons insisté sur une pratique de modélisation qui s'insère dans un cadre où on évalue explicitement le pouvoir prédictif de modèles d'estimation de la biomasse aérienne d'arbres tropicaux. Nous avons également voulu vérifier (1) si la structure des erreurs des modèles avait une distribution gaussienne ou gamma, (2) si une fonction lien dans

le cadre des modèles linéaires généralisés (MLG) était plus appropriée qu'une transformation et (3) si une fonction lien ou une transformation, tous deux avec la puissance 0.25, pouvait produire de meilleures prédictions que la transformation logarithme naturel qui est largement utilisée en foresterie. Si tous les modèles avaient été construits avec la même variable réponse, nous aurions pu les comparer à l'aide de l'AIC, comme l'ont fait Henry et al. (2010), mais des modèles avec fonction lien et avec différentes transformations de la variable réponse ne peuvent pas être comparés à l'aide de ce coefficient (Burnham & Anderson 2002). Nous avons donc dû nous résoudre à estimer l'erreur prévue pour une prédiction à l'aide du bootstrap. Cette pratique nous a permis de réaliser qu'une distribution des erreurs gaussienne est la plus appropriée dans la plupart des cas, sauf pour l'espèce d'arbre *Cedrela odorata* pour laquelle une structure d'erreur gamma a produit les meilleures prédictions. Nous avons également conclu à notre grande surprise que l'utilisation d'une fonction lien dans les MLG, au lieu d'une transformation, ne produit pas nécessairement les meilleures prédictions, même si, avec une fonction lien, nous optimisons la variable réponse dans son échelle originale. Ce n'est pas, au meilleur de notre connaissance, un résultat qui est souvent rapporté dans la littérature. Nous en concluons qu'il est intéressant d'employer les MGL pour développer des équations allométriques en utilisant une structure d'erreur qui diffère de la gaussienne et qu'il est nécessaire d'évaluer correctement leur pouvoir prédictif. Puisque les équations allométriques sont essentielles pour l'élaboration d'un système de crédits de carbone (Harmon 2001) et qu'elles s'insèrent donc dans les programmes de régulation des flux de carbone *via* les changements d'utilisation des terres, la précision des prédictions devient un facteur clef. En développant et rapportant des modèles dont le pouvoir

prédictif est le plus élevé possible, nous augmentons nos chances de mitiger l'impact des changements climatiques sur les écosystèmes en rendant les systèmes de crédit de carbone le plus précis possible.

CADRE EXPLICATIF ETUDIE : L'ARM ET LA DIVERSITE BETA

Comme nous l'avons mentionné plus haut, la comparaison entre modèles peut se faire à l'aide d'un coefficient de détermination (R^2) dans un cadre explicatif. Ce coefficient est une mesure du pourcentage de variation du tableau réponse expliqué par le modèle et donc de la force de la relation asymétrique entre les deux tableaux. Ainsi dans un cadre d'étude de la diversité bêta, qui est la variation de la composition spécifique entre les sites dans une région géographique donnée, le processus de modélisation devrait inclure une comparaison des modèles à l'aide d'un tel coefficient.

L'arbre de régression multivariable (ARM) est une méthode de plus en plus employée pour identifier des types d'habitat ou mettre en relation la composition spécifique de communautés avec d'autres types de variables explicatives. L'ARM permet de faire ressortir les changements les plus abrupts dans la composition de la communauté le long des gradients écologiques étudiés. Si ce type de patron de distribution est dominant dans le jeu de donné, on s'attend à ce que l'ARM explique plus de variation de la réponse que des méthodes basées sur des prémisses de continuité et de linéarité comme l'ACR. Comme nous l'avons montré dans l'un de nos exemples, l'ARM peut identifier ces patrons lorsqu'ils ne sont pas dominants et que l'ARC ne les met pas en évidence : dans ce cas, les conclusions de l'ACR et de l'ARM deviennent complémentaires. Dans une étude de la diversité bêta, toute information sur les facteurs déterminant cette diversité est pertinente et utile pour

définir des stratégies de conservation. De plus, les divisions binaires simples d'un modèle ARM devraient intéresser et guider les gestionnaires de l'environnement. Il devient alors important de définir une mesure de pouvoir explicatif qui soit adéquate pour cette méthode d'analyse. Le coefficient de détermination ajusté (R^2_{GDF}) développé au chapitre 3 de cette thèse permet d'obtenir une mesure non biaisée du pourcentage de variation de la réponse qui est expliqué par l'ARM. De plus, il procure un moyen de comparer son pouvoir explicatif à celui d'autres modèles construits pour la même réponse, comme les modèles construits *via* l'analyse canonique de redondance (ACR). Il est important de s'assurer que nous utilisons la même mesure de variation dans les deux cas, soit la somme des carrés des écarts à la moyenne, pour faire les calculs.

Éventuellement, il serait intéressant de pouvoir tester la signification statistique du R^2 comme on le fait en ACR. Ce test permettrait de tester l'hypothèse nulle suivante ($\boldsymbol{H_0}$) contre $\boldsymbol{H_1}$:

> $\boldsymbol{H_0}$ *: $\rho^2 = 0$, il n'y a pas de relation entre la réponse et les variables explicatives qui corresponde à la structure de l'ARM construit,*
>
> $\boldsymbol{H_1}$ *: $\rho^2 \neq 0$, il y a une relation entre la réponse et les variables explicatives qui corresponde à la structure de l'ARM construit,*

où ρ^2 est la valeur de la population. Nous pourrions ainsi tester si le R^2 est significativement différent de 0. Nous présumons que puisque l'ARM ne comporte pas de supposition de normalité ou d'homoscédasticité, nous devrons utiliser un test par permutation (Legendre & Legendre 1998) pour tester ces hypothèses (si la chose est possible). Nous devrons également définir une statistique à tester qui sera de

préférence pivotale et qui aura donc une distribution indépendante des paramètres du modèle. Une statistique pivotale est une statistique qui ne dépend pas de paramètres inconnus ; sa valeur est donc seulement fonction de paramètres quantifiables à partir des données observées. En analyse de régression par exemple, un coefficient de régression *b* n'est pas pivotal alors que la statistique *t* qui lui est associée est pivotale. Nous pourrons tenter d'utiliser la même statistique *F* qu'en RDA en modifiant les degrés de libertés (*df*) comme nous l'avons fait dans le chapitre 3. Des simulations supplémentaires seront nécessaires afin de tester l'erreur de type I de ce nouveau test statistique.

Même si dans ce travail nous nous sommes surtout penchés sur la valeur explicative de l'ARM et la comparaison entre le R^2_{GDF} du modèle d'arbre et le R^2_a de l'ACR, nous suggérons que l'emploi des deux méthodes sur un même jeu de données devrait faire partie des protocoles d'analyse de la diversité bêta. Même si parfois la RDA performe mieux que le MRT en termes explicatifs, comme c'est le cas dans l'exemple des araignées dans le chapitre 3, certaines relations qui ne sont pas linéaires ni linéarisables sont mieux représentées par un ARM. Dans une étude de la diversité bêta, toute information supplémentaire quant aux principaux déterminants des patrons est pertinente, même si cela implique de combiner deux modèles pour obtenir un portrait plus complet des liens qui existent. Si, comme dans notre étude des araignées, l'ACR performe clairement mieux que l'ARM en termes explicatifs mais qu'une relation claire apparaît dans l'ARM alors qu'elle n'est pas perceptible en ACR, on ne devrait pas l'ignorer. On peut toujours coder cette relation sous forme binaire et l'ajouter comme variable explicative dans l'ACR afin d'obtenir un modèle explicatif plus complet de cette analyse. Notons que ce type de codage est apparenté

au codage additif binaire utilisé en analyse phylogénétique afin de représenter un arbre phylogénétique (e.g. Brooks & McLennan 1991).

Dans un autre ordre d'idée, on notera que dans le chapitre 3 les arbres construits sont élagués par validation croisée, même ceux des populations simulées. Il serait intéressant de voir s'il est possible d'utiliser le R^2_{GDF} pour choisir la taille de l'arbre le plus *explicatif* au lieu de la validation croisée qui permet de sélectionner la taille d'arbre procurant les meilleures *prédictions*. Ainsi, tout le protocole de modélisation serait basé sur l'explication de la réponse, ce qui serait plus cohérent. Nous pourrions également tenter de concevoir un R^2_{GDF} non biaisé pour chacune des bipartitions et donc obtenir le pourcentage des contributions à la variation expliquée de l'arbre qui refléterait celle de la population étudiée.

L'intérêt des ARM et de l'ACR pour la modélisation de la diversité bêta nous a également amenés à nous interroger sur la possibilité d'utiliser deux tableaux explicatifs dans le cadre de l'ARM. L'ACR partielle est reconnue entre autres pour la partition de la variation qui permet d'évaluer le pouvoir explicatif de deux tableaux (e.g. Borcard et al. 1992, Legendre et al. 2005, Peres-Neto et al. 2006, Legendre 2008) dont l'effet est considéré comme additif et linéaire. Entre autres, cette analyse a servi à étudier les fondements de la diversité bêta en partitionnant la variation de la composition spécifique entre les dynamiques des communautés engendrant des patrons spatiaux (tableau spatial) et le contrôle environnemental (tableau environnemental). En ARM, la modélisation de deux tableaux explicatifs est rendue plus difficile par le fait que la nature des variables explicatives est modifiée et leur configuration finale n'est connue qu'*a posteriori* : ce sont des seuils (c'est-à-dire des points de coupure) de ces variables qui sont utilisés pour expliquer la variation de la

réponse. En conséquence, il n'est pas possible de procéder comme on le fait en ACR partielle, soit utiliser les résidus du modèle du premier tableau en fonction du deuxième afin d'ajuster la réponse pour obtenir la variation expliquée exclusivement par le premier jeu de données. Pour traiter deux tableaux explicatifs *via* l'ARM, il est nécessaire de se référer à un contexte similaire à la MANOVA hiérarchique. Pour ce faire, on hiérarchise les hypothèses explicatives pour faire en sorte qu'un des tableaux représente un effet principal et l'autre un effet subordonné. En d'autres termes, cette procédure permet de forcer l'ordre dans lequel les hypothèses explicatives sont considérées. Il est possible de partitionner la variation de la réponse, comme nous l'avons montré dans le chapitre 4, entre les modèles construits sur le tableau principal et le tableau subordonné. Cette explication doit être attribuée à la structure hiérarchique que l'on a choisie *a priori* pour construire le modèle. En pratique, ceci nous permet d'étudier des patrons hypothétiquement sous-jacents à la distribution de la composition spécifique qui ne sont pas représentés dans l'ARM simple et qui peuvent tout de même être informatifs. On se rappellera que par exemple dans le chapitre 4, nous avons pu étudier l'effet de variables explicatives après avoir utilisé pour la première partition la présence ou l'absence d'arbustes qui produisent un microclimat dans lequel les oribates se distribuent. La composition spécifique abritée par la présence ou l'absence d'arbustes est subséquemment régie par différentes variables explicatives : en l'absence d'arbustes, la densité du substrat est la variable la plus explicative, et en présence d'arbustes c'est le contenu en eau qui régit le plus fortement la composition spécifique. Même si les pouvoirs explicatif et prédictif de ces modèles sont moins élevés que ceux de l'ARM usuel, nous en retirons quand même des informations supplémentaires sur la structure des données.

Dans le même souci qu'au chapitre 3, c'est-à-dire d'être cohérent dans un cadre de méthodologie explicative, nous voudrions éventuellement pouvoir ajuster le R^2_{ARMC} pour chacun des arbres construits ainsi que pour la structure globale. De plus, des tests de significations de R^2_{ARMC} pour la procédure globale permettraient de tester l'hypothèse nulle suivante ($\boldsymbol{H_0}$) contre $\boldsymbol{H_1}$:

> $\boldsymbol{H_0}$: *$\rho^2_{ARMC} = 0$, il n'y a pas de relation entre la réponse et les variables explicatives qui corresponde à la structure de la ARMC,*
>
> $\boldsymbol{H_1}$: *$\rho^2_{ARMC} \neq 0$, il y a une relation entre la réponse et les variables explicatives qui corresponde à la structure de la ARMC,*

où ρ^2_{ARMC} est la valeur de la population. Ce test de signification permettrait d'établir si le modèle d'ARMC, malgré un pourcentage d'explication possiblement inférieur à celui de l'ARM, est significativement différent de 0 et donc que la relation dévoilée par l'ARMC peut être extrapolée au niveau de la population statistique. Les tests de signification du modèle principal et des modèles subséquents feraient appel au même raisonnement que ceux d'un ARM simple, en incorporant le fait que les modèles subséquents sont classés *a priori* par le modèle principal.

BIBLIOTHEQUES R ACCOMPAGNANT LA THESE

La bibliothèque R RSIMSSDCOMPAS utilisée pour simuler les données en vue de l'estimation du biais du R^2 ajusté sera rendue disponible sur le site R-Forge. On y combine les éléments de simulation de SIMSSD (Legendre et al. 2002, Legendre et al. 2004) et de JCOMPAS (Minchin 1987, De Cáceres 2003) que nous énumérons ici brièvement. La combinaison des méthodes SIMSSD et JCOMPAS permet de simuler

des gradients environnementaux déterministes auxquels sont associées des espèces qui y sont linéairement liées (en spécifiant un coefficient de régression linéaire β) ou en définissant une niche pour chaque espèce. Les données sont simulées sur une grille carrée de la taille choisie par l'utilisateur sur laquelle les gradients environnementaux peuvent prendre plusieurs formes, six au total. Lorsque les espèces sont simulées avec une niche, l'utilisateur fournit au programme les paramètres d'une distribution bêta généralisée (courbe en forme de cloche asymétrique ou non), la plage des valeurs du gradient environnemental où elle peut apparaître, ainsi que la valeur maximale de son abondance. Ainsi, si on veut simuler des compositions spécifiques qui changent le long du gradient environnemental, on simule plusieurs groupes d'espèces avec des plages de valeurs similaires qui se succèdent. Une espèce rare occupe une très petite plage des valeurs du gradient, une espèce généraliste est présente sur une très grande plage. Il est également possible de spécifier des paramètres d'interaction interspécifique en spécifiant un coefficient de compétition c négatif ou positif à appliquer entre deux espèces. Par exemple, une espèce dont l'abondance est A1 et une autre espèce dont l'abondance est A2 pourraient être liées de la manière suivante : $A1' = A1 - cA2$ où A1' est l'abondance de la première espèce modifiée. L'utilisateur peut également spécifier une capacité de support pour les sites (nombre maximum d'individus), ajouter du bruit qualitatif (une espèce peut être aléatoirement absente à un site) et du bruit quantitatif (ajout d'un nombre aléatoire à chaque valeur d'abondance). Finalement, pour chacune des variables simulées, soit l'environnement et les espèces, il est possible d'ajouter une structure d'autocorrélation qui est basée sur un variogramme dont les paramètres sont spécifiés par l'utilisateur.

La bibliothèque R MVPARTWRAP contient pour sa part des fonctions encapsulant la fonction ***mvpart()*** de la bibliothèque MVPART de De'ath (2010). La bibliothèque MVPARTWRAP réalise tous les nouveaux calculs développés dans cette thèse. Par exemple, la fonction ***CascadeMRT()*** permet de construire un modèle ARMC lorsqu'on lui fournit les tableaux explicatifs et le tableau réponse. Il y a également la fonction ***CasMRTR2()*** reliée à cette même analyse qui permet d'illustrer dans un tableau une partition de la variation expliquée par chacun des modèles ARM individuels, comme celui qui est présenté au chapitre 4 de cette thèse. On retrouve de plus dans cette bibliothèque la fonction ***R2AGDF()*** qui permet de calculer le R^2_{GDF}. Cette fonction requiert un objet de classe MRT. Ces objets sont créés à l'aide de la fonction ***MRT()*** de la même bibliothèque; celle-ci fournit entre autres une sortie graphique supplémentaire de l'ARM montrant les bipartitions situées à une hauteur proportionnelle à leur variation expliquée. De plus, le sommaire de cet objet (obtenu à l'aide de la fonction générique ***summary()***) montre les espèces qui contribuent le plus à la déviance du modèle ainsi que les espèces indicatrices au sens de IndVal (Dufrêne & Legendre 1997, De Cáceres & Legendre, 2010) pour la partition finale et chaque bipartition.

CONCLUSION GENERALE

Les mécanismes à l'étude dans cette thèse ont nécessité une modélisation explicative dans le cas de la diversité bêta et prédictive dans le cas de l'estimation de la biomasse aérienne d'arbres tropicaux. Nous avons étudié l'incertitude des deux types de modélisation dans un souci de cohérence avec les objectifs d'étude de chacun des mécanismes. Des simulations de Monte Carlo ont permis d'établir que le

R^2_{GDF} permet d'évaluer le pouvoir explicatif du modèle et de le comparer à celui de l'ACR. L'ARMC pour sa part permet d'identifier des patrons sous-jacents de la diversité bêta non disponible dans l'ARM usuelle. Nous avons relevé quelques améliorations qui pourraient être apportées aux analyses faites dans le cadre d'une ARM et d'une AMRC, entre autres l'élaboration de tests de signification statistique pour les coefficients de détermination calculés. Les conclusions tirées de l'étude de l'estimation de la biomasse aérienne d'arbre tropicaux conduisent à des recommandations concernant la structure des erreurs des modèles, en notant que l'erreur gaussienne n'est pas toujours la plus appropriée. Les MGL sont le cadre dans lequel il est possible de choisir différentes structures d'erreur. Ils devraient donc faire partie des protocoles de modélisation de ces jeux de données.

Références

Aart, P. J. M. van der and N. Smeek-Enserink (1975). "Correlations between distributions of hunting spiders (*Lycosidae*, *Ctenidae*) and environmental charateristics in a dune." Nethelands Journal of Zoology **25**:1-45.

Akaike, H. (1973). "Information theory and an extension of the maximum likelyhood principle", in: Petrov, B.N., Cáski, F. (Eds)., 2nd International Symposium on Information Theory. Akadémiai Kiadó Budapest 1973, Tokyo, Japan, pp. 267-281.

Akaike, H. (1974). "A new look at the statistical model identification". IEEE Transactions on automatic control **ac-19**: 716-723.

Angeler, D. G., O. Viedma, S. Cirujano, M. Alvarez-Cobelas, and S. Sánchez-Carrillo (2008). "Microinvertebrate and plant beta diversity in dry soils of a semiarid agricultural wetland complex". Marine and Freshwater Research **59**: 418-428.

Anderson, M.J. (2001a). "A new method for non-parametric multivariate analysis of variance." Austral Ecology **26**: 32-46.

Anderson, M.J. (2001b). "Permutation tests for univariate or multivariate analysis of variance and regression." Canadian journal of Fisheries Aquatic Science **58**: 626-639.

Anderson, M.J. and Cribble, N.A. (1998). "Partitioning the variation among spatial, temporal and environmental components in a multivariate data set." Austral Ecology **23**: 158-167.

Anderson-Sprecher, R. (1994). "Model comparisons and R^2." The American Statistician **48**: 113-117.

Auguet, J.-C., Barberan, A. and E.O. Casamayor (2010). "Global ecological patterns in uncultured Archaea." ISME Journal **4**: 182-190.

Balvanera, P., A. B. Pfisterer, N. Buchmann, J.-S. He, T. Nakashizuka, D. Raffaelli, and B. Schmid (2006). "Quantifying the evidence for biodiversity effects on ecosystem functioning and services." Review and synthesis **9**: 1146-1156.

Birks, H. J. B., Austin, H. A., Indrevaer, N. E., Peglar, S. M. and C. Rygh (1998). "An annotated bibliography of canonical correspondence analysis and related constrained ordination methods 1986-1996". Available from H. J. B. Birks, Botanical Institute, University of Bergen, Allégaten 41, N-5007 Bergen, Norway. Also available from the WWWeb page http:/www.bio.umontreal.ca/casgrain/cca_bib/.

Birks, H.J.B., S.M. Peglar and H.A. Austin (1996). "An annotated bibliography of canonical correspondence analysis and related constrained ordination methods 1986–1993." Abstracta Botanica **20**: 17–36.

Blanchet, F. G., P. Legendre, and D. Borcard (2008). "Forward selection of explanatory variables." Ecology **89**: 2623-2632.

Bojsen, B.H. and D. Jacobsen (2003). "Effects of deforestation on macroinvertebrate diversity and assemblage structure in Ecuadorian Amazon streams." Archiv für Hydrobiologie **158**(3): 317-342.

Bond-Lamberty, B., Wang, C., and S.T. Gower (2002). "Aboveground and belowground biomass and sapwood area allometric equations for six boreal

tree species of northern Manitoba." Canadian Journal of Forest Research **32**: 1441-1451.

Borcard, D., F. Gillet, and P. Legendre (2011). "Numerical ecology with R." Springer, New York.

Borcard, D. and P. Legendre (1994). "Environmental control and spatial structure in ecological communities: an example using oribatid mites (Acari, Oribatei)." Environmental and Ecological statistics, **1**: 37-61.

Borcard, D., Legendre, P. and P. Drapeau (1992). "Partialling out the spatial component of ecological variation." Ecology **73**: 1045-1055.

Boyce, M. S. and L. L. McDonald (1999). "Relating populations to habitats using resource selection functions." Trends in Ecology & Evolution **14**: 268-272.

Boyd, J. and S. Banzhaf (2007). "What are ecosystem services? The need for standardized environmental accounting units." Ecological Economics **63**(2-3): 616-626.

Breiman, L., Friedman, J.H., Olshen, R.A. and C.J. Stone (1984). "Classification and Regression Trees." Wadsworth International Group, Belmont, California, USA.

Brenden, T., L. Wang, and Z. Su (2008a). "Quantitative identification of disturbance thresholds in support of aquatic resource management." Environmental Management **42**: 821-832.

Brown, S. (2002). "Measuring carbon in forests: current status and future challenges." Environmental Pollution **116**: 363-372.

Brown, S., Gillespie, A.J.R. and A.E. Lugo (1989). "Biomass estimation methods for tropical forests with applications to forest inventory data." Forest Science **35**: 881-902.

Burnham, K.P. and D.R. Anderson (2002). "Model selection and multimodel inference : A Practical Information-Theoretic Approach". Second edition. Springer - Verlag, New York.

Canadell, J.G. and M.R. Raupach (2008). "Managing forests for climate mitigation." Science **320**: 1456-1458.

Carr, G.M. and P.A. Chambers (1998). "Macrophyte growth and sediment phosphorus and nitrogen in a Canadian prairie river." Freshwater Biology **39**: 525-536.

Chambers, J.M. (1992). "Linear models" in: Chambers, J.M., Hastie, T.J. (Eds), Statistical Models in S. Wadsworth & Brooks/Cole.

Chave, J., Andalo, C., Brown, S., Cairns, M.A., Chambers, J.Q., et al. (2005). "Tree allometry and improved estimation of carbon stocks and balance in tropical forests." Oecologia **145**: 87-99.

Chen, L., Mi, X., Comita, L.S., Zhang, L., Ren, H. and Keping Ma (2010). "Community-level consequences of density dependence and habitat association in a subtropical broad-leaved forest." Ecology Letters **13**: 695-704.

Cherwin, K.L., T.R. Seastedt, and K.N. Suding (2009). "Effects of nutrient manipulations and grass removal on cover, species composition, and invasibility of a novel grassland in colorado." Restoration Ecology **17**: 818-826.

Chust, G., Chave, J. R. Condit et al. (2006). "Determinants and spatial modeling of tree β-diversity in a tropical forest landscape in Panama." Journal of Vegetation Science **17**(1): 83-92.

Claudet, J., D. Pelletier, J. Y. Jouvenel, F. Bachet, and R. Galzin (2006). "Assessing the effects of marine protected area (MPA) on a reef fish assemblage in a northwestern Mediterranean marine reserve: Identifying community-based indicators." Biological Conservation **130**: 349-369.

Crossman, E.J. (1996). "Taxonomy and distribution." In: Pike biology and exploration (ed. J.F. Craig), pp. 1-11. Chapman and Hall, London.

Davidson, T.A., Sayer, C.D., Langdon, P.G., Burgess, A. and M. Jackson (2010). "Inferring past zooplanktivorous fish and macrophyte density in a shallow lake: application of a new regression tree model." Freshwater Biology **55**: 584-599.

Davies, P.T. and M.K.S. Tso (1982). "Procedures for Reduced-rank Regression." Journal of the Royal Statistical Society: Series C (Applied Statistics) **31**: 244-255.

de Càceres, M. (2003). "JCOMPAS". JCOMPAS 1.0 user's manual, Dept. Biologia Vegetal, Universitat de Barcelona.

de Càceres, M., P. Legendre, and M. Moretti (2010). "Improving indicator species analysis by combining groups of sites." Oikos **119**: 1674-1684.

de Groot, R. S., M. A. Wilson, and R. M. J. Boumans (2002). "A typology for the classification, description and valuation of ecosystem functions, goods and services." Ecological Economics **41**: 393-408.

de Nie, H.W. (1987). "The decrease in aquatic vegetarian in Europe and its consequences for fish populations." EIFAC/CECPI.

De'ath, G. (2002). "Multivariate regression trees: a new technique for modeling species-environment relationships." Ecology, **83**: 1105–1117.

De'ath, G. (2010). "mvpart: Multivariate partitioning. R package version 1.3-1." http://cran.r-project.org/package=mvpart.

Deelder, C.L. (1984). "Synopsis of biological data on the eel, Anguilla anguilla (Linnaeus, 1758)." FAO, Rome, Italy.

DeVantier, L., De'ath, G., Turak, E., Done, T. and K. Fabricius (2006). "Species richness and community structure of reef-building corals on the nearshore Great Barrier Reef." Coral Reefs **25**: 329-340.

Diniz-Filho, J.A.F. and L.M. Bini (1996). "Assessing the relationship between multivariate community structure and environmental variables." Marine Ecology Progress Series **143**: 303-306.

DORIS (25/2/2010) *Leuciscus leuciscus* (Linnaeus, 1758), http://doris.ffessm.fr/fiche2.asp?fiche_numero=2166.

DORIS (30/7/2010) *Phoxinus phoxinus* (Linnaeus, 1758) , http://doris.ffessm.fr/fiche2.asp?fiche_numero=1656.

Dufrêne, M. and P. Legendre (1997). "Species assemblages and indicator species: the need for a flexible asymmetrical approach." Ecological Monographs **67**: 345-366.

Efron, B. (1983). "Estimating the Error Rate of a Prediction Rule: Improvement on Cross-Validation." Journal of the American Statistical Association **78**: 316-331.

Efron, B. and R. Tibshirani (1993). "An Introduction to the Bootstrap." Chapman and Hall.

Efron, B. and R. Tibshirani (1997). "Improvements on Cross-Validation: The .632+ Bootstrap Method." Journal of the American Statistical Association **92**: 548-560.

Eisenhauer, J. G. (2008). "Degrees of Freedom." Teaching Statistics **30**: 75-78.

Elith, J. and J.R. Leathwick (2009). "Species distribution models : ecological explanation and prediction across space and time." Annual Review of Ecology and Systematics **40**: 677-697.

Elton, C. (1927). "Animal ecology." London.

Ezekiel, M. (1930). "Methods of Correlational Analysis." Wiley, New York.

Fisher, B., R.K. Turner, and P. Morling (2009). "Defining and classifying ecosystem services for decision making." Ecological Economics **68**: 643-653.

Fournier, R.A., Luther, J.E., Guindon, L., Lambert, M.-C., Piercey, D., et al. (2003). "Mapping aboveground tree biomass at the stand level from inventory information: test cases in Newfoundland and Quebec." Canadian Journal of Forest Research **33**: 1846-1856.

Godfrey, L.G. (2008). "Testing for heteroskedasticity and predictive failure in linear regression models." Oxford Bulletin of Economics and Statistics **70**: 415-429.

Godfrey, L.G., Orme, C.D. and J.M.C.S. Silva (2006). "Simulation-based tests for heteroskedasticity in linear regression models: Some further results." The Econometrics Journal **9**: 76-97.

Grinnell, J. (1917). "The niche-relationships of the California Thrashers." The Auk **34**: 427-433.

Guisan, A. and W. Thuiller (2005). "Predicting species distribution: offering more than simple habitat models." Ecology Letters **8**: 993-1009.

Guisan, A. and N. E. Zimmermann (2000). "Predictive habitat distribution models in ecology." Ecological Modelling **135**:147-186.

Hanski, I. (1991). "The functional response of predators: worries about scale." Trends in Ecology & Evolution **6**: 141-142

Harmon, M.E. (2001). "Carbon sequestration in forests : addressing the scale question." Journal of Forestry **99**: 24-29.

Helmer, O. and N. Rescher (1959). "On the Epistemology of the Inexact Sciences." Management Science **6**: 25-52.

Hepp, T.E. and G.H. Brister (1982). "Estimating crown biomass in loblolly pine plantations in the carolina flatwoods." Forest Science **28**: 115-127.

Henry, M., A. Besnard, W.A. Asante, J. Eshun, S. Adu-Bredu, R. Valentini, M. Bernoux, and L. Saint-Andre (2010). "Wood density, phytomass variations within and among trees, and allometric equations in a tropical rainforest of Africa." Forest Ecology and Management **260**: 1375-1388.

Hodgson, J.A., Moilanen, A. Wintle, B.A. and C.D. Thomas (2011). "Habitat area, quality and connectivity: striking the balance for efficient conservation." Journal of Applied Ecology **48**: 148-152.

Hodgson, J.A., Thomas, C.D., Wintle, B.A. and A. Moilanen (2009). "Climate change, connectivity and conservation decision making: back to basics." Journal of Applied Ecology **46**: 964-969.

Horppila, J. and Kairesalo, T. (1992). "Impacts of bleak (*Alburnus alburnus*) and roach (*Rutilus rutilus*) on water quality, sedimentation and internal nutrient loading." Hydrobiologia **243-244**: 323-331.

Houghton, R.A. (1999). "The annual net flux of carbon to the atmosphere from changes in land use 1850–1990*." Tellus B **51**: 298-313.

Houghton, R.A. (2005). "Aboveground forest biomass and the global carbon balance." Global Change Biology **11**: 945-958.

Houghton, R.A. and J.L. Hackler (2001). "Carbon Flux to the Atmosphere from Land-Use Changes: 1850–1990 Carbon Dioxide Information Analysis Center" Oak Ridge National Laboratory, Oak Ridge, TN 37831.

Huggett, A.J. (2005). "The concept and utility of ecological thresholds' in biodiversity conservation." Biological Conservation **124**: 301-310.

Hutchinson, G.E. (1957). "Concluding remarks." Cold Spring Harbour Symposium on Quantitative Biology **22**: 415–427.

Jackson, D.A. and H.H. Harvey (1993). "Fish and Benthic Invretebrates : Community concordance and Commmunity-Environment relationships." Canadian journal of Fisheries and Aquatic Sciences **50**: 2641-2651.

Jackson, D.A., Peres-Neto, P.R. and J.D. Olden (2001). "What controls who is where in freshwater fish communities - the roles of biotic, abiotic, and spatial factors." Canadian journal of Fisheries and Aquatic Sciences **58**: 157-170.

Jenkins, D.G. and A.L. Buikema (1998). "Do similar communities develop in similar sites? A test with zooplankton structure and function." Ecological Monographs **68**: 421-443.

Ketterings, Q.M., Coe, R., van Noordwijk, M., Ambagau, Y., and C.A. Palm (2001). "Reducing uncertainty in the use of allometric biomass equations for predicting above-ground tree biomass in mixed secondary forests." Forest Ecology and Management **146**: 199-209.

Kim, K. and N.H. Timm (2007). "Univariate and multivariate general linear models: theory and applications with SAS. 2" illustrated edition. Chapman & Hall.

Koivula, M. and H. Vermeulen (2005) Highways and Forest Fragmentation – Effects on Carabid Beetles (Coleoptera, Carabidae). Landscape Ecology **20**: 911-926.

Kottelat, M. and J. Freyhof (2007). "Handbook of European freshwater fishes." Publications Kottelat, Cornol, Switzerland.

Kraenzel, M., Castillo, A., Moore, T. and C. Potvin (2003). "Carbon storage of harvest-age teak (Tectona grandis) plantations, Panama." Forest Ecology and Management **173**: 213-225.

Kremen, C. (2005). "Managing ecosystem services: what do we need to know about their ecology?" Ecology Letters **8**: 468-479.

Kremen, C. and R.S. Ostfeld (2005). "A call to ecologists: measuring, analyzing, and managing ecosystem services." Frontiers in Ecology and the Environment **3**: 540-548.

Kromrey, J.D. and C.V. Hines (1995). "Use of empirical estimates of shrinkage in multiple regression: a caution." Educational and Psychological Measurement **55**: 901-925.

Lal, R. (2008). "Carbon sequestration." Philosophical Transaction of the Royal Society of London, B, Biological Sciences **363**: 815-830.

Laliberté, E., Paquette, A. Legendre, P. and A. Bouchard (2009). "Assessing the scale-specific importance of niches and other spatial processes on bêta diversity: a case study from a temperate forest." Oecologia **159**: 377-388.

Larsen, D.R. and P.L. Speckman (2004). "Multivariate Regression Trees for Analysis of Abundance Data." Biometrics, **60**: 543-549.

Legendre, P. (2008). "Studying bêta diversity: ecological variation partitioning by multiple regression and canonical analysis." Journal of Plant Ecology **1**: 3-8.

Legendre, P. and M.J. Anderson (1999). "Distance-based redundancy analysis: testing multispecies responses in multifactorial ecological experiments." Ecological Monographs **69**: 1-24.

Legendre, P., Borcard, D. and P.R. Peres-Neto (2005). "Analyzing beta diversity: partitioning the spatial variation of community composition data." Ecological Monographs **75**: 435-450.

Legendre, P., Dale, M.R.T., Fortin, M.-J. Casgrain, P. and J. Gurevitch (2004). "Effects of spatial structure on the results of field experiments." Ecology **85**: 3202-3214.

Legendre, P., Dale, M.R.T., Fortin, M.-J. J., Gurevitch, M. Hohn and D. Myers (2002). "The consequences of spatial structure for the design and analysis of ecological field surveys." Ecography **25**: 601-615.

Legendre, P., De Caceres, M. and D. Borcard (2010). "Community surveys through space and time: testing the space and time interaction in the absence of replication." Ecology **91**: 262-272.

Legendre, P. and M. J. Fortin (1989). "Spatial pattern and ecological analysis." Vegetatio **80**: 107-138.

Legendre, P. and E.D. Gallagher (2001). "Ecologically meaningful transformations for ordination of species data." Oecologia **129**: 271-280.

Legendre, P. and L. Legendre (1998). "Numerical Ecology." Second English Edition edition. Elservier.

Legendre, P. and L. Legendre (2012). "Numerial Ecology". Third English Edition edition. Elservier.

Legendre, P., Mi, X., Ren, H., Ma, K., Yu, M., Sun, I.-F. and F. He (2009). "Partitioning beta diversity in a subtropical broad-leaved forest of China." Ecology **90**: 663-674.

Legendre, P., Oksanen, J. and C.J.F. ter Braak (2011). "Testing the significance of canonical axes in redundancy analysis." Methods in Ecology & Evolution (in press).

Loreau, M. (2010). "The challenges of biodiversity science." Excellence in Ecology **17**. International Ecology Institute, Oldendorf/Luhe, Germany. xxviii + 120 pp.

Losi, C.J., Siccama, T.G., Condit, R. and J.E. Morales (2003). "Analysis of alternative methods for estimating carbon stock in young tropical plantations." Forest Ecology and Management **184**: 355-368.

MacArthur, R.H. (1972). "Patterns in the distribution of species." Harper & Row, New York.

Machado, J.A.F. and J.M.C.S. Silva (2000). "Glejser's test revisited." Journal of Econometrics **97**: 189-202.

Madgwick, H.A.I. and T. Satoo (1975). "On Estimating the Aboveground Weights of Tree Stands." Ecology **56**: 1446-1450.

Marrs, R.H., Galtress, K., Tong, C., Cox, E.S., Blackbird, S.J., Heyes, T.J., Pakeman, R. J. and M.G. Le Duc (2007). "Competing conservation goals, biodiversity or ecosystem services: Element losses and species recrutment in amanged moorlan-bracken model system." Journal of Environmental Management **85**:1034-1047.

McArdle, B.H. and M.J. Anderson (2001). "Fitting multivariate models to community data: a comment on distance-based redundancy analysis." Ecology **82**: 290-297.

McCullagh, P. and J.A. Nelder (1989). "Generalized Linear Models." Chapmall and Hall.

McCulloch, C.E. (2000). "Generalized Linear Models." Journal of the American Statistical Association **95**:1320-1324.

McGovern, S., Evans, C.D., Dennis, P., Walmsley, C. and M.A. McDonald (2011) "Identifying drivers of species compositional change in a semi-natural upland grassland over a 40-year period." Journal of Vegetation Science **22**: 346-356.

McQuarrie, A.D.R. and C.-L. Tsai (1998). "Regression and time series model selection." World Scientific Publishing Company, Singapore.

Menge, B.A. and A.M. Olson (1990). "Role of scale and environmental factors in regulation of community structure." Trends in Ecology & Evolution **5**: 52-57.

Minchin, P. R. (1987). "Simulation of multidimensional community patterns: towards a comprehensive model." Plant Ecology **71**: 145-156.

Moyle, P.B. and B. Vondracek (1985). "Persistence and Structure of the Fish Assemblage in a Small California Stream." Ecology **66**: 1-13.

Myers, R.H., Montgomery, D.C. and V.G. Goeffrey (2002). "Generalized linear models with applications in engineering and the sciences." New-York, John Wiley & Sons, inc.

Neter, J., Kutner, M.H., Nachtsheim, C.J. and W. Wasserman (1996). "Applied Linear Statistical Models." Fourth edition. IRWIN.

Ouellette, M.-H., DesGranges, J.-L., Legendre, P. and D. Borcard (2005). "L'arbre de régression multivariables: classification d'assemblage d'oiseaux fondée sur les caractéristiques de leur habitat." *in* Société Francophone de Classification, Montréal.

Ouellette, M.-H. and P. Legendre (2011). "An adjusted R^2 statistic for multivariate regression tree analysis." *Manuscript*.

Overman, M.P.J., Witte, L.J.H. and G.J. Saldarriaga (1994). "Evaluation of regression models for above ground biomass determination in Amazon rainforest." Journal of Tropical Ecology **10**: 207-218.

Page, L.M. and B.M. Burr (1991). "A field guide to freshwater fishes of North America north of Mexico." Houghton Mifflin Company, Boston.

Palumbi, S.R., Sandifer, P.A., Allan, J.D., Beck, M.W., Fautin, D.G., Fogarty, M.J., Halpern, B.S., Incze, L.S., Leong, J.-A., Norse, E., Stachowicz, J.J. and D.H. Wall (2009). "Managing for ocean biodiversity to sustain marine ecosystem services." Frontiers in Ecology and the Environment **7**: 204-211.

Pelletier, J., Kirby, K. and C. Potvin (2010). "Significance of carbon stock uncertainties on emission reductions from deforestation and forest degradation in developing countries." Forest Policy and Economics (Accepted for publication).

Peres-Neto, P.R., Legendre, P., Dray, S. and D. Borcard (2006). "Variation partitioning of species data matrices: estimation and comparison of fractions." Ecology **87**: 2614-2625.

Pihu, E. (1996). "Fishes, their biology and fisheries management in Lake Peipsi." Hydrobiologia **338**: 163-172.

Pinzón, J. and J. Spence (2010). "Bark-dwelling spider assemblages (Araneae) in the boreal forest: dominance, diversity, composition and life-histories." Journal of Insect Conservation **14**: 439-458.

Potvin, C. and N. Gotelli (2008). "Biodiversity enhances individual performance but does not affect survivorship in tropical trees." Ecology Letter **11**: 217-223.

Potvin, C., Mancilla, L., Buchmann, N., et al. (2011). "An ecosystem approach fo biodiversity effects : Carbon pools in a tropical tree plantation." Forest Ecology and Management **261**: 1614-1624.

Pulliam, H. R. (2000). "On the relationship between niche and distribution." Ecology Letters **3**: 349-361.

R Development Core Team (2010). "R: A language and environment for statistical computing." R Foundation for Statistical Computing, Vienna, Austria.

Rao, C. R. (1964). "The use and interpretation of principal component analysis in applied research." Sankhyaa Series A **26**: 329-358.

Rejwan, C., Collins, N.C., Brunner, L.J., Shuter B.J., and M.S. Ridgway (1999). "Tree Regression Analysis on the Nesting Habitat of Smallmouth Bass." Ecology **80**: 341-348.

Rodriguez, M.A. and W.M. Lewis Jr. (1997). "Structure of Fish Assemblages Along Environmental Gradients in Floodplain Lakes of the Orinoco River." Ecological Monographs **67**: 109-128.

Schlosser, I. J. (1982). "Fish Community Structure and Function along Two Habitat Gradients in a Headwater Stream." Ecological Monographs **52**: 395-414.

Segal, M. R. (1992). "Tree-Structured Methods for Longitudinal Data." Journal of the American Statistical Association **87**: 407-418.

Sheaves, M., Abrantes, K. and R. Johnston (2007). "Nursery ground value of an endangered wetland to juvenile shrimps." Wetlands Ecology and Management **15**: 311-327.

Shmueli, G. (2010). "To Explain or To Predict?" Statistical science **25**: 289-310.

Shvidenko, A., Barber, C.V. and R. Persson (2005). "Forest and Woodland Systems." Pages 585-621 *in* M. d. l. Angeles and C. Sastry, editors. Millennium Ecosystem Assessment : Ecosystem and human well-being. Island Press, Washington, DC.

Sonderegger, D.L., Wang, H., Clements W.H. and B.R. Noon (2009). "Using SiZer to detect thresholds in ecological data." Frontiers in Ecology and the Environment **7**: 190-195.

Southwood, T.R.E. (1977). "Habitat, the templet for ecological stategies?" Journal of Animal Ecology **46**: 337-365.

Southwood, T.R.E. (1988). "Tactics, strategies and templets." Oikos **52**: 3-18.

Sprugel, D.G. (1983). "Correcting for Bias in Log-Transformed Allometric Equations." Ecology **64**: 209-210.

Taylor, C.M., Winston, M.R. and W.J. Matthews (1993). "Fish species-environment and abundance relationship in a Great Plains river system." Ecography **16**:16-23.

ter Braak, C.J.F. (1988) "Partial canonical correspondence analysis." Classification and related methods of data analysis (ed. H.-H. Bock), pp. 551-558. North-Holland, Amsterdam.

Townsend, C.R. and A.G. Hildrew (1994). "Species traits in relation to a habitat templet for river systems." Freshwater Biology **31**:265-275.

Urban, M.C., Skelly, D.K., D. Burchsted et al. (2006). "Stream communities across a rural–urban landscape gradient." Diversity and Distributions **12**(4): 337-350.

Van, T.K., Rayachhetry, M.B. and T.D. Center (2000). "Estimating Above-ground Biomass of *Melaleuca quinquenervia* in Florida, USA." Journal of Aquatique Plant Management **38**: 62-67.

Vayssiéres, M.P., Plant, R.E. and B.H. Allen-Diaz (2000). "Classification Trees: An Alternative Non-Parametric Approach for Predicting Species Distributions." Journal of Vegetation Science **11**: 679-694.

Verneaux, J. (1973). "Cours d'eau de Franche-Comté (Massif du Jura). Recherches écologiques sur le réseau hydrographique du Doubs." Essai de biotypologie. Thèse d'état.

Walker, H.M. (1940). "Degrees of freedom." Journal of Educational Psychology **31**: 253-269.

Wang, C. (2006). "Biomass allometric equations for 10 co-occurring tree species in Chinese temperate forests." Forest Ecology and Management **222**(1-3): 9-16.

Wang, G.C.S. and C.L. Jain (2003). Regression analysis: modeling & forecasting, Graceway Pub.

Walker, H.M. (1940). "Degrees of freedom." Journal of Educational Psychology **31**(4): 253-269.

Wang, C. (2006). "Biomass allometric equations for 10 co-occurring tree species in Chinese temperate forests." Forest Ecology and Management **222**: 9-16.

Wang, G.C.S. and C.L. Jain (2003). "Regression analysis: modeling & forecasting." Institute of Business Forec.

Whittaker, R.H. (1960). "Vegetation of the Siskiyou Mountains, Oregon and California." Ecological Monographs **30**: 279-338.

Whittaker, R.H. (1972). "Evolution and Measurement of Species Diversity." Taxon **21**: 213-251.

Wiens, J.A., Addicott, J.F. Case T.J., and J. Diamond (1986). "Overview : the importance of spatial and temporal scale in ecological investigations." Pages 145-153 *in* J. Diamond and T. J. Case, editors. Community Ecology. Harper and Row, New-York.

Wiens, J.A., Rotenberry J.T. and B. Van Horne (1987). "Habitat occupancy patterns of North America shrubsteppe birds : The effects of spatial scale." Oikos **48**: 132-147.

Wood Density Database http://www.worldagroforestry.org/sea/Products/AFDbases/WD/.

Work, T.T., Shorthouse, D.P. Spence, J.R. Volney J.A. and D. Langor (2004). "Stand composition and structure of the boreal mixedwood and epigaeic arthropods of the Ecosystem Management Emulating Natural Disturbance (EMEND)

landbase in northwestern Alberta." Canadian Journal of Forest Research **34**: 417-430.

Ye, J. (1998). "On Measuring and Correcting the Effects of Data Mining and Model Selection." Journal of the American Statistical Association **93**: 120-131.

Zaman, A. (2000). "The inconsistency of the Breusch-Pagan Test." Journal Economical Social Research **2**: 1-11.