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

Marie-Hélène Ouellette[1,2,*], Pierre Legendre[1,2] and Daniel Borcard[2]

[1]*Groupe de recherche interuniversitaire en limnologie et en environnement aquatique (GRIL).*
[2]*Département de sciences biologiques, Université de Montréal, C.P. 6128, Succursale Centre-Ville, Montréal, QC H3C 3J7, Qc, Canada.*

[*]*Correspondence author. E-mail: marieheleneo@gmail.com*

Running head: Nested multivariate regression trees

Word count: 7304

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.

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, take 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 taken into account. This is useful when one of the sets is thought to have an effect that varies as a function of the other.

3. To achieve this type of assessment, we propose a *cascade of multivariate regression trees* (CMRT). We decompose the total dispersion of a response matrix between two explanatory data sets in a nested manner. By handling each leaf (group) resulting from the first-level 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 the users to impose a nested structure to their causal hypotheses in multivariate regression tree analysis.

5. To illustrate this new procedure, we use 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).

KEY-WORDS: cascade, multivariate regression tree, nested explanatory assessment, species composition drivers

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 2012). 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 (redundancy analysis, RDA, and canonical correspondence analysis, CCA) and MANOVA (Legendre & Anderson 1999; Anderson 2001; 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 RDA: Davies & Tso 1982; 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 & Gribble 1998; Peres-Neto *et al*. 2006). The effects of both explanatory sets are then hypothesized to be additive. Partial RDA and partial CCA both allow a constrained ordination of the response matrix **Y** on the explanatory variables **X** 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 can 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 when 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 such a case, thresholds better describe the relationship between the two data sets than linear models. Subsequently, the variation of each leaf (or 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 allocates objects of the response matrix to homogenous groups, with partition criteria imposed 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 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.

CMRT is a procedure modelling the response data by means of 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. 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, or the residuals normally distributed. 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.

MATERIALS AND METHODS

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 data set analysed at each level; see Fig. 1 for a diagram of the general structure. The number of waves is equal to 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.

Let **Y** be the response matrix, **A** and **S** be respectively the main and subordinate explanatory tables. First an MRT model is computed with **Y** as the response and **A** as the explanatory table. Variables in **A** may represent spatial scales: broad, medium and fine scales, or else landscape and microhabitat variables, which are another representation of scales. The hierarchy can also be based on the nature of the explanatory data sets, for example: morphometry of a river (main) and land use (subordinate). See the *Nested hypotheses in ecology* subsection of the Discussion for more examples. Cross-validation is carried out to prune the tree and complete the first wave of the cascade.

Pruning is achieved by a resampling method called *v*-fold cross-validation (Breiman *et al*. 1984). First, the response data set is randomly split into *v* test subsets of roughly equal numbers of objects. These test subsets correspond to randomly chosen response row vectors, e.g. sites, with their corresponding species abundances. Then, *v* trees are built from *v* learning sets constructed by removing each of the *v* test sets one at a time from the whole set of objects. All trees are fully grown, and for each tree size, the cross-validated relative error is calculated as follows:

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

where $y_{ij(k)}$ is one observation in test set *k*, $\hat{y}_{j(k)}$ is the predicted value of this observation in the *k*-th tree computed from the corresponding learning set, *n* is the number of observations, and *m* is the number of variables in the response matrix **Y**. Cross-validation for the *v* test subsets produces predicted values for all *n* observations, which are all included in the calculation of the *CVRE* statistic. If the response data contains species abundances, the predicted response for observation *i* is a particular species composition.

The first wave thus consists of analyzing a single drop containing all observations through an MRT model. The response set is hypothesized to vary as a function of **A** (main effect). The first wave will identify groups of sites with the most homogeneous species composition, split by the explanatory variables coded in **A**.

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 shapes the partition produced by this first wave by limiting the number of splits. The value is

131 determined by the user: a split will not be performed unless it explains at least as much variation
132 as the chosen R^2 value. For the first wave of analysis, it is important to set the complexity
133 parameter high enough to identify only the main factors determining variation in species
134 composition.

135 Let ***g*** be the number of leaves resulting from the first wave (see Fig. 1.b). In a second step,
136 the variation of the response data in each leaf (the leaf response tables are noted $\mathbf{Y_h}$ for h = 1, …,
137 g) is modelled independently with the **S** explanatory table to form the subordinate drops. For
138 these drops, the complexity parameter may be reduced to a small value since the second wave is
139 intended to model finer variation in species composition. The default value of the complexity
140 parameter is 0.01 in the ***mvpart()*** R function; it is passed from ***rpart.control()***; both functions are
141 found in the MVPART package.

142 The algorithm used to fit all MRT models is the standard recursive greedy splitting
143 algorithm described in Breiman *et al* (1984) and De'ath (2002). Each tree is fully grown and its
144 final size is chosen by v-fold cross validation, where v is often chosen to be 10 (Breiman *et al*.
145 1984). The user may choose between the "min" or "1se" rules (Breiman *et al* 1984, De'ath 2002)
146 for the validation. Breiman *et al.* (1984) suggested that both rules should lead to about the same
147 risk. The less complex model, obtained with the "1se" rule, should be chosen in most cases since
148 the aim is to minimise both risk and complexity. In the case of a regression tree, the risk is the
149 cross-validated relative error (the variation between true and predicted values of test objects
150 divided by the total variation of the response) and the complexity is the number of splits. Thus
151 the risk versus complexity assessment can be made by examining the plot of the cross-validated
152 relative error as a function of the size of the tree, to see if both rules lead to similar risks. In this
153 paper, the within-group sum-of-squares around the mean is used as the criterion to be minimized,
154 even though other criteria could be used.

155 The combined model, called *cascade*, is exactly that: a cascade of models, depicting in a
156 nested manner the partitioning of the response data by two sets of explanatory variables (Fig. 1).
157 Two general conclusions may emerge from a cascade: either the explanatory variables and splits
158 of the second wave are the same for all leaves identified in the first wave, which means that the
159 subordinate effect is the same over all subordinate data sets, or they are not. Therefore the
160 subordinate drops may be examined in turn to identify differences in splits and explanatory
161 variables among them. This approach is conceptually analogous to the search for interaction in
162 MANOVA.

163 ## R^2 PARTITION

164 We define the R^2 of a single MRT tree as 1 minus the relative error defined by De'ath
165 (2002). Thus a single coefficient of determination (R^2) can be computed for each drop.
166 Consequently, a coefficient of determination (R^2) can be obtained for the global analysis by
167 pooling the R^2 of the first wave with the weighted R^2 obtained in the second wave (Fig. 1b). The
168 CMRT procedure implies that the subordinate R^2 are computed as proportions of the variation of
169 their corresponding leaf as defined in the first wave. Each of these secondary drop R^2 must then
170 be reexpressed as a proportion of the total variation of the response data; the overall R^2 is finally
171 obtained by summing the R^2 of the main drop and those of the subordinate drops. This procedure
172 is valid because the subordinate drops are independent from one another.

173 The diagram provided by the ***CasMRTR2()*** function of the MVPARTWRAP package takes
174 the form of a square of unit area. The entire area represents the total variation of the response
175 data. The proportion of the total response variation explained by each drop is represented as a
176 shaded box of corresponding area. The box for the drop of the first wave is at the far left. Its
177 height is 1, so that its width represents its R^2. The partitioning of the remaining variation is

represented by drawing a box for each drop produced by the first wave. 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. The heights of the rectangles represent the R^2 of the subordinate drops *within* their leaf. Therefore, their areas represent their R^2 as ratios of the total response variation.

CASE STUDIES

We illustrate the CMRT procedure by using two data sets that were submitted to 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 is used for the first wave, and the usual 0.01 value is used for the second wave. Also, both community response matrices are Hellinger transformed prior to the analysis (Legendre & Gallagher 2001).

ORIBATID MITES

The first data set consists of three data tables (species abundances of oribatid mites, micro-environmental variables, and spatial coordinates) extracted from 70 peat moss cores collected 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 (Borcard & Legendre 1994). The sampling area is 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) determines the first split of the MRT (Fig. 2). Since oribatids are not aquatic, in this extremely wet environment some oribatids prefer 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. Other substrate and micro-environmental variables are available as explanatory variables, in particular 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 density' as the main effect because shrubs impose particular microclimate and microsubstrate changes 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 shrubs (indicator morphospecies *Tectocepheus velatus*, *Malaconothrus cf. egregius, Oppiella nova, Fuscozetes setosus, Hypochthoniella sp1 & sp2,* and *Galumnidae*).

In subordinate drops 2 and 3 (Fig. 3), different explanatory variables are identified 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 (0.163) 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. Fig. 4 shows the variation partitioning of the original response of the oribatid mite data by drops.

DOUBS RIVER FISH

The Doubs River fish data were collected by Verneaux (1973; see also Verneaux *et al.* 2003) who considered the fish species composition to be an ecological indicator of water quality 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's thesis, i.e., 30 sites described by three data tables: the fish species composition (abundance classes ranging from 0 to 5), 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 (Fig. 5), the distance from 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 CMRT procedure, we use the morphological variables 'mean discharge' and 'slope' as the main explanatory set. Note that these variables are likely to have been represented in the CMRT analysis by their proxy, i.e., distance from the source. We will comment this choice, made for demonstration purposes only, in the Discussion. The physical and chemical variables (calcium concentration (hardness), pH, phosphate, nitrate, ammonium, dissolved oxygen and biological oxygen demand) are selected as the subordinate explanatory set.

The resulting cascade is shown in Figure 6. 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

247 Salmonid region is characterized by the brown trout (*Salmo trutta fario,* a Salmonid) and the
248 common minnow (*Phoxinus phoxinus*, a Cyprinid). The Cyprinid region has the bleak (*Alburnus*
249 *alburnus*), the common nase (*Chondrostoma nasus*), the ruff (*Acerina cernua*), the pumpkinseed
250 sunfish (*Lepomis gibbosus*), the European bitterling (*Rhodeus amarus*), the European eel
251 (*Anguilla anguilla*), the roach (*Rutilus rutilus*), the spirlin (*Spirlinus bipunctatus*), the common
252 carp (*Cyprinus carpio*), the whitebream (*Blicca bjoerkna*), the common barbell (*Barbus barbus*),
253 the common bream (*Abramis brama*), the rudd (*Scardinius erythrophthalmus*) and the south-west
254 European nase (*Chondrostoma toxostoma*) as indicator species.

255 Within each zone identified by the first drop, the water quality variables are used in the
256 subordinate analyses to identify and explain finer differences in species composition. No further
257 splits are found in the Salmonid region (*v*-fold cross-validation pointed to one group). It is not the
258 case for the Cyprinid region (right-hand leaf of drop 1, called drop 3 in our analysis), which
259 showed three species assemblages responding to two explanatory variables: ammonium
260 concentration and dissolved oxygen; see Fig. 6 for a map of the sites along the river and the
261 cascade of analyses, and Fig. 7 for a summary of the explained variation.

262 Group 2 of the tree of drop 3 contains sites 23-25, characterized by large concentrations of
263 ammonium (>=0.45 mg/L) and, by correlation, by large concentrations of phosphorus (r =
264 0.9695) and high biological oxygen demand (r = 0.8858); these two variables, which would
265 produce the same split, are not shown in the tree. The bleak *Alburnus alburnus*, the chub
266 *Leuciscus cephalus cephalus*, and the roach *Rutilus rutilus* are the indicator species of this group
267 (sites 23-25). The bleak is present at sites 21-30 but particularly successful at the highly
268 eutrophized sites 23-25. This species feeds on zooplankton near the surface (Horppila &
269 Kairesalo 1992) which is, for this species, an important habitat for feeding (de Nie 1987) and to
270 lay eggs (Pihu 1996). Thus the indicator value of this species at sites 23–25 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 one is 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

295 silver bream (*Blicca bjoerkna)* still waters (Kottelat & Freyhof 2007), and the pumpkinseed
296 (*Lepomis gibbosus*) vegetated pools (Page & Burr 1991).

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

304 ## DISCUSSION

305 ### GENERAL REMARKS ON THE PROCEDURE

306 CMRT offers the opportunity to address ecological hypotheses in a preferential order,
307 allowing one to override the original explanatory order of the variables presented in MRT
308 analysis to explore specific avenues by testing the influence of precise variables on the response
309 data. Both MRT and CMRT produce a hierarchy; the peculiarity of the CMRT procedure resides
310 in the possibility to pre-select the explanatory set of variables that will be used to compute the
311 first few bipartitions. Therefore, when creatively applied with specific hypotheses in mind, the
312 cascade provides new insights on the data structure that would not have been available in simple
313 MRT analysis. In order to exploit the CMRT procedure to its full potential, the explanatory
314 variables selected for the first wave should be different from the first bipartition of the simple
315 MRT; if it was the same, the two analyses would depict the same pattern. Actually, if we had
316 used in CMRT the same first explanatory variables that were identified by the simple MRT
317 model, the resulting CMRT model would have been the same as the MRT result, but with a

318 smaller number of leaves because the independent cross-validations conducted in the drops
319 would have reduced the overall power. This is what happened with our second example (Doubs
320 fish data), where the variable selected among those chosen for the first wave, mean minimum
321 discharge (Fig. 3, drop 1), was represented by its proxy, distance from the source, in the non-
322 hierarchical MRT (Fig. 2). Furthermore, because distance from the source actually explains much
323 of the chemical variation along the river, it also appears in further splits of the original MRT.
324 This variable being absent from the CMRT, the corresponding splits have been identified by true
325 physical or chemical variables, leading to similar, if not completely identical results.

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

334 For the total sum of squares of Y to be meaningful, the response variables have to be
335 dimensionally homogeneous, so that the sums of squares of individual variables are additive. If
336 the variables are not dimensionally homogenous, e.g. environmental descriptors, they have to be
337 standardized prior to MRT and CMRT analysis.

338 **THE CASE STUDIES**

339 We propose two contrasting case studies to stress the importance of the choice of the
340 explanatory variables of the first wave. The oribatid mite example illustrates a case where a

hypothesis about the effect of shrub vegetation leads us to impose the corresponding variable as a first-level effect. This is clearly different from the spontaneous order of the variables, as revealed by the standard MRT, but it allows us to gain new insights about the hypothesized effect. Therefore, the application of CMRT adds to our knowledge of the ecological processes at work.

The Doubs river example, on the contrary, shows that the choice of a variable representing many of the important ecological drivers of the river system and strongly correlated to the first-rank variable identified in the classical MRT is not adequate. This choice not only leads to a result that closely resembles the one obtained by classical MRT, but this result is impoverished by the lack of power induced by the sequential nature of the CMRT method. Therefore, in this case, the CMRT has not added to our knowledge of the system, although, as shown in the Results, this simpler classification can be well explained in ecological terms.

NESTED HYPOTHESES IN ECOLOGY

CMRT allows for the first time users to impose a hierarchy 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 and land use impact variables as the subordinate effect. With the CMRT procedure, inherently, for each of the groups identified by the morphometry explanatory data, the subordinate effect of land use impact can be studied and identified in a fully independent manner.

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. This first step, corresponding to a clustering with chronological constraint, is followed by secondary analyses of each segment

using environmental variables, where different explanatory variables may express themselves in different time segments. 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 using the distance from the source variable, 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 clustering, spatially constrained by the geographic coordinates of the sites (see e.g. Legendre & Legendre 2012, Chapitre 13). .

Another possible application of CMRT is for space-time surveys. Legendre, De Cáceres & Borcard (2010) showed how one can 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 community data, has changed through time, or conversely, that the species composition has changed differently through time at the different sampling sites. In that case, CMRT could be used to analyse the multivariate time series from each site separately, or the multivariate data across sampling sites from each sampling time separately.

In some applications, the nested structure inherent in CMRT may be imposed by the researcher for heuristic reasons. 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 at n sites in a forest (space) over t time steps, and the study is concerned with 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 delineate regions of the forest, i.e. groups of geographically contiguous sites, that are the most similar through times. Each of these regions with similar species assemblages may respond differently in time to disturbances: for example a local drought could boost the invasive ability of a species. The secondary analysis, done separately on each region using time as the explanatory variable, would allow the identification of regions where the communities changed most over time, possibly as a consequence of the invasion. (2) Let us now suppose that our main interest is to study the effect of an unusually long drought affecting the whole forest. In this case we would use time as the main factor to first focus on the evolution of the overall species composition through time, pointing perhaps at main extinction events due to this drought. Subsequently, we could study each assemblage identified along the time line and see how they are structured in space, or with respect to environmental factors that may condition the structure of the community through space. The number of sites affected by the invasive species may vary greatly from time to time.

In any space-time study, spatial or temporal correlation between adjacent sites due to neutral processes of community dynamics (see e.g. Legendre & Legendre 2012) may be present along one or the other sampling axis, or both. Further simulations are needed to fully understand the effect and the extent of this effect on the CMRT modelling process, notably on the cross-validated estimation of the error made on a prediction, which is the basis to pick the size of the tree.

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

411 linear model, if the assumptions of such a procedure are met, may also be used to model the
412 subordinate drops. Thus a mixture of modelling procedures may be used in the framework. The
413 computation of an overall R^2 is still valid because the subordinate analyses are independently
414 conducted in each drop and the calculation of the R^2 is independent in each analysis. This
415 framework is also applicable to univariate CART classification or regression tree models.
416 Moreover, more than two waves could be used. This would require that the data set be large in
417 order to have a sufficient number of sites in the leaves of the second wave and some variation left
418 to be explained in the third wave of the analysis.

419 **RELATING CMRT TO NESTED MANOVA**

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

432 Finally, the possibility of fitting trees in an additive, non-nested way has yet to be
433 explored. But this approach would be conceptually very different from CMRT, it would answer

434 different questions, and its development would imply the resolution of several mathematical
435 issues about the nature and computation of residuals in a non-linear context.

436 ## CONCLUSION

437 The CMRT procedure is a framework where nested ecological hypotheses are privileged.
438 Users must choose in which order two (or more) explanatory sets are considered in an MRT
439 structure. It is also possible to partition the explained variation (R^2) among the sets and ultimately
440 obtain a coefficient of determination for the complete cascade of MRT analyses. The final CMRT
441 model may be subjectively assessed for interaction between the explanatory sets, to evaluate if
442 the effect of the subordinate set changed as a function of the group membership produced by the
443 first wave of analysis. The overall procedure is interesting for fundamental as well as applied
444 ecological studies, and may be applied in other fields such as geography, oceanography, soil
445 science, as well as outside the biological domain.

446 ## ACKNOWLEDGEMENTS

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

450 ## REFERENCES

451 Anderson, M.J. (2001a). A new method for non-parametric multivariate analysis of variance.
452 *Austral Ecology*, **26**, 32-46.
453 Anderson, M J. (2001b). Permutation tests for univariate or multivariate analysis of variance and
454 regression. *Canadian journal of Fisheries Aquatic Science*, **58**, 626-639.

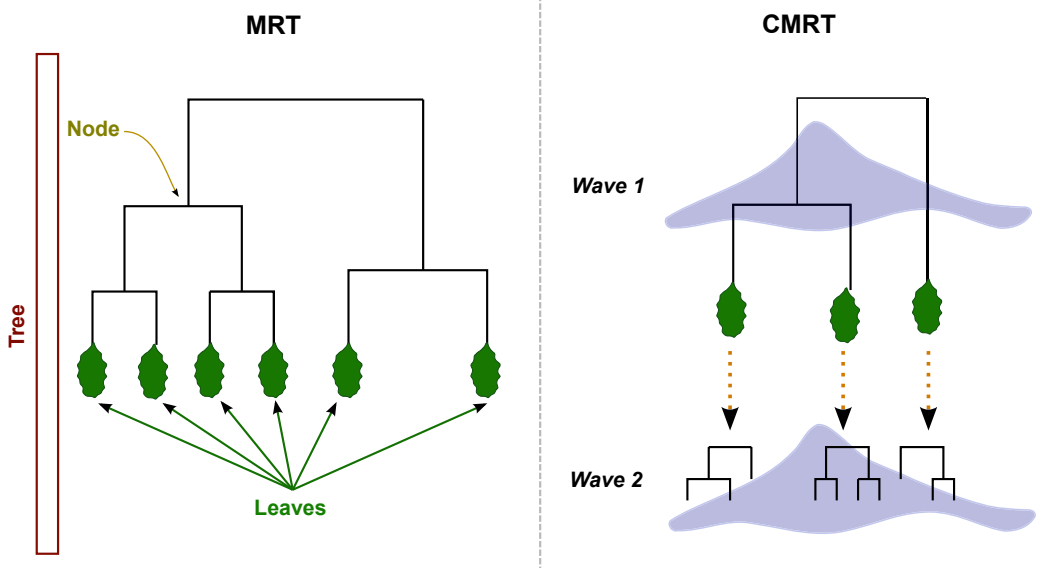
455 Anderson, M.J. and N.A. Gribble (1998) Partitioning the variation among spatial, temporal and
456 environmental components in a multivariate data set. *Austral Ecology*, **23**, 158-167.
457 Auguet, J.-C., Barberan, A. & E.O. Casamayor (2010) Global ecological patterns in uncultured
458 Archaea. *ISME J*, **4**, 182-190.
459 Borcard, D., Gillet, F. & P. Legendre (2011). *Numerical ecology with R*. New York, Springer.
460 Borcard, D. and P. Legendre (1994) Environmental control and spatial structure in ecological
461 communities: an example using oribatid mites (Acari, Oribatei). *Environmental and*
462 *Ecological statistics*, **1**, 37-61.
463 Borcard, D., Legendre, P. & P. Drapeau (1992) Partialling out the Spatial Component of
464 Ecological Variation. *Ecology*, **73**, 1045-1055.
465 Breiman, L., Friedman, J.H., Olshen, R.A. & C.J. Stone (1984) *Classification and Regression*
466 *Trees*. Belmont, California, USA, Wadsworth International Group.
467 Carr, G.M. and P.A. Chambers (1998) Macrophyte growth and sediment phosphorus and
468 nitrogen in a Canadian prairie river. *Freshwater Biology*, **39**, 525-536.
469 Chappaz, R., Brun, G. & G. Olivari (1989) Données nouvelles sur la biologie et l'écologie d'un
470 poisson Cyprinidé peu étudié *Chondrostoma toxostoma* (Vallot, 1836). Comparaison avec
471 *Chondrostoma nasus* (L. 1766). *Comptes rendus de l'académie des sciences, Paris, série*
472 *III*, **309**, 181-186.
473 Chen, L., Mi, X., Comita, L.S., Zhang, L., Ren, H. & K. Ma (2010) Community-level
474 consequences of density dependence and habitat association in a subtropical broad-leaved
475 forest *Ecology Letters*, **13**, 695-704.
476 Crossman, E.J. (1996) Taxonomy and distribution. *Pike biology and exploration*. J. F. Craig.
477 London, Chapman and Hall**:** 1-11.

478 Davidson, T.A., Sayer, C.D., Langdon, P.G., Burgess, A. & M. Jackson (2010) Inferring past
479 zooplanktivorous fish and macrophyte density in a shallow lake: application of a new
480 regression tree model. *Freshwater Biology*, **55**, 584-599.
481 Davies, P.T. & M.K.S. Tso (1982) Procedures for Reduced-rank Regression. *Journal of the Royal*
482 *Statistical Society: Series C (Applied Statistics)*, **31**, 244-255.
483 De'ath, G. (2002) Multivariate regression trees : a new technique for modeling species-
484 environment relationships. *Ecology*, **83**, 1105–1117.
485 Deelder, C.L. (1984) *Synopsis of biological data on the eel, Anguilla anguilla (Linnaeus, 1758)*.
486 Rome, Italy, FAO.
487 de Nie, H.W. (1987) The decrease in aquatic vegetarian in Europe and its consequences for fish
488 populations, EIFAC/CECPI. **Occasional paper No. 19**.
489 DeVantier, L., De'ath, G., Turak, E., Done, T. & K. Fabricius (2006) Species richness and
490 community structure of reef-building corals on the nearshore Great Barrier Reef *Coral*
491 *Reefs*, **25**, 329-340.
492 DORIS (25/2/2010). *Leuciscus leuciscus* (Linnaeus, 1758),
493 http://doris.ffessm.fr/fiche2.asp?fiche_numero=2166.
494 DORIS (30/7/2010). *Phoxinus phoxinus* (Linnaeus, 1758),
495 http://doris.ffessm.fr/fiche2.asp?fiche_numero=1656.
496 Dufrêne, M. & P. Legendre (1997) Species assemblages and indicator species: the need for a
497 flexible asymmetrical approach. *Ecological Monographs*, **67**, 345-366.
498 Horppila, J. & T. Kairesalo (1992). Impacts of bleak (*Alburnus alburnus*) and roach (*Rutilus*
499 *rutilus*) on water quality, sedimentation and internal nutrient loading. *Hydrobiologia*, **243-**
500 **244**, 323-331.

501 Koivula, M. & H. Vermeulen (2005) Highways and Forest Fragmentation – Effects on Carabid
502 Beetles (Coleoptera, Carabidae). *Landscape Ecology*, **20**, 911-926.
503 Kottelat, M. & J. Freyhof (2007). *Handbook of European freshwater fishes*. Cornol, Switzerland,
504 Publications Kottelat.
505 Larsen, D.R. and P.L. Speckman (2004) Multivariate Regression Trees for Analysis of
506 Abundance Data. *Biometrics*, **60**, 543-549.
507 Legendre, P. & M.J. Anderson (1999) Distance-based redundancy analysis: testing multispecies
508 responses in multifactorial ecological experiments. *Ecological Monographs*, **69**, 1-24.
509 Legendre, P., De Cáceres, M. & D. Borcard (2010) Community surveys through space and time:
510 testing the space-time interaction in the absence of replication. *Ecology*, **91**, 262-272.
511 Legendre, P. & E.D. Gallagher (2001) Ecologically meaningful transformations for ordination of
512 species data. *Oecologia*, **129**, 271-280.
513 Legendre, P., Oksanen, J. and C.J.F. ter Braak (2011) Testing the significance of canonical axes
514 in redundancy analysis. *Methods in Ecology & Evolution* **2**, 269-277.
515 Legendre, P. & L. Legendre (2012). *Numerical Ecology*, 3rd English edn. Elsevier Science BV,
516 Amsterdam.
517 McArdle, B.H. & M.J. Anderson (2001) Fitting multivariate models to community data: a
518 comment on distance-based redundancy analysis. *Ecology*, **82**, 290-297.
519 Ouellette, M.-H. & P. Legendre (*in prep.*) An adjusted R^2 statistic for multivariate regression tree
520 analysis. *Manuscript*.
521 Ouellette, M.-H., DesGranges, J.-L., Legendre, P. & D. Borcard (2005) L'arbre de régression
522 multivariables: classification d'assemblage d'oiseaux fondée sur les caractéristiques de
523 leur habitat. *Société Francophone de Classification*, Montréal.

524 Page, L.M. & B.M. Burr (1991) *A field guide to freshwater fishes of North America north of*
525 *Mexico*. Boston, Houghton Mifflin Company.
526 Peres-Neto, P.R., Legendre, P., Dray, S. & D. Borcard (2006) Variation partitioning of species
527 data matrices: estimation and comparison of fractions. *Ecology*, **87**, 2614-2625.
528 Pihu, E. (1996). Fishes, their biology and fisheries management in Lake Peipsi. *Hydrobiologia*,
529 **338**, 163-172.
530 Pinzón, J. & J. Spence (2010) Bark-dwelling spider assemblages (Araneae) in the boreal forest:
531 dominance, diversity, composition and life-histories. *Journal of Insect Conservation*, **14**,
532 439-458.
533 R Development Core Team (2010) *R: A language and environment for statistical computing*.
534 Vienna, Austria, R Foundation for Statistical Computing.
535 Sheaves, M., Abrantes, K. & R. Johnston (2007) Nursery ground value of an endangered wetland
536 to juvenile shrimps. *Wetlands Ecology and Management*, **15**, 311-327.
537 ter Braak, C.J.F. (1988) Partial canonical correspondence analysis. *Classification and related*
538 *methods of data analysis*. H.-H. Bock. North-Holland, Amsterdam, 551-558.
539 Verneaux, J. (1973). *Cours d'eau de Franche-Comté (Massif du Jura). Recherches écologiques*
540 *sur le réseau hydrographique du Doubs. Essai de biotypologie.* Besançon. **Thèse d'état**.
541 Verneaux, J., Schmitt, A., Verneaux, V. & C. Prouteau (2003). Benthic insects and fish of the
542 Doubs river system: typological traits and the development of a species continuum in a
543 theoretically extrapolated watercourse. *Hydrobiologia*, **490**, 63-74.
544
545
546
547

548 # FIGURES



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.

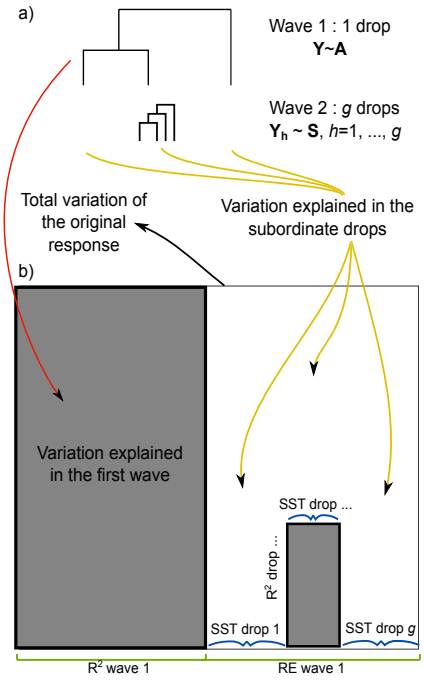
Subsequent drops: drops other than drop 1, subsequent from wave 1.

CMRT : A set of waves.

549

550 **Box 1.** Terminology review for MRT and CMRT analyses. There are four drops (four trees) in

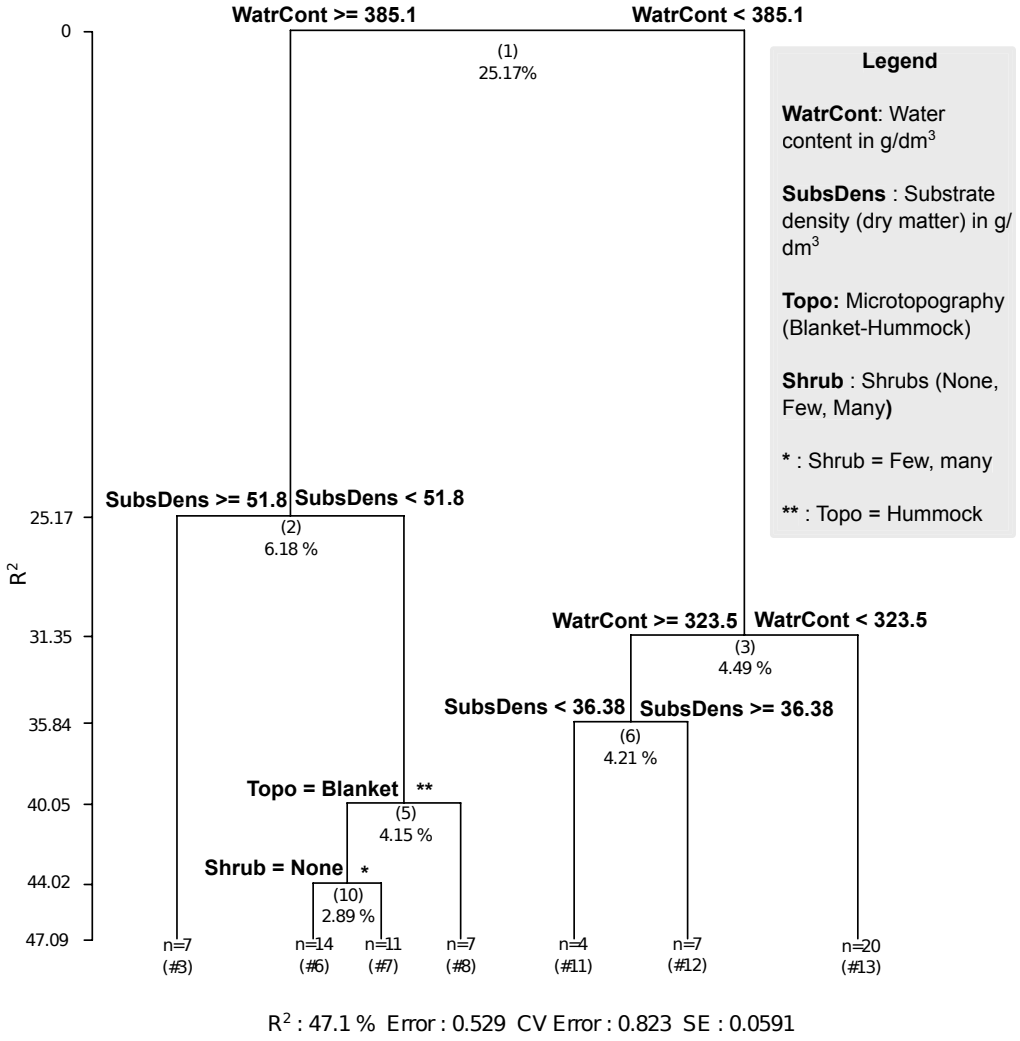
551 this diagram : one in wave 1 and three in wave 2.



552

553 **Fig. 1.** (a) Diagram of the CMRT procedure along with (b) a general R^2 diagram. In (b) partition
554 of the variation explained by the whole cascade in a square whose total area represents the total
555 variation in the response data (100%). The shaded area on the left represents the variation of the
556 response data explained by the first wave (main analysis). The shaded area or areas (there may be
557 more than one) on the right represent the variation explained by the subordinate drops of the
558 second wave. For each shaded rectangle in the white area on the right, its width represents the
559 proportion of the relative error (RE, unexplained variation) of the first wave while its height
560 represents the R^2 of the subsequent response explained by the subordinate drop. The white area is
561 the variation that remains unexplained at the end of the waves.

562



563

564 **Fig. 2.** MRT analysis for the oribatid mite data. Details: see legends of Figs. 2 and 3. Each node
565 bears its identification number in parentheses, e.g. (1), corresponds to the one found in the
566 `summary.MRT` function of the `MVPARTwrap`. Under the number is the percentage of explained
567 variation. For each leaf, the number in parentheses, e.g. (#3), is the one found in the
568 `summary.MRT` function of the `MVPARTwrap` package; the number of objects in the leaf is also
569 shown, e.g. n = 7.

570

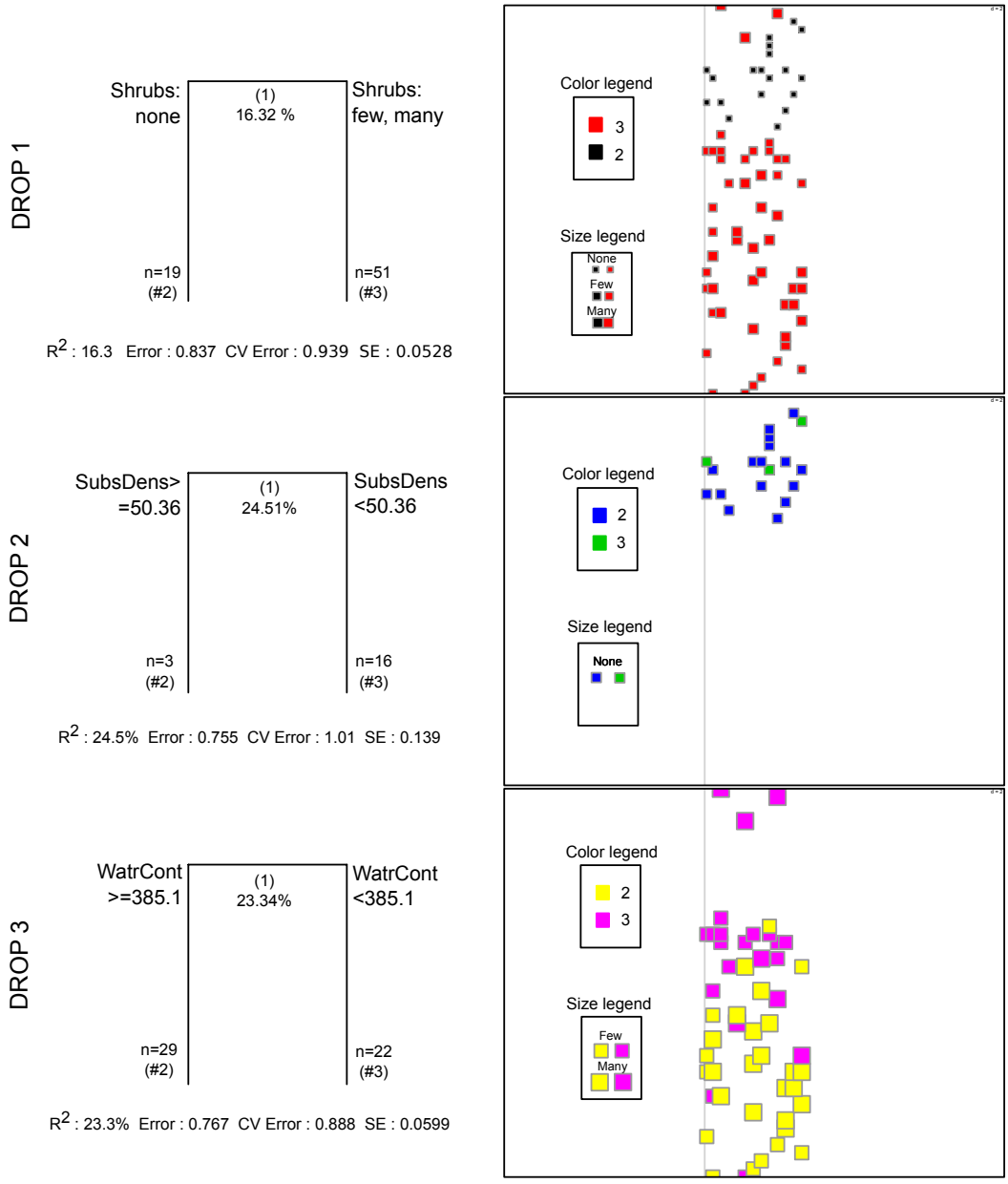
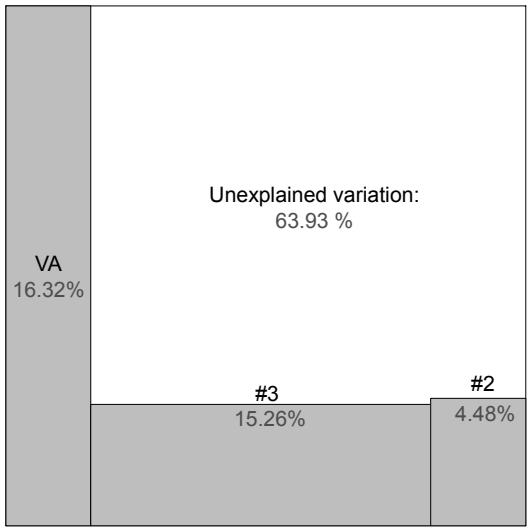


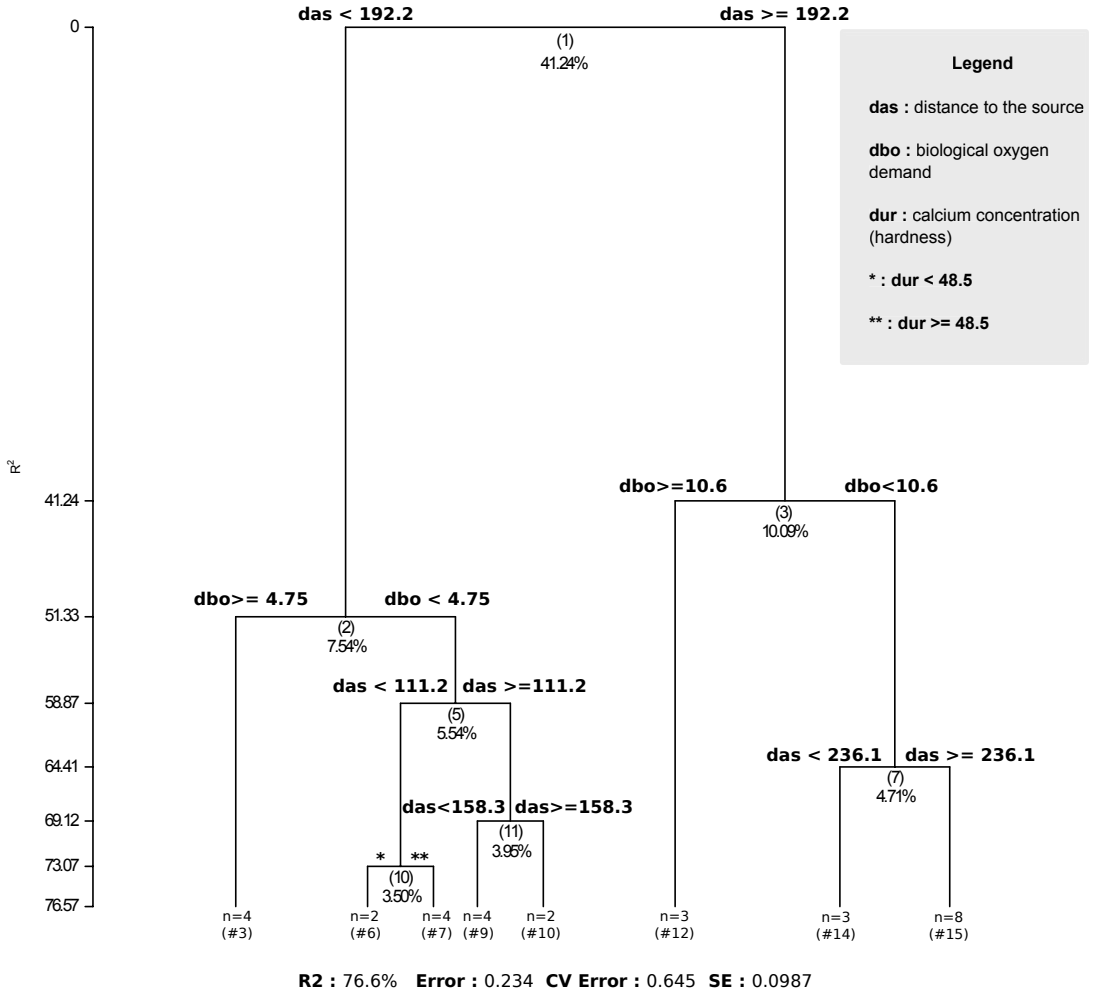
Fig. 3. Summary of the CMRT analysis results for the oribatid mite data with the explanatory variable shrub density as the primary (main) effect. Details: see legend of Figs. 1 and 4. The explanatory variables used to split the data were the shrub states (none, few, many; the variable is noted 'Shrubs'), the substrate density (dry matter) in g/dm^3 noted 'SubsDens', and the water content in g/dm^3 noted 'WatrCont'.



577
578

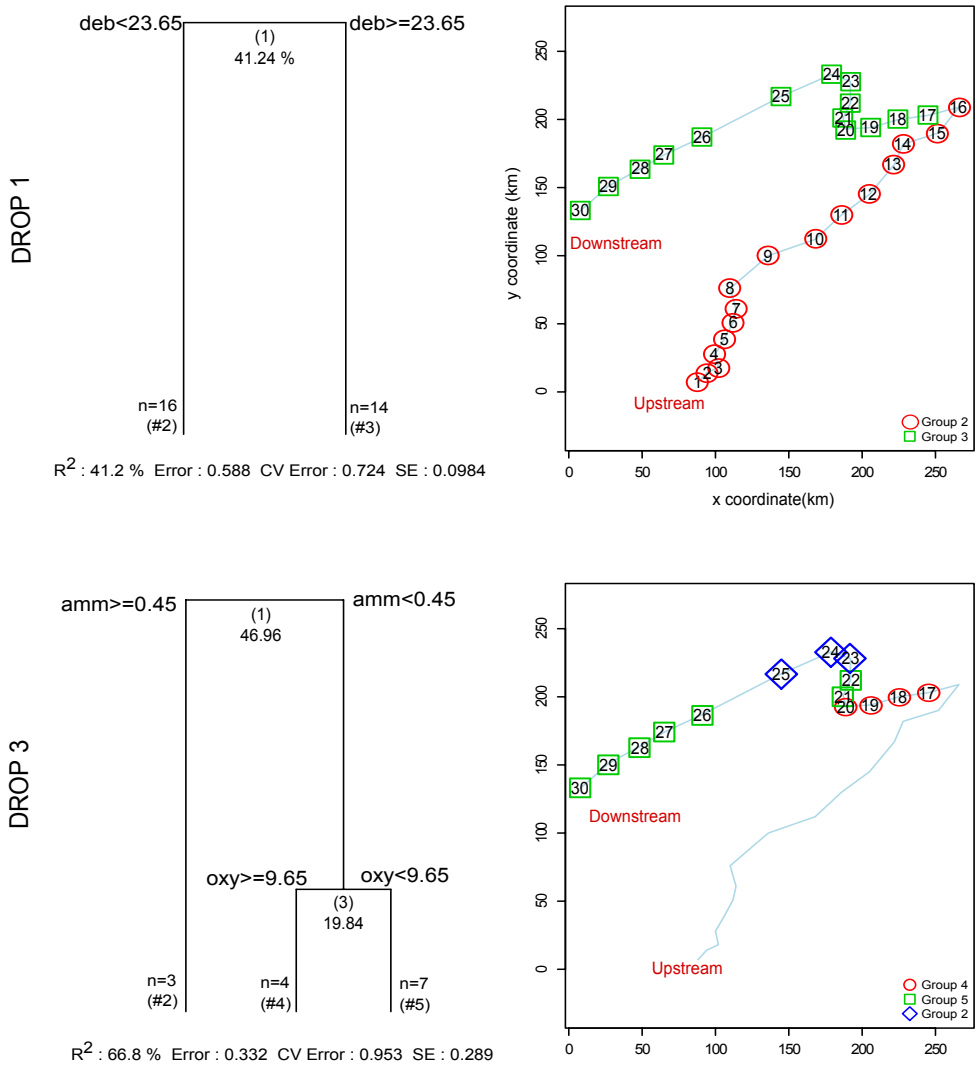
579 **Fig. 4.** Output of the ***CasMRTR2()*** function for the oribatid mite data. The global R^2 is 36.07%;
580 the portion of the global R^2 explained by subordinate drops 2 and 3 together is 19.74. The VA
581 percentage (16.32%) is the proportion of the response variation explained by the main
582 explanatory variable, here shrub density.

583



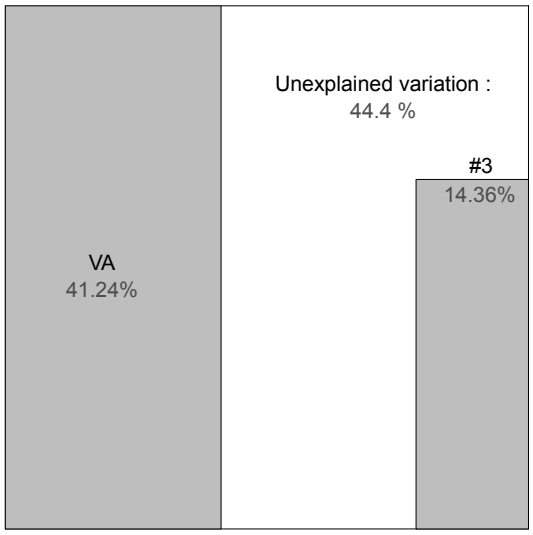
584

585 **Fig. 5.** Original MRT analysis of the Doubs River fish data.



586

587 **Fig. 6.** CMRT analysis results for the Doubs River data. Left: arborescence of the drop; right we
588 find: corresponding geographical map of the groups. The number (#) and size (n) of each leaf are
589 shown. The number and percentage of explained variation are given for each node. Three
590 explanatory variables appear in this figure: mean minimum discharge (deb), ammonium
591 concentration (amm) and dissolved oxygen (oxy).



592

593 **Fig. 7.** Output of the ***CasMRTR2()*** function for the Doubs River fish data. The global R^2 is
594 55.6%, the portion of the global R^2 explained by the subordinate drop 3 is 14.36%, and only that
595 one has explained any variation in the second wave. The drop number corresponds to the number
596 of the leaf in the tree of the first drop (Fig. 3). The VA percentage (41.24%) is the variation
597 explained by the main explanatory variable, here the 'mean discharge' variable.

598