Forward selection of traits and environmental variables in dc-CA using Canoco 5

Cajo J.F. ter Braak April 2022 cajo.terbraak@wur.nl

Abstract. This note explains how the analyses of the paper "Mauro et al 2022 Hay meadows' overriding effect shapes ground beetle functional diversity in mountainous landscapes, Ecosphere" were performed using Canoco 5.

1. Introduction

The analyses use double constrained correspondence analysis (dc-CA) which is part of Canoco since version 5.10 (March 2018). The analyses in the paper were performed using Canoco 5.12 (a minor update). If you use Canoco 5.15 and want to reproduce the p-values (approximately) you must uncheck each check box for "Permute residualized predictor(s)", so that Canoco permutes residualized responses, the only permutation version in Canoco 5.12. Canoco 5.15 uses an improved version of residualized response permutation and if you wish the old version you must click in Edit|Settings|Canoco5 options|Actions the box "Y-permutation type in legacy mode". Note that residualized predictor permutation outperforms residualized response permutation in analyses with potentially huge weight difference (e.g. difference in abundance totals of species). See ter Braak (2021), which focusses on weighted redundancy analysis (RD), and for correspondence analysis-type analyses (CCA and dc-CA) ter Braak and te Beest (2022). It is noted below where this leads to differences compared to the published text. Notably, with residualized predictor permutation, there is no statistical evidence that the last terms added in the forward selection of environmental variables and traits (pH and Body length, respectively) contribute to the explanatory power of the models constructed so far in the selection.

The file "Gobbi2022_PitfallSqrtLdivEffort512.c5p" is the Canoco 5.12 project that contains the analysis. A Canoco 5.12 project can be opened, inspected and modified using Canoco 5.15 but the reverse is not true.

2. Data

The initial data tables in the Canoco project are termed Abundance, Environment and Traits. For application of double constrained correspondence analysis a fourth table is added by clicking Data|Add new tables|Transpose compositional table, yielding:

Project: Gobbi2022_PitfallSqrtLdivEffort512.c5p

Data tables:

Table	Cases	Vars	Type
Abundance	385	84	compos.
Environment	385	39	general
Traits	84	19	general
Transposed Abundance	84	385	compos.

Here is a short description of the data tables. Recall from the main text that "In all analyses, categorical variables (factors) were coded as sets of indicator (0/1) variables as customary in regression analysis. Body length was log-transformed to make its distribution more symmetric."

The Abundance table in the Canoco project is a compositional table (i.e. with non-negative variables measured on the same scale) and contains the square root of the ratio of the count per species in each pitfall and the effort (number of pitfall days). The ratio was taken so as to give pitfall a more even weight in the analysis and the square root was taken to give the taxa a more even weight in the analysis.

The Traits table is a general table (i.e. with variables measured in different measurement units, e.g. mm. and kg.) and contains the traits data of the taxa of table S1 of the paper. Except for body length, the other four traits are categorical (nominal). For ease of use in forward selection, these traits are expanded into sets of dummy (1/0) variables, with each variable representing a single category of the categorical variable. For example, for the first trait, chorology, this is achieved in Canoco by clicking on cell C1, right click in the cell and selecting Expand into dummy variables (the last line in the context-dependent menu). Similarly, the other nominal variables are expanded, resulting in 19 variables in the Traits data table. Body length in the Traits table is in mm. It is log transformed by right clicking in the top-left cell of the Traits table and selecting Transformation and Standardization. Here, as you can check, Body length was set be being log(Ax+B) transformed with A = 1 and B = 0.

The Environment table is a general table and contains the variables of the sampling design, namely plot, transect, sector, and the geographic, habitat and environmental variables, and effort days. For the permutation tests based on plots (sets of 5 pitfalls), the sampling design needs to be balanced and therefore the missing pitfalls were imputed; the imputed pitfalls have a missing value for effort days. Their values of abundance and environmental variables were computed as the mean of the actually measured pitfalls in the corresponding plot. The categorial variable habitat was expanded into dummy variables.

Finally a note on the order of the pitfalls (sites) in the data tables. Note that pitfalls of the same plot are together (this is handy for the hierarchical permutation tests, i.e. that based on plots) and plots are in order of their elevation along the transect to which they belong (this is needed to allow for accounting for possible autocorrelation among neighboring plots by cyclic permutation).

3. Table 1: Partitioning of the trait-structured variation by levels of the hierarchical design of the study

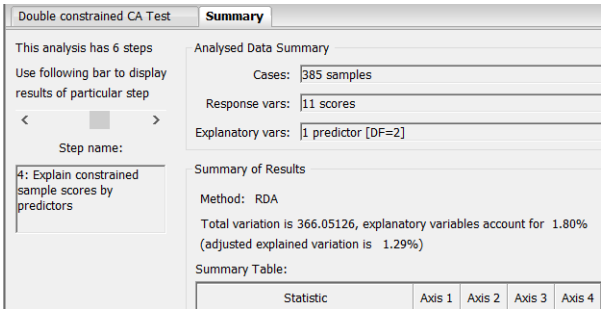
The main text says "The trait-structured variation was partitioned in the four parts that naturally follow from the sampling design, i.e., the variation 1) among sectors, 2) among transects within sectors, 3) among plots within transects and 4) between pitfalls within plots. The parts can be obtained by three dc-CA analyses, all of which use all trait variables but have sector, transect and plot as respective environmental predictor variables. The adjusted coefficients of determination (adj R^2) are obtained from these analyses following Peres-Neto et al (2006)" . Table 1 gives the result.

Table 1. Partitioning of the trait-structured variation by levels of the hierarchical design of the study (df = degrees of freedom; R^2 percentage of the total trait-structured inertia; adj R^2 = as R^2 but adjusted for degrees of freedom following Peres-Neto et al. 2006).

Between	**Within**	**df**	R^2	adjR^2
Sectors	-	2	2%	1%
Transects	Sectors	9	12%	10%
Plots	Transects	65	49%	43%
Pitfalls	Plots	308	37%	--

The first three analyses in the Canoco project file (dc-CA sector, dc-CA transect and dc-CA plots) are needed to construct Table 1. These are dc-CA analyses with all traits of table S1 and the environmental variables sector, transect and plots, respectively. These variables in the Environment table are factors with 3, 12 and 77 levels, one for each sampled sector, transect and plot in the Stelvio National Park, respectively. To replicate such an analysis: click Analysis|New analysis|Canoco Advisor and select the first four tables (i.e. inclusive the Transposed Abundance) and select Abundance as focal table and then select Double-constrained-CA from the list "Select the analysis to be created". dc-CA and other correspondence analysis-type analyses are only offered in Canoco if the focal table is set to being compositional.

Table 1 gives degrees of freedom (df) and percentages of the trait-structured variation. This variation is given in Step 4 of the 6 steps of the dc-CA algorithm. For example, for 'sector' under Summary we obtain:



yielding 1.80% and 1.29% for R2 and adj R2, which gives the rounded values of 2% and 1% in the first data row of Table 1. The results so obtained from the three analyses are:

Explanatory variable/Analysis	**df**	R^2	adjR^2
Sectors	2	1.80%	1.29%
Transects	11	13.64%	11.09%
Plots	76	62.78%	53.59%

From this, the values in Table 1 have been derived following Peres-Neto et al (2006). The values of the first row do not need an adjustment. The values for the second row (Transects) in Table 1 are, consecutively, 11-2 = 9 df, 13.64-1.80 = 11.84 (rounded 12%) and 11.09 - 1.29 = 9.8 (rounded 10%) for R^2 and adjusted R^2.

The values for the third row (Plots) in Table 1 are, consecutively, 76-11=65, 62.78-13.64= 49.14 (rounded 49%) and 53.59-11.09= 42.5 (rounded 43%).

The values of Pitfalls are 385-1-76=308 df and 100-62.78 = 37.22 (rounded37%).

Peres-Neto et al (2006) argued for use of permutation-based adjusted R^2 values; these are available in Canoco 5 by changing a check box in Settings. In my experience, the difference with the previous method is usually small.

4. Tables 2 and 3: Selection of environmental and trait variables.

The main text says "further analyses focused on the trait-environment variation within transects by two analyses, 1) analyzing all variation within transects and 2) analyzing the small-scale variation within plots. Statistical tests in both analyses used the hierarchical and spatial design. In the first analysis, sampling plots were permuted by random cyclic shifts within transects, so as to account for the spatial order of the plots, while keeping the sets of five pitfalls per plot together. For this analysis, each plot must have the same number of pitfalls; therefore, the design was made up of five pitfalls per plot. However, because some pitfalls were damaged by wildlife (*cf*. Results), in plots where the number was less than 5, we added 1-3 pitfalls (for a total of 23 across plots), for which values of abundance and environmental variables were computed as the mean of the actually measured pitfalls in the plot. In the second analysis, pitfalls were randomly permuted within their plot. The importance of the individual traits in explaining the environmentally structured variation was assessed by the explained inertia in dc-CAs on each trait using all environmental variables. Their joint importance was assessed using dc-CA with all environmental variables and a forward selection of trait variables. In this analysis, the importance of a trait is assessed by the explained inertia that the trait contributes on entry in the model (i.e., conditionally on the effect of the already selected traits). Analogously, the importance of the individual environmental variables in explaining the trait-structured variation was assessed by a forward selection of environmental variables using dc-CAs using all trait variables."

4.1 Within-transect trait- and environmentally structured variation

The results of "1) analyzing all variation within transects" are in Table 2 of the main text which is reproduced here:

Table 2. Importance of selected environmental variables (a) and functional traits (b) in explaining the within-transect trait- and environmentally structured variation, respectively, using double constrained correspondence analysis with covariate transect. Variables were selected by forward selection and tested using permutation tests based on the hierarchical design of the study (pitfalls within sampling plots along transects). Expl. % = fraction of trait-structured (a) and environmentally structured (b) variation explained by individual single variables (Simple term) and by individual variables during forward selection, i.e., after removing the effects of the terms included earlier (Forward selection). *p*-value (adj) = *p*-value after adjustment by False Discovery rate (using $n = 999$ permutations).

(a) Trait-structured variation (19 % of within-transect taxonomic variation)				
	Simple term		Forward selection	
Environmental variable	Expl. %	*p*-value (adj)	Expl. %	*p*-value (adj)
hay meadow	15.6	0.002	15.6	0.002

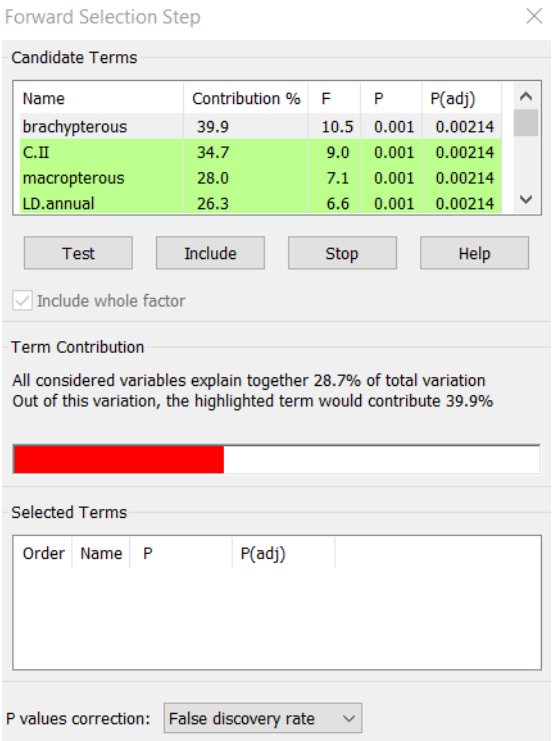
elevation	7.6	0.002	2.7	0.018
canopy	7.0	0.002	2.3	0.005
pH	7.1	0.002	1.0	0.030
siliceous alpine grassland	1.6	0.064	-	0.171
Combined (adj. R^2)			38.0	0.001[a]
(b) Environmentally structured variation (18 % of within-transect taxonomic variation)				
	Simple term		Forward selection	
Functional trait	Expl. %	*p*-value (adj)	Expl. %	*p*-value (adj)
brachypterous	11.4	0.002	11.4	0.002
specialised predator	3.4	0.027	5.7	0.008
body length	2.3	0.060	3.2	0.015
Chorology II	10.0	0.002	-	0.300
Combined (adj. R^2)			34.0	0.001

[a] based on cyclic shifts of sampling units within transects keeping together the five pitfalls per sampling plot.

In this section the within-transect variation is considered. That means that the differences between transects are removed from the analysis by making the factor transect a covariate. To be able to specify covariate in Canoco, the quick wizard mode must be switch off, by clicking  when it is 'on' (the default). To select environmental and trait variables, a new analysis is created using the analysis Double-constrained-CA-FS instead of the Double-constrained-CA in section 3.

The first dialog box that appears (sometimes only after a while) says which tables are used in Step 1 of the analysis. In the next box Covariate data is set to Predictors in 'Environment' table. The next box is to indicate which environmental variables must be used in the analysis. Here we move the design variables plot, transect and sector to the left box, and habitat (as it is represented by its categories as dummy variables) and the variables from n.species to and including weight. In the next box, use covariate transect only (select all variables on the right, move all to the left and move transect back to the right). In the next box (Constrain species composition by predictors)

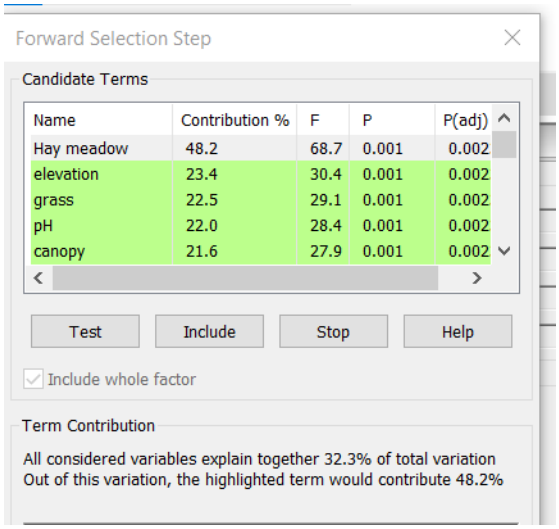
nothing needs to be changed; click Next. In the next box on “Test or Explore Predictor Effects” you may wish to change to “Not performed”. Then, in Step 2 (Explain constrained species scores by traits) the trait data are mentioned; click Next. No covariates are needed here; click Next. In the next box, traits to be used must be selected. Move the factors that are represented by their categories (chorology, w.morph, diet larv.devel) to the left and also the added trait factors brachypterous and specialized predator. Click Next till Step 4 (Perform stepwise selection of traits) Test or Explore Predictor Effects, where you can specify the permutation test for traits. We applied Unrestricted permutation with 999 permutations. Click Next till to arrive at a similar box in Step 6 (Perform stepwise selection of predictors) Test or Explore Predictor Effects in which the permutation test for environmental variables must be specified. Select here: Hierarchical design and 999 permutations. Also check Blocks defined by covariates and move transect to the right hand box. In the box Split-plot Layout the number of split-plots in each whole plot must be set to 5 and for split plots set TAKE 5 and SKIP NEXT 0, as pitfalls within plots are consecutive in the data file. This keeps the pitfalls of each plots together and thus permutes plots instead of pitfalls. Check Time series or linear transect under Whole-plot permutation and No permutation under Split-plot permutation, so as to account for any spatial autocorrelation between the plots along each transect. Click Next until the analysis starts and the first results appear



Select False discovery rate for P values correction and scoll through the top panel to the bottom until the last traits gets a P-value and scroll back. This ensures that all P-values are being computed (ones that are not visible may not have been computed yet). This is needed for precise calculation of the False discovery rate, which depends on all P-values in the list. The dummy variable “brachypterous” is the best trait (at the top) and explains most (39.9%) of what can be explained by all traits, which is 28.7% of the total variation. Here, the total variation is the environmentally structured variation, namely the variation in the species niche centroids (SNCs) of all orthonormalized environmental variables in the analysis The top trait is also

statistically significant (P_{adj} = 0.002). It indicates that brachypterous species react differently to the environment than other species. Select this trait by clicking Include and continue adding traits till the top trait is no longer significant at the 5% level (this trait is Chorology II in this analysis).

Next, the first step of the forward selection of environmental variable appears, showing:



(Scroll through the list down and up again as for traits to ensure that all P-values have been computed). The environmental variable at the top, Hay meadow, best explains the trait-structured variation; it explains 48.2% of what all environmental variables can explain which is 32.2% of the total variation. Here, the total variation is the trait-structured variation, i.e. the variation in the community weighted means (CWMs) of all orthonormalized traits in the analysis. The top variable is also significant. It indicates that the species in hay meadow differ in trait composition compared species in other habitats. Select this variable by clicking Include and continue adding environmental variables till the top one is no longer significant at the 5% level; this environmental variable is Siliceous alpine grassland in this analysis. At this point, the second variable was pH with P= 0.004 and P(adj) = 0.03 so this variable was included instead. At the next step Siliceous alpine grassland is the best variable with a P-value far above 0.05, so that Stop was clicked.

Table 2a says that the trait-structured variation is 19 % of the within-transect taxonomic variation and Table 2b says that the environmentally-structured variation is 18 % of the within-transect taxonomic variation. These values come from Steps 1 and 3 where the Abundance table is regressed on to all environmental and trait variables in the analysis, respectively, and are displayed as 19.48% and 17.78%, respectively. These are unadjusted R^2 values as they represent simply what is the total variation being analyzed in the table.

The results in the last two columns of Table 2 (Forward selection) are from Steps 4 and 6 in the Summary tab, where the selected terms are displayed. Click (the second copy) and paste it in Excel for easy reformatting and closer inspection. The results in the first two columns of Table 2 are from the first step in the selection. Here, we

computed the % Explained values from the Contribution values that were displayed; these term have a fixed ratio. For example, for elevation the contribution was 23.4 while it was 48.2 for Hay Meadow. Hay Meadow explained 15.6%. Elevation thus explained (15.6/48.2)*23.4 = 7.57 (rounded 7.6%). In Steps 6 and 4 the final models explained 21% and 17% (adj R2), respectively. The values given in Table 2 are 38% and 34%; these values come from Steps 10 and 8 and are fractions of the trait- and environmentally structured variation of the selected traits and environmental variables as opposed to all traits and environmental variables. The choice for the latter (higher) values is motivated by noting that the variation due to the non-selected traits and environmental variables may just represent noise and is thus irrelevant.

Table S5 contains the statistics of the final model. It can be viewed in the Canoco project of the analysis (which I renamed to Table 2: dc-CA within transects) by clicking  and going to the ExplVars (12) and SupplVars (12) tabs. By clicking again, all extra tabs are hidden again. Note that the regression coefficients can be plotted using the graph wizard (click Graph|Advise on graphs), by selecting "predictors+traits -canonical weights biplot".

Fig. 2 is the default graph offered with some post-editing using Canoco, which is a biplot of fourth-corner coefficients. You can add such a graph by clicking (with the analysis Table 2: dc-CA within transects being selected by clicking on its name) Graph|Advise on graphs and selecting predictors+traits biplot with optional species scores. Fig. 2 was obtained from this by flipping the first axis, so that the arrow for elevation points to the right: the elevation increases from left to right over the diagram. Graphing options can be changed as follows. The mnemonic is that a graph belongs to an analysis, so click Analysis and from there "Plot creation options" and, in the screen that comes, check Flip axes under Horizontal and 1. Click OK and recreate the graph by clicking  . If the shape of the figure changes, click  and change under Unimodal Methods Focus on… to Symmetric scaling. You can edit the graph to the texts in Fig. 2. Note that the first axis is much more important than the second.

The P-values are based on (legacy) residualized response permutation (i.e. without taking account of the intercept, see ter Braak 2021). With residualized predictor permutation, the last environmental and trait variable added in Table 2 (pH and Body length) are no longer significant (Padj =0.14 and 0.53, respectively). Note that the default graph is a biplot of fourth-corner correlation coefficients. As the simple effect of pH is fairly large (Table 2) and as it is this effect that is related to the fourth-corner correlation, pH has a relatively large arrow.

4.2 Within-plot trait- and environmentally structured variation

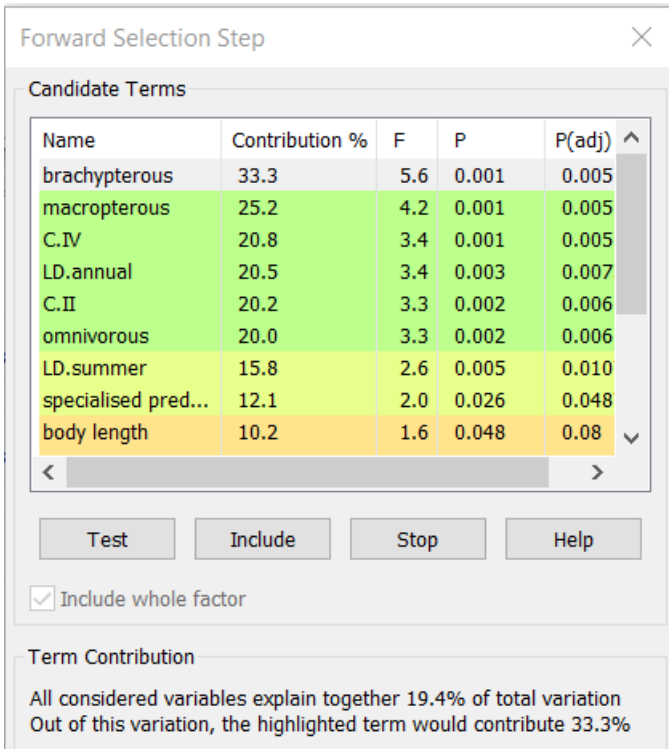
The results of "2) analyzing all variation within plots" are in Table 3 of the main text which is reproduced here:

Table 3. Importance of selected environmental variables (a) and functional traits (b) in explaining the within-plot trait- and environmentally structured variation, respectively, using double constrained correspondence analysis with covariate plot. Variables are selected by forward selection and tested using permutation tests based on the hierarchical design of the study (pitfalls within sampling units along transects). Explained % = fraction of trait-structured (a) and environmentally structured (b) variation explained by individual single variables (Simple term) and by individual variables during forward selection, i.e., after accounting for the effects of the terms included earlier (Forward selection). *p*-value (adj) = *p*-value after adjustment by False Discovery rate (using $n = 999$ permutations).

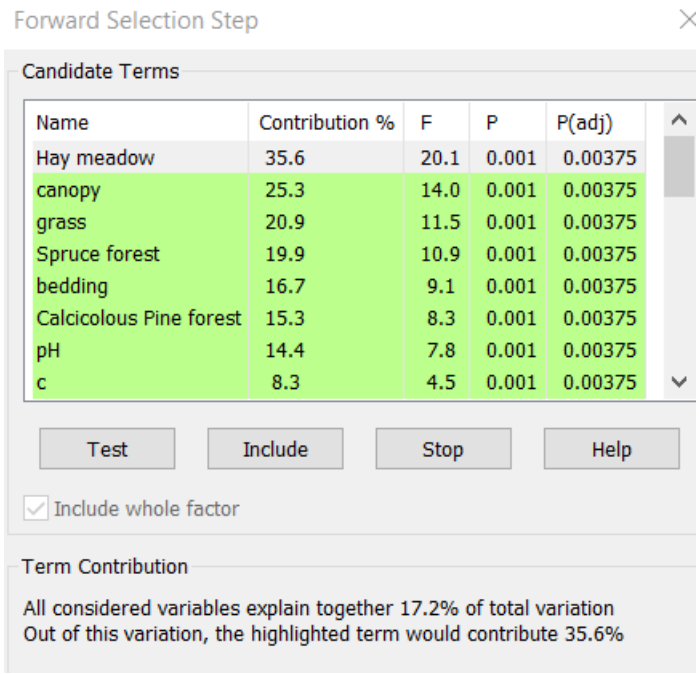
(a) Trait-structured variation (50% of within-plot variation)				
	Simple term		Forward selection	
Environmental variable	Expl. %	*p*-value (adj)	Expl. %	*p*-value (adj)
hay meadow	6.1	0.004	6.1	0.003
canopy	4.4	0.004	1.3	0.015
calcicolous pine forest	2.6	0.004	-	0.051
Combined (adj. R^2)			20.1	0.001[a]
(b) Environmentally structured variation (18 % of within-plot variation)				
	Simple term		Forward selection	
Functional trait	Expl. %	*p*-value (adj)	Expl. %	*p*-value (adj)
brachypterous	6.5	0.002	6.5	0.003
specialised predator	2.5	0.032	3.5	0.023
body length	2.0	0.077	-	0.745
Combined (adj. R^2)			35.1	0.001

[a] based on random permutation of pitfalls within sampling plots.

The analysis is obtained similar to that of section 4.1. The results of the first steps of the selection of traits is:



and the first step in the selection of environmental variables is



All numbers can be reproduced from this analysis. I renamed the analysis to “Table 3: dc-CA within plots”.

5. Useful links

www.canoco.com

www.canoco5.com

canoco5.com/index.php/resources

www.microcomputerpower.com

6. References

Peres-Neto, P.R., Legendre, P., Dray, S. & Borcard, D. (2006) Variation partitioning of species data matrices: Estimation and comparison of fractions. *Ecology,* **87,** 2614-2625. https://doi.org/10.1890/0012-9658(2006)87[2614:VPOSDM]2.0.CO;2

ter Braak, C.J.F. (2021) Predictor versus response permutation for significance testing in weighted regression and redundancy analysis. *Journal of statistical computation and simulation*. http://dx.doi.org/10.1080/00949655.2021.2019256

ter Braak, C.J.F. & te Beest, D.E. (2022) Improved permutation testing in canonical correspondence analysis, comparing Canoco, vegan and ade4. *Submitted (preprint available in the Canoco software)*.