R package douconca: multi-trait multi-environment ordination of ecological communities by double constrained correspondence analysis.

Cajo J. F. ter Braak1 and Bart-Jan van Rossum1

1 Biometris, Wageningen University & Research, Wageningen, the Netherlands

Email: cajo.terbraak@wur.nl

ORCID CtB: <http://orcid.org/0000-0002-0414-8745>

**Abstract**

1. Trait-based ecology aims to learn which species occur and prosper where on the basis of their traits and the environmental conditions. The study of species-environment relationships using (constrained) ordination and the like, thus becomes a study of trait-environment relationships. Focussing on inter-species trait and inter-site environmental variation, such a study requires three data tables: tables on abundance, traits and environmental variables of a set of species and sites. The oldest such method is the ordination method RLQ. The most commonly used method is community-weighted means (CWM) regression, while applications based on generalized linear mixed models (GLMM) are scarce. CWM-regression analyses each trait separately; multi-trait GLMMs do not currently allow for dimension reduction of the trait-environment relationships.
2. Here we propose double-constrained correspondence analysis (dc-CA). Like RLQ, dc-CA is an ordination method with associated visualizations (notably, biplots), but unlike RLQ, dc-CA allows for model comparison and prediction because dc-CA is regression-based. Also, dc-CA is a natural multi-trait extension of CWM-regression and fourth-corner analysis. The dc-CA method constructs trait and environmental composites (gradients) that sequentially best explain the abundance data and that have maximum fourth-corner correlation.
3. The R package douconca on CRAN implements dc-CA. It is based on the community ecology package vegan and contains functions for visualization, statistical testing and prediction. It has a formula-interface so as to allow non-linear trait and environmental modelling and interactions. Statistical testing proceeds using the ter Braak max test, so as to guard against overoptimism. The focal method of visualization is the one-dimension biplot as in principal response curve analysis. More traditionally ordination diagrams can be constructed as they are from analyses with vegan.
4. This application paper describes the main functionality of douconca by three case studies. In one case study, dc-CA summarized the main results of many CWM-regressions in a single diagram. In another case study, GLMM did not find the usual leave-traits economics spectrum, but dc-CA found it in its first axis. Moreover, the species-environment relations in this study were multi-dimensional, of which the traits explained a single dimension only.
5. Trait-environment analysis becomes easier and more insightful with dc-CA.

Keywords: community ecology; trait-environment relationships; constrained ordination; R package douconca; double constrained correspondence analysis

max 8 words

# Introduction

Species have environmental preferences with each species inhabiting a specific part (niche) of the multi-dimensional habitat space (Pocheville 2015). Founded on evolutionary adaption of species to their environment, trait-based ecology aims to understand the niche differences between species from their traits and so to help predict future changes in species composition and traits from the global or local changes in the environment (Rolhauser, Waller & Tucker 2021). Because of biodiversity, it is rather impractical to aim to learn from data about the ecological niche of each individual species. It is much more practical to learn from data about the traits determining the ecological niche of a species. The quest for the holy-grail (Funk *et al.* 2017) is to find such traits and starts with find traits with a strong trait-environment relationships.

The most commonly method for the analysis of trait-environment relationships is community-weighted means (CWM-) regression in which the average value of a single trait at a site is regressed upon a set of environmental variables (predictors in this context) and an recent advanced method is based on generalized linear mixed models (GLMM), both evaluated in ter Braak et al (2019). A multi-trait, multi-environment analysis is possible using the ordination method called RLQ (Dray *et al.* 2014) or using GLMM (Niku *et al.* 2021). A disadvantage of RLQ is that it is based on co-variance, which hampers prediction and comparison of models (Peng *et al.* 2021). A disadvantage of GLMM is its computational complexity when all relevant random terms are included (ter Braak 2019), and that it does not yet cater for the construction of composite traits and environmental gradients. Double-constrained correspondence analysis (dc-CA) overcomes both disadvantages (ter Braak, Šmilauer & Dray 2018b; Peng et al. 2021).

In this paper we describe dc-CA and its recent implementation in the R package douconca (on CRAN) followed by three case studies. The case studies each show a particular aspect of dc-CA. Some also pose challenges that go beyond the scope of this paper, although dc-CA may of help in tackling them.

# Double constrained correspondence analysis

RLQ, GLMM and CWM-regression have disadvantages that are overcome in dc-CA: dc-CA is based on regression, allows for dimension reduction, with each axis representing a composite trait and an environmental gradient that have maximum fourth-corner correlation given all earlier axes (ter Braak et al. 2018b). By combining a community-level test with a species-level test (ter Braak, Cormont & Dray 2012; ter Braak et al. 2018b), dc-CA overcomes the overoptimism (particularly, inflated type I error rate) which hampers CWM-regression (Lepš & de Bello 2023). Also, dc-CA is a natural extension of CWM-regression to the multi-trait case (ter Braak et al. 2018b). Without dimension reduction, dc-CA is equivalent with a series of CWM-regressions, one for each traits.

# The R package douconca

So far, dc-CA was only available in the Canoco software for ecological ordination (ter Braak & Šmilauer 2018) and in R-scripts in appendices to papers. Recent applications of dc-CA ; Sîrbu *et al.* 2022) used Canoco.

The R package douconca is an R implementation of dc-CA. The main function of is dc\_CA() with associated functions anova(), scores(), coef(), fitted(), predict() and plot() for statistical testing, ordination scores, coefficients, fitted, prediction and plotting, respectively. The dedicated dc-CA plot function plots a single dimension. The interpretation of the plot is easier than that of an ordination diagram and is warranted as the first axis of dc-CA is so dominant in many applications, that it makes sense to plot this axis and its definition separately.

The douconca douconca package is based on the vegan package and follows what is customary for vegan users were possible. In particular, dc\_CA() has a formula interface, *e.g.*

mod\_CCA <- dc\_CA(  
 formulaEnv = ~ Moisture + Condition(Management),  
 formulaTraits = ~ SLA + Condition(Seedmass),  
 response = abun, dataEnv = envData, dataTraits = traitData)

with trait and environmental variable names in the trait data and environmental data, respectively. The response (abundance or occurrence data) is specified separately. The formulas allow for polynomial terms and interactions.

A dc-CA analysis thus requires three data tables with at least, say, 15 species and 15 sites for statistical inference. Any missing values should be imputed first, for example by using the R package mice (van Buuren & Groothuis-Oudshoorn 2011).

Trait and environmental variables can be qualitative (factors), quantitative or a mixture thereof. The number of trait variables (with each category of a factor counting as single trait) should be less than the number of species and the number of environmental variables should similarly be less than the number of sites. A selection of variables is needed if the numbers are higher, possibly via forward selection in which at each step the best predictor is selected from the pool of trait and environmental variables (Gobbi *et al.* 2022). With a single trait and single quantitative environmental variable, dc-CA is an embellished fourth-corner correlation analysis.

By default, dc\_CA() divides the abundance data by the site totals, giving equal site weights. This accounts for possible differences in plot size and increases the similarity of dc-CA with CWM regression, as the latter is usually unweighted. Appendix S1 shows how a dc-CA is then calculated in package douconca.

# Applications

## Dune meadow data

The douconca example data set **data**("dune\_trait\_env") is essentially the vegan data set **data**("dune") without the two mosses but with two sets of traits added: five (LEDA) functional traits (taken from Jamil et al. (2013)) and four ecological traits (Ellenberg (1992) indicator values). Here these two sets are compared in explanatory power. Moisture and Management are sufficient to explain the two major dimensions of taxonomic composition. Constraining the species scores to be a linear combination of the four ecological traits (formulaTraits = ~F+R+N+L) decreases the explained variance (sum of the constrained eigenvalues) from 0.97 to 0.57. Constraining by the five functional traits instead (formulaTraits = ~ SLA + Height + LDMC + Seedmass + Lifespan) gives a decrease to 0.35. Moreover, a permutational ANOVA show that the first two axes are significant with the ecological traits (P<0.01 for both axes), whereas only the first axis is significant with the functional traits (P≈0.05). The above value of 0.97 can be obtained by canonical correspondence analysis (CCA, i.e. single constrained):

mod\_cca <-cca(Y~Moisture+Management, data = envData)

or by dc-CA, but with a trait formula that does not impose a constraint:  
  
mod\_CCA <- dc\_CA(formulaEnv = ~ Moisture+Management,  
 formulaTraits = ~Species,  
 response = abun, envData, traitData)

where Y is abun divided by the site total. This example serves to show the relationship between CCA and dc-CA and also that the eigenvalues and variances of different models can be compared as they are on the same footing: variance of fitted response values. Such comparisons are without sense in RLQ (Peng et al. 2021). As explained variance increases with the number of predictors (even random ones), differences in number of predictors need to accounted for (Peres-Neto *et al.* 2006). The example is simple, as there are even fewer ecological traits than functional traits.

## Wisconsin forest understory vegetation

Rolhauser et al. (2021) analysed trait-environment relationships in Wisconsin forest understory vegetation (185 species in 189 sites) using a GLMM, which included linear and quadratic main and interaction effects between four traits and six environmental variables, and all random site-dependent trait and species-dependent environmental variables effects for valid statistical inference (ter Braak 2019). The model had 43 fixed effects, 24 of which are trait-environment interactions.

The three data tables needed for dc-CA were extracted from the single online spreadsheet. Applying a dc-CA with linear trait and environmental formulas to this data (after data transformation as in the paper), showed a strong first axis, the eigenvalue of which (0.36) is a factor 10 larger than the second. The site-level permutation test showed two significant axes, but in the species-level test only the first axis is judged significant (P-values 0.001 and 0.515). The max test thus follows the species-level test.

The rank-1 dc-CA (i.e. first axis only) models the trait-environment interaction with ten regression coefficients. Like correspondence analysis, dc-CA does not model the trait and environmental main effects. Such effects require separate analysis, for example by regressing the species-totals on the traits and the site-totals on the environment (ter Braak 2019).

Figure 1 visualizes the rank-1 model. The two scatter plots on the left have the first dc-CA axis as abscissa: constrained site scores in the top figure and constrained species scores in the bottom figure. The ordinate in the top panel is the CWM of the trait composite, the best linear combination of traits found by dc-CA, the t-values of which are shown in top-middle panel. The ordinate in the bottom panel is the species niche centroid (SNC) with respect to the environmental gradient, the best linear combination of environmental variables found by dc-CA, the t-values of which are shown in bottom-middle panel. The right-hand panel shows the top-20 species with the highest contribution (Greenacre 2013)(species total time squared constrained species score).

|  |
| --- |
|  |
| **Figure 1. Dc-CA plot. All traits were log-transformed.** |

The trait composite and environmental gradient in Figure 1 are characterized by the t-values of the regression coefficients in the middle panels[[1]](#footnote-1). The default gradient description is by the correlations of the traits and environmental variables with the axes that they define (*i.e.* the intra-set correlations), as they avoid issues due to multi-collinearity (*e.g.* small t-values for two strong predictors because they happen to be highly correlated), but no such issues are important in this case study and regression coefficients and their t-values are thought to be more illustrative of how the cumulative effects of traits and environmental variables form a dc-CA axis.

From the bottom t-value plot in Figure 1, the first axis is a gradient from low to high mean annual temperature (MAT), combined with low to high nitrogen content (%N) and temperature seasonality (TSD) and high to low soil sand (% Sand). The additional influences of mean annual precipitation (MAP), and basal area (BA, reducing light availability) in the gradient are small. The trait composite has a fourth-corner correlation of 0.58 (the square-root of the first eigenvalue) with this gradient and runs from low to high maximum vegetative height (VH) and from high to low leaf mass per area (LMA) and leaf carbon content (LCC). This first axis is in perfect agreement with what the authors write in the introduction of their paper (trait abbreviations in square backets added for clarity) “We expect increasingly benign conditions for plant growth (e.g. higher MAT, MAP and %N, and lower %Sand) to generally favour acquisitive species, that is, taller plants [VH] with larger [LS] and cheaper leaves in terms of both mass [LMA] and carbon investments [LCC]”. The exception is LS, perhaps, in the sense that LS has a tiny effect (but of the expected sign) on dc-CA axis 1 (intra-set correlation 0.2).

The GLMM found a negative LMA‒MAT relationship that was only marginally significant and the authors express surprise of such weak effects “given the strong negative effect found under controlled growth conditions”. The negative LMA – MAT relationship is the strongest of all on the first dc-CA axis, and also in terms of fourth-corner correlation (-0.41).

A permutational ANOVA showed that the first two axes are significant in the site-level test (P <0.01 for both axes), whereas only the first axis is significant in the species level test (P <0.01 and 0.51 for axes 1 and 2, respectively). In a state-wide data set, two to four gradients (dimensions) are expected (e.g. longitude, latitude, height, soil fertility). Here, dc-CA discovered that the four traits form one interpretable gradient, that of leaf economics. What would be good traits for the other gradients? Whereas a dc-CA model can include polynomials and interactions, these were not needed in this illustration. Note that CWMs and SNCs perform best in unimodal niche models (ter Braak & Barendregt 1986; ter Braak & Looman 1986).

## Amazonian forest trees

Pinho et al. (2024) investigated the effect of human modification of tropical landscapes (forest loss, landscape configuration and local degradation) on the trait composition of Amazonian and Atlantic forests across 271 forest plots in six regions in Brazil using abundance data on 1207 tree species and five functional traits. Forest loss and landscape configuration (log number of fragments and of edge density) were measured at three spatial scales (using concentric buffers of 500-, 1,000- and 2,000-m radius around each forest plot). The inverses of forest cover and basal area were used as proxies of forest loss (FL) and local degradation (LD).

CWM-regression for each trait per region led to a 5 traits × 6 regions = 30 variable selection problems with 10 candidate variables as putative predictors of the functional composition. The selection task was reduced by taking a causal modelling approach, resulting in a few number of regressions per selection problem, which were visualized in a detailed graph. As a safeguard against the overoptimism of CWM regression, they were complemented with species-level analyses using SNC regression.

As it is regression method, dc-CA was able to adhere to the causal modelling approach. But, a forward selection starting from all 10 variables resulted in the same final model. For this data, dc-CA thus could do without the causal modelling assumptions.

The dc-CA was forced to analyse the common within-region differences by setting the factor ‘region’ as a condition in the environmental formula, for example Y ~ FL2000 + LD + Condition(region) in the final model. Such a within-region dc-CA approach was first used in Gobbi et al. (2022).

All CWM regressions were summarized in a single one-dimensional dc-CA diagram like Figure 1 (Figure 5 of Pinho et al. (2024)). The second dimension was a factor of 44 smaller than the first and non-significant, neither at the species-level nor at site-level (both P-values >.4).

# Summary and conclusion

The R package douconca makes dc-CA an attractive option for multi-trait multi-environment association and regression analysis.

**DATA AVAILABILITY STATEMENT**

No new data has been generated. The package can be installed from CRAN.

**SUPPORTING INFORMATION**

Additional supporting information can be found online in the Supporting Information section at the end of this article.

Appendix S1. Skeleton dc-CA algorithm

Appendix S2. R code of the three case studies

**REFERENCES**

Dray, S., Choler, P., Dolédec, S., Peres-Neto, P.R., Thuiller, W., Pavoine, S. & ter Braak, C.J.F. (2014) Combining the fourth-corner and the RLQ methods for assessing trait responses to environmental variation. *Ecology,* **95,** 14-21.http://dx.doi.org/10.1890/13-0196.1

Ellenberg, H. (1992) *Indicator values of plants in Central Europe*. Goltze.9783884525180

Funk, J.L., Larson, J.E., Ames, G.M., Wright, J., Butterfield, B.J., Cavender-Bares, J., Williams, L., Firn, J., Laughlin, D.C. & Sutton-Grier, A.E. (2017) Revisiting the Holy Grail: Using plant functional traits to understand ecological processes. *Biological Reviews,* **92,** 1156-1173.https://doi.org/10.1111/brv.12275

Gobbi, M., Corlatti, L., Caccianiga, M., ter Braak, C.J.F. & Pedrotti, L. (2022) Hay meadows’ overriding effect shapes ground beetle functional diversity in mountainous landscapes. *Ecosphere,* **e4193**. https://doi.org/10.1002/ecs2.4193

Greenacre, M.J. (2013) The contributions of rare objects in correspondence analysis. *Ecology,* **94,** 241–249.http://dx.doi.org/10.1890/11-1730.1

Jamil, T., Ozinga, W.A., Kleyer, M. & ter Braak, C.J.F. (2013) Selecting traits that explain species–environment relationships: a generalized linear mixed model approach. *Journal of Vegetation Science,* **24,** 988-1000.http://dx.doi.org/10.1111/j.1654-1103.2012.12036.x

Kleyer, M., Dray, S., Bello, F., Lepš, J., Pakeman, R.J., Strauss, B., Thuiller, W. & Lavorel, S. (2012) Assessing species and community functional responses to environmental gradients: which multivariate methods? *Journal of Vegetation Science,* **23,** 805-821.http://dx.doi.org/10.1111/j.1654-1103.2012.01402.x

Lepš, J. & de Bello, F. (2023) Differences in trait–environment relationships: Implications for community weighted means tests. *Journal of Ecology,* **111,** 2328-2341.https://besjournals.onlinelibrary.wiley.com/doi/abs/10.1111/1365-2745.14172

Niku, J., Hui, F.K.C., Taskinen, S. & Warton, D.I. (2021) Analyzing environmental-trait interactions in ecological communities with fourth-corner latent variable models. *Environmetrics,* **n/a,** e2683.https://doi.org/10.1002/env.2683

Peng, F.-J., ter Braak, C.J.F., Rico, A. & Van den Brink, P.J. (2021) Double constrained ordination for assessing biological trait responses to multiple stressors: A case study with benthic macroinvertebrate communities. *Science of the Total Environment,* **754,** 142171.https://doi.org/10.1016/j.scitotenv.2020.142171

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

Pinho, B.X., Tabarelli, M., ter Braak, C.J.F., Wright, S.J., Arroyo-Rodríguez, V., Benchimol, M., Engelbrecht, B.M.J., Pierce, S., Hietz, P., Santos, B.A., Peres, C.A., Müller, S.C., Wright, I.J., Bongers, F., Lohbeck, M., Niinemets, Ü., Slot, M., Jansen, S., Jamelli, D., de Lima, R.A.F., Swenson, N., Condit, R., Barlow, J., Slik, F., Hernández-Ruedas, M.A., Mendes, G., Martínez-Ramos, M., Pitman, N., Kraft, N., Garwood, N., Guevara Andino, J.E., Faria, D., Chacón-Madrigal, E., Mariano-Neto, E., Júnior, V., Kattge, J. & Melo, F.P.L. (2021) Functional biogeography of Neotropical moist forests: Trait–climate relationships and assembly patterns of tree communities. *Global Ecology and Biogeography,* **30,** 1430-1446.https://doi.org/10.1111/geb.13309

Pocheville, A. (2015) The Ecological Niche: History and Recent Controversies. *Handbook of Evolutionary Thinking in the Sciences* (eds T. Heams, P. Huneman, G. Lecointre & M. Silberstein), pp. 547-586.Springer Netherlands, Dordrecht.https://doi.org/10.1007/978-94-017-9014-7\_26

Rolhauser, A.G., Waller, D.M. & Tucker, C.M. (2021) Complex trait‒environment relationships underlie the structure of forest plant communities. *Journal of Ecology,* **109,** 3794-3806. https://besjournals.onlinelibrary.wiley.com/doi/abs/10.1111/1365-2745.13757

Sîrbu, I., Benedek, A.-M., Brown, B.L. & Sîrbu, M. (2022) Disentangling structural and functional responses of native versus alien communities by canonical ordination analyses and variation partitioning with multiple matrices. *Scientific Reports,* **12,** 12813.https://doi.org/10.1038/s41598-022-16860-6

Sîrbu, I., Benedek, A.M. & Sîrbu, M. (2021) Variation partitioning in double-constrained multivariate analyses: linking communities, environment, space, functional traits, and ecological niches. *Oecologia,* **197,** 43-59.https://doi.org/10.1007/s00442-021-05006-6

ter Braak, C.J.F. (2019) New robust weighted averaging- and model-based methods for assessing trait–environment relationships. *Methods in Ecology and Evolution,* **10,** 1962-1971.https://doi.org/10.1111/2041-210X.13278

ter Braak, C.J.F. & Barendregt, L.G. (1986) Weighted averaging of species indicator values: its efficiency in environmental calibration. *Mathematical Biosciences,* **78,** 57-72.http://dx.doi.org/10.1016/0025-5564(86)90031-3

ter Braak, C.J.F., Cormont, A. & Dray, S. (2012) Improved testing of species traits–environment relationships in the fourth-corner problem. *Ecology,* **93,** 1525-1526.http://dx.doi.org/10.1890/12-0126.1

ter Braak, C.J.F. & Looman, C.W.N. (1986) Weighted averaging, logistic regression and the Gaussian response model. *Vegetatio,* **65,** 3-11.http://dx.doi.org/10.1007/BF00032121

ter Braak, C.J.F. & Šmilauer, P. (2018) *Canoco reference manual and user's guide: software for ordination (version 5.10).* Microcomputer Power, Ithaca, USA

ter Braak, C.J.F., Peres-Neto, P.R. & Dray, S. (2018a) Simple parametric tests for trait–environment association. *Journal of Vegetation Science,* **29,** 801-811.https://doi.org//10.1111/jvs.12666

ter Braak, C.J.F., Šmilauer, P. & Dray, S. (2018b) Algorithms and biplots for double constrained correspondence analysis. *Environmental and Ecological Statistics,* **25,** 171-197.https://doi.org/10.1007/s10651-017-0395-x

van Buuren, S. & Groothuis-Oudshoorn, K. (2011) mice: Multivariate Imputation by Chained Equations in R. *Journal of Statistical Software,* **45,** 1 - 67.https://www.jstatsoft.org/index.php/jss/article/view/v045i03

Zelený, D. (2018) Which results of the standard test for community weighted mean approach are too optimistic? *Journal of Vegetation Science,* **29,** 953-966.https://doi.org/10.1111/jvs.12688

1. The t-values for traits are obtained by regressing the unconstrained species scores on to the traits. Similarly, the t-values for the environmental variables are obtained by regressing the unconstrained site scores on to the environmental variables. Small such t-values (less than 2 in absolute value) indicate non-significant effects, even if the associated standardized regression coefficient is large. But be aware that values larger than 2 in absolute value do not necessarily indicate significance, as they are not t-distributed in reduced-rank regression. [↑](#footnote-ref-1)