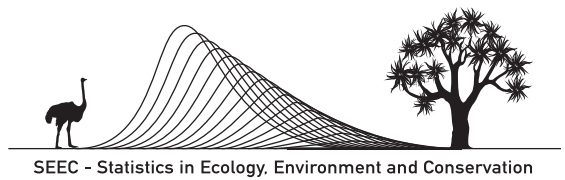
SEEC stats toolbox seminar series:

Constrained ordination

Mzabalazo Z. Ngwenya

Centre for Statistics in Ecology, Environment and Conservation (SEEC)



Department of Statistical Sciences
University of Cape Town

Outline

1. Introduction
2. Constrained ordination
3. R example

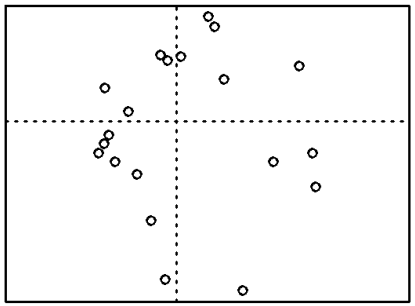
1. Introduction

- Community ecology data usually consists of one or more of the following data frames
  - **L** matrix: matrix of species composition (sample *x* species matrix); may contain abundances or presence/absence records
  - **R** matrix: matrix of environmental variables or more generally sample attributes (sample *x* environmental variables matrix)
  - **Q** matrix: matrix of species attributes or traits (species *x* traits matrix)
- These data are multivariate and can be analysed in two ways; either via cluster analysis or ordination.[1]

[1] See http://www.seec.uct.ac.za/introduction-multivariate-analyses for an introduction of multivariate analyses methods in ecology

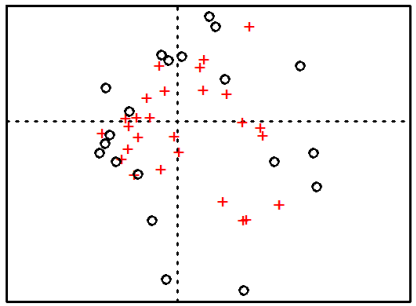
- Cluster analysis seeks to find discontinuities in a data set such that items in a cluster are similar and items between clusters are as different as possible.
- Ordination on the other hand extracts trends in a data set in the form of continuous axes.
- The aim of ordination methods is to represent the relationships in the data set as faithfully as possible using fewer dimensions then there initial number of dimensions in the data set. The data is then visualized using this reduced number of dimensions.

Scatterplot



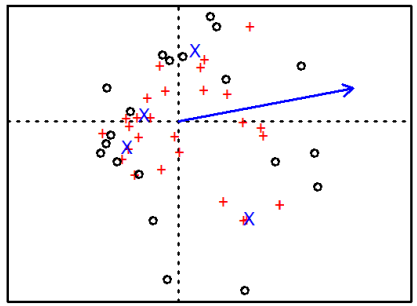
(o = sample scores)

Biplot



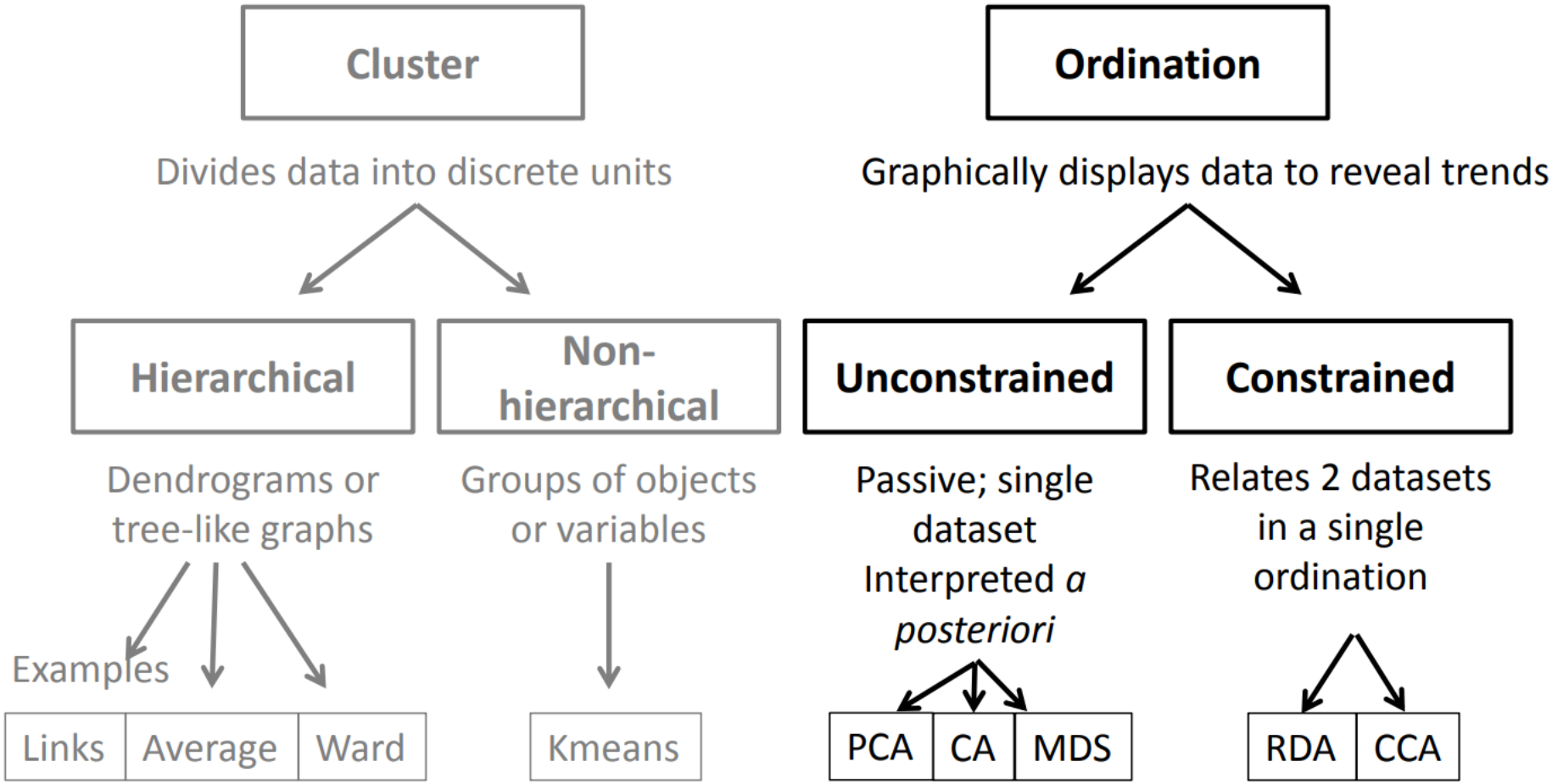
(o = sample, + = species scores)

Triplot



(o = sample, + = species scores,
vector & centroid = env. vars)

- From a practical point of view we seek to extract and visualize environmental gradients from ecological data-sets which are then be used as basis for describing the differential habitat preferences.
- Hence from a practical point of view with ordination we seek to either
  1. Describe the community pattern
  2. Explain and test changes in species composition via some variables
- The first task can be addressed via unconstrained ordination (indirect gradient descent) and the second task via constrained ordination (direct gradient descent).


Cluster
Divides data into discrete units
Hierarchical
Non-hierarchical
Dendrograms or tree-like graphs
Groups of objects or variables
Examples
Links
Average
Ward
Kmeans
Ordination
Graphically displays data to reveal trends
Unconstrained
Constrained
Passive; single dataset
Interpreted *a posteriori*
Relates 2 datasets in a single ordination
PCA
CA
MDS
RDA
CCA

- Constraining an ordination means that the ordination only shows community variation that can be attributed to some descriptor (environmental) variables.

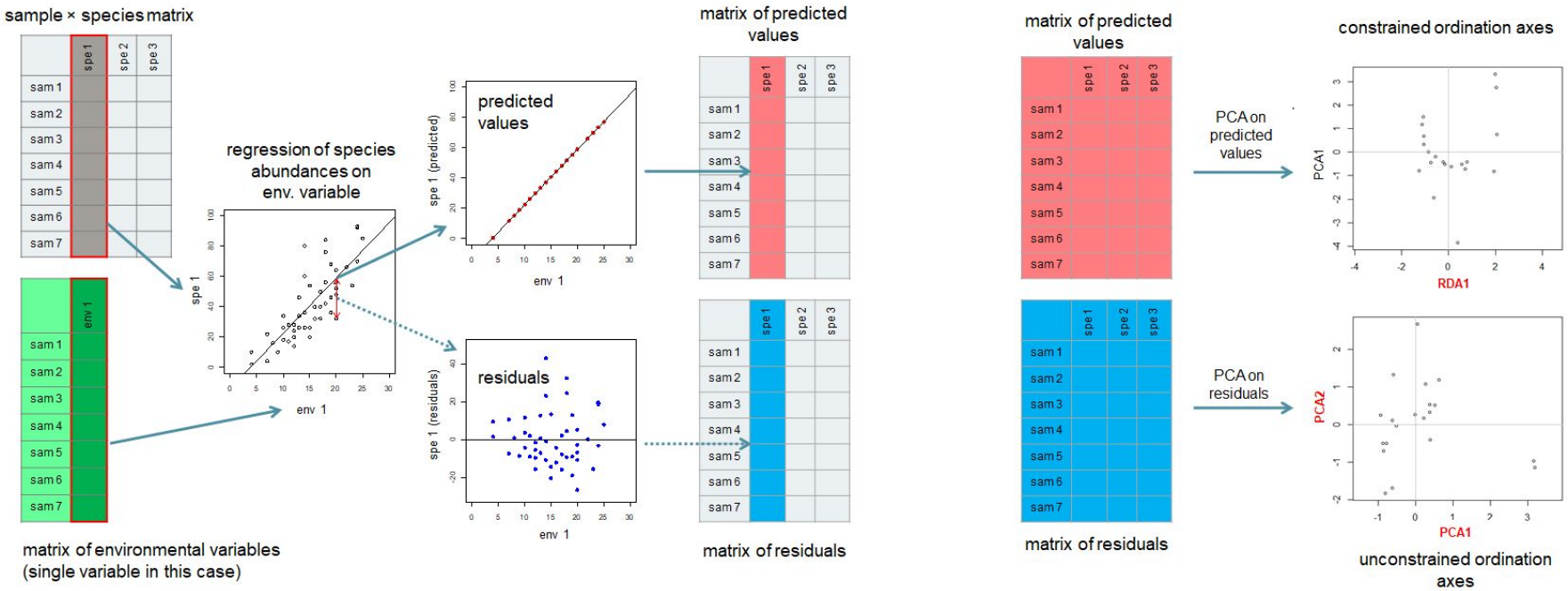
Unconstrained	Constrained
Operates on one data matrix	Explores the relationships between multiple data matrices
Descriptive; no statistical test to assess significance of detected patterns	Confirmatory analysis; able to test the significance of the relationships between explanatory (environmental) variables and species composition

2. Constrained ordination

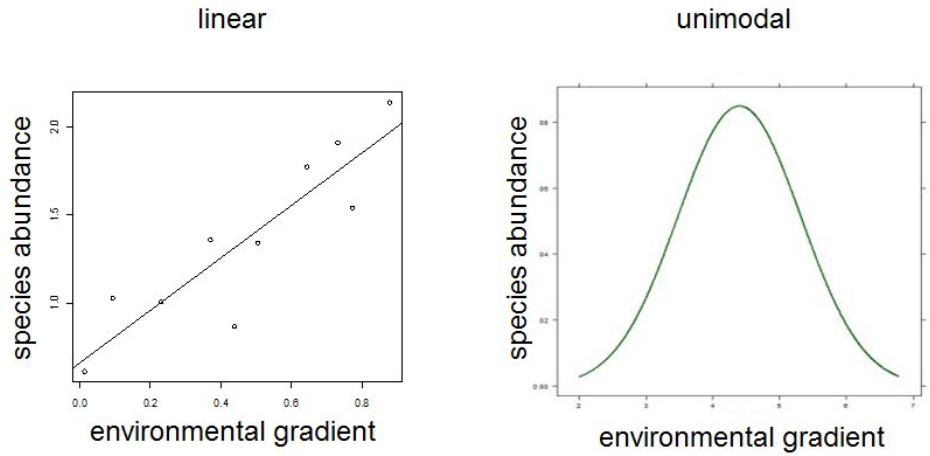
- Constrained ordination can be thought of as a set of multivariate regression analyses in which each set of variables is regressed against each sample.
- Because of the link to regression constrained ordination allows one to decompose the total variance in species composition data into a fraction explained by environmental variables (related to constrained ordination axes) and not explained by environmental variables (related to unconstrained ordination axes).

- Constrained ordination relates the species composition directly to the environmental variables and extracts the variance in species composition which is directly related to the environment.
- Hence constrained ordination axes correspond to the directions of the variability in data which is explained by environmental variables.
- Two of the most widely used constrained ordination methods are Redundancy analysis (RDA) and Canonical Correspondence Analysis (CCA).

- RDA is multiple linear regression followed by PCA of fitted values table



- Canonical correspondence analysis (CCA) combines multiple regression with CA.
- The algorithm of RDA is modified such that instead of raw species composition data, the set of regressions is done on the $\bar{Q}$ matrix, and weighted multiple regression is used instead of simple multiple regression,
- RDA is best for exploring linear responses to environmental gradients whilst CCA is best for investigating unimodal relationships.



3. R example

- Data[2] from package vegan
- **L** matrix: Reindeer pastures with 24 sites and 44 species; reported values are estimated cover of each species
- **R** matrix: 14 continuous environmental variables observed at the 24 reindeer pasture sites

[2] Väre, H., Ohtonen, R. and Oksanen, J. (1995) Effects of reindeer grazing on understorey vegetation in dry Pinus sylvestris forests. Journal of Vegetation Science 6, 523–530.

Preliminaries:

```
library(vegan) # load package

# load and check data
data("varespec")  #sample x species matrix (L)
str(varespec)
head(varepec)
```

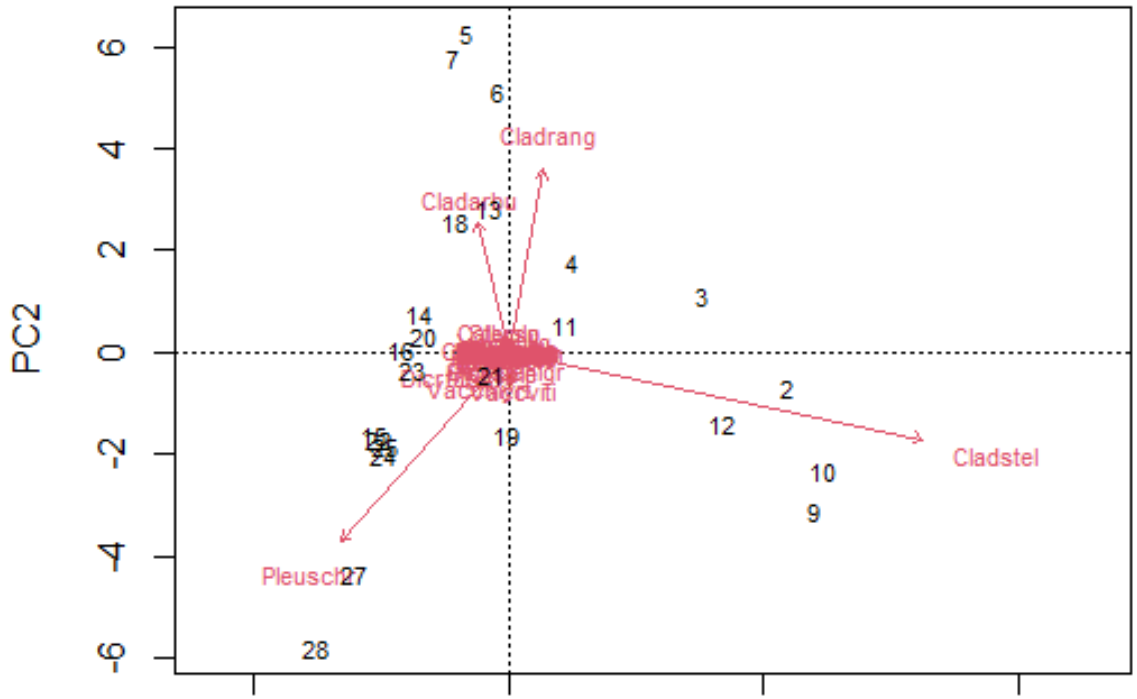
	Callvulg	Empenigr	Rhodtome	Vaccmyrt	Vaccviti	Pinusylv	Descflex	Betupube	Vacculig	Diphcomp	Dicrsp	Dicrfusc	Di
18	0.55	11.13	0.00	0.00	17.80	0.07	0.00	0	1.60	2.07	0.00	1.62	
15	0.67	0.17	0.00	0.35	12.13	0.12	0.00	0	0.00	0.00	0.33	10.92	
24	0.10	1.55	0.00	0.00	13.47	0.25	0.00	0	0.00	0.00	23.43	0.00	
27	0.00	15.13	2.42	5.92	15.97	0.00	3.70	0	1.12	0.00	0.00	3.63	
23	0.00	12.68	0.00	0.00	23.73	0.03	0.00	0	0.00	0.00	0.00	3.42	
19	0.00	8.92	0.00	2.42	10.28	0.12	0.02	0	0.00	0.00	0.00	0.32	

```
data("varechem") # sample x environmental variables matrix (R)
str(varechem)
```

```
'data.frame': 24 obs. of  14 variables:
$ N        : num  19.8 13.4 20.2 20.6 23.8 22.8 26.6 24.2 29.8 28.1 ...
$ P        : num  42.1 39.1 67.7 60.8 54.5 40.9 36.7 31 73.5 40.5 ...
$ K        : num  140 167 207 234 181 ...
$ Ca       : num  519 357 973 834 777 ...
$ Mg       : num  90 70.7 209.1 127.2 125.8 ...
$ S        : num  32.3 35.2 58.1 40.7 39.5 40.8 33.8 27.1 42.5 60.2 ...
$ Al       : num  39 88.1 138 15.4 24.2 ...
$ Fe       : num  40.9 39 35.4 4.4 3 ...
$ Mn       : num  58.1 52.4 32.1 132 50.1 ...
$ Zn       : num  4.5 5.4 16.8 10.7 6.6 9.1 7.4 5.2 9.3 9.1 ...
$ Mo       : num  0.3 0.3 0.8 0.2 0.3 0.4 0.3 0.3 0.3 0.5 ...
$ Baresoil: num  43.9 23.6 21.2 18.7 46 40.5 23 29.8 17.6 29.9 ...
$ Humdepth: num  2.2 2.2 2 2.9 3 3.8 2.8 2 3 2.2 ...
$ pH       : num  2.7 2.8 3 2.8 2.7 2.7 2.8 2.8 2.8 2.8 ...
```

plot(rda(varespec))

```
Call: rda(X = varespec)

Inertia Rank
Total            1826
Unconstrained    1826   23
Inertia is variance

Eigenvalues for unconstrained axes:
PC1   PC2   PC3   PC4   PC5   PC6   PC7   PC8
983.0 464.3 132.3  73.9  48.4  37.0  25.7  19.7
(Showing 8 of 23 unconstrained eigenvalues)
```



1. Model building:

```
RDA.all <- rda(varespec ~., data= varechem) # RDA <- rda(varespec, varechem)
RDA.all
```

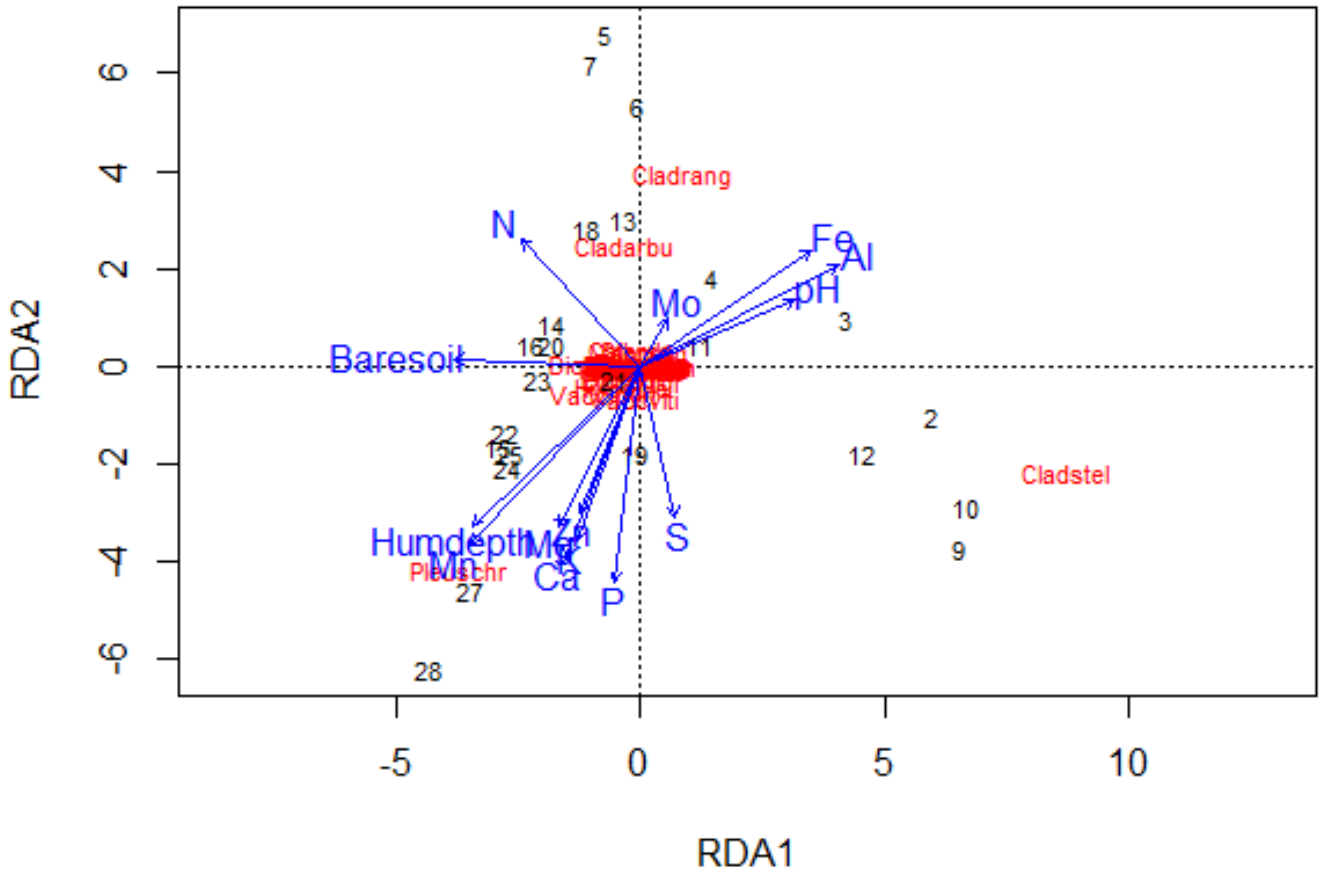
```
Call: rda(formula = varespec ~ N + P + K + Ca + Mg + S + Al + Fe + Mn + Zn +
Mo + Baresoil + Humdepth + pH, data = varechem)

Inertia Proportion Rank
Total         1825.6594     1.0000
Constrained   1459.8891     0.7997   14
Unconstrained  365.7704     0.2003    9
Inertia is variance

Eigenvalues for constrained axes:
RDA1  RDA2  RDA3  RDA4  RDA5  RDA6  RDA7  RDA8  RDA9 RDA10 RDA11 RDA12 RDA13 RDA14
820.1 399.3 102.6  47.6  26.8  24.0  19.1  10.2   4.4   2.3   1.5   0.9   0.7   0.3

Eigenvalues for unconstrained axes:
PC1    PC2    PC3    PC4    PC5    PC6    PC7    PC8    PC9
186.19  88.46  38.19  18.40  12.84  10.55   5.52   4.52   1.09
```

summary(RDA.all)

```
Call:
rda(formula = varespec ~ N + P + K + Ca + Mg + S + Al + Fe + Mn + Zn + Mo + Baresoil + Humdepth + pH,
     data = varechem)

Partitioning of variance:
Inertia Proportion
Total          1825.7     1.0000
Constrained    1459.9     0.7997
Unconstrained   365.8     0.2003

Eigenvalues, and their contribution to the variance

Importance of components:
                          RDA1     RDA2      RDA3     RDA4    RDA5     RDA6     RDA7
Eigenvalue            820.1042 399.2847 102.56168 47.63169 26.8382 24.04809 19.06438
Proportion Explained    0.4492   0.2187   0.05618  0.02609  0.0147  0.01317  0.01044
Cumulative Proportion   0.4492   0.6679   0.72409  0.75019  0.7649  0.77806  0.78850

Accumulated constrained eigenvalues
Importance of components:
                          RDA1     RDA2      RDA3     RDA4     RDA5     RDA6     RDA7
Eigenvalue            820.1042 399.2847 102.56168 47.63169 26.83822 24.04809 19.06438
Proportion Explained    0.5618   0.2735   0.07025  0.03263  0.01838  0.01647  0.01306
Cumulative Proportion   0.5618   0.8353   0.90551  0.93814  0.95653  0.97300  0.98606
```

```
Species scores

                RDA1       RDA2       RDA3       RDA4       RDA5       RDA6
Callvulg -0.1516557  0.3244635  8.016e-02  1.0775112 -0.8752695 -0.0895046
Empenigr -0.1601177 -0.2952352  6.174e-01 -0.2508970  0.6812008 -0.2528771
Rhodtome -0.1148931 -0.0305967  4.398e-02 -0.0351583  0.0390864  0.0098011

Site scores (weighted sums of species scores)

       RDA1    RDA2    RDA3     RDA4    RDA5    RDA6
18 -1.08199  2.8291  2.2645 -2.46209  1.9279 -4.8908
15 -2.88751 -1.6383 -1.1533  2.41906  0.1465 -1.4756

Site constraints (linear combinations of constraining variables)

       RDA1     RDA2     RDA3     RDA4    RDA5     RDA6
18 -2.9026  2.89996  2.96998 -1.69026  3.5479 -4.23983
15  0.2116 -1.77427  0.73050 -0.09712  0.1127 -3.55952

Biplot scores for constraining variables

             RDA1     RDA2     RDA3      RDA4      RDA5      RDA6
N        -0.37451  0.40005 -0.32028  0.120999 -0.021093  0.248110
P        -0.07887 -0.67736 -0.18434 -0.212893 -0.346547 -0.006476
K        -0.20146 -0.54606 -0.07107  0.248513 -0.530396 -0.100114
Ca       -0.23542 -0.59524  0.04435 -0.126410 -0.106923  0.179456
```

```
plot(RDA.all)
```


RDA2
RDA1
-6
-4
-2
0
2
4
6
-5
0
5
10
N
Fe
Al
pH
Mo
Baresoil
Humdepth
Mg
Mn
Ca
K
P
S
Cladrang
Cladarbu
Cladstel
5
7
6
18
13
4
3
14
16
20
11
23
21
2
22
15
25
24
19
12
10
9
27
28

2. Variation explained and significance:

i) R-squared for whole model

```
(adjR.rda <- RsquareAdj(RDA.all)) #adjusted R-squared for whole model
```

```
$r.squared
[1] 0.7996503

$adj.r.squared
[1] 0.4879952
```

ii) Significance of model

```
anova.cca(RDA.all)
```

```
Permutation test for rda under reduced model
Permutation: free
Number of permutations: 999

Model: rda(formula = varespec ~ N + P + K + Ca + Mg + S + Al + Fe + Mn + Zn + Mo + Baresoil + Humdepth + pH,
  data = varechem)
Df Variance      F Pr(>F)
Model    14  1459.89 2.5658  0.004 **
Residual  9   365.77
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

iii) Significance of each axis

```
anova.cca(RDA.all, by ="axis")
```

```
Permutation test for rda under reduced model
Forward tests for axes
Permutation: free
Number of permutations: 999

Model: rda(formula = varespec ~ N + P + K + Ca + Mg + S + Al + Fe + Mn + Zn + Mo + Baresoil + Humdepth + pH,
         data = varechem)
Df Variance       F Pr(>F)
RDA1      1   820.10 20.1792  0.008 **
RDA2      1   399.28  9.8246  0.105
RDA3      1   102.56  2.5236  0.969
RDA4      1    47.63  1.1720  1.000
RDA5      1    26.84  0.6604  1.000
RDA6      1    24.05  0.5917  1.000
RDA7      1    19.06  0.4691  1.000
RDA8      1    10.17  0.2502  1.000
RDA9      1     4.43  0.1090  1.000
RDA10     1     2.27  0.0559  1.000
RDA11     1     1.54  0.0378  1.000
RDA12     1     0.93  0.0228  1.000
RDA13     1     0.72  0.0176  1.000
RDA14     1     0.31  0.0077  1.000
Residual  9   365.77
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

iv) Significance of each explanatory variable (depends on order of terms)

```
anova.cca(RDA.all, by ="terms")
```

```
Permutation test for rda under reduced model
Terms added sequentially (first to last)
Permutation: free
Number of permutations: 999

Model: rda(formula = varespec ~ N + P + K + Ca + Mg + S + Al + Fe + Mn + Zn + Mo + Baresoil + Humdepth + pH,
           data = varechem)
Df Variance      F Pr(>F)
N          1   193.72 4.7667  0.016 *
P          1   181.88 4.4753  0.020 *
K          1    59.66 1.4681  0.222
Ca         1    80.83 1.9890  0.125
Mg         1    19.68 0.4842  0.716
S          1   241.87 5.9515  0.004 **
Al         1   173.36 4.2657  0.014 *
Fe         1    47.25 1.1626  0.335
Mn         1    22.58 0.5555  0.660
Zn         1    33.69 0.8291  0.477
Mo         1   169.55 4.1719  0.016 *
Baresoil   1    87.99 2.1650  0.103
Humdepth   1   106.06 2.6097  0.073 .
pH         1    41.75 1.0273  0.382
Residual   9   365.77
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

v) Significance of each explanatory variable (does not depend on order of terms)

```
anova.cca(RDA.all, by ="margin")
```

```
Permutation test for rda under reduced model
Marginal effects of terms
Permutation: free
Number of permutations: 999

Model: rda(formula = varespec ~ N + P + K + Ca + Mg + S + Al + Fe + Mn + Zn + Mo + Baresoil + Humdepth + pH,
           data = varechem)
          Df Variance      F Pr(>F)
N          1    45.29 1.1143  0.316
P          1    43.83 1.0785  0.375
K          1   110.96 2.7302  0.062 .
Ca         1    29.46 0.7250  0.532
Mg         1   101.97 2.5090  0.081 .
S          1   184.72 4.5451  0.008 **
Al         1    30.52 0.7511  0.517
Fe         1    17.19 0.4229  0.748
Mn         1    49.78 1.2249  0.282
Zn         1    21.83 0.5372  0.676
Mo         1   114.90 2.8272  0.053 .
Baresoil   1    83.17 2.0464  0.109
Humdepth   1   142.36 3.5029  0.030 *
pH         1    41.75 1.0273  0.376
Residual   9   365.77
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

3. Variable selection

```
RDA.null <- rda(varespec ~1, data= varechem)
RDA.sel <- ordiR2step(RDA.null, scope= formula(RDA.all),
            R2scope =adjR.rda$adj.r.squared, direction ="both", permutations = 999)
RDA.sel
```

```
Call: rda(formula = varespec ~ Al + K + N + Baresoil, data = varechem)

Inertia Proportion Rank
Total         1825.6594     1.0000
Constrained    923.5535     0.5059    4
Unconstrained  902.1059     0.4941   19
Inertia is variance

Eigenvalues for constrained axes:
RDA1  RDA2  RDA3  RDA4
594.2 235.8  74.2  19.4

Eigenvalues for unconstrained axes:
PC1   PC2   PC3   PC4   PC5   PC6   PC7   PC8
491.2 165.9  96.6  42.7  32.3  21.8  16.4  11.4
(Showing 8 of 19 unconstrained eigenvalues)
```

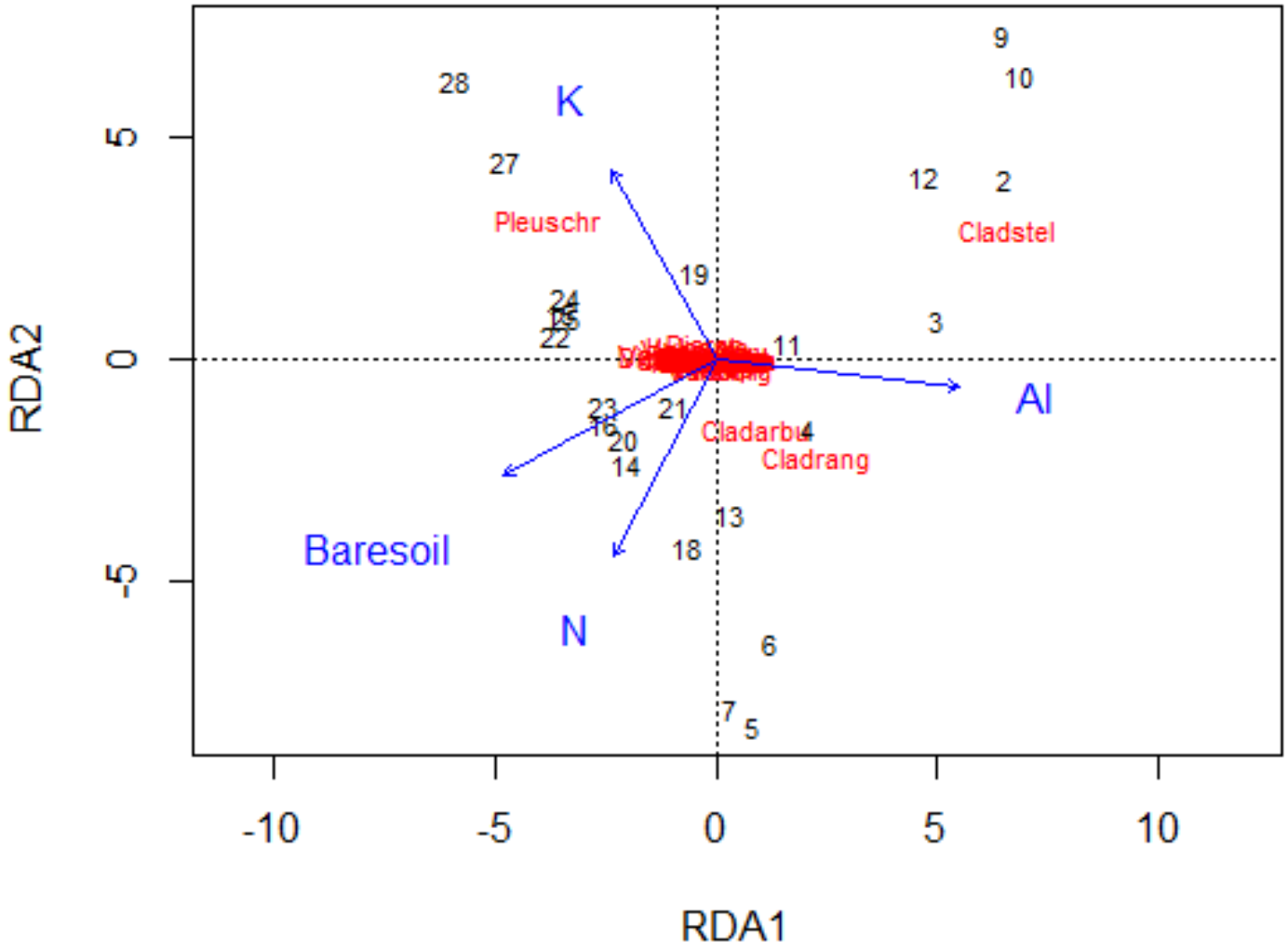
```
RDA.sel$anova  # adjusted R-squared values for individual terms
```

```
                 R2.adj Df    AIC      F Pr(>F)
+ Al             0.16836  1 177.72 5.6563  0.005 **
+ K              0.25607  1 175.93 3.5936  0.013 *
+ N              0.33672  1 174.00 3.5537  0.024 *
+ Baresoil       0.40185  1 172.29 3.1775  0.029 *
<All variables> 0.48800
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
anova.cca(RDA.sel, by ="axis")
```

```
Permutation test for rda under reduced model
Forward tests for axes
Permutation: free
Number of permutations: 999

Model: rda(formula = varespec ~ Al + K + N + Baresoil, data = varechem)
Df Variance       F Pr(>F)
RDA1      1   594.15 12.5139  0.002 **
RDA2      1   235.79  4.9662  0.020 *
RDA3      1    74.17  1.5622  0.421
RDA4      1    19.44  0.4094  0.836
Residual 19   902.11
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
plot(RDA.sel)
```


RDA2
RDA1
5
0
-5
-10
-5
0
5
10
K
Al
Baresoil
N
Pleuschr
Cladstel
Cladarbu
Cladrang
28
27
9
10
12
2
19
24
22
3
11
23
21
16
20
14
4
13
18
6
7
5

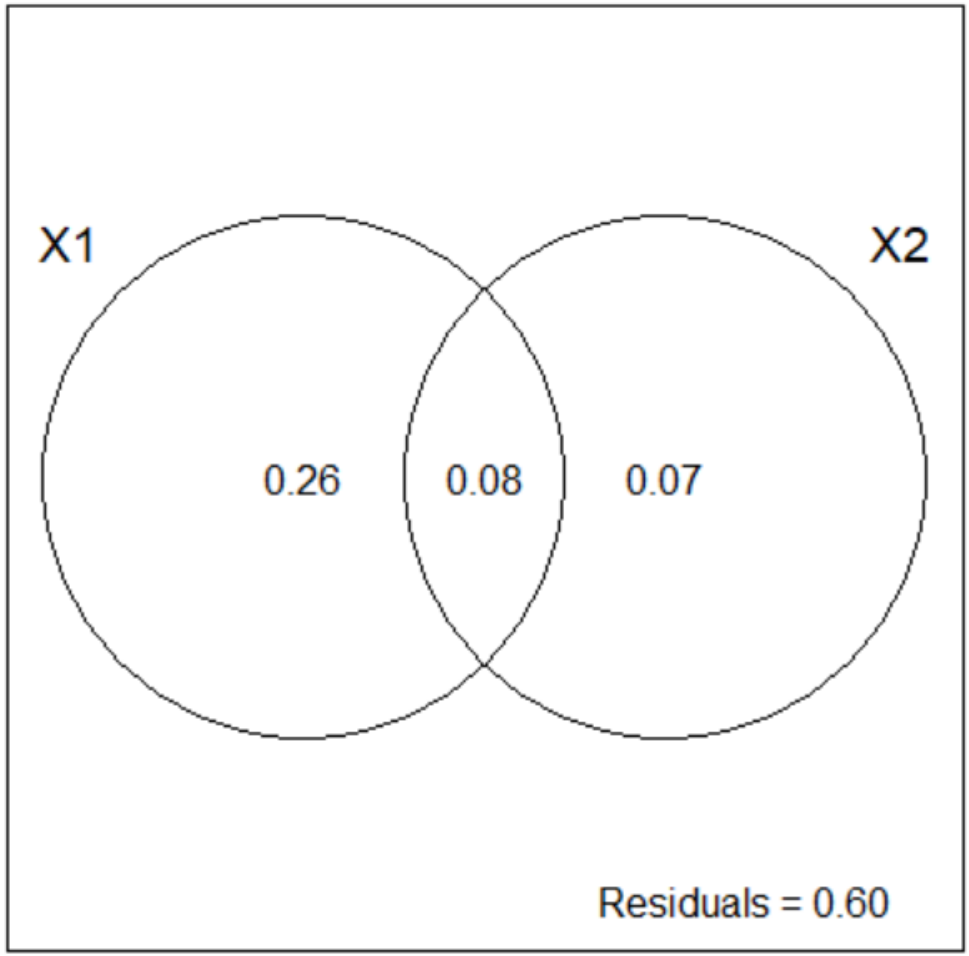
Variation partitioning

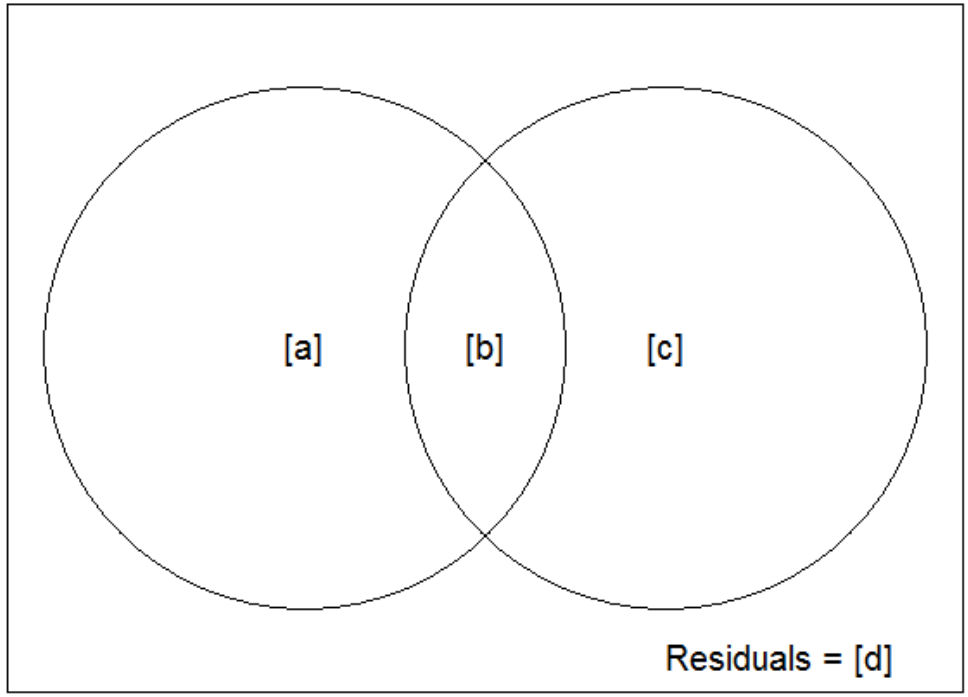
i) Partitioning variation to groups of terms

```
varpart.rda.grp <- varpart(varespec, ~ Al + K + N, ~ Baresoil, data = varechem)
varpart.rda.grp
```

```
Partition of variance in RDA

Call: varpart(Y = varespec, X = ~Al + K + N, ~Baresoil, data = varechem)

Explanatory tables:
X1:  ~Al + K + N
X2:  ~Baresoil

No. of explanatory tables: 2
Total variation (SS): 41990
Variance: 1825.7
No. of observations: 24

Partition table:
                     Df R.squared Adj.R.squared Testable
[a+b] = X1            3   0.42324       0.33672     TRUE
[b+c] = X2            1   0.18353       0.14642     TRUE
[a+b+c] = X1+X2       4   0.50587       0.40185     TRUE
Individual fractions
[a] = X1|X2           3                 0.25543     TRUE
[b]                   0                 0.08129    FALSE
[c] = X2|X1           1                 0.06513     TRUE
[d] = Residuals                         0.59815    FALSE
---
Use function 'rda' to test significance of fractions of interest
```

```
plot(varpart.rda.grp)
```





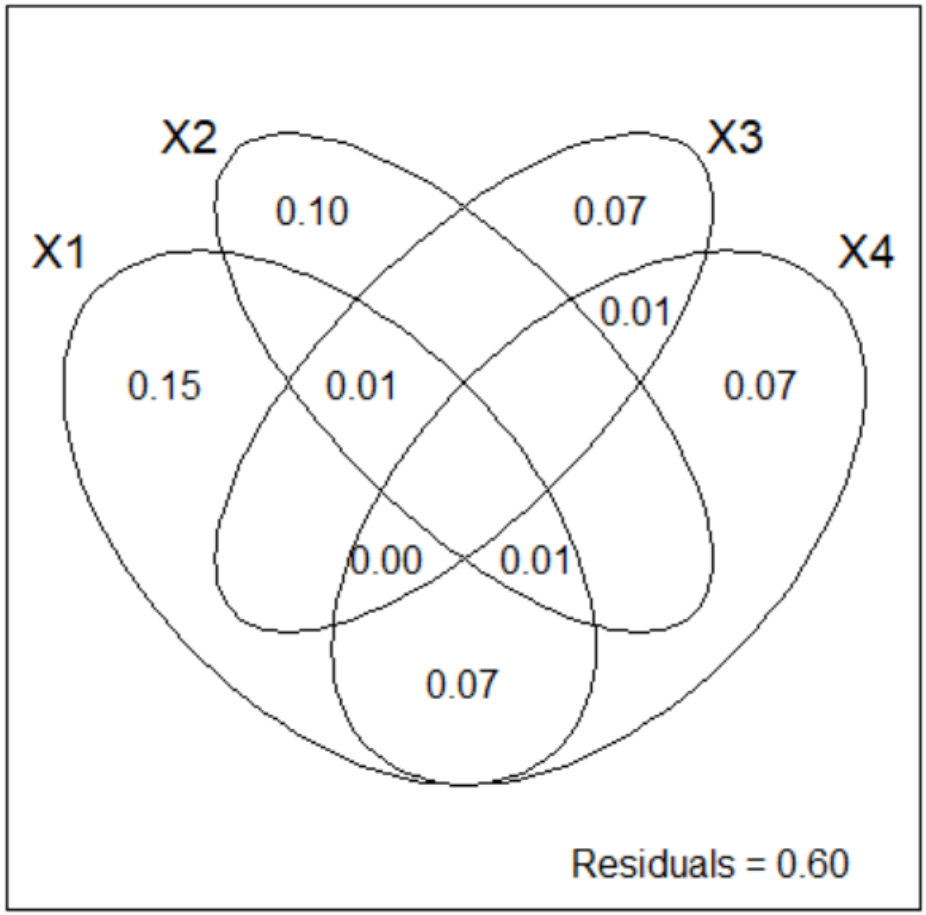
- a - variation explained by variable 1 (conditional (or partial) effect of variable 1, i.e. variation this variable would explain if putting variable 2 as covariable);
- c - variation explained by variable 2;
- b - shared variation explained by both variables (cannot be decided to which of them should be attributed, and is a result of correlation between both variables);
- a+b - variation explained by variable 1 (independent simple (or marginal) effect of variable 1, i.e. variation this variable would explain if it is as the only explanatory variable in the model);
- b+c - variation explained by variable 2;
- d - unexplained variation.

ii) Partitioning variation to individual terms

```
varpart.rda.ind <- varpart(varespec, ~ Al, ~ K, ~N, ~ Baresoil, data = varechem)
varpart.rda.ind
```

```
Partition of variance in RDA

Call: varpart(Y = varespec, X = ~Al, ~K, ~N, ~Baresoil, data = varechem)

Explanatory tables:
X1:  ~Al
X2:  ~K
X3:  ~N
X4:  ~Baresoil

No. of explanatory tables: 4
Total variation (SS): 41990
Variance: 1825.7
No. of observations: 24

Partition table
Df R.square Adj.R.square Testable
[aeghklno] = X1                1  0.20452      0.16836     TRUE
[befiklmo] = X2                1  0.09053      0.04919     TRUE
[cfgjlmno] = X3                1  0.10611      0.06548     TRUE
[dhijkmno] = X4                1  0.18353      0.14642     TRUE
[abefghiklmno] = X1+X2         2  0.32076      0.25607     TRUE
[acefghjklmno] = X1+X3         2  0.30487      0.23866     TRUE
[adeghijklmno] = X1+X4         2  0.30075      0.23415     TRUE
[bcefgijklmno] = X2+X3         2  0.19742      0.12099     TRUE
[bdefhijklmno] = X2+X4         2  0.26566      0.19572     TRUE
```

```
plot(varpart.rda.ind)
```



Bibliography

Borcard, D., Gillet, F. and Legendre, P. (2018).
Numerical Ecology with R. Second edition. Springer.

Legendre, P. and Legendre, L. (2018).
Numerical Ecology. Third English edition. Elsevier Science BV, Amsterdam.

Oksanen, J., Blanchet, F.G., Friendly, M., Kindt, R., Legendre, P., McGlinn, D., Minchin, P.R., O'Hara, R.B., Simpson, G.L., Solymos, P., Henry, M., Stevens, H., Szoecs, E. and Wagner, H. (2020).
vegan: Community Ecology Package. R package version 2.5-7. https://CRAN.R-project.org/package=vegan