Data Science Central

THE ONLINE RESOURCE FOR BIG DATA PRACTITIONERS

Modelling multiple response variables

General linear models (not Generalized linear model)

Linear Model	Common name
$Y = \mu + X$	Simple linear regression
$Y = \mu + A_1$	One-factorial (one-way) ANOVA
$Y = \mu + A_1 + A_2 + A_1{\times}A_2$	Two-factorial (two-way) ANOVA
$Y = \mu + A_1 + X\ (+A_1 \times X)$	Analysis of Covariance (ANCOVA)
$Y = \mu + X_1 + X_2 + X_3$	Multiple regression
$Y = \mu + A_1 + g + A_1 \times g$	Mixed model ANOVA
$Y_1 + Y_2 + \cdots Y_r = \mu + A_1 + A_2 + A_1{\times}A_2$ $(Y_1, Y_{,2}, \ldots Y_p) = \mu + X_1 + X_2 + \cdots X_p$	Multivariate ANOVA (MANOVA) and RDA (Redundancy Analysis)

Y (response) is a continuous variable

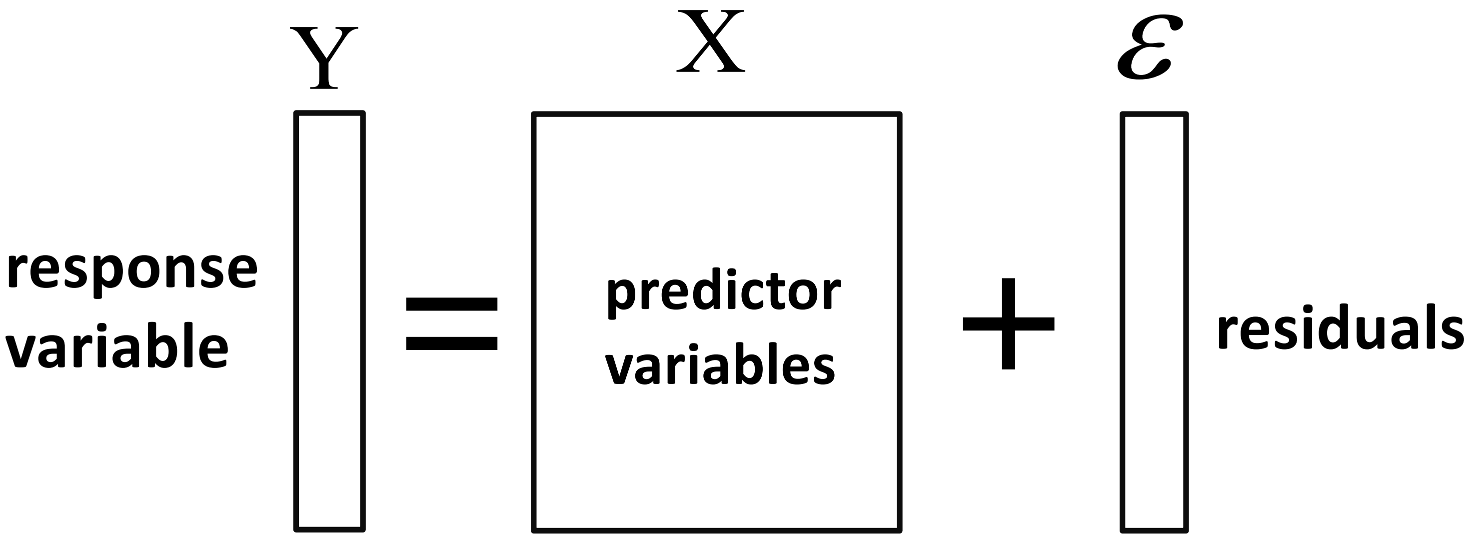
X (predictor) is a continuous variable

A represents categorical predictors (factors)

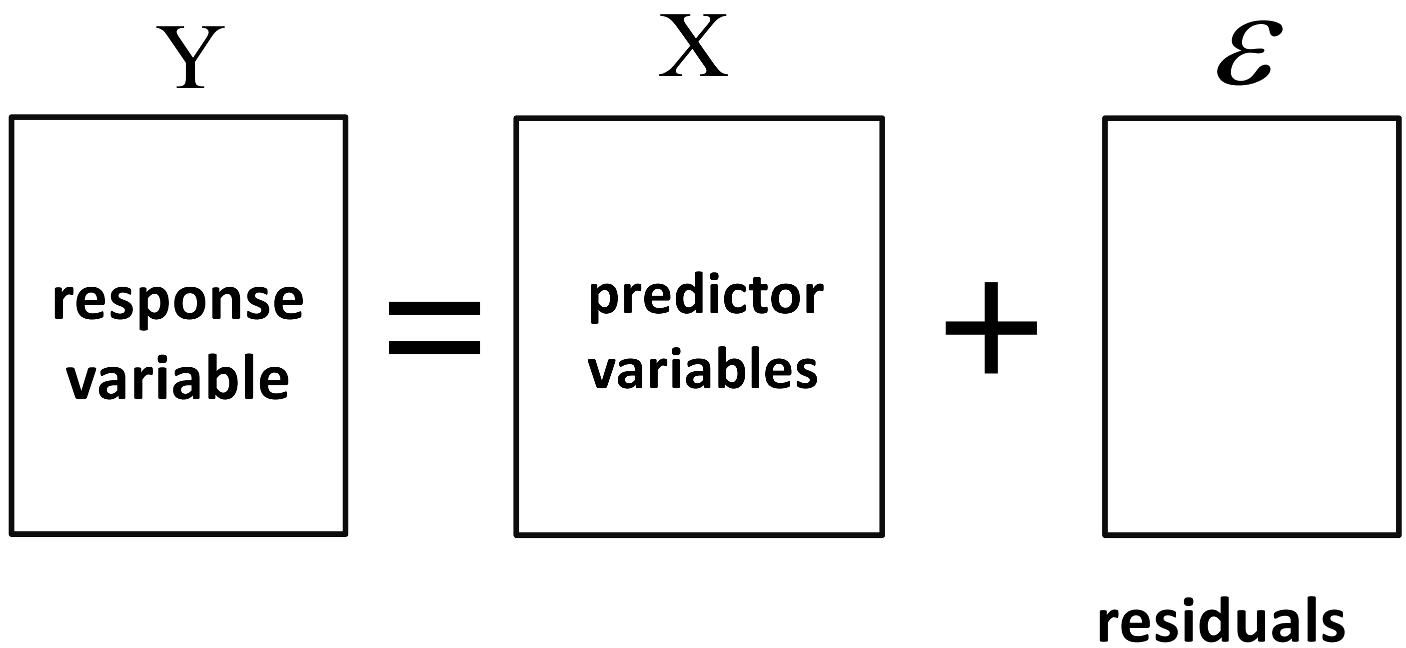
g represents groups of data

p represents the number of predictors

Classic multiple regression



Extending the classical multiple regression to multiple response variables



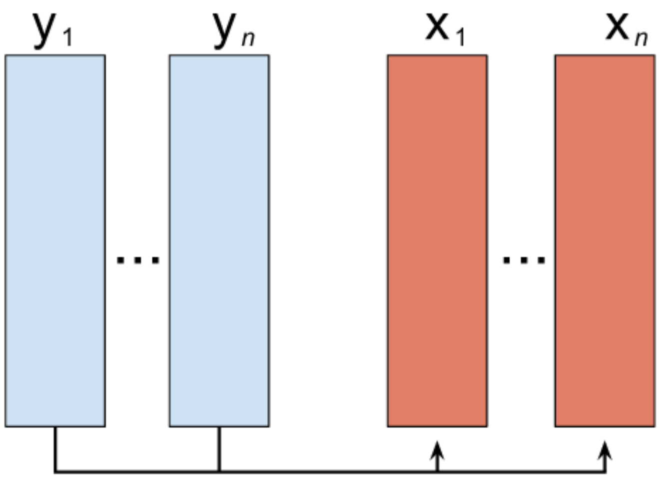
Modelling multiple response variables

Identify commonalities and differences among response variables in their relationships with predictors:

- Which response variables share common patterns of variation in relation to specific predictors?

- Which response variables exhibit distinct or unique variation with respect to certain predictors?

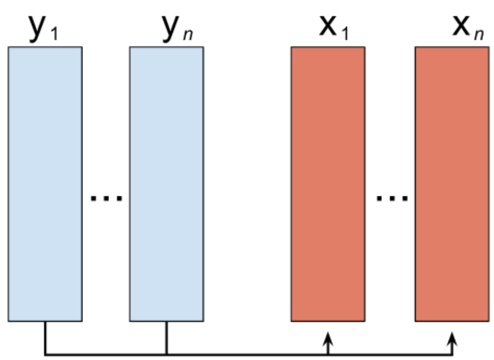
Redundancy Analysis



The basics -

1) Each response separately is regressed against all predictors.
2) Predicted values from each separate regression are then used in a Principal Component Analysis (PCA) so that common and unshared trends of variation in predicted values are described.

Redundancy Analysis

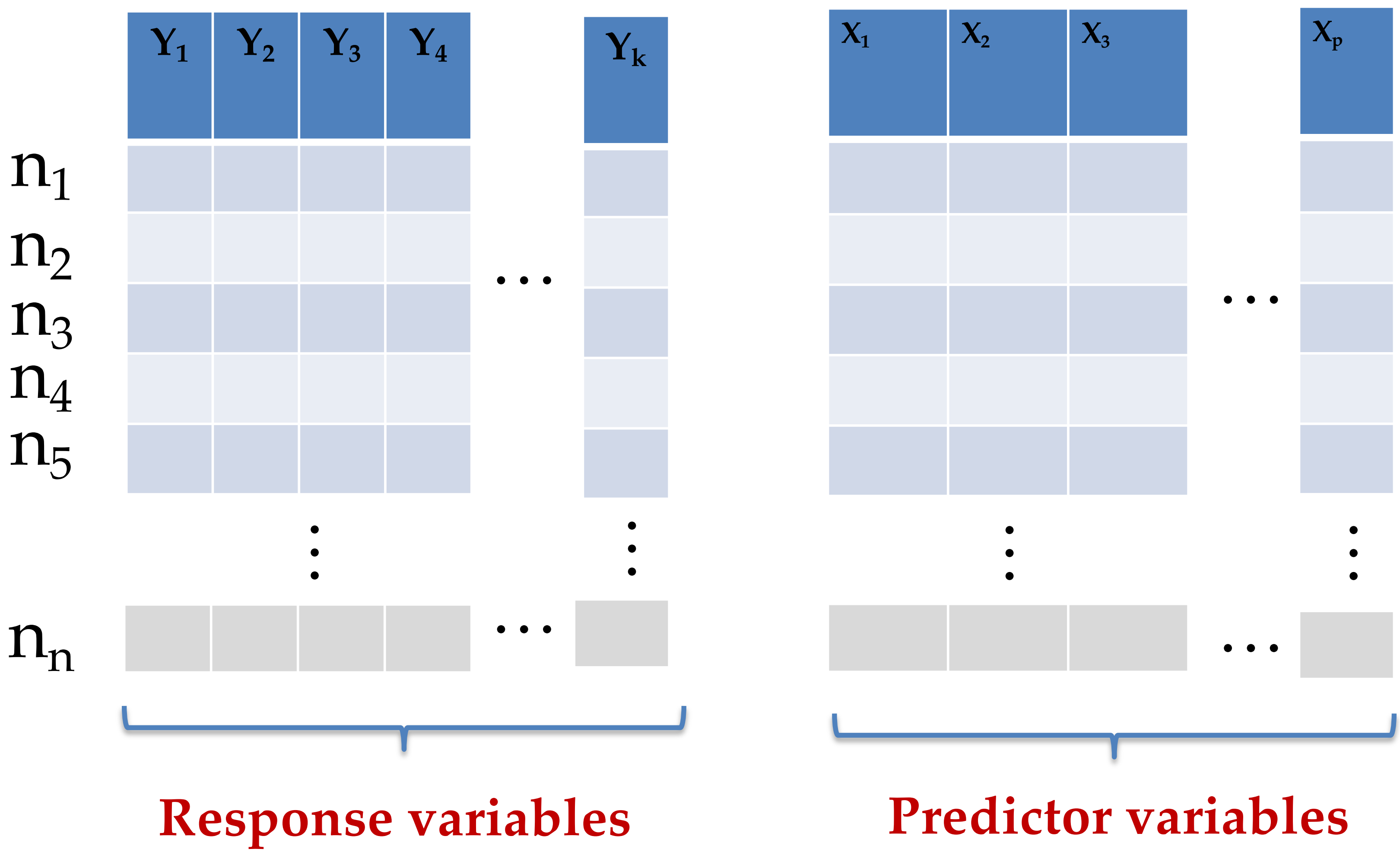


The basics -

1) Each response separately is regressed against all predictors.
2) Predicted values are used in a PCA so that common and unshared trends of variation are uncovered and described.

Because the PCA here is based on predicted Y values rather than the original Y values, the method is known as "constrained PCA"; since PCA is an ordination method, the general method is known as "constrained ordination".

The usual data format for Redundancy Analysis



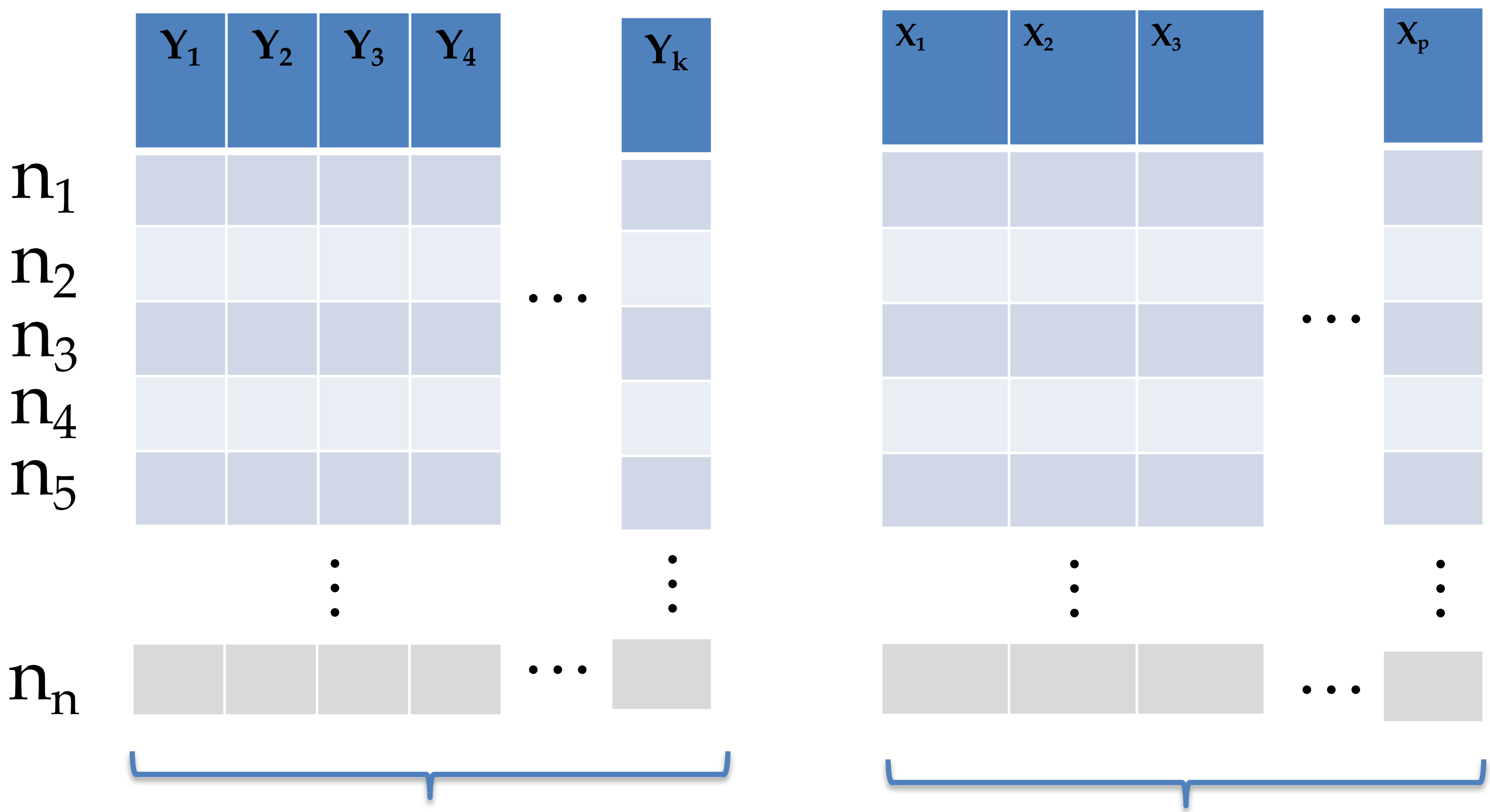
Redundancy Analysis – some examples
Ex. 1

Benthic diatom communities respond rapidly to environmental change. At four shallow sites in the Windmill Islands (Casey, East Antarctica), redundancy analysis showed that sediment grain-size, light availability, and water depth explained 30% of the variation in diatom relative abundances.

Sediment mud content (<63 µm) alone accounted for 18% of the variation across all sites, and over 25% within two sites. Location differences explained 28% of variation, largely driven by site-specific differences in grain-size, light, and depth.

Cunningham L. and McMinn A. 2004. The influence of natural environmental factors on benthic diatom communities from the Windmill Islands, Antarctica. PHYCOLOGIA 43: 744-755

The usual data format for Redundancy Analysis



Response variables

Diatoms

Predictor variables

sediment grain-size, light availability and water depth account

Redundancy Analysis – some examples
Ex. 2

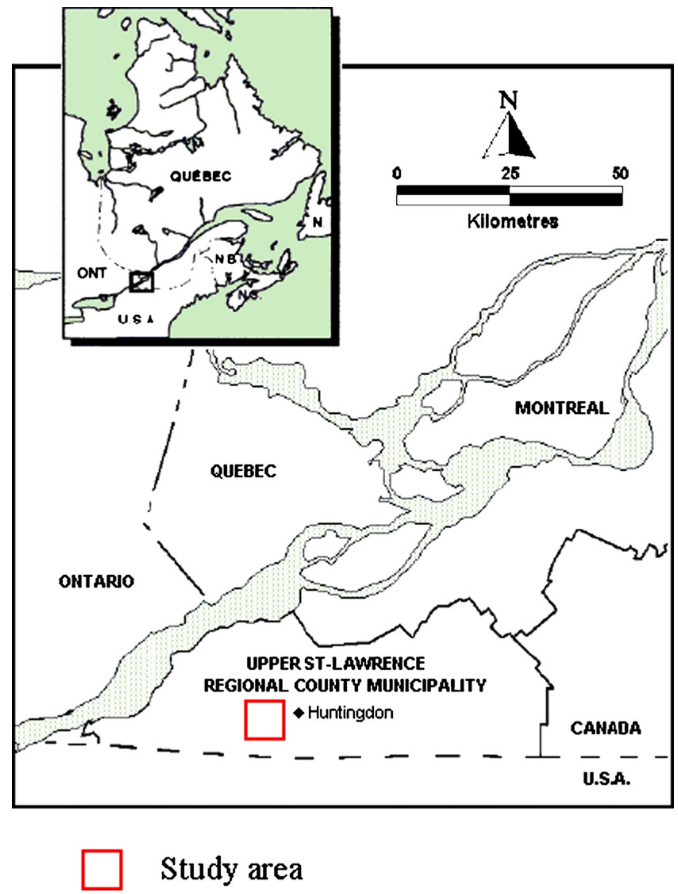
Available online at www.sciencedirect.com
ScienceDirect
Landscape and Urban Planning 83 (2007) 228–244

LANDSCAPE
AND
URBAN PLANNING
www.elsevier.com/locate/landurbplan

Abandoned farmlands as components of rural landscapes:
An analysis of perceptions and representations

Karyne Benjamin [a,b,*], André Bouchard [a,c], Gérald Domon [b,c]

In order to establish relationships between the ***10 perception criteria of a land use type and the socio-economic variables of the owners,*** canonical ***redundancy analyses (RDA)*** were done for each land use using the Canoco programme (ter Braak and Smilauer, 2002).



Land perception – criteria (response variables)

Y

Table 3

Tidy	7	6	5	4	3	2	1	Untidy
Beautiful	7	6	5	4	3	2	1	Ugly
Varied	7	6	5	4	3	2	1	Uniform
Pleasant	7	6	5	4	3	2	1	Unpleasant
Useful	7	6	5	4	3	2	1	Useless
Stressful	7	6	5	4	3	2	1	Relaxing
Cause shame	7	6	5	4	3	2	1	Cause pride
Rare	7	6	5	4	3	2	1	Common
Artificial	7	6	5	4	3	2	1	Natural
Productive	7	6	5	4	3	2	1	Unproductive
				Undecided				

Hay field

Woodlot

Corn field

Pasture

Shrub dominated abandoned farmland

Herbaceous abandoned farmland

In order to establish relationships between the ***<u>10 perception criteria of a land use type</u>*** and the socio-economic variables of the owners, canonical redundancy analyses (RDA) were done for each land use using the Canoco programme (ter Braak and Smilauer, 2002).

X

Table 1

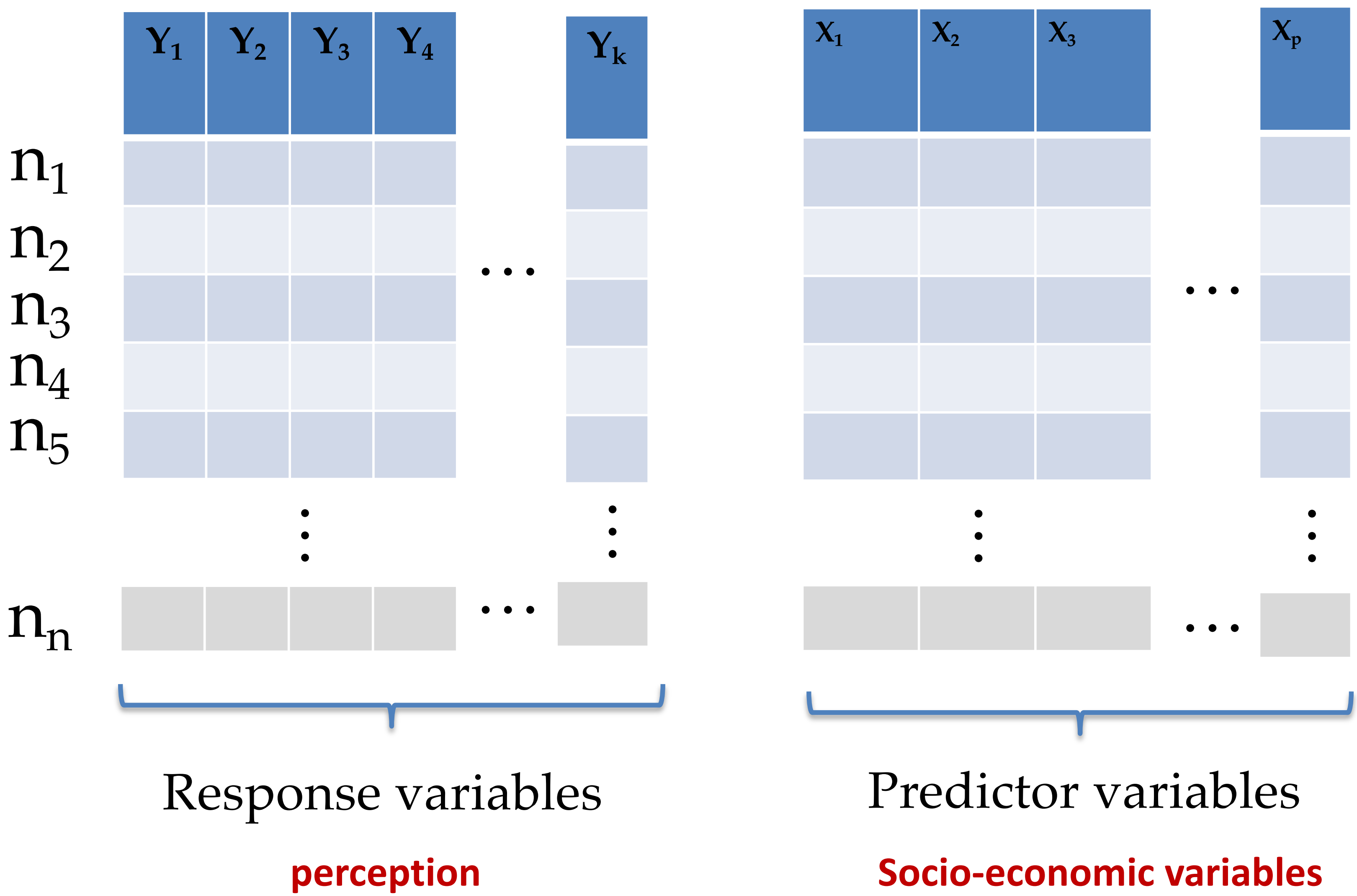
Origin of the owner	
Neo-rural	8
Rural	25
Age	
Between 30 and 40 years	4
Between 40 and 50 years	12
Between 50 and 60 years	9
Between 60 and 70 years	3
Between 70 and 80 years	3
More than 80 years	2
Occupation sector	
Primary sector (farming)	13
Secondary sector (labourer)	8
Tertiary sector	6
Retirees and pensioners	6
Education level	
Primary	6
Secondary-college	23
University	4
Number of children	
None	6
1	1
2	11
3	9
4	4
5	1
6 and more	1
Language spoken	
French	23
English	10
Stage of abandoned farmland	
Shrub dominated	23
Herbaceous	10
Time since acquisition of abandoned farmland	
Less than 10 years	9
10–19 years	12
20–29 years	6
30–39 years	3
More than 50 years	3

Table 1 (*Continued*)

Value of buildings	
0–25,000$	1
25,001–50,000$	10
50,001–75,000$	8
75,001–100,000$	2
10,0001–200,000$	6
200,001–300,000$	3
300,001–500,000$	3
Member of UPA	
No	17
Yes	16
Mean value	
Ecocentric	78.5%[a]
Anthropocentric	73.75%[a]
Apathetic	33.75%[a]

In order to establish relationships between the 10 perception criteria of a land use type and the ***socio-economic variables*** of the owners, canonical redundancy analyses (RDA) were done for each land use using the Canoco programme (ter Braak and Smilauer, 2002).

The usual data format for Redundancy Analysis



Step 1 – estimated predictive values

General multiple regression equation

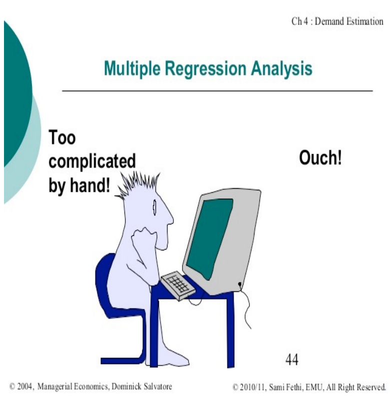
$$Y = b_o + b_1 X_1 + b_2 X_2 + b_3 X_3 ... + b_p X_p$$

Estimating slopes for all predictors

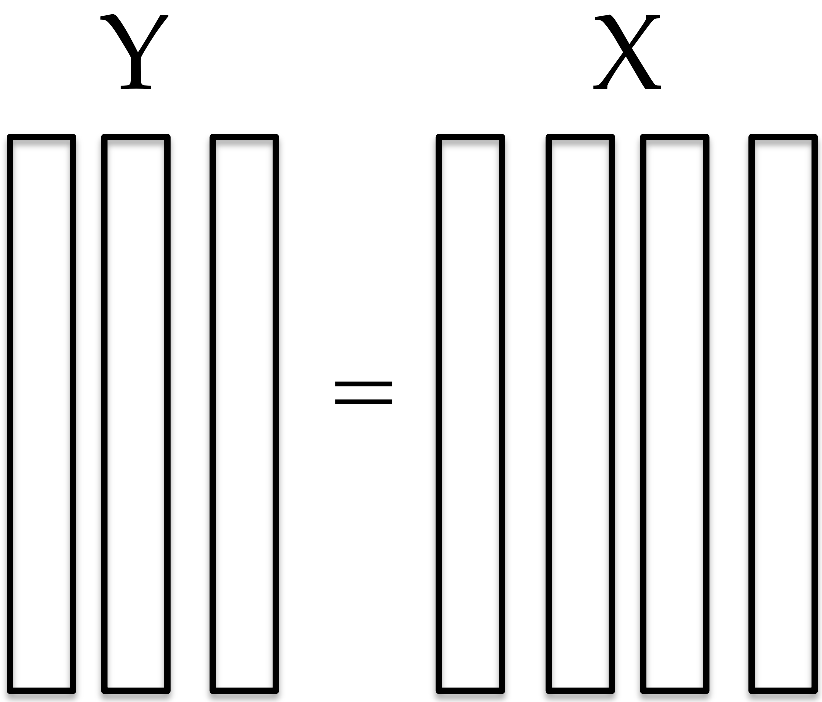
$$b = (\mathrm{X}^{\mathrm{T}}\mathrm{X})^{-1}\mathrm{X}^{\mathrm{T}}\mathrm{Y}$$

Estimating predicted values

$$\hat{\mathrm{Y}} = \mathrm{X}(\mathrm{X}^{\mathrm{T}}\mathrm{X})^{-1}\mathrm{X}^{\mathrm{T}}\mathrm{Y}$$



Step 1 – estimated predictive values

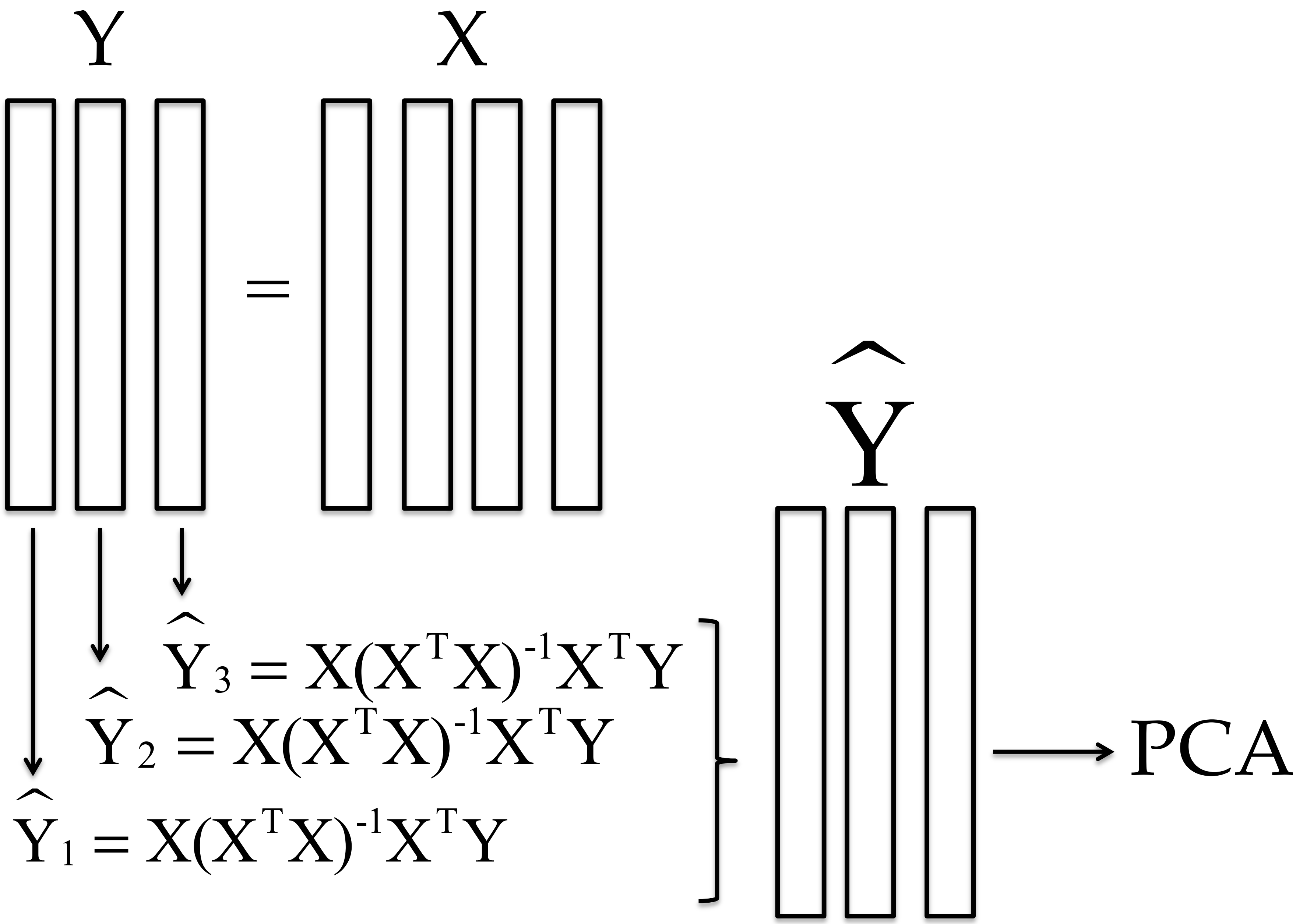


$$\hat{Y}_3 = X(X^TX)^{-1}X^TY$$

$$\hat{Y}_2 = X(X^TX)^{-1}X^TY$$

$$\hat{Y}_1 = X(X^TX)^{-1}X^TY$$

Step 2 – PCA on predictive values

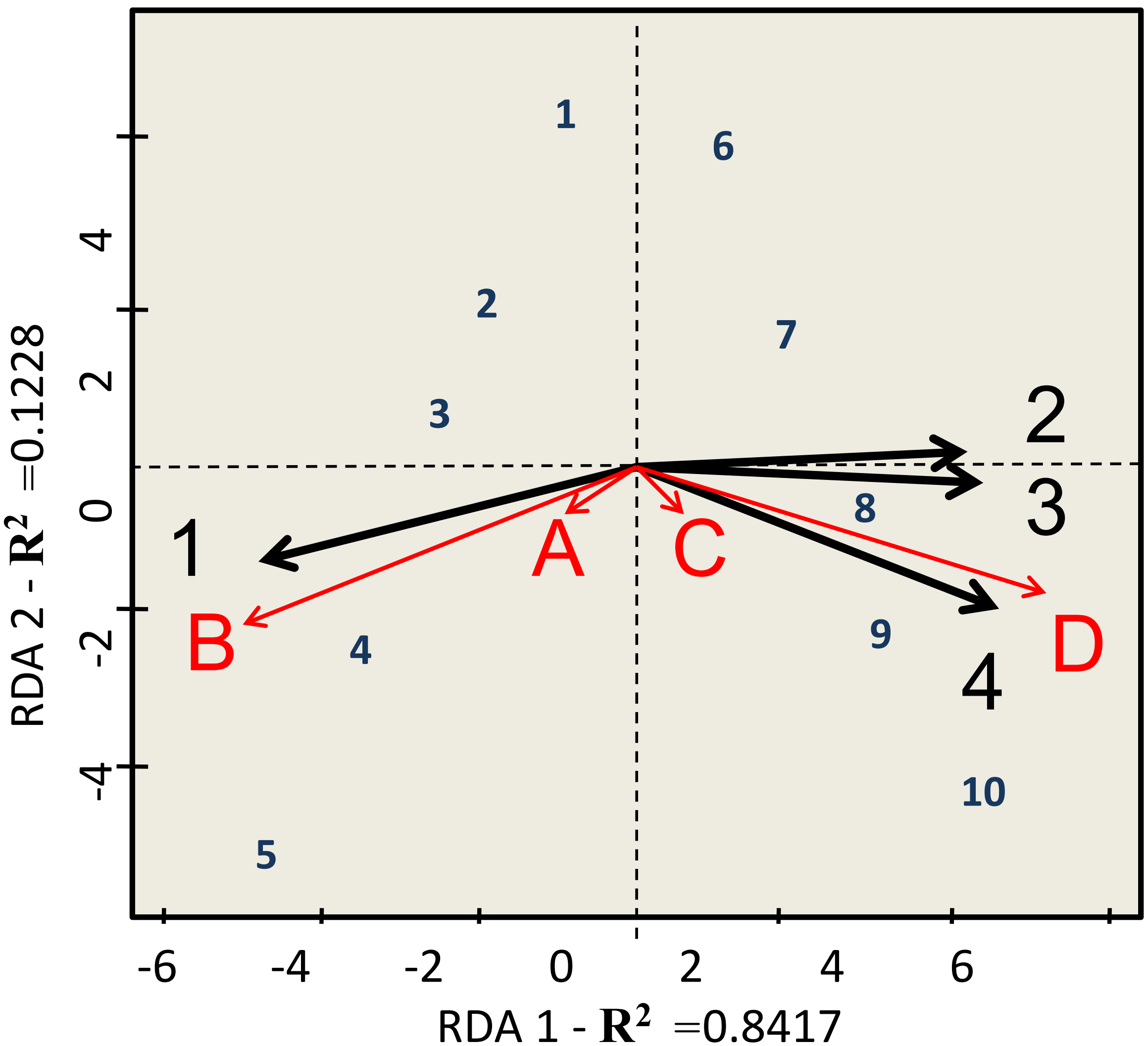


Fictional example (easy to understand)

What kinds of patterns do you observe?

	Y (species densities)				X (environmental predictors)			
sites	A	B	C	D	1	2	3	4
1	1.2	10.4	0	0	7.34	0.17	0.63	53.73
2	2.2	20.6	0	0	7.31	0.09	0.37	49.75
3	3.4	30.1	0	0	10.82	0.18	0.66	54.35
4	4.3	41.3	0	0	9.73	0.05	0.59	37.83
5	5.1	52.1	0	0	15.66	0.04	0.59	47.23
6	0	0	1.3	11.4	0.36	1.33	2.25	62.09
7	0	0	2.1	22.6	0.07	3.06	3.54	72.83
8	0	0	3.5	31.4	0.56	3.36	5.60	91.93
9	0	0	4.1	39.8	0.05	1.54	6.42	90.03
10	0	0	5.2	49.1	0.25	2.05	8.75	72.03

Fictional example
RDA biplot – PCA on predicted values



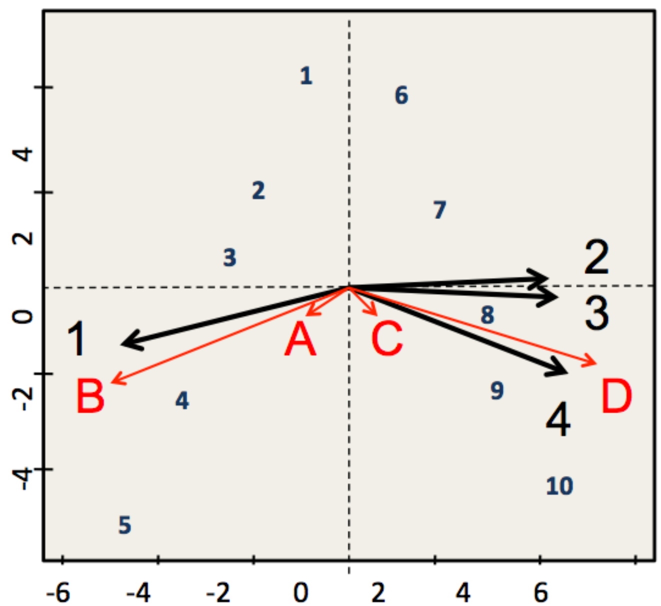
Variation	
Total	753.82
Constrained	727.08
Unconstrained	26.73

R^2 (total): 0.9645

727.08/753.82= 0.9645

(P<0.001) – permutation based test – more in the tutorial

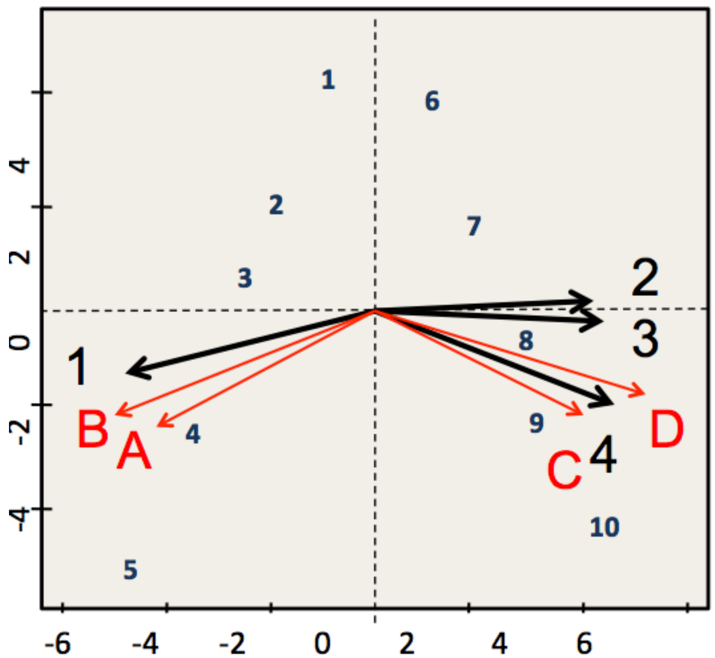
Fictional example (for ease of understanding)
RDA biplot – PCA on predicted values



	Y (species densities)				X (environmental predictors)			
sites	A	B	C	D	1	2	3	4
1	1.2	10.4	0	0	7.34	0.17	0.63	53.73
2	2.2	20.6	0	0	7.31	0.09	0.37	49.75
3	3.4	30.1	0	0	10.82	0.18	0.66	54.35
4	4.3	41.3	0	0	9.73	0.05	0.59	37.83
5	5.1	52.1	0	0	15.66	0.04	0.59	47.23
6	0	0	1.3	11.4	0.36	1.33	2.25	62.09
7	0	0	2.1	22.6	0.07	3.06	3.54	72.83
8	0	0	3.5	31.4	0.56	3.36	5.60	91.93
9	0	0	4.1	39.8	0.05	1.54	6.42	90.03
10	0	0	5.2	49.1	0.25	2.05	8.75	72.03

Response variables were mean-centered (mean = 0), while retaining their original variance (i.e., not standardized to unit variance), prior to running the regression model and calculating predicted values.

Fictional example (for ease of understanding)
RDA biplot – PCA on predicted values



	Y (species densities)				X (environmental predictors)			
sites	A	B	C	D	1	2	3	4
1	1.2	10.4	0	0	7.34	0.17	0.63	53.73
2	2.2	20.6	0	0	7.31	0.09	0.37	49.75
3	3.4	30.1	0	0	10.82	0.18	0.66	54.35
4	4.3	41.3	0	0	9.73	0.05	0.59	37.83
5	5.1	52.1	0	0	15.66	0.04	0.59	47.23
6	0	0	1.3	11.4	0.36	1.33	2.25	62.09
7	0	0	2.1	22.6	0.07	3.06	3.54	72.83
8	0	0	3.5	31.4	0.56	3.36	5.60	91.93
9	0	0	4.1	39.8	0.05	1.54	6.42	90.03
10	0	0	5.2	49.1	0.25	2.05	8.75	72.03

Response variables were standardized (mean=0, variance =1) prior to the regression model and calculating predicted values

Available online at www.sciencedirect.com
ScienceDirect
Estuarine, Coastal and Shelf Science 76 (2008) 45–56
ESTUARINE COASTAL AND SHELF SCIENCE
www.elsevier.com/locate/ecss

Detecting environmental change in estuaries: Nutrient and heavy metal distributions in sediment cores in estuaries from the Gulf of Finland, Baltic Sea

S. Vaalgamaa [a,*], D.J. Conley [b]

Redundancy analysis (RDA) is a multivariate direct gradient analysis method in which variables are presumed to have linear relationships to environmental gradients (i.e., linear species response curves) (Birks, 1995). Only the sediment geo-chemistry from years 1975 to 1998 from each site was used in order to determine potential relationships between the present land use and sediment geochemistry. The correlation structure between sediment geochemistry and catchment and basin variables is summarized as an RDA correlation biplot.

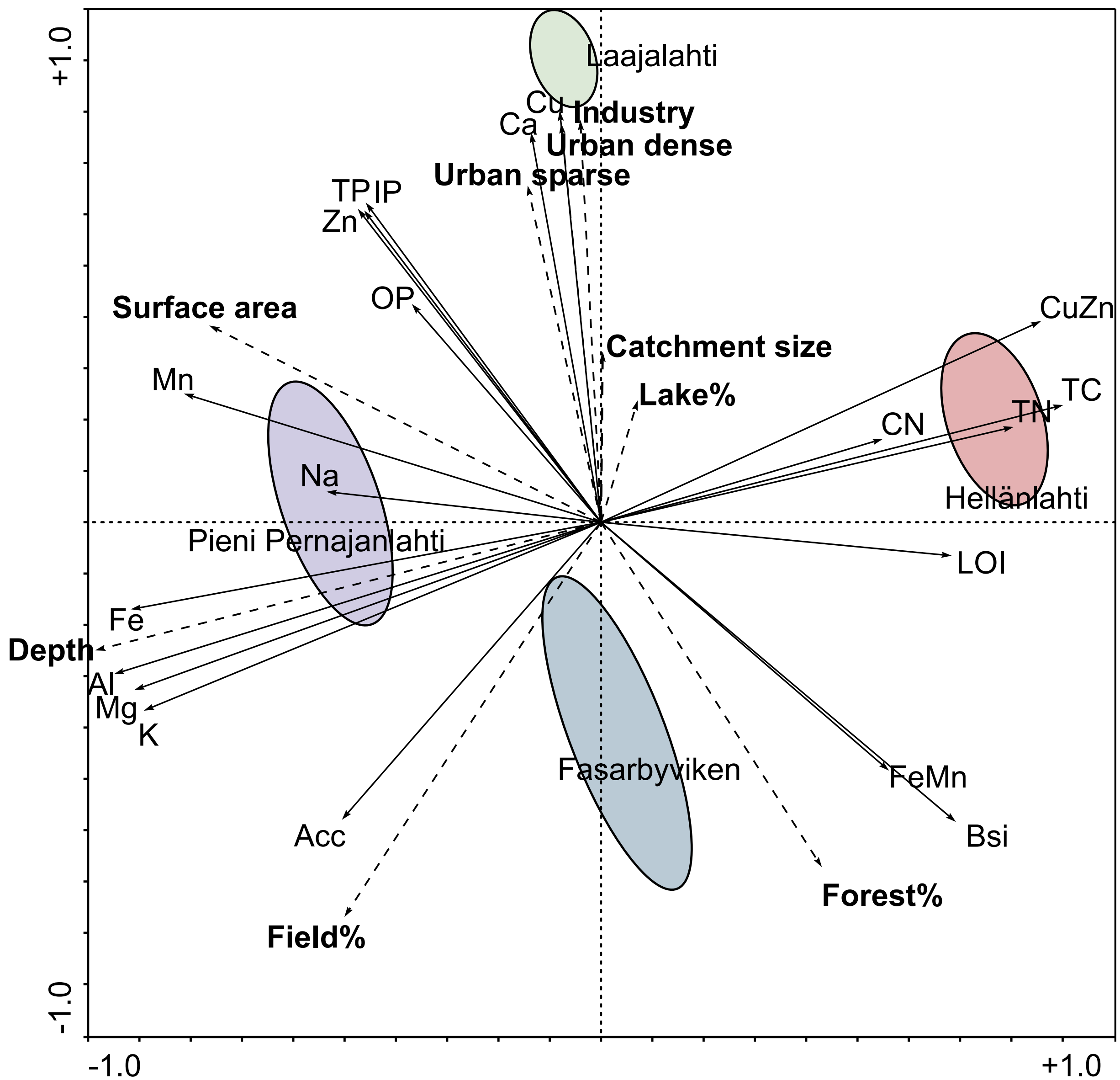
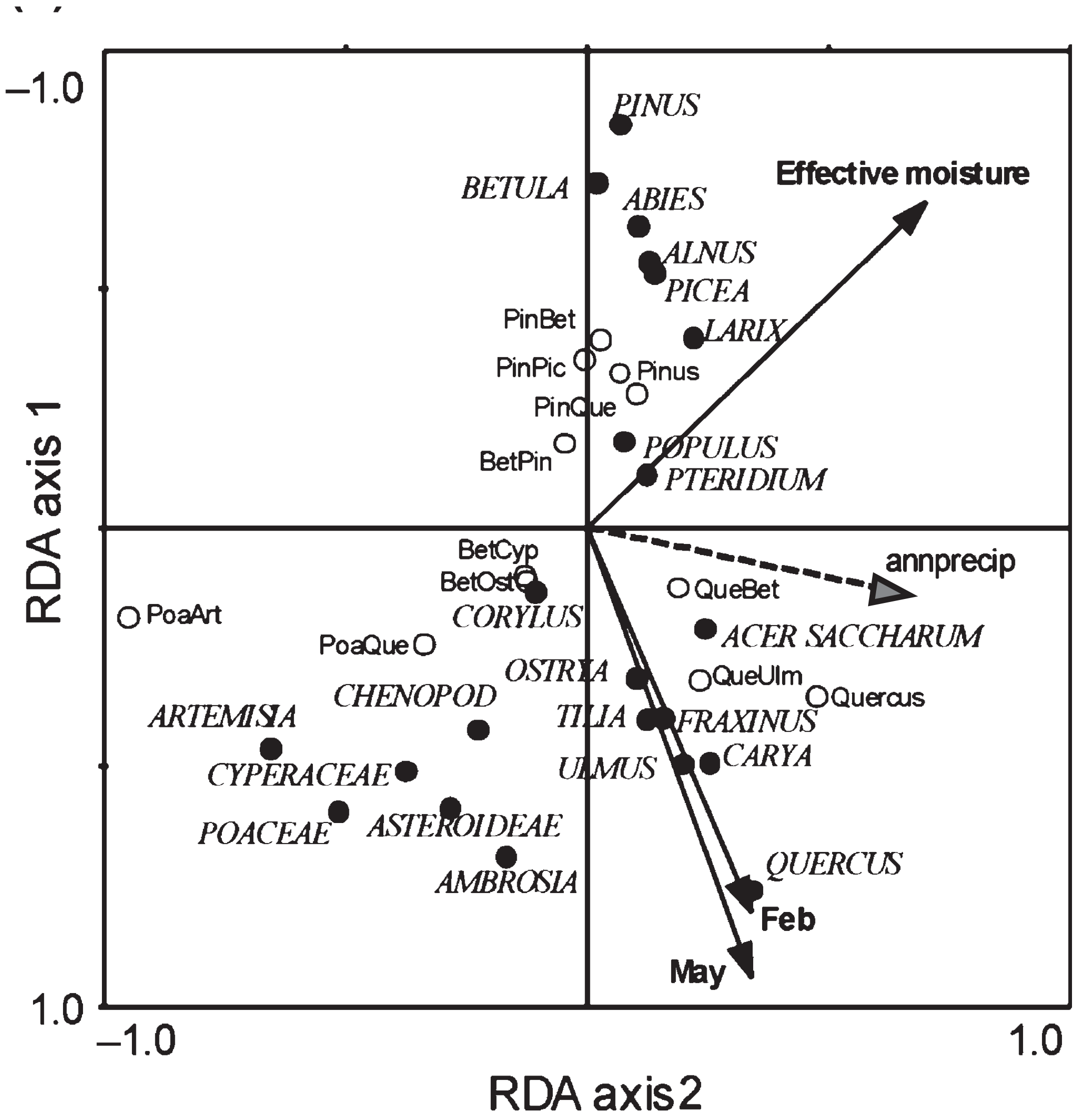


Fig. 6. Redundancy analysis (RDA) ordination diagram showing the relationship between different land-use types and measured sediment variables. Samples from different sites are located in oval-shaped areas.

Redundancy analysis (RDA) is a multivariate direct gradient analysis method in which variables are presumed to have linear relationships to environmental gradients (i.e. linear species response curves) (Birks, 1995). Only the sediment geo- chemistry from years 1975 to 1998 from each site was used in order to determine potential relationships between the present land use and sediment geochemistry. The correlation structure between sediment geochemistry and catchment and basin variables is summarized as an RDA correlation biplot.



Journal of Biogeography (J. Biogeogr.) (2008) **35**, 306–324

ORIGINAL ARTICLE

A pre-European settlement pollen–climate calibration set for Minnesota, USA: developing tools for palaeoclimatic reconstructions

Jeannine-Marie St. Jacques*, Brian F. Cumming and John P. Smol

Two-dimensional redundancy analysis (RDA) ordination diagram of the 1870 pollen and climate data sets showing species (solid circles), climate variables (arrows) and centroids of the 12 biogeographical zones derived from clustering (hollow circles). Analysis includes all 133 pre-Euro-American settlement samples from Minnesota, Iowa, Wisconsin and North and South Dakota. The length of the climate arrows of the RDA ordination plot indicates the importance of that variable in explaining the pollen distributions, whereas the direction of the arrows shows the approximate correlation to the ordination axes. Solid arrows represent forward-selected climate variables and dashed lines represent climate variables that were plotted passively in the ordination.