Multivariate Analysis of Ecological Data

Jan Lepš & Petr Šmilauer

Faculty of Biological Sciences, University of South Bohemia České Budějovice, 1999

Foreword

This textbook provides study materials for the participants of the course named Multivariate Analysis of Ecological Data that we teach at our university for the third year. Material provided here should serve both for the introductory and the advanced versions of the course. We admit that some parts of the text would profit from further polishing, they are quite rough but we hope in further improvement of this text.

We hope that this book provides an easy-to-read supplement for the more exact and detailed publications like the collection of the Dr. Ter Braak' papers and the Canoco for Windows 4.0 manual. In addition to the scope of these publications, this textbook adds information on the classification methods of the multivariate data analysis and introduces some of the modern regression methods most useful in the ecological research.

Wherever we refer to some commercial software products, these are covered by trademarks or registered marks of their respective producers.

This publication is far from being final and this is seen on its quality: some issues appear repeatedly through the book, but we hope this provides, at least, an opportunity to the reader to see the same topic expressed in different words.

Table of contents

1. INTRODUCTION AND DATA MANIPULATION
1.1. Examples of research problems7
1.2. Terminology
1.3. Analyses
1.4. Response (species) data10
1.5. Explanatory variables11
1.6. Handling missing values12
1.7. Importing data from spreadsheets - CanoImp program13
1.8. CANOCO Full format of data files
1.9. CANOCO Condensed format17
1.10. Format line
1.11. Transformation of species data
1.12. Transformation of explanatory variables
2. METHODS OF GRADIENT ANALYSIS
2.1. Techniques of gradient analysis
2.2. Models of species response to environmental gradients
2.3. Estimating species optimum by the weighted averaging method
2.4. Ordinations
2.5. Constrained ordinations
2.6. Coding environmental variables
2.7. Basic techniques
2.8. Ordination diagrams
2.9. Two approaches
2.10. Partial analyses
2.11. Testing the significance of relationships with environmental variables
2.12. Simple example of Monte Carlo permutation test for significance of correlation
3. USING THE CANOCO FOR WINDOWS 4.0 PACKAGE

3.1. Overview of the package Canoco for Windows 4.0	
CANOCO 4.0	
WCanoImp and CanoImp.exe.	
CEDIT	
CanoDraw 3.1	
CanoPost for Windows 1.0	35
3.2. Typical analysis workflow when using Canoco for Windows 4.0	36
3.3. Decide about ordination model: unimodal or linear ?	
3.4. Doing ordination - PCA: centering and standardizing	39
3.5. Doing ordination - DCA: detrending	40
3.6. Doing ordination - scaling of ordination scores	41
3.7. Running CanoDraw 3.1	41
3.8. Adjusting diagrams with CanoPost program	43
3.9. New analyses providing new views of our datasets	43
3.10. Linear discriminant analysis	44
4. DIRECT GRADIENT ANALYSIS AND MONTE-CARLO PERMUT TESTS	
4.1. Linear multiple regression model	46
4.1. Linear multiple regression model4.2. Constrained ordination model	
	47
4.2. Constrained ordination model	47 47
4.2. Constrained ordination model4.3. RDA: constrained PCA	47 47 47
 4.2. Constrained ordination model 4.3. RDA: constrained PCA 4.4. Monte Carlo permutation test: an introduction 	47 47 49 49
 4.2. Constrained ordination model	

	Divisive classifications Analysis of the Tatry samples	
	VISUALIZATION OF MULTIVARIATE DATA WITH CANODRAW 3.1 ID CANOPOST 1.0 FOR WINDOWS	72
6.1.	What can we read from the ordination diagrams: Linear methods	.72
6.2.	What can we read from the ordination diagrams: Unimodal methods	.74
6.3.	Regression models in CanoDraw	.76
6.4.	Ordination Diagnostics	.77
6.5.	T-value biplot interpretation	.78
	CASE STUDY 1: SEPARATING THE EFFECTS OF EXPLANATORY RIABLES	80
7.1.	Introduction	. 80
7.2.	Data	. 80
7.3.	Data analysis	. 80
	CASE STUDY 2: EVALUATION OF EXPERIMENTS IN THE NDOMIZED COMPLETE BLOCKS	84
8.1.	Introduction	.84
8.2.	Data	.84
8.3.	Data analysis	.84
SP OF	CASE STUDY 3: ANALYSIS OF REPEATED OBSERVATIONS OF ECIES COMPOSITION IN A FACTORIAL EXPERIMENT: THE EFFEC FERTILIZATION, MOWING AND DOMINANT REMOVAL IN AN IGOTROPHIC WET MEADOW	
9.1.	Introduction	. 88
9.2.	Experimental design	. 88
9.3.	Sampling	. 89
9.4.	Data analysis	. 89
9.5.	Technical description	.90
9.6.	Further use of ordination results	.93
10. ME	TRICKS AND RULES OF THUMB IN USING ORDINATION	94

10.1	Scaling options
10.2	Permutation tests
10.3	Other issues
11.	MODERN REGRESSION: AN INTRODUCTION
11.1	Regression models in general
11.2	General Linear Model: Terms
11.3	Generalized Linear Models (GLM) 99
11.4	Loess smoother
11.5	Generalized Additive Model (GAM) 101
11.6	Classification and Regression Trees 101
11.7	Modelling species response curves: comparison of models 102
12.	REFERENCES110

1. Introduction and Data Manipulation

1.1. Examples of research problems

Methods of multivariate statistical analysis are no longer limited to exploration of multidimensional data sets. Intricate research hypotheses can be tested, complex experimental designs can be taken into account during the analyses. Following are few examples of research questions where multivariate data analyses were extremely helpful:

• Can we predict loss of nesting locality of endangered wader species based on the current state of the landscape? What landscape components are most important for predicting this process?

The following diagram presents the results of a statistical analysis that addressed this question:



Figure 1-1 Ordination diagram displaying the first two axes of a redundancy analysis for the data on the waders nesting preferences

The diagram indicates that three of the studied bird species decreased their nesting frequency in the landscape with higher percentage of meadows, while the fourth one (*Gallinago gallinago*) retreated in the landscape with recently low percentage of the area covered by the wetlands. Nevertheless, when we tested the significance of the indicated relations, none of them turned out to be significant.

In this example, we were looking on the dependency of (semi-)quantitative response variables (the extent of retreat of particular bird species) upon the percentage cover of the individual landscape components. The ordination method provides here an extension of the regression analysis where we model response of several variables at the same time.

• How do individual plant species respond to the addition of phosphorus and/or exclusion of AM symbiosis? Does the community response suggest an interaction effect between the two factors?

This kind of question used to be approached using one or another form of analysis of variance (ANOVA). Its multivariate extension allows us to address similar problems, but looking at more than one response variable at the same time. Correlations between the plant species occurrences are accounted for in the analysis output.



Figure 1-2 Ordination diagram displaying the first two ordination axes of a redundancy analysis summarizing effects of the fungicide and of the phosphate application on a grassland plant community.

This ordination diagram indicates that many forbs decreased their biomass when either the fungicide (*Benomyl*) or the phosphorus source were applied. The yarrow (*Achillea millefolium*) seems to profit from the fungicide application, while the grasses seem to respond negatively to the same treatment. This time, the effects displayed in the diagram are supported by a statistical test which suggests rejection of the null hypothesis at a significance level $\alpha = 0.05$.

1.2. Terminology

The terminology for multivariate statistical methods is quite complicated, so we must spend some time with it. There are at least two different terminological sets. One, more general and more abstract, contains purely statistical terms applicable across the whole field of science. In this section, we give the terms from this set in italics, mostly in the parentheses. The other set represents a mixture of terms used in the ecological statistics with the most typical examples from the field of community ecology. This is the set we will focus on, using the former one just to be able to refer to the more general statistical theory. This is also the set adopted by the CANOCO program. In all the cases, we have a dataset with the **primary data**. This dataset contains records on a collection of observations - **samples** $(sampling units)^{+}$. Each sample collects values for multiple **species** or, less often, **environmental variables** (variables). The primary data can be represented by a rectangular matrix, where the rows typically represent individual samples and the columns represent individual variables (species, chemical or physical properties of the water or soil, etc).

Very often is our primary data set (containing the *response variables*) accompanied by another data set containing the *explanatory variables*. If our primary data represents a community composition, then the explanatory data set typically contains measurements of the soil properties, a semi-quantitative scoring of the human impact etc. When we use the explanatory variables in a model to predict the primary data (like the community composition), we might divide them into two different groups. The first group is called, somehow inappropriately, the **environmental variables** and refers to the variables which are of the prime interest in our particular analysis. The other group represents the so-called **covariables** (often refered to as *covariates* in other statistical approaches) which are also explanatory variables with an acknowledged (or, at least, hypothesized) influence over the *response variables*. But we want to account for (or subtract or partial-out) such an influence **before** focusing on the influence of the variables of prime interest.

As an example, let us imagine situation where we study effects of soil properties and type of management (hay-cutting or pasturing) on the plant species composition of meadows in a particular area. In one analysis, we might be interested in the effect of soil properties, paying no attention to the management regime. In this analysis, we use the grassland composition as the species data (i.e. primary data set, with individual plant species acting as individual response variables) and the measured soil properties as the environmental variables (explanatory variables). Based on the results, we can make conclusions about the preferences of individual plant species' populations in respect to particular environmental gradients which are described (more or less appropriately) by the measured soil properties. Similarly, we can ask, how the management style influences plant composition. In this case, the variables describing the management regime act as the environmental variables. Naturally, we might expect that the management also influences the soil properties and this is probably one of the ways the management acts upon the community composition. Based on that expectation, we might ask about the influence of the management regime **beyond** that mediated through the changes of soil properties. To address such question, we use the variables describing the management regime as the environmental variables and the measured properties of soil as the covariables.

One of the keys to understanding the terminology used by the CANOCO program is to realize that the data refered to by CANOCO as the **species data** might, in fact, be any kind of the data with variables whose values we want to **predict**. So, if we would like, for example, predict the contents of various metal ions in river water, based on the landscape composition in the catchment area, then the individual ions' concentrations would represent the individual "species" in the CANOCO terminology. If the **species data** really represent the species composition of a community, then we usually apply various abundance measures, including counts,

⁺ There is an inconsistency in the terminology: in classical statistical terminology, **sample** means a collection of sampling units, usually selected at random from the population. In the community ecology, sample is usually used for a description of a sampling unit. This usage will be followed in this text. The general statistical packages use the term **case** with the same meaning.

frequency estimates and biomass estimates. Alternatively, we might have information only on the presence or the absence of the species in individual samples. Also among the explanatory variables (I use this term as covering both the environmental variables and covariables in CANOCO terminology), we might have the quantitative and the presence-absence variables. These various kinds of data values are treated in more detail later in this chapter.

1.3. Analyses

If we try to model one or more response variables, the appropriate statistical modeling methodology depends on whether we model each of the response variables separately and whether we have any explanatory variables (predictors) available when building the model.

The following table summarizes the most important statistical methodologies used in the different situations:

Response	Predictor(s)				
variable	Absent	Present			
is one	• distribution summary	• regression models s.l.			
are many	 indirect gradient analysis (PCA, DCA, NMDS) cluster analysis 	 direct gradient analysis constrained cluster analysis discriminant analysis (CVA) 			

Table 1-1 The types of the statistical models

If we look just on a single response variable and there are no predictors available, then we can hardly do more than summarize the distributional properties of that variable. In the case of the multivariate data, we might use either the ordination approach represented by the methods of **indirect gradient analysis** (most prominent are the principal components analysis - PCA, detrended correspondence analysis - DCA, and non-metric multidimensional scaling - NMDS) or we can try to (hierarchically) divide our set of samples into compact distinct groups (methods of the cluster analysis s.l., see the chapter 5).

If we have one or more predictors available and we model the expected values of a single response variable, then we use the **regression models** in the broad sense, i.e. including both the traditional regression methods and the methods of analysis of variance (ANOVA) and analysis of covariance (ANOCOV). This group of method is unified under the so-called **general linear model** and was recently further extended and enhanced by the methodology of **generalized linear models** (**GLM**) and **generalized additive models** (**GAM**). Further information on these models is provided in the chapter 11.

1.4. Response (species) data

Our primary data (often called, based on the most typical context of the biological community data, the **species data**) can be often measured in a quite precise (quantitative) way. Examples are the dry weight of the above-ground biomass of plant species, counts of specimens of individual insect species falling into soil traps or the percentage cover of individual vegetation types in a particular landscape. We

can compare different values not only by using the "greater-than", "less-than" or "equal to" expressions, but also using their ratios ("this value is two times higher than the other one").

In other cases, we estimate the values for the primary data on a simple, semiquantitative scale. Good example here are the various semi-quantitative scales used in recording composition of plant comunities (e.g. original Braun-Blanquet scale or its various modifications). The simplest variant of such estimates is the presenceabsence (0-1) data.

If we study influence of various factors on the chemical or physical environment (quantified for example by concentrations of various ions or more complicated compounds in the water, soil acidity, water temperature etc), then we usually get quantitative estimates, with an additional constraint: these characteristics do not share the same units of measurement. This fact precludes use of the unimodal ordination methods and dictates the way the variable are standardized if used with the linear ordination methods.

1.5. Explanatory variables

The explanatory variables (also called predictors) represent the knowledge we have and which we can use to predict the values of tje response variables in a particular situation. For example, we might try to predict composition of a plant community based on the soil properties and the type of management. Note that usually the primary task is not the prediction itself. We try to use the "prediction rules" (deduced from the ordination diagrams in the case of the ordination methods) to learn more about the studied organisms or systems.

Predictors can be quantitative variables (like concentration of nitrate ions in soil), semiquantitative estimates (like the degree of human influence estimated on a 0 - 3 scale) or factors (categorial variables).

The factors are the natural way of expressing classification of our samples / subjects - we can have classes of management type for meadows, type of stream for a study of pollution impact on rivers or an indicator of presence or absence of settlement in the proximity. When using factors in the CANOCO program, we must recode them into so-called **dummy variables**, sometimes also called the **indicator variables**. There is one separate variable per each level (different value) of the factor. If a particular sample (observation) has certain value of the factor, there is value 1.0 in the corresponding dummy variable. All the other dummy variables comprising the factor have value of 0.0. For example, we might record for each our sample of grassland vegetation whether this is a pasture, a meadow or an abandoned grassland. We need three dummy variables for recording such factor and their respective values, for a meadow are 0.0, 1.0, and 0.0.

Additionally, this explicit decomposition of factors into dummy variables allows us to create so-called **fuzzy coding**. Using our previous example, we might include into our dataset site which was used as a hay-cut meadow until the last year, but it was used as a pasture this year. We can reasonably expect that both types of management influenced the present composition of the plant community. Therefore, we would give values larger than 0.0 and less than 1.0 for both first and second dummy variable. The important restriction here is (similarly to the dummy variables coding a normal factor) that the values must sum to a total of 1.0. Unless we can

quantify the relative importance of the two management types acting on this site, our best guess is to use values 0.5, 0.5, and 0.0.

If we build a model where we try to predict values of the response variables ("species data") using the explanatory variables ("environmental data"), we can often encounter a situation where some of the explanatory variables have important influence over the species data yet our attitude towards these variables is different: we do not want to interpret their effect, only take this effect into account when judging effects of the other variables. We call these variables **covariables** (often also **covariates**). A typical example is from a sampling or an experimental design where samples are grouped into logical or physical blocks. The values of response variables for a group of samples might be similar due to their proximity, so we need to model this influence and account for it in our data. The differences in response variables that are due to the membership of samples in different blocks must be extracted ("partialled-out") from the model.

But, in fact, almost any explanatory variable could take the role of a covariable - for example in a project where the effect of management type on butterfly community composition is studied, we might have the localities placed at different altitudes. The altitude might have an important influence over the butterfly communities, but in this situation we are primarily focused on the management effects. If we remove the effect of the altitude, we might get a much more clear picture of the influence the management regime has over the butterflies populations.

1.6. Handling missing values

Whatever precaution we take, we are often not able to collect all the data values we need. A soil sample sent to a regional lab gets lost, we forget to fill-in particular slot in our data collection sheet, etc.

Most often, we cannot get back and fill-in the empty slots, usually because the subjects we study change in time. We can attempt to leave those slots empty, but this is often not the best decision. For example, when recording sparse community data (we might have a pool of, say, 300 species, but average number of species per sample is much lower), we use the empty cells in a spreadsheet as absences, i.e. zero values. But the absence of a species is very different from the situation where we simply forgot to look for this species! Some statistical programs provide a notion of missing value (it might be represented as a word "NA", for example), but this is only a notational convenience. The actual statistical method must further deal with the fact there are missing values in the data. There are few options we might consider:

We can remove the samples in which the missing values occur. This works well if the missing values are concentrated into a few samples. If we have, for example, a data set with 30 variables and 500 samples and there are 20 missing values populating only 3 samples, it might be vise to remove these three samples from our data before the analysis. This strategy is often used by the general statistical packages and it is usually called the "case-wise deletion".

If the missing values are, on the other hand, concentrated into a few variables and "we can live without these", we might remove the variables from our dataset. Such a situation often occurrs when we deal with data representing chemical analyses. If "every thinkable" cation type concentration was measured, there is usually a strong correlation between them. If we know values of cadmium concentration in the air deposits, we can usually predict reasonably well the concentration of mercury (although this depends on the type of the pollution source). Strong correlation between these two characteristics then implies that we can usually do reasonably well with only one of these variables. So, if we have a lot of missing values in, say, Cd concentrations, it might be best to drop it from the data.

The two methods of handling missing values described above might seem rather crude, because we lose so much of our results that we often collected at a high expense. Indeed, there are various "imputation methods". The simplest one is to take the average value of the variable (calculated, of course, only from the samples where the value is not missing) and replace the missing values with it. Another, more sophisticated one, is to build a (multiple) regression model, using samples without missing values, for predicting the missing value of the response variable for samples, where the selected predictors' values are not missing. This way, we might fill-in all the holes in our data table, without deleting any sample or variable. Yet, we are deceiving ourselves - we only duplicate the information we have. The degrees of freedom we lost initially cannot be recovered. If we then use such supplemented data with a statistical test, this test has erroneous idea about the number of degrees of freedom (number of independent observations in our data) supporting the conclusion made. Therefore the significance level estimates are not quite correct (they are "overoptimistic"). We can alleviate this problem partially by decreasing statistical weight for the samples where missing values were estimated using one or another method. The calculation is quite simple: in a dataset with 20 variables, a sample with missing values replaced for 5 variables gets weight 0.75 (=1.00 - 5/20). Nevertheless, this solution is not perfect. If we work with only a subset of the variables (like during forward selection of explanatory variables), the samples with any variable being imputed carry the penalty even if the imputed variables are not used, at the end.

1.7. Importing data from spreadsheets - CanoImp program

The preparation of the input data for the multivariate analyses was always the biggest obstacle to their effective use. In the older versions of the CANOCO program, one had to understand to the overly complicated and unforgiving format of the data files which was based on the requirements of the FORTRAN programming language used to create the CANOCO program. The version 4.0 of CANOCO alleviates this problem by two alternative mechanisms. First, there is now a simple format with a minimum requirements as to the file contents. Second, probably more important improvement is the new, easy way to transform data stored in the spreadsheets into the strict CANOCO formats. In this section, we will demonstrate how to use the WCanoImp program, serving for this purpose.

We must start with the data in your spreadsheet program. While the majority of users will use the Microsoft Excel program, the described procedure is applicable to any other spreadsheet program running under Microsoft Windows. If the data are stored in a relational database (Oracle, FoxBase, Access, etc.) we can use the facilities of our spreadsheet program to first import the data there. In the spreadsheet, we must arrange our data into rectangular structure, as laid out by the spreadsheet grid. In the default layout, the individual samples correspond to the rows while the individual spreadsheet columns represent the variables. In addition, we have a simple heading for both rows and columns: the first row (except the empty upper left corner) contains names of variables, while the first column contains names of the individual samples. Use of heading(s) is optional, WCanoImp program is able to generate simple names there. If using the heading row and/or column, we must observe limitation imposed by the CANOCO program. The names cannot have more than eight characters and also the character set is somewhat limited: the most safe strategy is to use only the basic English letters, digits, hyphen and space. Nevertheless, WCanoImp replaces prohibited characters by a dot and also shortens names longer than the eight character positions. But we can lose uniqueness (and interpretability) of our names in such a case, so it's better to take this limitation into account from the very beginning.

In the remaining cells of the spreadsheet must be only the numbers (whole or decimal) or they must be empty. No coding using other kind of characters is allowed. Qualitative variables ("factors") must be coded for CANOCO program using a set of "dummy variables" - see the section 2.6 for more details.

After we have our data matrix ready in the spreadsheet program, we select this rectangular matrix (e.g. using the mouse pointer) and copy its contents to the Windows Clipboard. WCanoImp takes this data from the Clipboard, determines its properties (range of values, number of decimal digits etc) and allows us to create new data file containing these values but conforming to the one of two CANOCO data file formats. It is now hopefully clear that the above-described requirements concerning format of the data in spreadsheet program apply only to the rectangle being copied to the Clipboard. Outside of it, we can place whatever values, graphs or objects we like.

After the data were placed on the Clipboard or even a long time before that moment, we must start the WCanoImp program. It is accessible from the Canoco for Windows program menu (**Start/Programs**/[*Canoco for Windows folder*]). This import utility has easy user interface represented chiefly by one dialog box, displayed below:

Bg WCanoImp
HOW TO USE THIS PROGRAM
1) In your spreadsheet:
* Copy your data table to the Clipboard
* any labels must be in Row 1 / Column 1
2) Confirm the options below and Save
 OPTIONS Each column is a Sample Generate labels for: □ Samples (Samp0001 Samp0002 etc.) □ Species / Env. Variables (Var0001 Var0002 etc.)
Save in Condensed Format
<u>Save</u> Exit <u>H</u> elp

Figure 1-3 The main window of the WCanoImp program.

The upper part of the dialog box contains a short version of the instructions provided here. As we already have the data on the Clipboard, we must now look at the WCanoImp options to check if they are appropriate for our situation. The first option (**Each column is a Sample**) applies only if we have our matrix transposed in respect to the form described above. This might be useful if we do not have many samples (as for example MS Excel limits number of columns to 256) but we have a high number of variables. If we do not have names of samples in the first column, we must check the second checkbox (i.e. ask to **Generate labels for:** ... **Samples**), similarly we check the third checkbox if the first row in the selected spreadsheet rectangle corresponds to the values in the first sample, not to the names of the variables. Last checkbox (**Save in Condensed Format**) governs the actual format used when creating data file. Unless we worry too much about the hard disc space, it does not matter what we select here (the results of the statistical methods should be identical, whatever format we choose here).

After we made sure the selected options are correct, we can proceed by clicking the **Save** button. We must first specify the name of the file to be generated and the place (disc letter and directory) where it will be stored. WCanoImp then requests a simple description (one line of ASCII text) for the dataset being generated. This one line appears then in the analysis output and remind us what kind of data we were using. A default text is suggested in the case we do not care about this feature. WCanoImp then writes the file and informs us about the successfull creation with a simple dialog box.

1.8. CANOCO Full format of data files

The previous section demonstrated how simple is to create CANOCO data files from our spreadsheet data. In an ideal world, we would never care what the data files created by the WCanoImp program contain. Sadly, CANOCO users often do not live in that ideal world. Sometimes we cannot use the spreadsheet and therefore we need to create data files without the WCanoImp assistance. This happens, for example, if we have more than 255 species and 255 samples at the same time. In such situation, the simple methodology described above is insufficient. If we can create the TABseparated values format file, we can use the command-line version of the WCanoImp program, named CanoImp, which is able to process data with substantially higher number of columns than 255. In fact, even the WCanoImp program is able to work with more columns, so if you have a spreadsheet program supporting a higher number of columns, you can stay in the realm of the more user-friendly Windows program interface (e.g. Quattro for Windows program used to allow higher number of columns than Microsoft Excel).

Yet in other cases, we must either write the CANOCO data files "in hand" or we need to write programs converting between some customary format and the CANOCO formats. Therefore, we need to have an idea of the rules governing contents of these data files. We start first with the specification of the so-called **full format**.

WCanoImp produced data	file						
(I5,1X,21F3.0)							
21							
1 0	1 0 0	0 0 0	0 0	0 0 0	0 0 0	0 0	
2 1 0 0 1 0	0 1 0	0 0 0	0 0	0 0 0	0 0 0	0 0	
3 0 1 0 1 0	0 0 1	0 0 0	0 0	0 0 0	0 0 0	0 0	
•••							
48 1 1 0 0 1	0 0 0	0 0 0	0 0	0 0 0	0 0 0	0 1	
0 0 0 0 0 0	0 0 0	0 0 0	0 0	0 0 0	0 0 0	0 0	
PhosphatBenlate Year94	Year95	Year98	B01	B02	в03	в04	B05
B06 B07 B08	B09	B10	B11	B12	B13	B14	B15
B16							
PD01 PD02 PD03	PD04	PD05	PD06	PD07	PD08	PD09	PD10
PD11 PD12 PD13	PD14	PD15	PD16	C01	C02	C03	C04
•••							

Figure 1-4 Part of a CANOCO data file in the full format. The hyphens in the first data line show the presence of the space characters and should not be present in the actual file

The first three lines in the CANOCO data files have a similar meaning for both the full and condensed formats. The first line contains a short textual description of the data file, with the maximum length of 80 characters. Second line contains the exact description of the format for the data values that occur in the file, starting from the fourth line. The format line is described in more detail in the section 1.10. The third line contains a single number, but its meaning differs between full and condensed formats. In the full format, it gives the total number of variables in the data matrix.

Generally, a file in the full format displays the whole data matrix, including the zero values as well. Therefore, it is more simple to understand when we look at it, but it is much more tedious to create, given that majority of the values for community data will be zeros.

In full format, each sample is represented by a fixed number of lines - one line per sample is used in the above example. There we have 21 variables. First sample (on the fourth row) starts with its number (1) followed by another 21 values. We note that number of spaces between the values is identical for all the rows, the data fields are well aligned on their right margins. Each field takes a specified number of positions ("columns") as specified in the format line. If the number of variables we have would not fit into one line (which should be shorter than 127 columns), we can use additional lines per sample. This is then indicated in the format description in the format line by the slash character. The last sample in the data is followed by a "dummy" sample, identified by its number being zero.

Then the names ("labels") for variables follow, which have very strict format: each name takes exactly eight positions (left-padded or right-padded with spaces, as necessary) and there are exactly 10 names per row (except the last row which may not be completely filled). Note that the required number of entries can be calculated from the number of variables, given at the third row in the condensed format. In our example, there are two completely full rows of labels, followed by a third one, containing only one name.

The names of the samples follow the block with variable names. Here the maximum sample number present in the data file determines necessary number of entries. Even if some indices between 1 and this maximum number are missing, the corresponding positions in the names block must be reserved for them.

We should note that it is not a good idea to use TAB characters in the data file - these are still counted as one column by the CANOCO program reading the data, yet they are visually represented by several spaces in any text editor. Also we should note that if creating the data files "by hand", we should not use any editor inserting format information into the document files (like Microsoft Word or Wordperfect programs). The Notepad utility is the easiest software to use when creating the data files in CANOCO format.

1.9. CANOCO Condensed format

The condensed format is most useful for sparse community data. The file with this format contains only the nonzero entries. Therefore, each value must be introduced by the index specifying to which variable this value belongs.

```
WCanoImp produced data file
(15, 1X, 8(16, F3.0))
8
  --1----23--1----25-10----36--3
                                                53 5
                                                                            85
___
                                      41 4
                                                         57 3
                                                                   70
                                                                       5
б
          89 70
                  100 1
    1
                            102 1
                                     115 2
                                               121
                                                    1
    2
                   26
                       1
                             38 5
                                      42 20
                                                50
                                                   1
                                                         55 30
                                                                   57
                                                                       7
                                                                            58
          11
             1
5
    2
          62 2
                    69 1
                             70 5
                                      74 1
                                                77
                                                   1
                                                         86 7
                                                                   87
                                                                       2
                                                                            89
30
         131 15
   79
    0
{\tt TanaVulgSeneAquaAvenPratLoliMultSalxPurpErioAnguStelPaluSphagnumCarxCaneSalx}
Auri
. . .
SangOffiCalaArunGlycFlui
PRESEK SATLAV CERLK
                       CERJIH CERTOP
                                          CERSEV ROZ13
                                                          ROZ24
                                                                   ROZC5
ROZR10
. . .
```

Figure 1-5 Part of a CANOCO data file in the condensed format. The hyphens in the first data line show the presence of the space characters and should not be present in the actual file

In this format, the number of rows needed to record all values varies from sample to sample. Therefore, each line starts with a sample index and also the format line describes the format of one line only. In the example displayed in the Figure 1-5, the first sample is recorded in two rows and this sample contains eight species. For example, a species with the index 23 has the value 1.0, while a species with the index 25 has value 10. By checking the maximum species index, we can find that there is a total of 131 species in the data. The value in the third line of the file with condensed format does not specify this number, but rather the maximum number of the "variable index"-"variable value" pairs ("couplets") in a single line. The last sample is again followed by a "dummy" sample with zero index. The format of the two blocks with names of variables and samples is identical to that of the full format files.

1.10. Format line

The following example contains all the important parts of a format line specification and refers to a file in the condensed format.

```
(I5,1X,8(I6,F3.0))
```

First, note that the whole format specification must be enclosed in the parentheses. There are three letters used in this example (namely I, F, and X) and generally, these are sufficient for describing any kind of contents a condensed format might have. In the full format, the additional symbol for line-break (new-line) is the slash character (*I*).

The format specifier using letter I is used to refer to indices. These are used for sample numbers in both condensed and full formats and for the species numbers, used only in the condensed format. Therefore, if you count number of I letters in the format specification, you know what format this file has: if there is just a one I, it is a full format file. If there are two or more Is, this is a condensed format file. If there is no I, this is a wrong format specification. But this might also happen for the free format files or if the CANOCO analysis results are used as an input for another analysis (see section 10.2). The I format specifier has the Iw form, where w is followed by a number, giving width of the index field in the data file, reserved for it. This is the number of columns this index value uses. If the number of digits needed to describe the integral value is shorter than this width, the number is right-aligned, padded with space characters on its left side.

The actual data values use the Fw.d format specifiers, i.e. the F letter followed by two numbers, separated with a dot. The first number gives the total width of the data field in the file (number of columns), while the other gives the width of the part after the decimal point (if larger than zero). The values are in the field of specified width right-aligned, padded with the spaces to their left. Therefore, if the format specifier says F5.2, we know that the two rightmost columns contain the first two decimal digits after the decimal point. In the third column from the right side is the decimal point. This leaves up to two columns for the whole part of the value. If we have values larger than 9.99, we would fill up the value field completely, so we would not have any space visually separating this field from the previous one. We can either increase the w part of the F descriptor by one or we can insert a X specifier.

The nX specifier tells us that n columns contain spaces and should be, therefore, skipped. An alternative way how to write it is to revert the position of the width-specifying number and the X letter (Xn).

So we can finally interpret the format line example given above. The first five columns contains the sample number. Remember that this number must be right-aligned, so a sample number 1 must be written as four spaces followed by the digit '1'. Sixth column should contain space character and is skipped by CANOCO while reading the data. The next value preceding included pair of parentheses is a **repeat specifier**, saying that the format described inside the parentheses (species index with a width of six columns followed by a data value taking three columns) is repeated eight times. In the case of the condensed format there might be, in fact, fewer than eight pairs of "species index" - "species value" on a line. Imagine that we have a sample with ten species present. This sample will be represented (using our sample format) on two lines with the first line completely full and the second line containing only two pairs.

As we mentioned in section 1.8, a sample in a full format data file is represented by a fixed number of lines. The format specification on its second line therefore contains description of all the lines forming a single sample. There is only one I field referring to the sample number (this is the I descriptor the format specification starts

with), the remaining descriptors give the positions of individual fields representing the values of all the variables. The slash character is used to specify where CANOCO needs to progress to the next line while reading the data file.

1.11. Transformation of species data

As we show in the Chapter 2, the ordination methods find the axes representing regression predictors, optimal in some sense for predicting the values of the response variables, i.e. the values in the species data. Therefore, the problem of selecting transformation for these variables is rather similar to the one we would have to solve if using any of the species as a response variable in the (multiple) regression method. The one additional restriction is the need to specify an identical data transformation for all the response variables ("species"). In the unimodal (weighted averaging) ordination methods (see the section 2.2), the data values cannot be negative and this imposes further restriction on the outcome of a potential transformation.

This restriction is particularly important in the case of the log transformation. Logarithm of 1.0 is zero and logarithms of values between 0 and 1 are negative values. Therefore, CANOCO provides a flexible log-transformation formula:

$$\mathbf{y}' = \log(\mathbf{A}^*\mathbf{y} + \mathbf{C})$$

We should specify the values of A and C so that after these are applied to our data values (y), the result is always greater or equal to 1.0. The default values of both A and C are equal to 1.0 which maps neatly the zero values again to zeros and other values are positive. Nevertheless, if our original values are small (say, in range 0.0 to 0.1), the shift caused by adding the relatively large value of 1.0 dominates the resulting structure of the data matrix. We adjust the transformation here by increasing the value of A, e.g. to 10.0 in our example. But the default log transformation (i.e. log(y+1)) works well for the percentages data on the 0-100 scale, for example.

The question when to apply a log transformation and when to stay on the original scale is not easy to answer and there are almost as many answers as there are statisticians. Personally, I do not think much about distributional properties, at least not in the sense of comparing frequency histograms of my variables with the "ideal" Gaussian (Normal) distribution. I rather try to work-out whether to stay on the original scale or to log-transform using the semantics of the problem I am trying to address. As stated above, the ordination methods can be viewed as an extension of the multiple regression methods, so let me illustrate this approach in the regression context. Here we might try to predict the abundance of a particular species in a sample based on the values of one or more predictors (environmental variables and/or ordination axes in the context of the ordination methods). Now, we can formulate the question addressed by such a regression model (let us assume just a single predictor variable for simplicity) like "How the average value of the species Y changes with the change of the value of the environmental variable X by one unit?". If neither the response variable nor the predictors are log transformed, our answer can take the form "The value of species Y increases by B if the value of environmental variable X increases by one measurement unit". Of course, B is the regression coefficient of the linear model equation $Y = B_0 + B^*X + E$. But in the other cases we might prefer to see the appropriate style of the answer to be "If value of environmental variable X increases by one, the average abundance of the species increases by ten percent". Alternatively, we can say, "the abundance increases 1.10 times". Here we are thinking on a multiplicative scale, which is not assumed by the linear regression model. In such a situation, I would log transform the response variable.

Similarly, if we tend to speak about an effect of the the environmental variable value change in a multiplicative way, this predictor variable should be log-transformed. As an example, if we would use the concentration of nitrate ions in soil solution as a predictor, we would not like our model to address a question what happens if the concentration increases by 1 mmol/l. In such case, there would be no difference in change from 1 to 2 compared with a change from 20 to 21.

The plant community composition data are often collected on a semiquantitative estimation scale and the Braun-Blanquet scale with seven levels (r, +, 1, 2, 3, 4, 5) is a typical example. Such a scale is often quantified in the spreadsheets using corresponding ordinal levels (from 1 to 7, in this case). Note that this coding already implies a log-like transformation because the actual cover/abundance differences between the successive levels are more or less increasing. An alternative approach to use of such estimates in the data analysis is to replace them by the assumed centers of the corresponding range of percentage cover. But doing so, we find a problem with the r and + levels because these are based more on the abundance (number of individuals) of the species rather than on its estimate cover. Nevertheless, using the very rough replacements like 0.1 for r and 0.5 for + rarely harms the analysis (compared to the alternative solutions).

Another useful transformation available in CANOCO is the square-root transformation. This might be the best transformation to apply to the count data (number of specimens of individual species collected in a soil trap, number of individuals of various ant species passing over a marked "count line", etc.) but the log transformation is doing well with these data, too.

The console version of CANOCO 4.0 provides also the rather general "linear piecewise transformation" which allows us to approximate the more complicated transformation functions using a poly-line with defined coordinates of the "knots". This general transformation is not present in the Windows version of CANOCO, however.

Additionally, if we need any kind of transformation which is not provided by the CANOCO software, we might do it in our spreadsheet software and export the transformed data into the CANOCO format. This is particularly useful in the case our "species data" do not describe community composition but something like the chemical and physical soil properties. In such a case, the variables have different units of measurement and different transformations might be appropriate for different variables.

1.12. Transformation of explanatory variables

Because the explanatory variables ("environmental variables" and "covariables" in CANOCO terminology) are assumed not to have an uniform scale and we need to select an appropriate transformation (including the frequent "no transformation" choice) individually for each such variable. But CANOCO does not provide this feature so any transformations on the explanatory variables must be done before the data is exported into a CANOCO compatible data file.

Nevertheless, after CANOCO reads in the environmental variables and/or covariables, it transforms them all to achieve their zero average and unit variance (this procedure is often called *normalization*).

2. Methods of gradient analysis

Introductory terminological note: The term **gradient analysis** is used here in the broad sense, for any method attempting to relate the species composition to the (measured or hypothetical) environmental gradients. The term **environmental variables** is used (traditionally, as in CANOCO) for any explanatory variables. The quantified species composition (the explained variables) is in concordance with the Central-European tradition called relevé. The term **ordination** is reserved here for a subset of methods of gradient analysis.

Often the methods for the analysis of species composition are divided into gradient analysis (ordination) and classification. Traditionally, the classification methods are connected with the discontinuum (or vegetation unit) approach or sometimes even with the Clemensian organismal approach, whereas the methods of the gradient analysis are connected with the continuum concept, or with the individualistic concept of (plant) communities. Whereas this might (partially) reflect the history of the methods, this distinction is no longer valid. The methods are complementary and their use depends mainly on the purpose of the study. For example, in the vegetation mapping the classification is necessary. Even if there are no distinct boundaries between the adjacent vegetation types, we have to cut the continuum and to create distinct vegetation units for mapping purposes. The ordination methods can help to find repeatable vegetation patterns, discontinuities in the species composition, or to show the transitional types etc. and are now used even in the phytosociological studies.

2.1. Techniques of gradient analysis

The Table 2-1 provides an overview of the problems with try to solve with our data using one or another kind of statistical methods. The categories differ mainly by the type of the information (availability of the explanatory = environmental variables, and of the response variables = species) we have available.

Further, we could add the *partial* ordination and *partial* constrained ordination entries to the table, where we have beside the primary explanatory variables the so-called *covariables* (=covariates). In the partial analyses, we first extract the dependence of the species composition on those covariables and then perform the (constrained) ordination.

The environmental variables and the covariables can be both quantitative and categorial ones.

Data, I	have	Apriori				
no. of envir. var	no. of species	knowledge of species- environment relationships	I will use	I will get		
1, <i>n</i>	1	NO	Regression	Dependence of the species on environment		
none	п	YES	Calibration	Estimates of environmental values		
none	п	NO	Ordination	Axes of variability in species composition (can be – should be - aposteriori related to measured environmental variables, if available) Giechoms P1 P2 .P3 P4 . Cirsium Rubus		
1, <i>n</i>	п	NO	Constrained ordination	Variability in species composition explained by environmental variables Relationship of environmental variables to species axes Giechoma P1 P2 P3 P4 Cirsium PH		

Table 2-1

2.2. Models of species response to environmental gradients

Two types of the model of the species response to an environmental gradient are used: the model of a linear response and of an unimodal response. The linear response is the simplest approximation, the unimodal response expects that the species has an optimum on an environmental gradient.



Figure 2-1 Linear approximation of an unimodal response curve over a short part of the gradient

Over a short gradient, a linear approximation of any function (including the unimodal one) works well (Figure 2-1).



Environmental variable (e.g.,pH, elevation)

Figure 2-2 Linear approximation of an unimodal response curve over a long part of the gradient

Over a long gradient, the approximation by the linear function is poor (Figure 2-2). It should be noted that even the unimodal response is a simplification: in reality, the response is seldom symmetric, and also more complicated response shapes are found (e.g. bimodal ones).

2.3. Estimating species optimum by the weighted averaging method

Linear response is usually fitted by the classical methods of the (least squares) regression. For the unimodal response model, the simplest way to estimate the species optimum is by calculating the weighted average of the environmental values where the species is found. The species importance values (abundances) are used as weights in calculating the average:

$$WA = \frac{\sum Env \times Abund}{\sum Abund}$$

where Env is the environmental value, and *Abund* is abundance of the species in the corresponding sample.⁺ The method of the weighted averaging is reasonably good when the whole range of a species distribution is covered by the samples (Figure 2-3).



Figure 2-3 Example of the range where the complete response curve is covered

Complete range covered:					
Environmental value Species pro					
abundance					
0.1	0				
0.5	10				
2.0	80				
4.2	252				
2.0	160				
0.5	50				
0.1	12				
9.4	564				
	Species abundance 0.1 0.5 2.0 4.2 2.0 0.5 0.1				

$$WA = \frac{\sum Env \times Abund}{\sum Abund} = 564 / 9.4 = 60$$

On the contrary, when only part of the range is covered, the estimate is biased:

Only part of the range covered:

Environmental. value	Species abundance	product	
	abundance		
60	4.2	252	
80	2.0	160	
100	0.5	50	
120	0.1	12	
Total	6.8	472	

$$WA = \frac{\sum Env \times Abund}{\sum Abund} = 472/6.8 = 69.4$$

The longer the axis, the more species will have their optima estimated correctly.

⁺ Another possibility is to estimate directly the parameters of the unimodal curve, but this option is more complicated and not suitable for the simultaneous calculations that are usually used in the ordination methods.

The techniques based on the linear response model are suitable for homogeneous data sets, the weighted averaging techniques are suitable for more heterogeneous data.

2.4. Ordinations

The problem of an unconstrained ordination can be formulated in several ways:

1. Find the configuration of samples in the ordination space so that the distances of samples in this space correspond best to the dissimilarities of their species composition. This is explicitly done by the non-metric multidimensional scaling (NMDS).

2. Find "latent" variable(s) (= ordination axes), for which the total fit of dependence of all the species will be the best. This approach requires the model of species response to the variables to be explicitly specified: linear response for linear methods, unimodal response for weighted averaging (the explicit "Gaussian ordinations" are not commonly used for computational problems). In linear methods, the sample score is a linear combination (weighted sum) of the species scores. In the weighted averaging methods the sample score is a weighted average of the species scores (after some rescaling).

Note: the weighted averaging contains implicit standardization by both samples and species. On the contrary, for the linear methods, we can select standardized and non-standardized versions.

3. Let us consider the samples to be points in a multidimensional space, where species are the axes and position of each sample is given by the corresponding species abundance. Then the goal of ordination is to find a projection of the multidimensional space into a space with reduced dimensionality that will result in minimum distortion of the spatial relationships. Note that the result is dependent on how we define the "minimum distortion".

It should be noted that the various formulations could lead to the same solution. For example, the principal component analysis can be formulated in any of the above manners.

2.5. Constrained ordinations

The constrained ordinations can be best explained within the framework of the ordinations defined as a search for the best explanatory variables (i.e. the problem formulation 2 in the previous paragraph). Whereas in the unconstrained ordinations we search for any variable that explains best the species composition (and this variable is taken as the ordination axis), in the constrained ordinations the ordination axes are weighted sums of environmental variables. Consequently, the less environmental variables we have, the stricter is the constraint. If the number of environmental variables is greater than the number of samples minus 1, then the ordination is unconstrained.

The unconstrained ordination axes correspond to the directions of the greatest variability within the data set. The constrained ordination axes correspond to the directions of the greatest variability of the data set that can be explained by the environmental variables. The number of constrained axes cannot be higher than the number of environmental variables.

2.6. Coding environmental variables

The environmental variables can be either quantitative (pH, elevation, humidity) or qualitative (categorial or categorical). The categorial variables with more than two categories are coded as several *dummy variables*; the dummy variable' values equal either one or zero. Suppose we have five plots, plots 1 and 2 being on limestone, plots 3 and 4 on granite and plot 5 on basalt. The bedrock will be characterized by three environmental variables (limestone, granite, basalt) as follows:

	limestone	granite	basalt
Plot 1	1	0	0
Plot 2	1	0	0
Plot 3	0	1	0
Plot 4	0	1	0
Plot 5	0	0	1

The variable basalt is not necessary, as it is a linear combination of the previous two: basalt = 1 - limestone - granite. However, it is useful to use this category for further graphing.

2.7. Basic techniques

Four basic ordination techniques exist, based on the underlying species response model and whether the ordination is constrained or unconstrained (Ter Braak & Prentice, 1998):

	Linear methods	Weighted averaging		
unconstrained	Principal Components	Correspondence Analysis		
unconstrained	Analysis (PCA)	(CA)		
	Redundancy Analysis	Canonical		
constrained	•••	Correspondence Analysis		
	(RDA)	(CCA)		

Table 2-2

For the weighted averaging methods, the detrended versions exist (i.e. Detrended Correspondence Analysis, DCA, the famous DECORANA, and Detrended Canonical Correspondence Analysis, DCCA, see section 3.5). For all the methods, the partial analyses exist. In partial analyses, the effect of covariables is first partialled out and the analysis is then performed on the remaining variability.

2.8. Ordination diagrams

The results of an ordination are usually displayed as the ordination diagrams. Plots (samples) are displayed by points (symbols) in all the methods. Species are shown by the arrows in the linear methods (the direction, in which the species abundance

increases) and by the points (symbols) in the weighted averaging methods (the species optimum). The quantitative environmental variables are shown by arrows (direction, in which the value of environmental variable increases). For qualitative environmental variables, the centroids are shown for individual categories (the centroid of the plots, where the category is present).



Figure 2-4: Examples of typical ordination diagrams. Analyses of data on the representation of *Ficus* species in forests of varying successional age in Papua New Guinea. The species are labeled as follows: *F. bernaysii* - BER, *F. botryocarpa* - BOT, *F. conocephalifolia* - CON, *F. copiosa* - COP, *F. damaropsis* - DAM, *F. hispidoides* - HIS, *F. nodosa* - *NOD*, *F. phaeosyce* - PHA, *F. pungens* -PUN, *F. septica* - SEP, *F. trachypison* - TRA, *F. variegata* - VAR, and *F. wassa* - WAS. The quantitative environmental variables are the slope and successional age, the qualitative is the presence of a small stream (NoStream, Stream). Relevés are displayed as open circles.

2.9. Two approaches

If you have both the environmental data and the species composition (relevés), you can both calculate the unconstrained ordination first and then calculate regression of ordination axes on the measured environmental variables (i.e. to project the

environmental variables into the ordination diagram) or you can calculate directly the constrained ordination. **The approaches are complementary and should be used both!** By calculating the unconstrained ordination first you surely do not miss the main part of the variability in species composition, but you could miss the part of variability that is related to the measured environmental variables. By calculating the constrained ordination, you surely do not miss the main part of the variability explained by the environmental variables, but you could miss the main part of a variability that is not related to the measured environmental variables.

Be carefull to always specify the method of the analysis. From an ordination diagram you can tell whether a linear or unimodal analysis was used but you cannot distinguish between the constrained and unconstrained ordinations.

The hybrid analyses represent a "hybrid" between the constrained and the unconstrained ordination methods. In the standard constrained ordinations, there are as many constrained axes as there are independent explanatory variables and only the additional ordination axes are uncostrained. In the hybrid analysis, only a prespecified number of canonical axes is calculated and any additional ordination axes are unconstrained. In this way, we can specify the dimensionality of the solution of the constrained ordination model.

2.10. Partial analyses

Sometimes, we need to partial-out first the variability, which can be explained by one set of explanatory variables, and then to analyse the remaining variability (i.e. to perform the analysis on the residual variation). This is done by the **partial analyses**, where we extract first the variability, which can be explained by the **covariables** (i.e. the variables effect of which should be partialled out); then we perform the (constrained) ordination. The covariables are often (continuous or categorical) variables, effect of which is uninteresting, e.g. the blocks in the experimental designs. When we have more explanatory variables, then performing several analyses, with one of the variables being the explanatory variable and the rest acting as the covariables, enables us to test the partial effects (analogously to the effects of partial regression coefficients in a multiple regression).

2.11. Testing the significance of relationships with environmental variables

In the ordinary statistical test, the value of the statistics calculated from the data is compared with the expected distribution of the statistics under the null hypothesis tested and based on this comparison, we estimate the probability of obtaining results as different from the null hypotheses or even more extreme than our data are. The distribution of the test statistics is derived from the assumption about the distribution of the original data (i.e. why we expect the normality of the response residuals in least square regressions). In CANOCO, the distribution of the test statistics (F-ratio in the latest version of CANOCO is a multivariate counterpart of the ordinary F-ratio, the eigenvalue was used in the previous versions) under the null hypothesis of independence is not known; the distribution depends on the number of environmental variables, on their correlation structure, on the distribution of the species abundances etc. However, the distribution can be simulated and this is used in the **Monte Carlo permutation test.**

In this test, the distribution of the test statistic under the null hypothesis is obtained in the following way: The null hypothesis is that the response (the species composition) is independent of the environmental variables. If this is true, then it does not matter which set of explanatory variables is assigned to which relevé. Consequently, the values of the environmental variables are randomly assigned to the individual relevés and the value of the test statistics is calculated. In this way, both the distribution of the response variables and the correlation structure of the explanatory variables remain the same in the real data and in the null hypothesis simulated data. The resulting significance level (probability) is calculated as $1 \pm m$

follows: $P = \frac{1+m}{1+n}$; where *m* is the number of permutations where the test statistics

was higher in random permutation than in the original data, and n is total number of permutations. This test is completely distribution free: this means that it does not depend on any assumption about the distribution of the species abundance values. The permutation scheme can be "customized" according to the experimental design used. This is the basic version of the Monte Carlo permutation test, more sophisticated approaches are used in CANOCO, particularly with respect to the use of covariables – see the Canoco for Windows manual (Ter Braak & Šmilauer, 1998).

2.12. Simple example of Monte Carlo permutation test for significance of correlation

We know the heights of 5 plants and content of the nitrogen in soil, where they were grown. The relationship is characterized by a correlation coefficient. Under some assumptions (two-dimensional normality of the data), we know the distribution of the correlation coefficient values under the null hypothesis of independence. Let us assume that we are not able to get this distribution (e.g. the normality is violated). We can simulate this distribution by randomly assigning the nitrogen values to the plant heights. We construct many random permutations and for each we calculate the correlation coefficient with the plant height. As the nitrogen values were assigned randomly to the plant heights, the distribution of the correlation coefficients corresponds to the null hypothesis of independence.

Plant height	<u>Nitrogen (in</u> <u>data)</u>	1-st permutation	2-nd permutation	3-rd permutation	4-th permutation	5-th etc
5	<u>3</u>	3	8	5	5	
7	<u>5</u>	8	5	5	8	
6	<u>5</u>	4	4	3	4	
10	<u>8</u>	5	3	8	5	
3	<u>4</u>	5	5	4	3	
Correlation	<u>0.878</u>	0.258	-0.568	0.774	0.465	0.###

Significance of correlation=

1 + no. of permutations where (r>0.878) 1 + total number of permutations

for the one-tailed test or

$\frac{1 + \text{no. of permutations where (|r|>0.878)}}{1 + \text{total number of permutations}}$

for the two-tailed test.

Note, that the F-test as used in ANOVA (and similarly the F-ratio used in the CANOCO program) are the one-sided tests.

3. Using the Canoco for Windows **4.0** package

3.1. Overview of the package

The Canoco for Windows package is composed of several separate programs and their role during the process of the analysis of ecological data and the interpretation of the results is summarized in this section. Following sections then deal with some typical usage issues. As a whole, this chapter is not a replacement for the documentation, distributed with the Canoco for Windows package.

Canoco for Windows 4.0

This is the central piece of the package. Here we specify the data we want to use, specify the ordination model and testing options. We can also select subsets of the explained and explanatory variables to use in the analysis or change the weights for the individual samples.

Canoco for Windows package allows us to analyse data sets with up to 25 000 samples, 5000 species, and 750 environmental variables plus 1000 covariables. There are further restrictions on the number of data values. For species data, this restriction concerns non-zero values only, i.e. the absences are excluded, as these are not stored by the program.

Canoco for Windows allows one to use quite a wide range of the ordination methods. The central ones are the linear methods (PCA and RDA) and unimodal methods (DCA and CCA), but based on them, we can use CANOCO to apply other methods like the discriminant analysis (CVA) or the **metric** multi-dimensional scaling (principal coordinates analysis, PCoA) to our data set. Only the non-metric multidimensional scaling is missing from the list.

CANOCO 4.0

This program can be used as a less user-friendly, but slightly more powerful alternative to the Canoco for Windows program. It represents non-graphical, console (with the text-only interface) version of this software. The user interface is identical to the previous versions of the CANOCO program (namely versions 3.x), but the functionality of the original program was extended and in few places exceeds even the user friendly form of the version 4.0.

The console version is much less interactive than the Windows version - if we make a mistake and specify an incorrect option, there is no way back to the wrongly answered question. We can only terminate the program.

Nevertheless, there are few "extras" in the console version functionality. In my opinion, the only one worth of mentioning is the acceptance of "irregular" design specifications. You can have, for example, data repeatedly collected from the permanent plots distributed over three localities. If the data were collected different number of years, there is no way to specify this design to the Windows' version of the package so as to assure correct permutation restrictions during the Monte Carlo permutation test. The console version allows to specify the arrangement of samples (in terms of special and temporal structure and / or of the general split-plot design) for each block of samples independently.

Another advantage of the console version is its ability to read the analysis specification (normally entered by the user as answers to individual program' questions) from a "batch" file. Therefore, it is possible to programatically generate such batch files and run few to many analyses at the same time. This option is obviously an advantage only for experienced users.

WCanoImp and CanoImp.exe

The functionality of the WCanoImp program was already described in the section 1.7. The one substantial deficiency of this small, user-friendly piece of software is its limitation by the capacity of the Windows' Clipboard. Note that this is not such a limitation as it used to be for the Microsoft Windows 3.1 and 3.11. More importantly, we are limited by the capacity of the sheet of our spreadsheet program. For the Microsoft Excel, we cannot have more than 255 columns of data, so either we must limit ourselves to at most 255 variables or at most 255 samples. The other dimension is more forgiving -65536 rows in the Microsoft Excel 97 version.

If our data does not fit into those limits, we can either fiddle around with splitting the table, exporting parts and merging the resulting CANOCO files (not a trivial exercise) or we can use the console (command line) form of the WCanoImp program – program canoimp.exe. Both programs have the same purpose and the same functionality, but there are two important differences. The first difference is that the input data must be stored in a text file. The content of the file is the same as what the spreadsheet programs place onto the Clipboard. This is a textual representation of the spreadsheet cells, with transitions between the columns marked by the TAB characters and the transition between rows marked by the new-line characters. So the simplest way to produce such input file for the canoimp.exe program is to proceed as if using the WCanoImp program, up to the point the data were just copied to the Clipboard. From there, we switch to WordPad program (in Windows 9x) or to Notepad program (in Windows NT 4.x and Windows 2000), create a new document and select the Edit/Paste command. Then we save the document as an ASCII file (cannot be done otherwise in Notepad, but WordPad supports other formats, as well). Alternatively, we can save our sheet from the spreadsheet program using the File/Save as... command and selecting format usually called something like *Text file (Tab separated)*. Note that this works flawlessly only if the data table is the only contents of the spreadsheet document.

The second difference between the WCanoImp utility and the canoimp.exe program is that the options we selected in the WCanoImp main window must be passed (together with the name of the input file and of the desired output file) on the command line used to invoke the canoimp program. So, a typical execution of the program from the command prompt looks similarly to this example:

d:\canoco\canoimp.exe -C -P inputdta.txt output.dta

where the -C option means output in the condensed format, while the -P option means a transposition of the input data (i.e. rows represent variables in the input text file). The TAB-separated format will be read from the **inputdta.txt** and the CanoImp will create a new data file (and overwrite any existing file with the same name) named **output.dta** in the CANOCO condensed format.

If you want to learn about the exact format of the command line when calling the canoimp.exe program, you can invoke it without any further parameters (that means, also without the names of input and output files). Program then provides a short output describing the required format of the parameters' specification.

<u>CEDIT</u>

Program CEDIT is available with the Canoco for Windows installation program as an optional component. It is not recommended for installation on the Windows NT (and Windows 2000) platform, where its flawless installation needs an intimate knowledge of the operating system, but it is supposed to work from the first start when installing on the Windows 9x, at least if you install into the default **c:\canoco** directory.

Availability of that program is by a special arrangement with its author and, therefore, no user support in case of any problems is available. If you install it, however, you get program documentation in a file in the installation directory, including instructions for its proper setup.

Another disadvantage (in eyes of many users) is its terse, textual interface, even more cryptic than that available with the console version of the CANOCO program. But if you enjoy using text editors under UNIX with such kind of interface, where commands are executed by entering one or few letters from your keyboard (Emacs being the most famous one), then you will love CEDIT.

Now, for the rest of us, what is the appeal of such program? It is in its extreme power for performing quite advanced operations on your data that are already in the CANOCO format. No doubt that most of these operations might be done (almost) as easily in the Windows spreadsheet programs, yet you do not always have your data in the appropriate format (particularly the legacy data sets). CEDIT can transform the variables, merge or split the data files, transpose the data, recode factors (expanding factor variable into a set of dummy variables) and much more.

CanoDraw 3.1

The CanoDraw 3.1 program is distributed with the Canoco for Windows package and it is based on the original 3.0 version which was available, as an add-on for the CANOCO 3.1x software (a "lite" version of the CanoDraw 3.0 was distributed with each copy of the CANOCO 3.1x program for the PC platform).

There were only few functional changes between the versions 3.0 and 3.1 and as the original one was published in 1992, it is reflected by its user interface feeling clumsy by today' standards. First, while CanoDraw does not have a textual (console-like) user interface, it's graphics mode is limited to the standard VGA resolution (640x480 points) and it runs usually only in the full screen mode. But it can be usually started directly from the Windows environment, so that we can interactively switch between the Canoco for Windows and CanoDraw on one side, and between CanoDraw and CanoPost program, when finalizing the look of the produced diagrams.

CanoDraw concentrates lot of functionality on a small foothold. This is the reason it is sometimes difficult to use. Besides displaying the simple scattergrams of ordination scores and providing appropriate mutual rescaling of scores when preparing so-called biplots and triplots, CanoDraw enables further exploration of our data based on the ordination results. It provides to this aim a palette of methods, including generalized linear models, loess smoother model and portraing results of

these methods with the contour plots. Further, we can combine the ordination data with geographical coordinates of the individual samples, classify our data into separate classes and visualize the resulting classification, compare sample scores in different ordination methods and so on.

As for the output options, CanoDraw supports direct output to several types of printers, including HP LaserJet – compatible printers, but today users of CanoDraw 3.1 are advised to save their graphs either in the Adobe Illustrator (AI) format or in the PostScript (PSC) format which can be further enhanced with the CanoPost program. While the Adobe Illustrator program provides powerfull platform for further enhancement of any kind of graphs, here lays its limitation, too. This program has no idea what an ordination method is: it does not know that the symbols and arrows in an ordination plot cannot be moved around, in contrast to the labels, or that the scaling of the vertical axis might not be changed independently of the horizontal one. Last, but not least, using Adobe Illustrator needs further software license, while CanoPost is provided with the Canoco for Windows package. Additionally, AI files can be exported even from the CanoPost, so users do not miss the handsome features of the Adobe Illustrator program.

CanoPost for Windows 1.0

This program reads files produced by the CanoDraw program and saved in the PostScript format (usually with the **.psc** extension). Note that these are valid files in the PostScript language, so you might print them on a laser printer supporting that language. But to use them with the CanoPost, you do not need a PostScript printer! Also, CanoPost is able to read only the PostScript files produced by CanoDraw program, not any other kind of PostScript files.

CanoPost allows further modification of the graphs, including change of the text, style for the labels, symbols, line or arrows. Positions of labels can be adjusted by dragging them around the symbols or arrows their label. The adjustments made to particular plots can be saved into style sheets, so they can be easily applied to any other ordination diagrams. Beside the work on individual graphs, CanoPost allows us to combine several graphs into a single plot.

Adjusted plots may be saved in the CanoPost own format (with the **.cps** extension), printed on any raster output device supported by our Windows' installation or exported as a bitmap (.BMP) file or in the Adobe Illustrator format.

3.2. Typical analysis workflow when using Canoco for Windows 4.0



Figure 3-1 The simplified workflow in using the Canoco for Windows package
The Figure 3-1 shows a typical sequence of actions taken when analyzing multivariate data. We first start with the data sets recorded in a spreadsheet and export them into CANOCO compatible data files, using the WCanoImp program. In the Canoco for Windows program, we either create a new CANOCO project or clone an existing one using the *File/Save as...* command. Cloning retains all the project settings and we can change only those that need to be changed. Of course, changing names of the source data files invalidates choices dependent on them (like the list of environmental variables to be deleted).

Each project is represented by two windows (views). The **Project view** summarizes the most important project properties (e.g. type of the ordination method, dimensions of the data tables and names of the files the data are stored in). Additionally, the Project view features a column with buttons providing shortcuts to the commands most often used when working with the projects: running analysis, modifying project options, starting the CanoDraw program, saving the analysis log etc. The **Log view** records the users' actions on the project and the output provided during the project analysis. Some of the statistical results provided by the CANOCO are available only from this log. Other results are stored in the "SOL file", containing the actual ordination scores. The content of the Log view may be extended by entering a new text (comments) into the log: the Log view works as a simple text editor.

We can define the project settings using the Project Setup wizard. This wizard can be invoked for example by clicking the **Options** button in the Project view. CANOCO displays the first page from a sequence of pages containing various pieces of information the program needs to know to apply an appropriate type of the ordination method. This sequence is not a static one, the page displayed at a certain time depends on the choices made in the preceding pages. For example, some of the options are specific for the linear ordination methods, so these pages are displayed only if a linear method (PCA or RDA) was chosen. We proceed between the pages using the **Next** button. But we might return to the preceding pages using the **Back** button. Some of the critical choices to be made with the Setup Wizard are discussed in more detail later in this chapter. On the last page, the **Next** button is replaced by the **Finish** button. After we click this button, the changes in the options are applied to the project. If we were defining a new project, CANOCO asks for the name of the file where the project will be saved.

After the project is defined, the analysis might be performed (the data analyzed) by clicking the **Analyze** button in the project view (or, alternatively, by using the shortcut button from the toolbar or using the menu command). On success, the results are stored in the solution file (its name was specified on the second Project Setup wizard page) and additional information is placed into the Log view, where it might be inspected. In the Log view, we can find a statistical summary for the first four ordination axes, information on the correlation between the environmental variables and the ordination axes, indication of the outliers, and the results of the Monte Carlo permutation tests. Part of this information is essential for performing certain tasks, but nothing needs to be retained for plotting the ordination diagrams with the CanoDraw program. CanoDraw needs only the results stored in the solution file.

With the CanoDraw program, we can explore the ordination results and combine them with the information from the original data. Here we define the basic contents of the ordination diagrams (range of axes, which items are plotted, contents of the attribute plots etc.). The resulting diagrams can be further adjusted (change of symbol type, size and colors, change of label font and position, change of line type, etc.) and combined in the CanoPost for Windows program, providing publicationready graphs.

3.3. Decide about ordination model: unimodal or linear?

This section provides a simple-to-use "cookbook" for deciding whether we should prefer the ordination methods based on the model of linear species response to the underlying environmental gradient or the weighted-averaging (WA) ordination methods, corresponding to the model of unimodal species response. Inevitably, the presented recipe is somewhat simplistic, so it should not be followed blindly.

In the Canoco for Windows project that we use to decide between the unimodal and linear methods, we try to match as many choices we will make in the final analysis, as possible. If we have covariables, we use them here as well, if we use only a subset of the environmental variables, we subset them here too. If we log-transform (or square-root-transform) our species data, we do it here as well.

For this trial project, we select the weighted averaging method with detrending. This means either the DCA for the indirect gradient analysis or DCCA for the constrained analysis. Then we select detrending by segments (which also implies the Hill's scaling of ordination scores) and then we select the other options as in the final analysis and run the analysis. We then look into the analysis results stored in the Log view. At the end of the log, there is the Summary table and in it is a row starting with "Lengths of gradient", looking similarly to the following example:

Lengths of gradient :	2.990 1.324	.812 .681
-----------------------	-------------	-----------

Now we locate the largest value (the longest gradient) and if that value is larger than 4.0, we should use the unimodal method (DCA, CA, or CCA). Use of the linear method would not be appropriate, as the data are too heterogeneous and too many species deviate from the assumed model of linear response. On the other hand, if the longest gradient is shorter than 3.0, the linear method is probably a better choice (not necessarily, see Ter Braak et Šmilauer 1998, section 3.4 on page 37).

Centering and Standardization ? × SAMPLES SPECIES ⊙ <u>N</u>one O None C Center by sample Center by species O Standardize by norm Standardize by norm C Center and standardize Center and standardize ndardize by error <u>v</u>ariance <u>N</u>ext > < <u>B</u>ack Cancel Help

3.4. Doing ordination - PCA: centering and standardizing

Figure 3-2 Centering and standardization options in the Project Setup wizard

This Project Setup wizard page is displayed for the linear ordination methods (PCA and RDA) and refers to the manipulations with the species data matrix before the ordination is calculated.

The centering by samples (the option in the left half of the wizard page) results in the average of each row to be equal to zero. Similarly, the centering by species (in the right half of the wizard page) results in the average of each column to be equal to zero. Centering by species is obligatory for the constrained linear method (RDA) or for any **partial** linear ordination method (i.e. where covariables are used).

Standardization (by samples or by species) results in the norm of each row or column being equal to one. The **norm** is the sum of squares of the row / column values. If we apply both the centering and the standardization, the centering is done first. Therefore, after centering **and** standardizing by species, the columns represent variables with zero average and unit variance. As a consequence, PCA performed on the species data then correspond to the "PCA on a matrix of correlations" (between the species).

If we have environmental variables available in the ordination method (always in the RDA and optionally in the PCA), we can select the standardization by the **error variance**. In this case, CANOCO estimates for each species separately the variance in the species values left unexplained after fitting that species to the selected environmental variables (and covariables, if any). The inverse of that variance is then used as the species weight. Therefore, the better is a species described by the provided environmental variables, the higher weight it has in the analysis.

3.5. Doing ordination - DCA: detrending



Figure 3-3: Detrending method selection in the Project Setup wizard

The original method of correspondence analysis suffers often by the so-called arch effect. With such effect in place, the scores of samples (and of the species) on the second ordination axis are a quadratic function of the scores on the first axis. *Hill et Gauch (1980)* proposed a heuristic, but often well working method of removing this arch effect, called **detrending by segments**. This method was criticized by several authors (see for example *Knox, 1989*), yet there are essentially no better ways of dealing with this artifact. Use of detrending by segments is not recommended for unimodal ordination methods where either covariables or environmental variables are present. In such case, if detrending procedure is needed, the **detrending by polynomials** is the recommended choice. The reader is advised to check the Canoco for Windows manual for more details on deciding between the polynomials of second, third, or fourth degree.

The whole detrending procedure is usually not needed for the constrained unimodal ordination. If an arch effect occurs in CCA, this is usually a sign of some redundant environmental variables being present. There may be two or more environmental variables strongly correlated (positively or negatively) with the each other. If we retain only one variable from such a group, the arch effect disappears. The selection of a subset of environmental variables with a low extent of crosscorelation can be performed using the forward selection of the environmental variables in the Canoco for Windows program.

3.6. Doing ordination - scaling of ordination scores

Scaling: Linear Methods Focus scaling on: C Inter-sample distances C Inter-species <u>correlations</u> C Symmetric	Species scores: Do not post-transform	Scaling: Unimodal Methods Focus scaling on: C Inter-sample distances C Inter-species distances C Symmetric	Scaling type: • biplot scaling (L^a) • Hill's scaling (L^a) / (1 - L)
< Back Next	Cancel Help	<back nex<="" td=""><td>t > Cancel Help</td></back>	t > Cancel Help

Figure 3-4 Scaling options for the linear and for the unimodal methods in the Project Setup wizard

The most important result of an ordination method applied to our data is the ordination diagram. Theoretically, we are able to reconstruct (with certain level of error) not only the primary data table (the species data), but also the matrix of (dis)similarities between our samples and correlation matrix between the species, all based on the ordination diagram. We usually do not attempt to recover such data from the ordination diagram (because we have the measured values available), but we definitively go part of this way when interpreting contents of the ordination diagram and generating interesting research hypotheses. The precision of conclusions we make about similarity of samples, relations between species and/or environmental variables, etc. depends partly on the relative scaling of the scores on individual ordination axes. One kind of scaling is more favorable if we focus our attention onto relation between our sampling units, another in the case we interpret the relation between the species.

The options are somewhat similar both in the linear and the unimodal ordination methods. First, we should decide whether we will focus, during the interpretation of the resulting ordination diagrams, on the samples (this includes comparison of classes of samples, as portrayed by the nominal environmental variables) or on the species.

Then, for the linear model, we should decide whether the differences in the abundances of individual species in the data should be mirrored by the lengths of their arrows (with dominant species having generally longer arrows than those with small abundace values) or whether each species should be relativized. The relativization is appropriate for the so-called correlation biplots.

In the case of the unimodal ordination, we should decide about the method of the interpretation of the ordination diagrams. For the data with very long composition gradients (with the large beta diversity across the samples), the **distance rule** is more appropriate and so is the Hill' scaling. In the other cases, the **biplot** scaling provides ordination diagrams that can be interpreted in a more quantitative way.

3.7. Running CanoDraw 3.1

The new version of the CANOCO package (Canoco for Windows 4.0) provided users with an additional program - CanoPost, that allows to substantially improve diagrams created with the program CanoDraw, which is available since the introduction of the CANOCO 3.x. When using program CanoDraw (version 3.1) in this new context, we should focus on creating **contents** of the diagrams and postpone the improvements of the appearance till we open the diagrams in the CanoPost program (see the section 3.8).

Program CanoDraw can be started directly from the Canoco for Windows program, using the **CanoDraw** button placed in the project view window. Nevertheless, this button is not always enabled because CANOCO observes the limitation on the data dimensions, imposed by the CanoDraw program. CanoDraw 3.1 is still a DOS program and so it cannot use the memory available for the Windows programs. Therefore, we can work only with the analyses with up to 500 samples, 500 species and 30 or less environmental variables. As an additional restriction, CanoDraw does not permit opening CANOCO analyses where the forward selection of environmental variables was explicitly performed. ⁺

If we invoke CanoDraw 3.1 via Canoco for Windows program, it is called with several command line parameters that specify:

- name of the CANOCO project file. After the CanoDraw is opened (it works only in a full-screen, DOS graphic mode), we do not need to navigate it to the CANOCO project file (the file with .CON extension), because this was already opened. CanoDraw also already parsed the file with the ordination results (the "solution" file) - of course unless some error was encountered during that process
- type of the output device as a PostScript printer. This setting is important for saving graphs compatible with the CanoPost program. It is not passed to CanoDraw only in the case the file path to the CANOCO project file is too long. The command line used to start CanoDraw is limited, in its extent to 127 characters, including full path to the CanoDraw program. The only change in CanoDraw settings that must be done explicitly is to change output from first parallel port to a file.

As it was already mentioned above, when using CanoDraw we should concentrate on the contents of the graphs: what is the range of the diagram' axes, which items are plotted, whether all the items appear with appropriate labels etc. We do not need to care about color and type of the symbols or lines, about the exact positions of the labels - all that can be changed later on in the CanoPost for Windows program.

First and foremost, we must decide what kind of diagram we want to produce. When summarizing CANOCO project where both species (community composition) and environmental data were used, the biplot diagram including species and environmental variables is the most appropriate default choice. Before creating that diagram, we should make two important decisions about its content:

• the qualitative explanatory ("environmental") variables are best represented by the centroid scores for the individual dummy variables. The symbols then represent the centroids of the scores of samples belonging to the particular class (level) of that qualitative variable. CanoDraw uses these scores for variables

⁺ Here, Canoco for Windows provides a workaround: after you run an analysis where "manual" (interactive) forward selection was employed, Canoco offers you to make the selection results explicit. This means that forward selection process is removed from the Canoco project settings and the environmental variables that were not selected during this process are marked as explicitly deleted from the analysis. We must then re-run the analysis (producing the ordination results identical to those from the analysis with forward selection) and then we can use the CanoDraw program.

marked as nominal. We can select them using the menu command File/Nominal env. variables.

• for a typical community composition data, it often does not make sense to display scores of all the species. Some are so rare in the data, that no relevant information is provided on their ecological preferences. Other species might not be characterized well by the explanatory variables used in the analysis. We usually try to define the subset of species appearing in an ordination diagram using a combination of two criteria, both accessible from the dialog box invoked by the **Options/Restrictions/Rules of visibility** menu command: **Minimum fit** states the minimum percentage of variability in species values explained by the ordination axes); this characteristic is not available for the unimodal-model based analyses where detrending by segments was applied; **minimum weight** is available only for the unimodal ordination model and specifies the minimum weight (relatively to the species with the highest weight in the data) a species must have, to be plotted.

3.8. Adjusting diagrams with CanoPost program

We start our work with a diagram (typically an ordination diagram) in the CanoPost for Windows program by importing the file produced by the CanoDraw 3.1 program. This file typically has the **.psc** extension in its name. Then, we can change the appearance of the diagram and save the modified diagrams in the CanoPost own file format (using the **.cps** extension). This file can be reopened later in the CanoPost program.

When adjusting the diagrams, it is better to start with their global features before adjusting the details. First, we should decide about the size of the resulting printout of the diagram. This paper printout metaphor is used even if the final result should be an image file included into an electronic document. CanoPost bases its expectation of the paper size on the default printer settings in the Windows environment, but we can change this using the **File/Page setup...** menu command. The size of this output media is displayed in CanoPost as a white rectangle upon which the diagram is drawn. Most often, we should try to fill up this output page with our graph as much as possible. To achieve this, we use combination of two commands (**Edit/Scale graph** and **Edit/Shift graph**), also accessible using the toolbar shortcuts.

After that, we might modify size of symbols and typeface and size of labels. Only after these settings are fixed makes sense to adjust positions of the labels. We try to minimize their overlap so as to improve the readability of the graph contents.

3.9. New analyses providing new views of our datasets

The work with the multivariate data has usually an iterative character. We might find useful to modify the explanatory variables, based on our exploration of the results we get from the initial analyses. This includes a selection of subset of environmental (explanatory) variables based on the knowledge we gain during the forward selection of environmental variables in CANOCO.

A special case of this subset selection is the recoding of the qualitative variables. Because the individual levels of a qualitative explanatory variable are

judged independently (as individual dummy variables), we can see which levels differ significantly from the others in terms of the community composition (the species data).

Based on the exploration of the relation of the explanatory variables to the ordination axes and to individual species, we can transform the environmental variables to maximize their linear relation with the gradients being recovered by the ordination model.

In some cases, the ordination results reveal two or more distinct groups of samples, based on their different composition. In that case, it is often useful to ramify the results by further separate analyses for each of those groups. In a direct gradient analysis, this often leads to a change of the picture: the environmental variables identifiable with the differences **between** these groups are frequently different from those acting **within** such groups.

Another type of a follow-up analysis is used when we identify some environmental variables important for explaining part of the structure in the species data. In such case, we can turn those variables into covariables and either test the additional effects of other, potentially interesting explanatory variables (partial direct gradient analysis) or just inspect the "residual" variation and try to informally interpret any pattern found in that way (partial indirect gradient analysis).

3.10. Linear discriminant analysis

(Fisher) linear discriminant analysis (LDA), also called the Canonical Variate Analysis (CVA), is a method allowing us to find the scores for samples, expressed as the linear combinations of the explanatory variables that (in a particular sense) optimally separate the *a priori* defined groups. The method can be performed using the Canoco for Windows software with some additional features, not available in the standard implementations of this method.

To perform LDA in CANOCO, the classification of samples must be used as the species data. Each variable represents a single class and the samples belonging to this class have value 1.0 for this variable and zero values for the other variables ("species"). This is the appropriate coding for the classical discriminant analysis, where the classification is "crisp". CANOCO allows us to perform a discriminant analysis based on a "fuzzy" classification, where (some) samples can belong to several classes. This situation might represent a true partial membership in several classes or our uncertainty about the true membership for those samples. The only requirement for fuzzy coding is that the sum of sample' values in species data is equal to one.

The variables we want to use for the discrimination enter the CANOCO analysis as the environmental variables. We then select a canonical correspondence analysis (CCA) using Hill's scaling with a focus on the species distances (value -2 in the console version of CANOCO).

The one distinct advantage of doing LDA in CANOCO is that we might perform a **partial** discriminant analysis. In such an analysis, we can look for explanatory variables allowing us to discriminate between given classes in addition to other known discriminatory variables. Another two distinct features of LDA in the CANOCO program is the ability to select a subset of the discriminating (explanatory) variables by means of the forward selection of environmental variables and the ability to test the discriminatory power of the variables by the non-parametric Monte-Carlo permutation test.

When plotting the results, the species scores represent the means (centroids) of the individual classes in the discriminatory space. Scores of samples (the **SamE** scores) are then the discriminant scores for the individual observations. A biplot containing biplot scores of environmental variables (**BipE**) presents the table of averages of individual explanatory variables within individual classes, while the regression / canonical coefficients (**Regr** scores) of environmental variables represent the loadings of the individual variables on the discriminant scores.

4. Direct gradient analysis and Monte-Carlo permutation tests



4.1. Linear multiple regression model

Figure 4-1 Graphical representation of a simple linear regression model

We start this chapter with a simple resume of the classical linear regression model, because its knowledge is very important if we want to grasp the meaning of the direct gradient analysis.

The Figure 4-1 presents a simple linear regression model used to model dependence of the values of a variable Y on the values of variable X. The figure shows difference between the true (observed) values of the response variable Y and the fitted ones (Y "hat"). The difference between those two is called a regression residual and is named \mathbf{e} in the Figure 4-1.

An important feature of all the statistical models (including the regression models) is that these models have two main components:

- the **systematic component** describes the variability in the values of the response variable(s) that can be explained by one or several explanatory variables (predictors), using a parameterized function. The simplest is the linear combination of the explanatory variables, the function used by the (general) linear regression models.
- the **stochastic component** is a term referring to the remaining variability in the values of the response variable, not cast by the systematic part of the model. The stochastic component is usually defined using its probabilistic, distributional properties.

We judge the quality of the **fitted** model by the amount of the variability of response variable described by the systematic component (usually comparing this amount to the amount of the unexplained variability - represented by the stochastic component). We try to present only the regression models where all the predictors significantly contribute to their quality. We can select such a subset of the predictors using the stepwise selection. Its most frequent type is the so-called forward selection.

Here we start with a null model with no predictors, assuming that the variability in the response variable cannot be predicted and that it represents only a stochastic variability. We then pick a single predictor from the set of the available variables - it is the one leading to a regression model with the highest amount of the variability explained. Even when adding the best predictor, its contribution might be just due to a chance: if we would randomly swap that predictor's values, such a non-sense variable would explain a positive amount of the variability in the values of the response variable, so we must test the contribution of the considered candidate predictor. If the contribution of the selected predictor is judged to be significant, we can repeat the process and try to look-up another good predictor in the pool of the remaining variables. We can again test its contribution and stop this selection as soon as the "best" among the remaining predictors is not "good enough".

4.2. Constrained ordination model

In the chapter 2, the linear and unimodal methods of **indirect gradient analysis** (PCA and CA, respectively) were defined as methods looking for one or more (mutually independent) gradients representing "optimal" predictors for fitting the regression models of linear or unimodal species response. The optimality is restricted by the assumptions of those methods and judged over all the species in the primary data.

The methods of **direct gradient analysis** (also called the constrained or canonical ordination methods) have an identical task, but the gradients these methods "are allowed to find" are further restricted. These gradients must be linear combinations of the submitted explanatory (environmental) variables. Therefore, we try to explain abundances of (all) the individual species using synthetic variables, but those variables are defined based on the values of the observed characteristics.

In that way, the methods of direct gradient analysis resemble the model of the multivariate multiple regression. But in such a regression with \mathbf{m} response variables (species in CANOCO) and \mathbf{p} predictors (environmental variables in CANOCO), we must estimate $\mathbf{m}*\mathbf{p}$ parameters (regression coefficients) from the data. This is not the case in constrained ordination, where the effect of predictors upon response variables is "focused" through a low number of intermediate gradients - the ordination axes, called here the **canonical** (or constrained) **axes**. There are as many canonical axes as there are **independent** explanatory variables.

In CANOCO, we quite often use the **partial** analyses where we specify, beside the environmental (explanatory) variables, the so-called **covariables** which represent effects we want to account for and, in that way, remove from the solution of the ordination model. Covariables (or covariates) are used in a similar context in the analysis of variance. Here we often use quantitative covariables in addition to the factors. In a regression analysis sensu stricto, the notion of covariates is not frequently used, but the difference among them and the "true" explanatory variables is only in the way we look at them. Both are explanatory variables in a regression model (and in the ordination model) and only the role we attribute to them difference.

4.3. RDA: constrained PCA

We will illustrate the topics from the previous section on the case of the redundancy analysis (RDA), a constrained form of the linear ordination method - principal

components analysis (PCA). We will use a very simple setup, where we try to determine only the first ordination axis (the first principal component) and we have two environmental variables (z_1 and z_2) available, constraining the ordination axis of RDA.

Both PCA and RDA methods try to find values of a new variable (we will name it \mathbf{x}), which represents an "optimum" predictor for the values of all the species. The value of that variable for the **i**-th sample is \mathbf{x}_i and we use it to predict values of the **k**-th species in the **i**-th sample using the following equation:

$$y_{ik} = b_{0k} + b_{1k}x_i + e_{ik}$$

Here both PCA and RDA must estimate two sets of parameters: the values \mathbf{x}_i which are the **sample scores** on the first ordination axis and the regression coefficients for each species (\mathbf{b}_{1k}) which are the **species scores** on the first ordination axis. The additional parameter for each species (\mathbf{b}_{0k}) represents the "intercept" of the fitted regression line and we can get rid of its estimation by centering the primary data by the species (see section 3.4).

Here the similarity between PCA and RDA ends, because in the constrained method, the values of the sample scores are further constrained. They are defined as the linear combination of the explanatory variables. In our example, we have two such variables:

$$x_i = c_1 z_{i1} + c_2 z_{i2}$$

Note that the parameters estimated here (c_j for j-th environmental variables) do not represent the scores of the environmental variables that are usually plotted in the ordination diagrams. Rather, they represent the **Regr** scores in the output from a CANOCO analysis. The normally plotted scores (**BipE** - biplot scores of environmental variables) are the regression coefficients from a regression model where the sample scores (x_i) are fitted separately for the individual explanatory variables. Therefore, they represent the marginal (independent) effects of individual environmental variables.

We may combine the two above equations into a single one, further illuminating the relation of the constrained ordination to the multiple multivariate regression:

$$y_{ik} = b_{0k} + b_{1k}c_1z_{i1} + b_{1k}c_2z_{i2} + e_{ik}$$

The expressions $\mathbf{b_k}^* \mathbf{c_j}$ represents the actual coefficient of a multiple multivariate regression model, but these coefficients are constrained by their definition: they are based on the two smaller sets of the species scores and of the environmental variables scores.

 Permutation Typ C Unrestricted p 		🗖 Blocks defi	ned by covariables	
		oral structure or spl	·	
C Read from file	e:			Browse
- Random Numbe	er Generator —			
⊙ <u>S</u> et seeds	23239	945	Randomize	
O Le <u>v</u> erage co	rrected residual			

4.4. Monte Carlo permutation test: an introduction

Figure 4-2 Introductory page of the Project Setup wizard for selecting the permutation type

CANOCO has the ability to statistically test the significance of the constrained ordination models described in the preceding section, using the **Monte-Carlo permutation tests**. These statistical tests relate to the general null hypothesis, stating the independence of the primary (species) data on the values of the explanatory variables. But CANOCO provides a rich toolbox of specific setups for the tests applied to the datasets with a particular spatial, temporal or logical internal structure, related to the experimental or sampling design. The Figure 4-2 illustrates the first page of a sequence of pages available in the Project Setup Wizard to specify the permutation test properties.

4.5. Null hypothesis model

The null model of the independence between the corresponding rows of the species data matrix and the environmental data matrix (referring to the same sample) is the basis for the permutation test in the CANOCO program. While the actual algorithm used in CANOCO does not use the approach described here, we might use it to illustrate better the meaning of the permutation test.

- We start by randomly re-shuffling (permuting) the samples (rows) in the environmental data table, while keeping the species data intact. Any combination of the species and environmental data obtained in that way is as much probable as the "true" data sets, if the null hypothesis is true.
- For each of the data sets permuted in this manner, we calculate the constrained ordination model and express its quality in a way similar to that used when judging the regression model quality. In a regression model, we use a F statistics which is a ratio of the variability of the response variable explained by the regression model (divided by the number of the model parameters) and of the residual variability (divided by the number of residual degrees of freedom). In the case of constrained ordination methods, we use similar statistics, described in more detail in the following section.
- We record the value of such a statistics for each permutation. The distribution of these values defines the distribution of this statistics under the null model. If the test statistics value obtained from the actual data (with no permutation of the rows of the environmental data table) is highly improbable to come from that distribution (supposedly being much larger, corresponding to a higher-quality

ordination model), we reject the null hypothesis. The probability that the "dataderived" value of the test statistics originates from the calculated null model distribution then represents the probability of Type I error, i.e. a probability of rejecting a correct null hypothesis.

The actual algorithm does not permute the actual environmental data table, it rather permutes the residuals from the fitted dependence of the species data on the environmental variables and based on each such permutation, constructs a "new data table". That algorithm is even more complicated in the case the covariables enter the picture.

4.6. Test statistics

The previous section described the general principle of the permutation test and here we discus the possible choices for the test statistics used in that test. We mentioned that such a statistics corresponds to the F statistics used in the parametric significance test of the regression model. But the choice of the definition for this statistics in constrained ordination is difficult, because of the multidimensionality of the obtained solution. In general, the variability in the response variables (of the species data) described by the explanatory (environmental) variables is cast in more than one canonical axis. The relative importance of the canonical axes decreases from the first up to the last canonical axis, but we can rarely ignore all the canonical axes beyond the first one. Therefore, we can either express the accounted variance using the sum of all canonical axes or focus on one canonical axis, typically the first one. This corresponds to two test statistics available in CANOCO version 3.x and 4.x and the two corresponding permutation tests:

• Test on the first canonical axis uses statistics defined in the following way:

$$F_{\lambda} = \frac{\lambda_1}{RSS_{X+1}/(n-p-q)}$$

The variance explained by the first (canonical) axis is represented by its eigenvalue (λ_1) . The residual sum of squares (RSS) term corresponds to the difference between the total variance in the species data and the amount of variability explained by the first canonical axis (plus by the covariables, if any are present in the analysis). The number of covariables is q.

• Test on the sum of the canonical axes. Here we refer to the overall effect of *p* explanatory variables, revealed on (up to) *p* canonical axes:

$$F_{trace} = \frac{\sum_{i=1}^{p} \lambda_i / p}{RSS_{X+Z} / (n-p-q)}$$

The RSS term in this formula represents the difference between the total variability in the species data and the sum of eigenvalues of all the canonical ordination axes.

As the previous section explained, the value of the one of the two test statistics originating from the original data is compared with the distribution of that statistics under the null model assumption. This is illustrated by the Figure 4-3.



Figure 4-3: The distribution of the F-statistics values from a Monte Carlo permutation test, compared with the F-statistics value of the "true" datasets

In this figure, we see a histogram approximating the shape of the distribution of the test statistics. The position of the vertical arrow marks the value calculated from the "real" data. The permutations where the corresponding F-like statistics value is above this level represent the evidence in favour of a non-rejection of the null hypothesis. The Type I error of this test is then calculated as:

$$P = \frac{n_x + 1}{N + 1}$$

where $\mathbf{n}_{\mathbf{x}}$ is the number of the permutations yielding F statistics as large or larger than that from the "real" data and **N** is the total number of permutations. The value of 1 is added to both numerator and denominator because (under the assumption of the null model) the statistics based on the actually observed data is also considered to come from the null model distribution and, therefore, "vote against" the null hypothesis rejection. The usual choices of the number of permutations (like 99, 199, 499) follow from that specific property of the formula given above.

4.7. Spatial and temporal constraints

The way the permutation test was described in the preceding section is correct only in the case the collected set of samples does not have any implied internal structure, namely if the samples are picked up randomly, independently of each other. In this case, we can fully randomly reshuffle the samples, because under the null model, each sample' environmental data (explanatory variables values) can be matched with any other sample' species data with an equal probability.

This is no longer true when the "relatedness" between different samples is not equal across the whole dataset. The basic three families of the internal dataset structure recognized by the Canoco for Windows are represented by the choices in the Setup Wizard page illustrated in the Figure 4-4.

Permutation Restrictions
Select type of restriction on the permutations
 Time series or line transects Disable random shifts of mirror image Rectangular spatial grid
C <u>S</u> plit-plot design
< <u>B</u> ack <u>N</u> ext > Cancel Help

Figure 4-4 Project Setup wizard page for selecting restrictions for a permutation test

The samples may be arranged along a linear transect through the space or along a time axis. Then the samples cannot be permuted randomly, because we must assume an autocorrelation between the individual observations in both the species and the environmental data. We should not disturb this correlation pattern during the test because our hypothesis concerns the relations between the species and environmental data, not within these datasets. To respect the autocorrelation structure, CANOCO (formally) bends the sequence of samples in both the species and the environmental data into a circular form and re-assignment is performed by "rotating" the environmental data band in respect to the species data band. The Canoco for Windows manual should be discussed for more details of this and other described permutation restrictions.

A similar spatial autocorrelation occurs when we generalize location of sample on a linear transect into general positioning of samples in space. However, CANOCO does not allow this and provides support only for the placement of samples on a rectangular, homogeneously spaced grid.

The most general model of the internal structure of the datasets is provided by the last item in the Figure 4-4, named the **split-plot design**. This type of permutation restrictions is described in more detail in the following section.

All these restrictions can be further nested within another level of blocks. Blocks are usually defined in the CANOCO program using (a subset of) nominal covariables and represent groups of samples that are similar to each other more than to the samples from the other blocks. In the permutation test, the samples are permuted (with eventual further restrictions) only within those blocks, never across the blocks. If we compare a constrained ordination model with a model of the analysis of variance, the blocks can be seen as a random factor with an effect not interesting for the interpretation.

4.8. Design-based constraints

Split-Plot Design - II	? ×						
Split-plots permutation scheme							
-Whole-plot level	Split-plot level						
C No permutation C Time series or linear transect C Spatial grid C Freely exchangeable	 No permutation Time series or linear transect Spatial grid Ereely exchangeable Disable shifts from mirror image 						
Disable chifts from mirror image	 Independent across whole plots Dependent across whole plots 						
<back next=""> Cancel Help</back>							

Figure 4-5 Project Setup wizard page for specifying permutation restrictions for a split-plot design

The split-plot design restriction available in Canoco for Windows 4.0 allows us to describe a structure with two levels of variability (with two "error levels").⁺

The upper level is represented by the so-called **whole-plots**. Each of the whole-plots contains the same number of **split-plots**, representing the lower level of the design. CANOCO provides a great extent of flexibility in permuting samples at the whole-plot and/or the split-plot levels, from no-permutation, through spatialy or temporaly restricted permutations up to a free exchangeability at both levels. In addition, CANOCO provides notion of dependence of the split-plots structure across the individual whole-plots. In this case, the particular permutations of the split-plots within the whole-plots are identical across the whole-plots on each independent permutation test.

4.9. Stepwise selection of the model

At the end of the section 4.1, the forward selection of the explanatory variables for a regression model was described in some detail. The forward selection available in the CANOCO program has the same purpose and methodology, using the partial Monte Carlo permutation test to assess the quality of each potential candidate predictor for extending the subset of the explanatory variables used in the constrained ordination model.

If we select an interactive ("manual") forward selection procedure in the Canoco for Windows program, CANOCO presents the following window during the analysis (Figure 4-6).

⁺ Another level can be added, in some cases, using the permutation within blocks defined by the covariables (see the previous section).



Figure 4-6 Dialog box for the Forward selection of environmental variables. The question-marks for the variable *BrLeaf* correspond to the variable which was not tested by the permutation test during the forward selection.

The Figure 4-6 illustrates the state of the forward selection procedure where three best explanatory variables (Forest, BrLeaf, E2) were already selected (they are displayed in the lower part of the window). The values in the window top show that the three selected variables account for approximately 72% of the total variability explained by all the environmental variables (i.e. 0.320 of 0.447).

The list of variables in the upper part of the window shows the remaining "candidate predictors" ordered by the decreasing contribution that the variable would provide when added to the set of already selected variables. We can see that the variable "ForDens" is a hot candidate. It would increase the amount of explained variability from 0.320 to 0.352 (0.320 + 0.032).

To judge whether such an increase is larger than a random contribution, we can use a partial Monte Carlo permutation test. In this test, we would use the candidate variable as the only explanatory variable (so we would get an ordination model with just one canonical axis) and the already selected environmental variables (Forest, BrLeaf, and E2 in this case) as the covariables, together with any *a priori* selected covariables. If we reject the null hypothesis for that partial test, we can include that variable into the subset.

The effect of the variable tested in such context is called its conditional, **partial effect** and its value is strictly dependent on the exact selection sequence. But on the start of the forward selection process, when no environmental variable entered the selected subset yet, we can test each variable separately, to estimate its independent, **marginal effect**. This is the amount of variability in the species data that would be explained by a constrained ordination model using that variable alone as an explanatory variable. The discrepancy between the order of variables sorted based on their marginal effects and the order corresponding to a "blind" forward selection (when alway picking the best candidate) is caused by the correlations between the explanatory variables. If these variables would be completely linearly independent, both these orders would be identical.

If the primary purpose of the forward selection is to find a sufficient subset of the explanatory variables that represents the relation between the collected species and environmental data, then we have a problem with the "global" significance level related to the whole selected subset, if treated as a single entity. If we proceed by selecting the environmental variables as long as the best candidate has Type I error estimate (P) lower than some preselected significance level α , then the "collective"

Type I error probability is in fact higher than this level. We do not know how large is the Type I error probability, but we know that the upper limit is $N_c^* \alpha$, where N_c is the maximum number of tests (comparisons) made during the selection process. The appropriate adjustment of the significance threshold levels on each partial test (selecting only variables with achieved Type I error probability estimate less than α/N_c) is called **Bonferoni correction**. Here the value of N_c reprents the maximum possible number of steps during the forward selection (i.e. the number of **independent** environmental variables). Use of the Bonferoni correction is a controversial issue.

Another difficulty we might encounter during the process of the forward selection of environmental variables occurs if we have one or more factors coded as dummy variables and used as the environmental variable. The forward selection procedure treats each dummy variable as an independent predictor so we cannot evaluate contribution of the whole factor at once. This is primarily because the whole factor contributes more than one degree of freedom to the constrained ordination model (similarly, a factor with **K** levels contributes **K** - 1 degrees of freedom in a regression model). In the constrained ordination, there are **K** - 1 canonical axes needed to represent the contribution of such a factor. On the other hand, the independent treatment of the factor levels provides an opportunity to evaluate the extent of differences between the individual classes of samples defined by such a factor. This is partly analogical to the multiple comparisons procedure in the analysis of variance.

4.10. Variance partitioning procedure

In the previous section, we explained the difference between the conditional and marginal effects of individual explanatory (environmental) variables upon the species data (response variables). We also stated that the discrepancy in the importance of the explanatory variables as judged by their marginal effects and their conditional effects is caused by the correlations between those explanatory variables. Any two explanatory variables that are correlated share part of their effect exercised (in the statistical, not necessarily the causal sense) upon the species data. The shared amount of the explanatory power for a pair of explanatory variables is equal to a difference between the variable A' marginal effect and its conditional effect evaluated in addition to the effect of the variable B.

This is the basis of the so-called **variance partitioning** procedure (Borcard, Legendre & Drapeau, 1992) where we, most often, do not quantify the effects of just two explanatory variables: rather, we attempt to quantify the effects (and their overlapp) of two or more **groups** of environmental variables representing some distinct, ecologically interpretable phenomena. A separation of the variability in space and in time might serve as a typical example.

We will describe the variance partitioning procedure using the simplest example with two groups of explanatory variables (X_1 and X_2). Each group contains one or more individual environmental variables. The diagram in Figure 4-7 presents the break-down of the total variability in the species data in respect to those two groups of variables.



Figure 4-7 Partitioning of the variance in the species data into the contributions of two subsets of the environmental variables (A, B, and shared portion C) and the residual variance (D).

The area marked as D corresponds to the residual variability, i.e. the variability not explained by the ordination model including environmental variables from both the X_1 and the X_2 groups. The fraction **A** represents the partial effect of the variables from group X_1 , similarly the fraction **B** represents the partial effect of the group X_2 . The effect shared by both groups is the fraction **C**. It is clear that the amount of the variability explained by the group X_1 , when ignoring variables from the group X_2 is equal to **A**+**C**. We proceed with the estimation by using the partial analyses.

We estimate **A** from the analysis where variables from X_1 were used as the environmental variables, while variables from X_2 as the covariables. Similarly, we estimate **B** as the sum of the eigenvalues of canonical axes from analysis where X_2 acted as the environmental variables and X_1 as the covariables. We then calculate the size of **C** by subtracting sum of **A** and **B** from the amount of variability explained by an ordination model with the variables from both X_1 and X_2 acting as the explanatory variables.

5. Classification methods

The aim of classification is to obtain groups of objects (samples, species) internally homogeneous and distinct from the other groups. When the species are classified then the homogeneity means their similar ecological behavior manifested by their distributional similarity. The classification methods are usually categorized as in the Figure 5-1.



Figure 5-1 Type of classification methods

Historically, the numerical classifications were considered an objective alternative to the subjective classifications (like Zuerich-Montpellier phytosociological system). It should be noted that the results of numerical classifications are objective in the sense that the same method gives (usually) the same results. However, it should be kept in mind that the results of numerical classifications are dependent on the methodological choices.

5.1. Sample data set

The various possibilities of classification will be demonstrated using data set of 14 relevés from an altitudinal transect in Nizke Tatry Mts, Slovakia. The relevé 1 was recorded in altitude 1200 m a.s.l., relevé 14 in 1830 m a.s.l. Relevés were recorded using the Braun-Blanquet scale (r, +, 1 to 5). For calculations, the scale was converted into numbers 1 to 7 (ordinal transformation of Van der Maarel, 1979). Data were entered as a classical phytosociological table (file tatry.xls) and data on the species composition was imported using the WCanoImp into the condensed Cornell (CANOCO) format, to enable use of CANOCO and TWINSPAN, and the data were also imported into a Statistica file.

First, let us have a look on the similarity structure displayed by the DCA analysis. The following species-samples biplot was obtained (Figure 5-2).



Figure 5-2: Species-samples biplot of results of DCA of the altitudinal transect in the Tatry Mts. Species with highest weight (the most frequent ones) were selected for the display.

The figure demonstrates that there is a linear variation in the data, corresponding to the altitudinal gradient. The samples 1 to 5 are the spruce–forest samples (characterized by e.g. *Picea abies, Dryopteris dilatata, Avenella [=Deschampsia] flexuosa*), and samples (11-) 12 to 14 are typical alpine grasslands, and there is a dwarf-pine zone in between (Figure 5-2). We will now have a look how particular classification methods will divide this gradient into vegetation types, and will also demonstrate technically, how to use particular procedures.



Figure 5-3: Response curves of the important species on the elevation gradient, fitted in CanoDraw by a second order linear predictor (GLM procedure).- see chapter 11.

5.2. Non-hierarchical classification (K-means clustering)

The goal of the method is to form (predetermined number of) groups of objects (clusters); the groups should be internally homogenous and different from each other. All the groups are of the same level, there is no hierarchy. Here, we will use the K-means clustering as a representative of the non-hierarchical classifications.

For the computation, the iterative relocation procedure is applied. The procedure starts with k (desired number of clusters) clusters, and then moves the objects to minimize within-cluster variability and maximize the between-cluster variability. When the clusters are different, then the ANOVA for particular species shows significant results, so the procedure can be seen as "ANOVA in reverse", or forming groups to get the most significant ANOVA results for most species (variables).

The relocation procedure stops when no move of a object (sample) improves the criterion. You should be aware, that in this way you get local extreme, but can be never sure about the global extreme. It is advisable to start with different initial clusters and check if the results are the same for all of them.

We will use Statistica program (procedure *K-means clustering*). In Statistica, call the *Cluster Analysis* procedure and select *K-means clustering*. On the panel, select:

Variables: All - alternatively, you can select a subset and calculate the classification using tje selected species only, e.g. only the herbs

1	📇 Cluster Analysis: K-N	leans Clustering	? 2
	💁 Variables: AL	L	
	<u>C</u> luster:	Cases (rows)	Cancel
	<u>N</u> umber of clusters:	3	
	Number of iterations:	10	
	Missing <u>d</u> ata:	Casewise deleted	
	-Initial cluster center	\$	
	• So <u>r</u> t distances an	ons to <u>m</u> aximize initial between-clu d take observations at constant ir (Number of clusters) observation	ntervals
	Batch processing a	nd printing	SELECT S

Figure 5-4 Dialog box for specifying a K-means clustering in the Statistica for Windows program

In the next panel, we should first ask for *Members of each cluster & distances*. Here we learn that in the first cluster, there are samples 1 to 6, in the second samples 7 to 11, and in the third 12 to 14. That the samples in each cluster form a row is caused by the linear character of variation in our data; otherwise, we can get one group containing samples 1, 3, and 5, and the other 2 and 4. For each cluster, we see the distances to the center of corresponding cluster, e.g. for cluster 1:

Members of Cluster Number 1 (tatry.sta) and Distances from Respective Cluster Center Cluster contains 6 cases Case No. Case No. Case No. Case No. Case No. Case No. C_1 C_2 C_3 C_4 C_5 C_6 Distance 0.49637 0.577225 0.456277 0.600805 0.520278 1.017142

The samples seem to be homogeneous, with the exception of the sample 6, being an outlier (similarly, sample 10 is outlier in cluster 2). The result is in a good agreement (including the determination of outliers) with the DCA analysis.

Now, we can ask for *Cluster means and Euclidean distances*. Table of the Euclidean distances provides an information on the similarity among the individual clusters and the table of cluster means provides an information on the mean values of the species in the particular clusters.

We can see that the most similar are the clusters 1 and 2, the most dissimilar 1 and 3 (as expected).

Euclidean Distances between Clusters (tatry.sta) Distances below diagonal Squared distances above diagonal

	No. 1	No. 2	No. 3
No. 1	0	1.616661	3.178675
No. 2	1.27148	0	2.212292
No. 3	1.782884	1.487377	0

Further, the representation of species in particular clusters is clear from the means:

Cluster Means (tatry.sta)							
	Cluster	Cluster	Cluster				
	No. 1	No. 2	No. 3				
PICEABIE	5.416667	1.6	0				
SORBAUCU	1.75	1.4	0				
PINUMUGO	0.333333	6.4	1.333333				
SALISILE	0	0.8	0				
etc							

Clearly, *Picea abies* is common in the cluster 1 (spruce forests) and missing in the cluster 3 (alpine meadows), dwarf pine, *Pinus mugo*, is rare outside the middle "krumbholz" zone, etc. Useful information is provided by the *Analysis of variance:*

	Between		V	Vithin	1			signif.
	SS	df	S	S	df	l	FI	C
PICEABIE	71.68095	2	2	21.40831		11	18.41552	0.000309
SORBAUCU	6.3	2	2	23.575		11	1.469778	0.271826
PINUMUGO	107.6571	2	2	15.2		11	38.9549	1.02E-05
SALISILE	2.057143	2	2	12.8		11	0.883929	0.440566
JUNICOMM	6.547619	2	2	4.666667		11	7.716836	0.00805
VACCMYRT	0.32381	2	2	20.03333		11	0.0889	0.915588
OXALACET	16.16667	2	2	19.33333		11	4.599139	0.035353
HOMOALPI	1.414286	2	2	4.3		11	1.808971	0.209308
SOLDHUNG	18.66666	2	2	7.333336		11	13.99999	0.000948
AVENFLEX	16.89524	2	2	4.033331		11	23.03897	0.000117
CALAVILL	0.928572	2	2	40.875		11	0.124945	0.88378
GENTASCL	7.295238	2	2	5.633332		11	7.122571	0.010368
DRYODILA	8.914286	2	2	4.8		11	10.21428	0.003107
etc.								

For each species, an ANOVA comparing means in the three clusters is calculated. Note that you should not interpret the p-values as ordinary Type I error probabilities, because the clusters were selected to maximize the differences between clusters, nevertheless, the p-values provide useful lead which species differ considerably between clusters (or, on which species the classification into clusters is based). *Picea abies, Pinus mugo,* or *Soldanella hungarica* differ considerably among particular clusters, whereas *Vaccinium myrtillus* does not (it was relatively common along the whole transect).

5.3. Hierarchical classifications

In hierarchical classifications, the groups are formed that contain subgroups, so there is a hierarchy of levels. When the groups are formed from the bottom (i.e. the method starts with joining the two most similar objects), then the classifications are called **agglomerative**. When the classification starts with the whole data set, which is first divided into two groups and those are further divided, the classification is called **divisive**. The term "Cluster analysis" is often used for the agglomerative methods only.

Agglomerative hierarchical classifications (Cluster analysis)

The aim is to form a hierarchical classification (i.e. groups, containing subgroups) which is usually displayed by a **dendrogram**. The groups are formed "from the bottom", i.e. the most similar objects are first joined to form the first cluster, which is then considered an object, and the joining continues until all the objects are joined in

the final cluster, containing all the objects. The procedure has two basic steps: in the first step, the similarity matrix is calculated for all the pairs of the objects (the matrix is symmetric, and on the diagonal there are either zeroes – for dissimilarity – or the maximum possible similarity values). In the second step, the objects are clustered (joined, amalgamated) so that after each amalgamation, the newly formed group is considered to be an object, and the similarities of the remaining objects to the newly formed one are recalculated. The individual procedures (algorithms) differ in the way they recalculate the similarities.

You should be aware that there are several methodological decisions affecting the final result of classification:



Figure 5-5 Our decisions affecting the results of an agglomerative hierarchical classification

Agglomerative hierarchical classifications are readily available in most statistical packages. We will demonstrate their use in the Statistica package, however the rules are similar in most packages. Statistica, similarly to most programs, allows for a direct input of the similarity matrix. This is useful because Statistica contains only a very limited number of dis/similarity measures. Those often used in ecology (like the percentage similarity) are not included. It is possible to prepare a simple macro, e.g. in the Excel spreadsheet which will calculate the similarities and then import the similarity matrix into Statistica.⁺

In its basic form, the procedure used in the Statistica program is quite simple:

Select *Joining (tree clustering)* in the startup panel: for plot clustering, use option as in the Figure 5-6.

⁺ You need to prepare the file according to rules for similarity matrices in Statistica: a symmetric square matrix, with names of columns being the same as names of rows, and surplus four rows, containing averages and standard deviations of variables (necessary only for the correlation and covariance type), and in first column of the third row number of items compared and in the fourth row the type of the matrix: 1 =correlation, 2 =similarities, 3 =dissimilarities, and 4 =covariance.

Cluster Analysis: Joining (Tre	e Clustering)	? X
➡ Variables: ALL		E OK
Input: Raw data		Cancel
<u>C</u> luster: Cases (rows)		
<u>Amalgamation (linkage) rule:</u>	Complete Linkage	•
Dis <u>t</u> ance measure:	Euclidean distances	•
	<u>p:</u> 2 <u> </u>	
Missing <u>d</u> ata:	Casewise deleted	
<u>Batch processing and printin</u>	9	CRSES S

Figure 5-6 Dialog box for hierarchical, agglomerative clustering in Statistica for Windows

All the variables are selected (you can do the analysis with a subset of variables only). *Raw data* means that in your data file are the raw data, not the similarity matrix. We selected the *Complete Linkage* amalgamation (clustering) procedure. Here are other feasible possibilities and the decision affects the resulting dendrogram. There are so called "short hand" methods (e.g. *Single Linkage*) where the distance between the clusters is defined as the distance between the closest points in the two clusters. Those methods produce dendrograms characterized by a high "chaining". The "long hand" methods, e.g. complete linkage, where the distance of two clusters is defined as the distance of the furthest points, tend to produce compact clusters of roughly equal size (Figure 5-7).



Figure 5-7: Distance of two clusters, A and B, as defined in the single linkage (full line) and in the complete linkage (broken line).

There are many methods, usually somewhere between the two cases cited: the average linkage method used to be very popular. However, this term was used inconsistently. The term unweighted-pair groups method (UPGMA) describe the most common variant. In ecology, the "short hand" methods are usually of little use. Compare the two resulting dendrograms (Figure 5-8). The results of the complete linkage are better ecologically interpretable, there are groups formed (and, based on our external knowledge of the nature of the data, they can be judged as better reflecting the elevation gradient).

We selected the *Euclidean distance* as a measure of the sample similarity. There are not many possibilities in the Statistica (PCORD, for example, provides a much wider selection). However, the matrix with any kind of indices can be entered manually (i.e. pre-calculated in any spreadsheet). Also, one of the important decisions affecting greatly the similarity values are the standardizations and data transformations: according to our experience (Kovář & Lepš 1986), the

transformations are more influential to the resulting classification than the selection of the clustering rule.

Also note that unlike the TWINSPAN (see below) the orientation of subgroups in the dendrogram is arbitrary and, therefore, should not be interpreted.



Figure 5-8: Comparison of single linkage and complete linkage clustering of the samples. Note the higher degree of chaining (samples 14 and 10 do not belong to any cluster, but are chained to the large cluster containing all the other samples. The complete linkage results are better interpretable.

Similarly, we can calculate the cluster analysis of variables (i.e. of the species). In this case, the correlation coefficient will be probably a reasonable measure of species distributional similarity (note, that the appropriate similarity measures differ according to whether we cluster samples or species).

The results of the clustering of the species are as follows:



Figure 5-9: The classification of species. Note that the procedure distinguished reasonably the group of the alpine grassland species (*Primula minima, Salix herbacea, Carex sempervirens, Campanula alpina*) on the left side of the diagram. Similarly, on the right side of the diagram there are the species of the spruce forests.

Divisive classifications

In **divisive classifications**, the whole set of objects is divided "from the top": first, the whole data set is divided into two parts, each part is then considered separately, and is divided further. When the division is based on a single attribute (i.e. on a single species), then the classification is called **monothetic**, when based on multiple species, then the classification is **polythetic**. The significance of monothetic methods is mostly a historical one. The classical "association analysis" was a monothetic method. The advantage of the divisive methods is that each division is accompanied by a rule used for the division, e.g. by a set of species typical for either part of the dichotomy.

The most popular divisive method (and program of the same name) TWINSPAN (Two Way INdicator SPecies ANalysis) was partially inspired be the classificatory methods of the classical phytosociology (use of indicators for definition of vegetation types). The idea of an indicator species is basically a qualitative one. Consequently, the methods works with qualitative data only. In order not to loose the information and the quantity of species, the concept of pseudo-species and pseudo-species cut levels was introduced. Each species can be present as several pseudospecies, according to its quantity in the sample. The pseudo-species is present, if the species quantity exceeds the corresponding cut level. Imagine, that we selected pseudo-species cut levels: 0, 1, 5, 20. Then the original table is translated to the table used by the TWINSPAN as follows:

	Species	Sample 1	Sample 2
Original	Cirsium oleraceum	0	1
table	Glechoma hederacea	6	0
	Juncus tenuis	15	25
Table with pseudo-	Cirsoler1	0	1
species used in	Glechede1	1	0
TWINSPAN	Glechede2	1	0
	Junctenu1	1	1
	Junctenu2	1	1
	Junctenu3	1	1
	Junctenu4	0	1

In this way, the quantitative data are translated into qualitative (presenceabsence) data.

In TWINSPAN, the dichotomy (division) is constructed on the basis of a correspondence analysis ordination. The ordination is calculated and the samples are divided into the left (negative) side and the right (positive) side of the dichotomy according to their score on the first CA axis. The axis is divided at the center of gravity (centroid). However, usually there are many samples at the center of gravity, Consequently, many samples are close to the border, and their classification would depend on many random factors. Therefore the new ordinations are constructed, which give a higher weight to the "preferrentials", i.e. species preferring one or the other side of the dichotomy. The algorithm is rather complicated, but the aim is to get polarized ordinations, i.e. ordinations where most of the samples are not close to the center of the gravity. In this way, the classification of the samples is not based so much on the species common in both parts of the dichotomy, but mostly on the species typical for one part of the dichotomy, and consequently (in concordance with the phytosociological tradition) these species can be considered to be good indicators of the ecological conditions. In the first division, the polarity (i.e. which part of the dichotomy will be negative and which positive) is arbitrary. In the subsequent divisions, the polarity is determined according to the similarity of the parts of the dichotomy to the "sister" group in the higher division. For example, in the dendrogram in the Figure 5-10, the group 01 is more similar to the group 1 than the group 00. Thanks to this, the samples are ordered in the final table, and we get table that is similar to the ordered phytosociological table.



Figure 5-10 A sample splits in a TWINSPAN classification

Also, the "rule", i.e. the set of species which were important for particular division, is printed by the program at each step; this highly increases the interpretability of the final results. The classification of samples is complemented with the classification of species and the final table is based on this two-way classification.

Analysis of the Tatry samples

In the following text we will demonstrate the use of TWINSPAN for the analysis of a set of 14 relevés from an altitudinal transect in Nizke Tatry Mts, Slovakia. TWINSPAN is useful for large data set, the small data set is used here for simplicity. We used the file tatry.spe, i.e. the file imported using the WCanoImp into the condensed Cornell (CANOCO) format. In the example, we asked for the long output, to show what information can be obtained from the program (usually, only short output is requested – even the short output is pretty long).

First, the headings are printed and the program lists the options selected. The important ones are

Cut levels: .00 1.00 2.00 3.00 4.00 5.00

Because the data were converted into numeric values using the ordinal transformation, there is no reason to further "downweight" the higher values. When the data are in the form of cover estimated, it is reasonable to use the default cut levels, i.e. 0 2 5 The values 0 1 10 100 1000 give results corresponding to a logarithmic transformation and are useful when the data are numbers of individuals, differing in order of magnitude.

From further options note the following (all are defaults here):

1. Minimum group size for division:52. Maximum number of indicators per division:73. Maximum number of species in final tabulation:1004. Maximum level of divisions:6

The option 1 means that groups containing less than 5 relevés are "final", i.e. are not further divided. For small data sets, it might be reasonable to decrease this value to 4. **2.** Number of indicators per division: usually, the default is fine. **3.** If you have more than 100 species, the 100 most common species will be included in the final tabulation. You can increase the value, if you need. **4.** This is another way how to control when to stop divisions (control by the size of the group – in option **1.** is better solution; for data set of reasonable size, however, 6 is usually enough).

```
Then, the first division is described:

DIVISION 1 (N= 14) I.E. GROUP *

Eigenvalue .565 at iteration 1

INDICATORS, together with their SIGN

Oreo Dist1(+)
```

The indicator for the first division is *Oreochloa disticha* (the 1 at the end means that the first pseudospecies cut level was used, i.e. presence of the species is enough to be considered indicator present).

```
Maximum indicator score for negative group 0
                                               Minimum indicator score for positive
group
       1
 Items in NEGATIVE group
                        2 (N=
                                10)
                                             i.e. group *0
Samp0001 Samp0002 Samp0003 Samp0004 Samp0005 Samp0006 Samp0007 Samp0008
Samp0009 Samp0010
 BORDERLINE negatives
                             1)
                       (N=
 Samp0009
 Items in POSITIVE group 3 (N= 4)
                                             i.e. group *1
 Samp0011 Samp0012 Samp0013 Samp0014
```

The division of samples. Note the warning for sample 9, that is was on the border between the two groups (this warning appears only when you ask for long output).

Now, the species preferring one side of the dichotomy (preferentials) are listed, with number of occurrences in each of the groups (e.g. *Picea abies* was present in 7 samples from the negative group and 1 sample from the positive group. Note, that preferentials are determined with respect to the number of samples in each group, and also for each pseudospecies cut level separately). Preferentials are provided in the long output only (it is a long listing, a large part is deleted here also).

NEGATIVE PREFERENTIALS Pice Abiel(7, 1) Sorb Aucul(7, 0) Oxal Acetl(7, 0) Sold Hung1(5, 0) Aven Flex1(10, 1) Gent Ascll(8, 0) Dryo Dilal(8, 0) Pheg Dryol(6, 0) Pren Purpl(2, 0) Poly Vertl(3, 0) SoAu cJu 1(3, 0) Luzu Pilol(2, 0) etc POSITIVE PREFERENTIALS Juni Comml(0, 2) Ligu Mutel(4, 4) Sold Carpl(2, 4) Ranu Montl(1, 2) Hupe Selal(3, 3) Geun Montl(2 21 0, 4) etc. NON-PREFERENTIALS Pinu Mugol(5, 2) Vacc Myrtl(10, 4) Homo Alpil(10, 4) Cala Villl(8, 3) Rume Arifl(4, 1) Vera Lobel(3) Pinu Mugo2(5, 2) Vacc Myrt2(10, 4) Homo Alpi2(10, 4) Cala Vill2(8, 3) Rume Arif2(4, 1) Vera Lobe2(5, 3) etc.

End of level 1

We can start now to draw the dendrogram. The part that is clear now is:



Figure 5-11 The first division in the TWINSPAN example

The positive group is final, because it is smaller than 5, which is the minimum size for division. The next level follows (without the preferentials):

DIVISION 2 (N= 10) I.E. GROUP *0 Eigenvalue .344 at iteration 1 INDICATORS, together with their SIGN Pice Abiel(-) Maximum indicator score for negative group -1 positive group 0 Items in NEGATIVE group 4 (N= 7) i.e. group *00

Samp0001	Samp0002	Samp0003	Samp0004	Samp0005	Samp0006	Samp0007
	POSITIVE g Samp0009	-	(N= 3)	i.	e. group *	01
			I.E. GRO oo few item			
End of	level 2					

In a similar way, we can continue in the construction of the dendrogram (one indicator per division is rather an exception, not a rule).



Figure 5-12 The second-level division in the TWINSPAN example

The next level is:

DIVISION 4 (N= 7) I.E. GROUP *00 Eigenvalue .279 at iteration 1 INDICATORS, together with their SIGN Pinu Mugol(+) Maximum indicator score for negative group 0 Minimum indicator score for positive group 1 Items in NEGATIVE group 8 (N= 5) i.e. group *000 Samp0001 Samp0002 Samp0003 Samp0004 Samp0005 Items in POSITIVE group 9 (N= 2) i.e. group *001 Samp0006 Samp0007

Note the meaning of indicator species *Pinus mugo* here. It is an indicator within the group 00 (containing 7 samples), where it is present just in two of the samples, 6 and 7 (forming positive group 001). However, it is also common the in samples 8, 9, 10, 11 and 12. So, the indicators are determined and should be interpreted just for a particular division, not for the total group. The group 001 is characterized by the presence of *Pinus mugo* in contrast to group 000, and not within the whole data set. Also, here is clear the advantage of TWINSPAN, where the orientation of groups (i.e. which group will be negative and which positive) depends on which of the groups is more similar to the group 01 (and this group, among others, has *Pinus mugo* in all its samples).

We will not show further divisions, (the group 000 contains 5 samples and still can be divided further), and will finish the dendrogram at this level:



Figure 5-13 Final stage of the TWINSPAN classification example

In a similar way, we can construct the classification (dendrogram) for the species. Note that the indicators are those used in the divisions, and often are limited in number. If you want to have characterized the division by more species, you should use preferentials.

Finally, the table is printed. It resembles the classical ordered phytosociological table and is accompanied by the classification of both samples and species.

			SSSSSSSSSSSSSS	
			aaaaaaaaaaaaaaa	
			mmmmmmmmmmmmm	
			ppppppppppppppppppppppppppppppppppppppp	
			000000000000000	
			000000000000000	
			0000000011111	
			21345678901234	
4	Sali	SIle	5	0000
29	Anth	Alpi	3	0000
30	Нуре	Macu	23	0000
31	Rubu	Idae	23	0000
28	Aden	Alli	22	0001
1	Pice	Abie	66666655	001000
7	Oxal	Acet	55344-43	001001
9	Sold	Hung	43444	001001
18	Luzu	Pilo	2-2	001001
20	Luzu	Sylv	3243	001001
12	Gent	Ascl	23333333	001010
14	Pheg	Dryo	4333-33	001010

$\begin{array}{cccccccccccccccccccccccccccccccccccc$	Pren Poly Stel Sorb Aven Dryo SoAu Athy Vacc Homo Cala Rume Pinu Vera Hupe Soli Vacc Gent Luzu Ligu Sold Ranu Geun Juni Puls Oreo Fest Care Junc Luzu Hier Care Pinu	Purp Vert Nemo Aucu Flex Dila CJu Dist Myrt Alpi Vill Arif Mugo Lobe Sela Virg Viti Punc Luzu Mute Carp Mont Comm Alba Dist Supi Alpi Supi Alpi Supi Alpi Punc Luzu Mont Carp Mont Comm Alba Dist Supi Alpi Supi Alpi Punc Luzu Mont Comm Alba Dist Miti Alpi Not Carp Mont Comm Alba Dist Miti Alpi Not Carp Mont Comm Alba Dist Miti Alpi Punc Luzu Mute Carp Mont Comm Alba Dist Alpi Alpi Not Comm Alba Dist Miti Alpi Not Comm Alba Dist Miti Alpi Alpi Alpi Not Comm Alba Dist Miti Alpi Alpi Mont Comm Alba Dist Miti Alpi Alpi Alpi Alpi Mont Comm Alba Dist Miti Alpi Alpi Alpi Alpi Alpi Alpi Alpi Alp	$\begin{array}{c} 23$	001011 001011 00100 00110 00110 00110 00110 00110 01 01
			00000000001111	

Note that from the bottom three lines, the three levels of division and memberships of all the samples in corresponding groups can be read: e.g., samples 11 to 14 are in the group 1 (which is not further divided).

6. Visualization of multivariate data with CanoDraw 3.1 and CanoPost 1.0 for Windows

The primary device for presenting results of an ordination model is the ordination diagram. The contents of an ordination diagram can be used to approximate the species data table, matrix of the distances between individual samples and/or matrix of the correlations or dissimilarities between the individual species. In an ordination analysis with environmental variables, we can use an ordination diagram to approximate contents of the environmental data table, correlation between species and environmental variables, correlation among the environmental variables etc. What can we read from an ordination diagram is summarized in the following two sections, separately for the linear and the unimodal ordination methods.

6.1. What can we read from the ordination diagrams: Linear methods

An ordination diagram based on a linear ordination method (PCA or RDA) might display scores for samples (represented by points), species (represented by arrows), quantitative environmental variables (represented by arrows), and nominal dummy variables (represented by points - centroids - and corresponding to the individual levels of a factor variable). The following Table 6-1 summarizes (after Ter Braak, 1994) what can be deduced from the ordination diagrams based on these scores.

	Scaling 1	Scaling 2
Compared entities	Focus on sample distances	Focus on species correlations
species X samples	(fitted) abundance values in species data	(fitted) abundance values in species data
samples X samples	Euclidean distances among samples	XXX
species X species	XXX	linear correlations among species
species X env.variables	linear correlations between species and environmental variables	linear correlations between species and environmental variables
samples X env.variables	XXX	values of in environmental variablesdata
env.variables X env.variables	marginal effects of environmental variables on ordination scores	correlations between among environmental variables
species X nominal env.vars.	mean species abundances within classes	mean species abundances within classes
samples X nominal env.vars.	membership of samples in the classes	membership of samples in the classes
nominal env.vars. X nom. env.vars.	Euclidean distances between sample classes	XXX
env.vars. X nominal env.vars.	XXX	averages of environmental variables' averages within classes

Table 6-1
If we project sample points perpendicularly onto a species' arrow, we approximate the ordering of that species values across the projected samples. If we use the sample scores that are linear combination of the environmental variables (**SamE** scores, typically in RDA), we approximate the **fitted**, not the observed values of those abundances. This interpretation holds true for both kinds of scaling. If a centering by species was performed, a sample point projecting onto the beginning of the coordinate system (perpendicularly to a species arrow) is predicted to have an average value of the corresponding species. The samples projecting further from the zero point in the direction pointed by the arrow are predicted to have above-theaverage abundances, while the sample points projecting onto the opposite direction are predicted to have below-the-average values.

Only in the scaling focused on the sample distances does the distance between the sample points approximate their dissimilarity, expressed using the Euclidean distance.

Only in the scaling focused on the species correlations do the angles among the species arrows approximate the (linear) correlation coefficients among the species. The approximated correlation between two variables is equal to the cosine of the angle between the corresponding arrows: arrows pointing in the same direction correspond to the species that are predicted to have a large positive correlation, the species with a large negative correlation are predicted to have arrows pointing in the opposite direction. Similarly, the pair of species with the arrows meeting at a right angle are predicted to have a low (linear) correlation.

We can apply a similar approximation method when comparing the species and environmental variables arrows. For example, if an arrow for a quantitative environmental variable points in a similar direction as a species arrow, the values of that species are predicted to be positively correlated with the values of that environmental variable. This interpretation holds true for both types of scaling of ordination scores.

We can perpendicularly project the sample points onto an arrow of an environmental variable. This gives us the approximate ordering of the samples in the order of increasing value of that environmental variable (if we proceed towards the arrow tip and beyond it). The environmental variables (as well as the covariables) are always centered (and standardized) before the ordination model is fitted, so similarly to projecting the sample points on the species arrows, a projection point near the zero point (the coordinate system origin) corresponds to an average value of that particular environmental variable in that sample.

The angle among the arrows of the environmental variables can be used to approximate correlation among those variables in the scaling focused on the species correlations. Note however, that this approximation is not as good as that we would achieve if analyzing environmental data table as the primary data in a PCA. If the scaling is focused on the intersample distances, we can interpret each arrow independently as pointing in the direction the sample scores would shift with an increase of that environmental variable' values. The length of the arrows allows us to compare the size of that effect across the environmental variables (note again that all the environmental variables enter the analysis with a zero average and a unit variance). CANOCO output allows a different interpretation of the scores of the nominal environmental variables. These are usually the dummy (indicator) variables with only 0 and 1 values, created by the expansion of a factor variable. These nominal environmental variables are represented by the points that are centroids of sample scores for the samples that have value of 1 for that particular dummy variable. We can view the original factor variable as a classification variable and then the individual nominal (dummy) variables correspond to the individual sample classes. So we can say that the centroid score for that nominal variable represents the average of the scores of samples belonging to that class.

If we project the centroids of the nominal environmental variables onto a species arrow, we can approximate the average values of that species in the particular classes of the samples. Similarly, the distance between the centroids of nominal variables approximates (in the scaling focused on the intersample distances) the dissimilarity of their species composition, expressed using the Euclidean distance.

In both types of scaling, the distance between the centroids of individual sample classes and a particular sample point allows us to predict membership of that sample. A sample belongs with the highest probability into the class with its centroid most close to that sample point.

If we project centroids of the nominal environmental variables on an arrow of a quantitative environmental variable, we can deduce the approximate ordering of that variable' average values in the individual sample classes.

6.2. What can we read from the ordination diagrams: Unimodal methods

The interpretation of the ordination diagrams originating from the unimodal ordination models is summarized in the Table 5-2. It has many similarities with the interpretation we discussed in detail for the linear ordination model in the preceding section, so we will point the reader to that section as needed.

The main difference in interpreting ordination diagrams from the linear and from the unimodal ordination methods lays in the different model of the species response along the constructed gradients (ordination axes). While a linear (monotone) change was assumed in the preceding section, here (many of) the species are assumed to have an optimum position along each of the ordination axes with their abundances (or probability of occurence for presence-absence data) decreasing symetrically in all directions from that point. The estimated position of that species optimum is displayed as the species score, i.e. as a point. These positions are calculated as the weighted averages of the sample positions with weights being that species' abundances in the respective samples.

Another important difference is that the dissimilarity between samples is based on the chi-square metrics, implying that any two samples with identical **relative** abundances (say two samples with three species present and their values being 1 2 1 and 10 20 10, respectively) are judged to be identical by an unimodal ordination method. The dissimilarity of the distribution of different species is judged using the same kind of metric, being applied to a transposed primary data matrix.

	Scaling 1	Scaling 2		
Compared entities	Focus on sample distances and Hill's scaling	Focus on species distances and biplot scaling		
species X samples	relative abundances of the species data table	(fitted) relative abundances of the species data table		
samples X samples	turnover distances among samples	χ^2 distances among samples (if λ s comparable)		
species X species	xxx	χ^2 distances among species distributions (from fitted abundances)		
species X env.variables	weighted averages - the species optima in respect to particular environmental variable	weighted averages - the species optima in respect to particular environmental variable		
samples X env.variables	xxx	values of the env.ironmental variables in the samples		
env.variables X env.variables	marginal effects of env. ironmental variables	correlations of among the env. ironmental variables		
species X nominal env.vars.	relative total abundances in sample classes	relative total abundances in sample classes		
samples X nominal env.vars.	membership of samples in the classes	membership of samples in the classes		
nominal env.vars. X nominal env.vars.	turnover distances between the sample classes	χ^2 distances (if λ s comparable) between the sample classes		
env.variables X nominal env.vars.	xxx	averages of environmental variables within the sample classes		

Table 6-2

The mutual position of sample and species points allows us to approximate relative abundances in the species data table. The species scores are nearby the sample points they occur in, with the highest relative abundance and similarly, the sample points are scattered near the positions of species that tend to occur in those samples. This way of interpretation is called the **centroid principle**. A more quantitative form of it works directly with the distances between the points. If we order the samples according to their distance to a point for a particular species, this ordering approximates ordering based on the decreasing relative abundance of that species.

For shorter gradient lengths, we can interpret the positions of species and samples in the ordination plot using the **biplot rule**, similarly to the interpretation used in the ordination diagrams based on the linear ordination methods. We simply connect the species points with the origin of the coordinate system and project the sample points perpendicularly on this line.

The distance among the sample points approximates the chi-square distances between the samples in the biplot scaling with the focus on species, but only if the ordination axes used in the ordination diagram explain a similar amount of variability (if they have comparable eigenvalues).

If we use Hill's scaling with a focus on the sample distances, then the distance between the samples is scaled in the "species turnover" units (also named **SD units**,

from the "standard deviation" of the species response curves) that correspond to the labelling of the ordination axes. The samples that are 4 units (or more) apart have a very low probability of sharing any species. There is a "half change" in the species composition along one SD unit.

The distance between the species points in the biplot scaling (with a focus on the species distances) approximates the chi-square distance between the species distributions.

If we project species points onto an arrow of a quantitative environmental variable, we get an approximate ordering of those species' optima in respect to that environmental variable. Similarly, we can project sample points on the arrow for a quantitative environmental variable to approximate values in the environmental data table, but only if the biplot scaling was used.

Interpretation of relation among environmental variables arrows (either using the angle between the arrows or comparing relative directions and size of the impact) is similar to that in the linear methods, described in the preceding section.

For centroids of nominal environmental variables (see previous section for explanation of their meaning), we can use the distance between species points and those centroids to approximate relative **total** abundances of the species in the samples of that class (summed over all the samples in the particular class).

Comparison between sample points and centroids of nominal variables and between the centroids and the arrows for the quantitative environmental variables proceeds as in the linear models, described in the previous section.

The distance among the particular centroids of nominal environmental variables is to be interpreted similarly as if measuring distance (dissimilarity) between sample points. But the distance refers to the average chi-square (or turnover, depending on the scaling) distances between the samples of the corresponding two sample classes.

6.3. Regression models in CanoDraw

We use the program CanoDraw to create ordination diagrams with the ordination scores of a selected combination of entities (e.g. samples and species). But beside that, this program can be used as a tool to explore the multivariate data in the framework of the ordination space, to review the trends indicated by the ordination diagrams, and to check the assumptions made by the ordination model we used in CANOCO. These facilities are discussed in this and the following sections. The technical aspects of its use are discussed in more detail in the CanoDraw 3.0 User's Guide (Šmilauer 1992).

Majority of the methods for visual exploration is concentrated in the CanoDraw submenu **Attributes**. From there, we can select for example a plot displaying values of a particular environmental variable in the ordination space. The actual value of that variable in a sample is displayed at its position using a symbol of the size proportional to that value. If we expect the value of the environmental variable to change monotonically (or even linearly, as in the constrained ordination model) across the ordination (sub)space, we should recognize such a pattern from this **symbol plot**. But often, plotting individual sample values for even medium-sized

data set does not allow for an efficient abstraction of the revealed pattern. But we can replace the original data points with a simpler statistical model. CanoDraw provides three broad classes of regression models we can use interchangeably in the attribute plots, in addition to the symbol plots.

Generalized linear models (GLM) represent an extension of the traditional linear regression model, accommodating for various distributional properties of the response variable. More information can be found in the section 11.3 of this book. CanoDraw allows us to fit GLMs with one of the four available distribution families (with the assumed Poisson, binomial, Gamma or Gaussian statistical distribution of the response variable). There is no choice of the link function - the canonical link functions for the selected distribution family are used. We can specify the systematic part of the model either exactly by fitting so called **fixed model** (so we can request, for example, to fit second order polynomials of the predictor(s)) or we can let the CanoDraw to select the model complexity using a stepwise selection based on the analysis of deviance tests. A feature specific to the CanoDraw implementation of the GLMs is that if the quadratic polynomial model with one predictor is fitted with the assumed Poisson distribution, CanoDraw recognizes this to represent an unimodal model of the species response and estimates the response curve optimum value and its width (using the tolerance parameter), if that is possible (see Ter Braak & Looman, 1986).

Generalized loess smoother is an extension of the now-classical loess smoother (Cleveland & Devlin, 1988), which is a locally weighted regression method. The difference between the standard loess smoother and the one implemented here is that CanoDraw uses generalized linear model plugged into the smoother algorithm. This way, the fitted response surfaces are more appropriate for the binary data than those from the classical loess method.

Generalized kriging is a regression method derived from an ordinary leastsquares regression model but taking into account the (modelled) spatial autocorrelation between samples collected in the space. The generalized kriging (also called the *universal kriging* by others) estimates, in addition to the standard method, the linear or polynomial trend in the response variable values across the samples' space. The predictors used in the kriging are assumed to be the spatial coordinates of the samples. This method might be appropriate not only if fitting species abundances or environmental variable values onto the spatial coordinates of samples (which is also possible with CanoDraw 3.x), but for fitting those into the ordination space, as there the sample scores are also cross-correlated.

Whatever regression model we choose, the results are summarized either by a fitted line or curve (in the case we have a single predictor, as if plotting the values of an environmental variable against the scores on one of the ordination axes) or by a contour plot, when there are two predictors (most often the sample coordinates on two ordination axes). CanoDraw allows us to modify the proposed levels of the fitted response variable, to be drawn in the contour plot.

6.4. Ordination Diagnostics

The diagrams (ordination diagrams and attribute plots) summarized in the preceding sections can be used to check how far the data analyzed with an ordination method fulfill the assumptions of the underlying model.

The first obvious check is that concerning our assumption about the shape of the species response along the ordination axes, representing the "recovered" gradients of the community composition change (with an underlaying environmental interpretation in the case of constrained ordination methods). Both in the case of the linear and of the unimodal ordination methods, we should probably let the CanoDraw select the regression model specification (i.e. null model vs. linear model vs. polynomial regression model), in the case of the generalized linear models. Note, however, that the restricted family of the polynomial specification of the response curves excludes the possibility to find that the response curve is, in fact, assymetric or that it has a bimodal shape (but we might interpret in this way the results where a second order polynomial with a minimum in the middle of the gradient is selected). The loess smoother provides a more general family of regression models with no particular assumption about the shape of the fitted curve.

Another assumptions of the constrained ordination methods (both RDA and CCA) is that the canonical ordination axes are the **linear** combinations of the submitted environmental variables. For a quantitative environmental variable, we can plot its value against the sample scores on one of the ordination axes to see whether the change along the ordination axis is the linear one. While this is often the case and it is partly a consequence of the definition of the sample scores (at least in the constrained ordination methods), we can sometimes find a monotonic change in the values, but far from the linear one. This often corresponds to a non-uniform response of community composition along the gradients such as the successional time or the altitude and it might suggest an appropriate transformation (for example a log transformation) for such explanatory variables.

CANOCO manual refers to the **ordination diagnostics** in a slightly different way, namely it means the statistics calculated with the ordination method. These statistics describe, in general terms, how well are the properties of individual samples and of individual species characterized by the fitted ordination model. This statistics is named **CFit** in the case of species (see Ter Braak et Šmilauer 1998, p. 174) and can be used in the CanoDraw program to select which species occur in the ordination diagram. Additionally, we can plot these values into the ordination space.

6.5. T-value biplot interpretation

T-value biplot is a diagram containing lines for the species and the environmental variables. It is primarily used to reveal statistically significant pairwise relations between species and environmental variables (i.e. which species depend on which environmental variables). We interpret the T-value biplot using the **biplot rule** (by means of the perpendicular projections) to approximate a table of T-values of the regression coefficients for multiple (weighted) regressions of the individual species values onto all the environmental variables.

To interpret the plot, we project tips of the environmental variables arrows onto a line overlaying the arrow of a particular species. The resulting projection points approximate the T values of the regression coefficients of the individual environmental variables, when used as predictors for values of that species. If the environmental variables' arrows tips project on the line further from the origin than is the species' arrowhead, we deduce that the corresponding regression coefficient has a value larger than 2.0. This holds true also for projection onto line segment going in the opposite direction, with the T-value being less than -2.0. Projection points between these limits (including the coordinate system origin) correspond to T-values between -2 and +2, with the predicted value of a T statistics being 0.0 for the coordinate origin' position.

If we want to find which species react significantly **positively** to a particular environmental variable, we can draw a circle with its center in the half way from the arrowhead for that variable and the origin of the coordinate system and its diameter being equal to the length of that variable' arrow. Species lines that end in that circle have a positive regression coefficient for that environmental variable with the corresponding T-value larger than 2.0. We can draw a similar circle in the opposite direction, corresponding to a significant negative regression coefficient.

An additional information on the T-value biplots can be found in Ter Braak & Šmilauer (1998), pp. 177 - $180.^+$

⁺ The T-value plots created by the CanoDraw 3.x program have two problems: first, for the lines representing the environmental variables, only the solid part is correct. The dashed segments should be removed in the CanoPost program. But also the default range of axes is not correct. Before creating the diagram (on the "virtual device" representing the PostScript output), we should use the popup menu to **Zoom-out**. This creates a range of axes more appropriate for the T-value plot.

7. Case study 1: Separating the effects of explanatory variables

7.1. Introduction

In many cases, we need to separate the effects of several explanatory variables, even when the explanatory variables are correlated. The example below comes from a field fertilization experiment (Pyšek and Lepš 1991). A barley field was fertilized with three nitrogen fertilizers (ammonium sulphate, calcium-ammonium nitrate, and liquid urea) and two different total nitrogen doses. For practical reasons, the experiment was not established in a correct experimental design; i.e. plots are pseudo-replications. The experiment was designed by hydrologists to assess the nutrient runoff and, consequently, smaller plots were not usable. We used this large-plot experimental setup.^{*} In 122 plots, species composition of weeds was characterized by classical Braun-Blanquet relevés (for the calculations, the ordinal transformation was used – i.e. numbers 1 to 7 were substituted for grades of the Br.-Bl. scale: r, +, 1, ..., 5). The barley cover was estimated in all the relevés.

We expected that the weed community was influenced both directly by the fertilizers and indirectly through the effect of crop competition. Based on experimental manipulations we can assess the overall fertilizer effect. However, barley cover was highly correlated with the fertilizer dose. As barley cover was not manipulated, there was no direct evidence of the effect of barley cover on the weed communities. However, the data enable us to partially separate the direct effects of fertilization from indirect effects of barley competition. This is done in a way similar to the way we separate the effects of correlated predictors on the univariate response in a multiple regression. The separation can be done using the variable of interest as an explanatory (environmental) variable and the others as covariables.

7.2. Data

For simplicity, we will work with simplified data, ignoring fertilizer type, and taking into account only the total fertilizer dose. Data are in the files fertil.spe (relevés) and fertil.env (fertilizer dose and barley cover), or in Excel files fertspe.xls and fertenv.xls (so that we can try creating our own files for CANOCO using WCanoImp). The dose of the fertilizer is 0 for unfertilized plots, 1 for 70 kg of N/ha and 2 for 140 kg N/ha.

7.3. Data analysis

We recommend the following steps:

1. Calculate the unconstrained ordination, i.e. DCA. This shows total variability by the length of the axes, which gives the total heterogeneity of the vegetation data. The length of the first axis in Hill's scaling is 3.8, which is in the "gray zone", where both linear and unimodal methods should perform reasonably well. Then use the passive analysis of the environmental data (the environmental data are projected on the ordination axes *ex post*).

^{*} We will just forget this drawback in the following text

2. Calculate a constrained ordination with all of the environmental variables available. The length of the gradient from the first analysis serves as a lead for selecting between CCA and RDA. Another lead should be whether we expect the majority of weed species to respond to fertilization and/or crop cover in a linear way or whether we expect some (all) species to have an optimum on this gradient. Roughly, if the species change their proportions on the gradient then the linear approximation is correct and if we expect qualitative changes in the species composition (many species appearing and disappearing) then the weighted averaging is better.

In the paper (Pyšek & Lepš 1991) we used CCA. In this example, we will use RDA. RDA enables use of standardized/non-standardized analyses; by standardization by samples we can separate differences in total cover from the differences in species composition. Use both a RDA (with fertilizer dose and barley cover are environmental variables) nonstandardized by samples (the results reflect **both** the differences in total cover **and** differences in relative species composition) and a RDA standardized by samples (the results reflect **only** differences in relative representation of particular species). However, you should note that the interpretation of standardized data based on species data values estimated on an ordinal scale may be problematic.

Test the significance by a Monte Carlo permutation test - unconstrained permutations. This is a test of the H_0 : there is no effect of environmental variables on species composition, i.e. the effect of both variables is zero. The rejection of the null hypothesis means that at least one of the variables has some effect on the species composition of the plant community. The meaning of this test is analogous to an overall ANOVA on the total model in a multivariate regression.

3. Calculate two separate partial constrained ordinations. In the first, use the fertilizer dose as the environmental variable and barley cover as a covariable. The analysis will reflect the effect of fertilization which cannot be explained by the effect caused by the increased barley cover.

In the second analysis, use barley cover as the environmental variable and fertilizer dose as a covariable. This analysis reflects the variability caused by the differences in the barley cover which cannot be attributed to the fertilization effects. Use both standardized and non-standardized by samples' analyses. In the Monte Carlo permutation test, it is possible to use the constrained permutations (permutations within blocks) when the dose is used as a covariable: because the dose has three separate values and the plots with the same value of the dose can be considered to form a block. The meaning of these tests is analogous to the test of significance of particular regression coefficients in the multiple regression.

Note that it may sometimes happen (not in this data set) that the analysis with both variables used as explanatory variables is (highly) significant, whereas none of the analyses with one of the variables used as an explanatory variable and the other as a covariable is significant. This case is equivalent to the situation in multiple linear regression where the ANOVA on the total regression model is significant and none of the regression coefficients differ significantly from zero. This happens when the predictors are highly correlated, we are then able to say that the predictors together explain a significant portion of total variability, but we are not able to say which of them is the important one. 4. Variation partitioning (see the section 4.10). We can be interested to know what part of the variability can be explained solely by the dose and what part solely by the cover. As the two variables are highly correlated, it is natural that there is some part of the variability, for which it is possible to say that it is caused by one of the two variables, but it is impossible to say by which one. We can do the variation partitioning using the following procedure:

1. Calculate the RDA with both variables being explanatory variables. We will get:

Axes variance		1	2	3	4 Total
Eigenvalues	:	.097	.046	.200	.131
1.000					
Species-environment correlations	:	.705	.513	.000	.000
Cumulative percentage variance					
Cumulative percentage variance					
of species data	:	9.7	14.3	34.4	47.5
of species-environment relation	on:	67.8	100.0	.0	.0
Sum of all unconstrained eigenval	ues				
1.000					
Sum of all canonical eigenval	ues				
.143					

There is a mistake in CANOCO output, the sum of all unconstrained eigenvalues should read sum of all eigenvalues.

The results show that 14.3% of the total variation in species data can be explained by both variables together.

2. Now calculate partial analysis, with cover being explanatory (environmental) variable and dose covariable:

Axes variance		1	2	3	4 Total
Eigenvalues 1.000	:	.074	.200	.131	.102
Species-environment correlations Cumulative percentage variance	:	.590	.000	.000	.000
of species data of species-environment relation	: on:	8.0 100.0	29.5 .0	• •	54.5 .0
Sum of all unconstrained eigenval	ues				
Sum of all canonical eigenval .074	ues				

The sum of all unconstrained eigenvalues is after fitting covariables Percentages are taken with respect to residual variances i.e. variances after fitting covariables

This shows that cover explains 7.4% of the total variability in species data which cannot be explained by the dose. Note, that we do not use the percentage presented in the cumulative percentage variance of species data (8%), because it is taken with respect to the residual variation (i.e. after fitting the covariable), but directly the corresponding canonical eigenvalue.

3. In a similar manner, we now calculate partial ordination with dose being the explanatory variable and cover a covariable:

Axes variance		1	2	3	4 Total
Eigenvalues 1.000	:	.047	.200	.131	.102
Species-environment correlations Cumulative percentage variance	:	.500	.000	.000	.000
of species data of species-environment relatio	: n:	5.2 100.0	27.4 .0		
Sum of all unconstrained eigenval .904	ues				
Sum of all canonical eigenval .047	ues				

The dose explains 4.7% of species variability that cannot be accounted for by the cover.

Now we can calculate:

Total explained variability is 14.3%. Of those, 7.4% can be explained solely by the effect of barley cover, 4.7% by the effect of fertilizer dose; this gives together 12.1%. For the remaining 2.2% it is impossible to decide which of the environmental variables is responsible for their explanation.

The variation partitioning can be displayed by the diagram:



8. Case study 2: Evaluation of experiments in the randomized complete blocks

8.1. Introduction

For an univariate response (e.g. number of species, total biomass) are the results of experiments set in randomized complete blocks evaluated using the two-way ANOVA without interactions (the interaction mean square (MS) is used as an error term - denominator in calculation of F-statistics). In the following tutorial, we will use CANOCO to evaluate in a similar way the community response (i.e. a multivariate response: species composition of the community). In this example, the experiment was established in four randomized complete blocks, the treatment had four levels and the response was measured once in time. The experiment is described in full by Špačková, Kotorová & Lepš (1998), here is the description of the experiment slightly simplified.

Experiment was established in March 1994, shortly after snow melt, in four randomized complete blocks. Each block contained four treatments, i.e. (1) removal of litter, (2) removal of litter and bryophytes, (3) removal of the dominant species *Nardus stricta*, and a (4) control plot where the vegetation remained undisturbed. Each plot was 1x1 m square. The original cover of *Nardus stricta* was about 25%. Its removal was very successful with nearly no re-growth. The removal in the spring causes only minor soil disturbance that was not visible in the summer.

8.2. Data

In each 1 m² plot adult plants and bryophytes cover was visually estimated in August 1994. At that time, a new square (0.5 x 0.5 m) was marked out in the center of each plot and divided into 25 0.1 x 0.1 m subplots. In each subplot adult plant cover and numbers of seedlings were recorded. In the following example, we will use the totals of seedlings in the 0.5 x 0.5 m plots. Data are in CANOCO format file seedl.spe or in Excel file seedlspe.xls.

8.3. Data analysis

Question: Is it necessary to use a multivariate method? Would not it be better to test effect on each species separately by an univariate method (i.e. either by ANOVA, or by Analysis of Deviance¹)?

Answer: It is much better to use a multivariate method. There is a danger in use of many univariate tests. If we perform several tests at nominal significance level $\alpha = 0.05$, the probability of Type I error is 0.05 in each univariate test. This means that when testing, say, 40 species, we could expect two significant results just as a result of Type I error. This can lead to "statistical fishing", when one picks up just the results of significant tests and tries to interpret them. We could overcome this by using Bonferroni correction (dividing the nominal significance level by the number of tests performed and performing the particular tests on the resulting significance

¹ Analysis of Deviance is a method in the Generalized Linear Models, a generalization of ANOVA, able to handle data having some non-Normal distribution (e.g. the Poisson distribution)

level; this assures that the overall probability of the Type I error in at least one of the tests is equal or smaller than α), but this leads to an extremely weak test. I consider it quite correct (but some statisticians probably do not) to use the univariate methods for particular species when we found the community response significant, however, we should keep in mind that the probability of type I error is α in each separate test. We should be aware that if we select the most responsive species according to the results of ordination, those species will very probably give significant differences, and the univariate test does not provide any further information in addition to the results of ordination.



Figure 8-1: The design of the experiment

The design of the experiment is shown in the Figure 8-1. Each quadrat is characterized by (1) the type of the treatment, and (2) the number of the block. Both are categorial variables and consequently have to be coded as a set of "dummy" variables for CANOCO. The variables characterizing the block structure will be used as covariables and the treatment variables will be the environmental variables.

The corresponding data set is in the Figure 8-2. CANOCO asks separately for the environmental data and for the covariables, so each of those can be in a separate file. If all the environmental variables and covariables are in the same file, you have to specify the same file name (seedl.env in this case; in Excel as seedlenv.xls) for both the environmental variables and the covariables and the nomit the first four variables from the covariable list and the last four variables from the list of the environmental variables. It is very convenient to have the environmental variables and the covariables in the same file - you can select various subsets of environmental variables and covariables in various partial analyses.

The fourth environmental variable is not necessary here: three ",dummy" variables are sufficient for coding one categorial variable with four categories. The fourth variable will be omitted from calculations, as it is a linear combination of the previous three variables (lit+moss = 1 - control - litter-r - nardus-r). However, it is useful to include it into the data, as after calculations, we will need to plot the results using CanoDraw and we will need to plot centroids of all the four categories. Similarly, the separate variable for the fourth block is not necessary.

```
Ohrazeni 1994-seedlings, design of experiment
   (i3,8(f3.0))
  8
  1
      1
           0
                                    0
               0
                   0
                       1
                           0
                               0
  2
      0
           1
               0
                   0
                       1
                           0
                               0
                                   0
  3
      0
           0
                   0
                       1
                               0
                                   0
               1
                           0
  4
      0
           0
               0
                   1
                       1
                           0
                               0
                                    0
  5
      1
           0
                   0
                       0
                           1
               0
                               0
                                    0
  6
      0
           1
               0
                   0
                       0
                           1
                               0
                                    0
  7
      0
           0
               1
                   0
                       0
                           1
                               0
                                   0
  8
      0
           0
               0
                   1
                       0
                           1
                               0
                                    0
                       0
  9
      1
           0
               0
                   0
                           0
                               1
                                    0
 10
      0
           1
               0
                   0
                       0
                           0
                               1
                                    0
           0
                               1
 11
      0
               1
                   0
                       0
                           0
                                    0
 12
           0
               0
                   1
                       0
                           0
                               1
                                   0
      0
      1
           0
               0
                       0
                               0
                                   1
 13
                   0
                           0
 14
      0
           1
               0
                   0
                       0
                           0
                               0
                                   1
                                   1
 15
      0
           0
               1
                   0
                       0
                           0
                               0
 16
      0
           0
               0
                   1
                       0
                           0
                               0
                                   1
 control litter-rnardus-rlit+mossblock1
                                                        block2 bl.
                       rel3
                                  rel4
                                                        rel6
 rel1
            rel2
                                             rel5
                                                                  re.
 rel11
            rel12
                       rel13
                                  rell4
                                             rel15
                                                        rel16
Figure 8-2: Environmental data characterizing the design of the experiment (CANOCO file in full
format). File seedl.env
```

As the vegetation in the plot is very homogeneous and we also have only categorial explanatory variables, we will use the Redundancy analysis, the method based on the linear model. Now, we could test (at least) two hypotheses.

The first null hypothesis can be stated as follows: **There is no effect of the manipulation on the seedlings**. To reject this hypothesis, it is enough if the total number of seedlings differs between treatments, even if the proportion of seedling species remains constant. When the proportion of seedlings changes also or when both the proportion and total number of seedling changes, the null hypothesis will be obviously also rejected.

The second hypothesis can be stated as: **The proportions of the seedlings differ between the treatments.** This means that the seedlings of different species differ in their reaction to the treatment. The first hypothesis can be tested only when we use no standardization by samples (the default in CANOCO). When we use a standardization by samples (usually by the sample norm), then we test the second null hypothesis that the proportions of the species remain constant between the individual treatments. Note that the standardization is part of the averaging in the Canonical Correspondence Analysis: consequently, CCA will not find any difference between the plots differing in the total numbers of seedlings, but with constant proportions of individual species. The first test is more powerful, however the significant result of the second test is more ecologically interesting: the fact that seedlings of different species respond in different ways to particular treatments is a good argument for the importance of regeneration niche for maintenance of species diversity.

The calculation of RDA follows the standard way (see the Canoco for Windows manual). Just take care of the following issues:

(1) When you have environmental variables and covariables in the same file do not forget to omit appropriate variables. Omitting of covariables is more important, because if the same variable is in both covariables and environmental variables, it will be automatically omitted from the environmental variables, as it does not explain any variability.

	U					
original	block	perml	perm2	perm3	perm 4	perm 5
1	1	2	4	1	etc.	
2	1	4	3	4	etc.	
3	1	3	2	2	etc.	
4	1	1	1	3	etc.	
5	2	7	5	7	etc.	
6	2	8	8	6	etc.	
7	2	5	7	8	etc.	
8	2	б	6	5	etc.	
9	3	11	9	9	etc.	
10	3	9	12	12	etc.	
11	3	10	10	11	etc.	
12	3	12	11	10	etc.	
13	4	14	16	14	etc.	
14	4	15	13	15	etc.	
15	4	16	15	13	etc.	
16	4	13	14	16	etc.	
Figure 8-3: Permu	tations within	n blocks				

(2) If you want to standardize by samples, you have to ask for a long dialog, when running the console version of the CANOCO.

(3) When performing the Monte Carlo test, you can ask for permutation within blocks, conditioned by all the three covariables (the fourth is collinear and is omitted from the analyses). Each permutation class will then correspond to a block in the experiment. The permutation within blocks is shown in the Figure 8-3. This was recommended in the previous versions of CANOCO (under the null hypothesis, the treatments are freely exchangable within a block). In the newer versions of CANOCO, the residuals after subtracting the effect of covariables are permuted, when the reduced model is selected (the residuals should be freely exchangable under the null hypothesis). Ask for the reduced model and the unrestricted permutations.

9. Case study 3: Analysis of repeated observations of species composition in a factorial experiment: the effect of fertilization, mowing and dominant removal in an oligotrophic wet meadow

9.1. Introduction

Repeated observations of experimental units are typical for many areas of ecological research. In experiments, the units (plots) are sampled first **before** the experimental treatment is imposed on some of them. In this way we obtain the "baseline" data, i.e. data where the differences between sampling units are caused solely by the random variability. After the treatment is imposed, the units are sampled once or several times, to reveal the difference between **development** (dynamics) of experimental and control units. This design is sometimes called replicated BACI – Before After Control Impact. Analyzing univariate response (e.g. number of species, or total biomass) in this design, we usually apply the repeated measurements model of ANOVA.

There are two possibilities, how to analyze the data: we can use the split-plot ANOVA with time, i.e. repeated measures factor being the "within plot" factor (this is called univariate repeated measurements ANOVA), or we can analyze the data using MANOVA; although the theoretical distinction between those two is complicated, the first option is usually used because it provides a stronger test. Interaction between time and the treatment is a term reflecting the difference in development of the units between treatments. CANOCO can analyze the repeated observations of the species composition in a way equivalent to the univariate repeated measurements ANOVA. Whereas in ANOVA all the terms are tested simultaneously, in CANOCO we must run a separate analysis to test each of the terms. We will illustrate the approach with an analysis of a factorial experiment applying fertilization, mowing and dominant removal to an oligotrophic wet meadow. The description of the experiment is simplified, full description is in Lepš (1999).

9.2. Experimental design

The experiment was established in a factorial design in three replications of each combination of treatments in 1994 (Fig. 9-1). The treatments were fertilization, mowing, and removal of the dominant species (*Molinia caerulea*); i.e. 8 combinations in three replications yielding 24 $2x^2m$ plots. The fertilization treatment is an annual application of 65 g/m² of commercial NPK fertilizer. The mowing treatment is the annual scything of the quadrats in late June or early July. The cut biomass is removed after mowing. *Molinia caerulea* was manually removed by screwdriver in April 1995 with a minimum of soil disturbance. New individuals are removed annually.

1 mown M-rem	2 unmown	3 mown M-rem	4 unmown	
5 unmown M-rem	6 mown	7 unmown M-rem	8 mown	
9 mown	10 unmown M-rem	11 mown	12 unmown M-rem	
13 unmown	14 mown M-rem	15 unmown	16 mown M-rem	
17 unmown M-rem	18 mown	19 unmown M-rem	20 mown	
21 mown M-rem	22 unmown	23 mown M-rem	24 unmown	
unfertilized				
fertilized				

Figure 9-1: Design of the experiment

9.3. Sampling

Plots were sampled in the growing season (June, July) each year, starting in 1994. Note that initial sampling was conducted before the first experimental manipulation in order to have baseline data for each plot. The cover of all vascular species and mosses was visually estimated in the central $1m^2$ of the $2m \times 2m$ plot.

9.4. Data analysis

Data are in the form of repeated measurements; the same plot was sampled four times. For univariate characteristic (=number of species) the corresponding repeated measurements ANOVA model was used (von Ende 1993). For species composition, I used Redundancy Analysis (RDA) in the CANOCO package with the Monte Carlo permutation test. CanoDraw and CanoPost programs from CANOCO package (Ter Braak & Šmilauer 1998) were used for the graphical presentations of the ordination results. RDA, a method based on a linear species response, was used because the species composition in the plots was rather homogeneous and the explanatory variables were categorical. Because *Molinia* cover was manipulated, it was passive in the analyses. This is very important because otherwise we would show (with a high significance) that Molinia has higher cover in the plots where it was not It should be noted that by using the various combinations of the removed. explanatory (environmental in CANOCO terminology) variables and covariables in RDA, together with the appropriate permutation scheme in the Monte Carlo test, we are able to construct tests analogous to the testing of significance of particular terms in ANOVA models (including repeated measurements). As the data form repeated observations that include the baseline (before treatment) measurements, the interaction of the treatment and the time is of the greatest interest and corresponds to the effect of the experimental manipulation. When we test for the interaction, the plot identifiers (coded as many dummy variables) are used as covariables. In this way we subtract the average (over years) of each plot, and the changes in the particular plot only are analyzed. Values of time were 0, 1, 2, 3 for the years 1994,

1995, 1996 and 1997 respectively. This corresponds to a model where the plots of various treatments do not differ in 1994 and the linear increase in difference is fitted by the analysis (this approach is analogous to the linear polynomial contrasts rather than the ordinary effect testing in a repeated measurement ANOVA). The other possibility is to consider time to be a categorial (=nominal) variable (each year would be a separate category) and to code it as several dummy variables.

9.5. Technical description

The species data are in the Excel file ohrazspe.xls, the design of experiment is in ohrazenv.xls. Using WCanoImp, we prepare condensed CANOCO species file ohraz.spe, and environmental data file ohraz.env. In both files, the samples (i.e. species and their cover values) are in the following order: samples from 1994 are samples no. 1 to 24, samples from 1995 are samples no. 25 to 48 etc. The order will be important for description of the permutation scheme. The names of samples are as r94p1, which means recorded 1994, plot 1. In the environmental data file, the first three variables describe the which treatments were applied: 1 – treated, 0 – non-treated. The next variable, Year is the time from the start of experiment, i.e. time as a quantitative variable. Next four variables, Yr0, Yr1, Yr2 and Yr3 are dummy variables, describing year as a categorical variable (for all the records done in 1994, Yr0=1, and Yr1=0, Yr2=0 and Yr3=0). The next variables, P1 to P24 are plot identifiers (i.e. for all the records from plot one, P1=1 and P2 to P24 are zero). We will use this file as both environmental variables and covariables, with selecting (deleting) proper variables from each of them.

We used following combinations of environmental variables and covariables to test selected hypotheses:

Table 9-1: Results of the RDA analyses of cover estimates of the species in $1m \times 1m$ plots. Data are centered by species. No standardization by samples was applied (Standardization N). Explanatory variables are environmental variables in CANOCO terminology. % expl. 1-st axis: percent of species variability explained by the first axis, measure of the explanatory power of explanatory variables. r 1-st axis: species-environment correlation on the first axis. F-ratio: the F-ratio statistics for the test on the trace. P- corresponding probability value obtained by the Monte Carlo permutation test, 499 random permutations (i.e., Type I error probability in testing the hypothesis that the effect of all the explanatory variables is zero). Yr - serial number of the year, M - mowing, F - fertilization, R - *Molinia* removal, PlotID - identifier of each plot. The asterisk (*) between two terms means an interaction.

Analysis	Explanatory variables	Covariables	Stand ardiza	% expl. 1-st axis	r 1-st axis	F- ratio	Р
<u></u>	X7 X7 43 F	DI JD	tion	160	0.060	5.00	0.000
C1	Yr, Yr*M, Yr*F, Yr*R	PlotID	Ν	16.0	0.862	5.38	0.002
C2	Yr*M, Yr*F, Yr*R	Yr, PlotID	N	7.0	0.834	2.76	0.002
C3	Yr*F	Yr, Yr*M, Yr*R, PlotID	N	6.1	0.824	4.40	0.002
C4	Yr*M	Yr, Yr*F, Yr*R, PlotID	N	3.5	0.683	2.50	0.002
C5	Yr*R	Yr, Yr*M, Yr*F, PlotID	N	2.0	0.458	1.37	0.040

Null hypotheses of the tests for the particular analyses are:

- C1: there are no directional changes in time in the species composition, neither common to all the treatments, nor specific for particular treatments (corresponds to the test of all within-subject effects in a repeated measures ANOVA).
- C2: The temporal trend in the species composition is independent of the treatments.
- C3 (C4, C5): Fertilization (or removal, mowing, respectively) has no effect on the temporal changes in the species composition (corresponds to the tests of particular terms in a repeated measures ANOVA).

Note that when the Plot ID is used as covariable, the main effects (i.e. M, F, and R) do not explain any variability, and it is meaningless to use them, either as covariables, or as explanatory variables.

In this analysis, we consider time to be a quantitative (continuous) variable. This means that we will use Year, and delete all Yr0 to Yr3. This corresponds to search for linear trend in the data. If we look for general differences in dynamics, we will consider time to be a categorial variable, and use Yr0 to Yr3 (on of them is redundant, but useful for ordination diagrams) and delete Year. In this case, however, the interaction between treatment and time must be defined as interaction with all the dummy variables.

The null hypothesis C1 is a little complicated and difficult to interpret ecologically. But this analysis is useful for a comparison with the other analyses in terms of the explained variability and of the species-environment correlation on the first axis. Also, the permutation scheme for this analysis is not unequivocal. For the

other analyses, under the null hypotheses, the dynamics is independent of the treatments imposed. This means that if the null hypothesis is true, then the plots are interchangable; however, still the records from the same plot should be kept together (technically speaking, the records done in different years in the same plot are subplots of the same main plot, and main plots are permuted). To do this, following options in the dialog should be used:

Perform test: I usually use "Both above tests". However, the procedure when one performs both tests and then uses that one which gives better result is not correct. It may be expected that the test of first ordination axis is stronger when there is a single dominant gradient, the test of all constrained axes together is stronger when there are several independent gradients in the data. Note that with single explanatory variable, the first axis and all the axes are the same. Use of the permutation "under reduced model" is recommended. Depending on your time available, computer speed and size of the problem, the number of permutations can be increased. The permutations should be Restricted for spatial or temporal structure or split-plot design, then in the next window Split-plot design, Number of split-plots in whole-plot is 4 (i.e. 4 records from each plot), and they are selected by the rule *Take 1 Skip 23*. This corresponds to the order of records in the file: in our case, the records from the same plot are separated by 23 record from another plots.⁺ The whole-plots are freely exchangable and "no permutation" is recommended on the split-plot level.* After running the analysis, it is advisable to check in the log file the permutation scheme. In our case, it should be:

```
Sample arrangement in the permutation test ***
Whole plot
                1 :
    1
          25
                49
                       73
                2:
Whole plot
                       74
     2
          26
                50
Whole plot
                 3 :
     3
          27
                51
                       75
etc...
```

showing that the whole-plots are composed of the records from the same plot, which is correct.

The relationship of particular species to experimental manipulations can be vizualized by ordination diagrams. Probably the best possibility is to display the results of analysis C2 by environmental variables – species biplot (Figure 9-2). Because we used the year as a continuous variable, the interaction of the time with the treatments is also a continuous variable and is shown by arrows. Because time was used as a covariable, the trends should be seen as relative to the average trend in the community. For example, the concordant direction of a species arrow with the **FER*YEAR** arrow means either that the species cover increases in the fertilized plots or that it decreases in the non-fertilized plots (or both).

⁺ Take care: if you use permutation within blocks – which is not our case, then the number of plots to skip is the number of plots to skip within a block.

^{*} According to the manual: Optionally, the time point could be permuted also, but this has a less secure basis than the permutation of sites.





9.6. Further use of ordination results

The results of the analyses can be further used. One of the possibilities is to use the species scores. Species scores on the constrained axis of analyses, where time*treatment was the only explanatory variable and the other factors were covariables (C3 to C5), were considered characteristic of the species response to the treatment. Then the following biological characteristics of species were tested as possible predictors of this response:

1. Species height, taken as the middle of the range given in the local flora.

2. Presence of arbuscular mycorrhizae (AM), based on data from Grime et al. (1988) and from the ecological flora database.

3. Seedling relative growth rate (RGR) from Grime et al. (1988).

Because I expected that species similar to *Molinia* would benefit most from the *Molinia* removal, I also used a fourth (ordinal) variable, dissimilarity of a species from *Molinia*, for predicting the effects of *Molinia* removal. Similarity 1 was assigned to graminoids higher than 50 cm, 2 to broad-leaved graminoids smaller than 50 cm, 3 to narrow-leaved graminoids smaller than 50 cm, and 4 to forbs. Spearman correlation was used for analysis of the relationship of this value with the RDA score of *Molinia* removal.

Technically, we can get the results of analyses from the .sol file into a spreadsheet (use Excel and read the file as delimited). Then you can use the species scores and add the information you need (plant height, mycorrhizal status, etc). In this case, it is reasonable to omit species which have a low frequency in the data (those species show no response in majority of sampling units, simply because they were not present neither before nor after the treatment).

10. Tricks and rules of thumb in using ordination methods

This chapter collects some practical advice we found useful when analyzing multivariate data, namely when using the ordination methods. These issues are sorted into several sections.

10.1. Scaling options

- We can use the CanoDraw 3.x program or some other, more general statistical package to fit the unimodal response curves for a selected subset of species, in relation to a selected ordination axis. In doing so, we will fit a generalized linear model (GLM) with an assumed Poisson distribution and the log link function, using abundances of a particular species as a response variable and the second order polynomial of the sample scores on the selected axis. When doing the analysis, we achieve the best results by selecting (in DCA or CCA) the Hill's scaling with a focus on the inter-sample distances.
- If we select in a linear ordination method (PCA, RDA) that the species scores are to be divided by their standard deviation, then the length of each species arrow expresses how well are the values for that species approximated by the ordination diagram. If we decide, on the other hand, that the scores are **not** divided by the standard deviation, the arrows length shows the variability in that species' values across the displayed ordination (sub)space.

10.2. Permutation tests

- If we need to test the significance of the second canonical axis, we should setup a new CANOCO project, similar to the original one, but specifying the sample scores on the first canonical axis as an (additional) covariable. If we did not have any covariables originally, this is easily achieved by specifying the solution (.sol) file of the original analysis as the file with the covariable data. CANOCO is able to parse the solution file, looking for the sample scores. Note, however, that the appropriate sample scores are those that are linear combinations of the environmental variables. Therefore, to let the CANOCO find the correct set of scores, we must change the text preceding the Samp section, replacing the "Sample Scores" text with something different (e.g. "Xample Scores" as suggested in the Canoco for Windows manual). In this way, CANOCO "misses" the first section when parsing the solution file and reads the scores from the **SamE** section. After modifying the file we should specify in the CANOCO project that only the AX1 scores are to be used as a covariable and variables AX2, AX3, and AX4 should be ignored. If we test the significance of the third canonical axis, we should retain both AX1 and AX2 scores as the covariables.
- When fitting a constrained ordination model, we should reduce the crosscorrelations among the environmental (explanatory) variables. Keeping two or more strongly correlated explanatory variables makes the resulting constrained ordination model unreliable, similarly to the same situation in the regression models. As an approximate criterion, we should reduce the set of the environmental variables so that no variable has the VIF (variance inflation factor, see Ter Braak et Šmilauer, 1998, p. 119) larger than 20.
- To keep a reasonable power (relatively low error Type II probability given the probability of the Type I error) of the Monte Carlo permutation test, the

suggested number of permutations can be estimated by specifying the desired precision of the significance level estimate - P_0 (say 0.01) and calculating the number of permutations as $N = (10/P_0)-1$, i.e. 999 for our example.

10.3. Other issues

- Users of the ordination methods too often ignore third and higher ordination axes. While these ordination axes are often not interpretable and, in the case of a direct gradient analysis, their contribution to the canonical subspace is not significant, this must be explored and/or tested using the Monte Carlo permutation test.
- We should not fall into a despair if the variability explained by the submitted environmental variables in a constrained ordination model is low, particularly if we analyze presence-absence data or data with many zeros. In such cases, we can often find a well-interpretable structure, even if the amount of the explained variability is less than 10%. Also, we should differentiate between a statistically significant relation and the strength of such a relation.
- The easiest way to calculate the matrix of correlation coefficients between the species and the environmental variables is to define CANOCO project using the redundancy analysis (RDA) and to center and standardize by the species. After we analyze the project, the output directory contains a file named **spec_env.tab** which contains these coefficients in the format of the CANOCO input data file.
- To compare results of two analyses that were using the same set of samples, we can specify the solution file from the first project as the file with the supplementary (passive) environmental variables within the other CANOCO project. Note however, that CanoDraw 3.x does not recognize supplementary environmental variables, so it does not display diagrams containing their scores.
- We should be careful when naming environmental variables and covariables that are used to define the interaction terms. We should try to concentrate meaning of the variable in the first four characters of their labels, as only these characters are used when defining the names of the first-order interaction terms. Otherwise, we might end up with several non-distinguishable names that prevent an easy interpretation of the ordination diagrams.

11. Modern regression: an introduction

The regression models allow us to model dependence of (usually) one response variable (which might be quantitative or qualitative) on one or more predictor variables. These can be, again, the quantitative variables and/or the factors. In this sense, the regression models also include methods like the analysis of variance (ANOVA) or the analysis of the contingency tables.

Durign the eighties, many new types of regression models were suggested, usually extending in some way the well established ones. This chapter provides a short summary of those methods that we found useful for study of the ecological data.

11.1. Regression models in general

All the regression models share some basic properties of the response variable(s) and the predictor(s). For simplicity, we will restrict our attention to the most frequent kind of a regression models, where exactly one response variable is modeled using one or more predictors.

The simplest way to describe any type of such regression model is following:

 $\mathbf{Y} = \mathbf{E}\mathbf{Y} + \mathbf{e}$

where **Y** refers to the values of the response variable, **EY** is the value of the response variable expected for particular values of the predictor(s) and \mathbf{e} is the variability of the true values around that expected value EY. The expected value of the response can be formally described as a function of the predictor values:

 $EY = f(X_1, ..., X_p)$

The EY part is also named the **systematic component** of the regression model, while the e part is named the **stochastic component** of the model. Their general properties and different roles they have when we apply regression models to our data are compared in the table 11-1.

Stochastic component	Systematic component
Mirrors a priori assumptions of the model	Is determined by our research hypothesis
Its parameter(s) are estimated during or	Its parameters (regression coefficients) are
after fitting (variance of response variable)	estimated by fitting the model
We use it to estimate model quality	We interpret it and perform (partial) tests
(regression diagnostics)	on it / on its individual parameters

Table 11-1

When we fit a particular regression model to our data, our assumptions about its stochastic component (usually the assumptions about the distributional properties and about the independence or particular type of cross-dependence between individual observations) are fixed, but we manipulate the complexity of the systematic component. In the perhaps simplest example of the classical linear regression model with one response variable Y and one predictor X, where the systematic component can be specified as:

 $EY = f(X) = \beta_0 + \beta_1 * X$

we can, in fact, achieve a larger complexity in modeling the dependence of the variable Y on X by a polynomial model. We can shift the model complexity along the complexity gradient starting with a **null model** EY = β_0 , through the linear dependency given above, and the quadratic polynomial dependency EY = $\beta_0 + \beta_1 * X + \beta_2 * X^2$, up to a polynomial of the **n**-th degree, where **n** is the number of data observations we have available, minus one. This most complex polynomial model (representing so called **full model** here) goes exactly through all the data points, but provides no simplification of the reality (which is one of the basic tasks of a model). We just simply replaced the *n* data points with *n* regression parameters ($\beta_0, ..., \beta_{n-1}$). The null model, on the other hand, simplifies the reality so much we do not learn anything new from such a model (and advancing our knowledge is another essential service provided by the models).

But from our discussion of those extreme cases, we can clearly see that the selection of model complexity moves us along a gradient from simple, not so precise models to (overly) complex models. Also, the more complicated models have another undesired property: they are too well fitted to our data sample, but they provide biased prediction for the non-sampled part of the statistical population. Our task is, in the general terms, to find a compromise point on this gradient, a model as simple as possible for it to be useful, but neither more simple nor more complex. Such a model is often referred as a **parsimonious** one.

11.2. General Linear Model: Terms

The first important stop on our tour over the families of modern regression methods is the general linear model. Note the word **general** - another type of regression method uses the word **generalized** in the same position, but meaning something different. In fact, the generalized linear model (GLM), discussed in the next section, is based on the general linear model (discussed here) and represents its generalization.

What makes the general linear model different from the traditional linear regression model is, from the user point of view, mainly that both the quantitative variables and the qualitative variables (factor) can be used as predictors. Therefore, the methods of analysis of variance (ANOVA) belong into the family of general linear models. For simplicity, we can imagine that any such factor is replaced by **k** "dummy", quantitative variables, if the factor has **k+1** different levels. Note that the method used by CANOCO (using dummy variables with values 0 or 1, with value 1 indicating membership of a particular observation in the corresponding class) is not the only possible one.

In this way, we can represent the general linear model by the following equation:

$$Y_i = \beta_0 + \sum_{j=1}^p \beta_j * X_{ji} + \varepsilon$$

but we must realize that a factor is usually represented by more than one predictor X_j and, therefore, by more than one regression coefficient. The symbol ϵ refers to the random, stochastic variable representing the stochastical component of the regression model. In the context of the general linear models, this variable is assumed to have a zero mean and a constant variance.

This model description immediately shows one very important property of the general linear model - its **additivity**. The effects of tje individual predictors are mutually independent. If we increase, for example, the value of one of the predictors by one unit, this has a constant effect (expressed by the value of the regression coefficient corresponding to that variable) independent of the value the other variables have and even independent of the original value of the variable we are incrementing.

The above-given equation refers to the (theoretical) population of all the possible observations, which we sample when collecting our actual dataset. Based on such a finite sample, we estimate the true values of the regression coefficients β_j , usually labelled as b_j . Estimation of the values of the regression coefficients is what we usually mean when we refer to **fitting** a model. When we take the observed values of the predictors, we can calculate the **fitted** (predicted) values of the response variable as:

$$\hat{Y} = b_0 + \sum_{j=1}^p b_j * X_j$$

The fitted values allow us to estimate the realizations of the random variable representing the stochastic component - such a realization is called the **regression** residual and labeled as e_i :

$$e_i = Y_i - \hat{Y}_i$$

Therefore, the residual is the difference between the observed value of the response variable and the corresponding value predicted by the fitted regression model.

The variability in the values of the response variable can be expressed by the **total sum of squares**, defined as

$$TSS = \sum_{i=1}^{n} (Y_i - \overline{Y})^2$$

In the context of the fitted regression model, this amount of variability can be further divided into two parts - the variability of the response variable explained by the fitted model - the **model sum of squares,** defined as

$$MSS = \sum_{i=1}^{n} \left(\hat{Y}_i - \overline{Y} \right)^2$$

and the so-called residual sum of squares defined as

$$RSS = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n} e_i^2$$

Obviously TSS = MSS + RSS. We can use these statistics to test the model significance. Under the global null hypothesis ("the response variable is independent of the predictors") the MSS is not different from the RSS, if both are relativized by their respective number of degrees of freedom.⁺

⁺ See any decent statistical textbook for more details, e.g. Sokal & Rohlf (1995)

11.3. Generalized Linear Models (GLM)

Generalized linear models (McCullagh & Nelder, 1989) extend the general linear model in two important, mutually related aspects. First, the **expected** values of the response variable (**EY**) are not supposed to be always directly equal to the linear combination of the predictor variables. Rather, the scale of the response depends on the scale of the predictors through some simple, parametric function, named **link function:**

$$g(EY) = \eta$$

where η is called the **linear predictor** and is defined similarly as the whole systematic component of the **general** linear model, namely:

$$\eta = \beta_0 + \sum \beta_j X_j$$

The use of the link function has the advantage that it allows us to map from the whole real-valued scale of the linear predictor (reaching, generally, from the $-\infty$ to $+\infty$) to a specific interval making more sense for the response variable (like non-negative values for counts or values between 0 and 1 for the probabilities).

Second, the generalized linear models do not have so specific assumptions about the stochastic part as the general linear models do. Its variance needs not to be constant (but, rather, it can depend on the expected value of the response variable, EY) and there is some variety in the assumed statistical distributions for the stochastic part (and, therefore, for the response variable).

The options we have for the link functions and for the assumed type of the distribution of the response variable cannot be fully independently combined, however. For example, the **logit** link function maps the real scale onto a range from 0 to +1, so it is not a good link function for, say, an assumed Poisson distribution. The table 11-2 lists some typical combinations of the assumed link functions and expected distributions of the response variable, together with a characteristic for the response variables matching these assumptions.

Variable type	"Typical" link function	Reference distribution
counts (frequency)	log	Poisson
probability	logit	binomial
(relative frequency)	or probit	binofinal
dimensions, ratios	inverse or log	gamma
quite rare type of measurements	identity	Gaussian ("normal")

Table 11-2

With this knowledge, we can summarize what kinds of regression models are embraced by the GLMs:

- "classical" general linear models (including most types of the analysis of variance)
- extension of those classical linear models to variables with non-constant variance (counts, relative frequencies)
- analysis of contingency tables (using log-linear models)
- models of survival probabilities used in toxicology (probit analysis)

The generalized linear models do not use the concept of a residual sum of squares. The extent of discrepancy between the true values of the response variable and those predicted by the model is expressed by the model' **deviance**. Therefore, to assess quality of a model, we use the statistical tests based on the **analysis of deviance**, quite similar in its concept to an analysis of variance of a regression model.

An important property of the general linear model, namely its **linearity**, is retained in the generalized linear models on the scale of the linear predictor. The effect of a particular predictor is expressed by a single parametr - linear transformation coefficient (the regression coefficient). Similarly, the model additivity is kept on the linear predictor scale. On the scale of the response variable, things might look differently, however. For example with a logarithmic link function, the additivity on the predictors' scale corresponds to a multiplicative effect on the scale of the response.

11.4. Loess smoother

The term **smoother** is used for a regression model attempting to describe (usually in a non-parametric way) the expected values of the response for a particular value(s) of the predictor variable(s). The values predicted by a smoother are less variable than the observed ones (hence the name 'smoother').

There are several types of the smoothers, some of them not very good, but simple to understand. A good example of such a smoother is the moving average smoother. An example of a better smoother is the so-called loess smoother (earlier also named lowess). This smoother is based on a locally weighted linear regression (Cleveland & Devlin 1988, Hastie & Tibshirani 1990). It uses the linear regression model to estimate the value of the response variable for a particular combination of the predictors' values, fitting the model using only the observations having their values of predictor variables sufficiently close to the estimation point. The area (band for a single predictor) around the estimation point used to select data for the local regression model fit is called the **bandwidth** and it is specified as a fraction of the total available dataset. Therefore, bandwidth value $\alpha=0.5$ specifies that at each estimation point, the half of the observations (those closest to the considered combination of the predictors' values) is used in the regression. The complexity of the local linear regression model is specified by the second parameter of the loess smoother, called its **degree** (λ). Typically, only two values are used: 1 for a linear regression model and 2 for a second-order polynomial model.

Furthermore, the data points used to fit the local regression model do not enter it with the same weight. Their weights depend on their distance from the considered estimation point in the predictors' space. If a data point has exactly the required values of the predictors, its weight is equal to 1.0 and the weight smoothly decreases to 0.0 at the edge of the smoother bandwidth.

An important feature of the loess regression model is that we can express its complexity using the same kind of units as in the traditional linear regression models - the number of degrees of freedom (DF) taken from the data by the fitted model. These are, alternatively, called the **equivalent number of parameters**. Further, because the loess model produces predicted values of the response variable (like the other models), we can work-out the variability in the values of the response accounted for by the fitted model and compare it with the residual sum of squares.

As we have the number of DFs of the model estimated, we can calculate the residual DFs and calculate the sum of squares per one degree of freedom (corresponding to the mean square in an analysis of variance of a classical regression model). Consequently, we can compare regression models using an analysis of variance the same way we do this for the general linear models.

11.5. Generalized Additive Model (GAM)

The generalized additive models (GAMs, Hastie & Tibshirani, 1990) provide an interesting extension to the generalized linear models (GLMs). The linear predictor of a GLM is replaced here by the so-called **additive predictor**. It is also represented by a sum of independent contributions of the individual predictors, nevertheless the effect of a particular predictor variable is not expressed using a simple regression coefficient. Instead, it is expressed - for the **j**-th predictor variable by a smooth function s_j , describing the transformation from the predictor values to the (additive) effect of that predictor on the expected values of the response variable.

$$\eta_{\rm A} = \beta_0 + \sum s_j(\mathbf{X}_j)$$

The additive predictor scale is again related to the scale of the response variable via the link function.

We can see that the generalized additive models include the generalized linear models as a special case, where for each of the predictors the transformation function is defined as:

$$s_j(X_j) = \beta_j \, * \, X_j$$

In the more general case, however, the smooth transformation functions (usually called "smooth terms") are fitted using a loess smoother or a cubic spline smoother. When fitting a generalized additive model, we do not prescribe the shape of the smooth functions of the individual predictors, but we must specify the complexity of the individual curves, in terms of their degrees of freedom. We usually need to select the type of smoother used to find the shape of the smooth transformation functions for the individual predictors. Typically, either the cubic spline smoother or the loess smoother are used.

With generalized additive models, we can do a stepwise model selection including not only a selection of the predictors used in the systematic part of the model but also selection of complexity of their smooth terms.

Generalized additive models cannot be easily summarized numerically, in contrast to the generalized linear models where their primary parameters - the regression coefficients - summarize the shape of the regression model. The fitted additive model is best summarized by plotting the estimated smooth terms representing the relation between the values of a predictor and its effect on the particular response variable.

11.6. Classification and Regression Trees

The tree-based models are probably the most non-parametric kind of regression models one can use to describe a dependence of the response variable values on the values of the predictor variables. They are defined by a recursive binary partitioning of the dataset into subgroups that are successively more and more homogenous in the values of the response variable. At each partitioning step, exactly one of the predictors is used to define the binary split. A split maximizing homogeneity and difference of the resulting two subgroups is then selected. Each split uses exactly one of the predictors - these might be either quantitative or qualitative variables.

The response variable is either quantitative (in the case of the **regression trees**) or qualitative (for the **classification trees**). The results of fitting are discribed by a "tree" portraying the successive splits. Each branching is described by the specific splitting rule: this rule has the form of an inequality for quantitative predictors (e.g. "VariableX2 < const") or the form of an enumeration of possible values for a qualitative variable - factor (e.g. "VariableX4 has value a, c, or d"). The two subgroups defined by such a split rule are further subdivided, until they are too small or sufficiently homogeneous in the values of the response. The terminal groups (leaves) are then identified by a predicted value of the response value (if this is a quantitative variable) or by a prediction of the object membership in a class (if the response variable is a factor).

When we fit a tree-based model to a dataset, we typically create ("grow") an overcomplicated tree based on our data. Then we try to find an optimum size of the tree for the prediction of the response values. A **cross-validation procedure** is used to determine the "optimum" size of the tree. In this procedure, we create a series of progressively reduced ("pruned") trees using only a subset of the available data and then we use the remaining part of the dataset to assess performance of the created tree: we pass these observations through the hierarchical system of the splitting rules and compare the predicted value of the response variable with its known value. For each size ("complexity") of the tree model, we do it several times. Typically, we split the dataset into 10 parts of approximately same size and, for each of these parts, we fit a tree model of given complexity using the remaining nine parts and use the current one for performance assessment. A graph of the dependency of the tree model "quality" on its complexity (model size) has typically a minimum corresponding to the optimum model size. If we use a larger tree, the model is overfitted - it provides a close approximation of the collected sample, but a biased description of the sampled population.

11.7. Modelling species response curves: comparison of models

In this section, we will use the example of search of appropriate description of the response of a selected tree species (beech - *Fagus sylvatica*) to the mesoclimatic factors, namely the altitude and the annual amount of precipitation. The Figure 11-1 displays the original data in the form of two scatterplots. Note that the response variable (frequency of the beech tree) is in a form of counts, being strictly non-negative. The regression models presented further in this section were fitted using the S-Plus for Windows program, version 4.5. The commands used to fit the models are not shown here - learning to use this advanced statistical package is not an easy task and would not fit into the scope of this book.



Figure 11-1 Relation of the frequency of Fagus sylvatica on the altitude and the annual precipitation amount

The frequency of *Fagus* seems to decrease with the increasing altitude, but there are also some very low values for the lowest altitude values. It is nearly impossible to judge whether is there any dependency of the *Fagus* frequency on the amount of annual precipitation.

The first model we will apply is a traditional linear model, fitted using an algorithm based on the least-squares criterion. The complexity of the model was selected using a stepwise selection procedure, where the tried model specifications were a null model (no dependence of the Fagus frequency on the considered variable), a linear model, a second-order polynomial (a quadratic dependence), and a third order polynomial. The selected models are displayed in the Figure 11-2. The dependence of the Fagus occurrence on the altitude is best fitted by a linear model. There we see one of the limitations the classical linear models have: the frequency for the altitude of 3000 meters a.s.l. is expected to be negative. The model for the annual precipitation influence over the Fagus occurrence is a second-order polynomial, indicating a local minimum for the precipitation around the 1500 mm per year.



Figure 11-2 Two linear models describing, separately, the dependency of *Fagus* frequency upon the altitude and the annual precipitation



Figure 11-3 Shape of two generalized linear models describing, separately, the dependency of *Fagus* frequency upon the altitude and the annual precipitation

The Figure 11-3 summarizes similar regression models as those in Figure 11-2, this time fitted as the generalized linear models. An assumed distribution of the response variable was the Poisson distribution and the logarithmic link function was

used. We again used a forward selection, this time based on the analysis of deviance tests (those are approximate tests based on the assumed chi-square distribution of the residual deviances of the compared models). This time, a second order polynomial model was selected for both the environmental factors, clearly showing a steep decrease in the occurrence frequency for the very low altitudes. For the high altitudes, on the other hand, the expected counts decrease gradually to the zero values.

The last family of regression model where we will fit two separate models for the two considered predictors are the generalized additive models. As for GLMs, we assumed the Poisson distribution and selected a log link function. Again, we have used a stepwise selection to decide about the model complexity, yet this time the selection was guided by a statistics addressing the model parsimony directly.⁺ During the stepwise selection, the candidate right sides were the null model, a (generalized) linear model with a linear term, a (generalized) linear model with a second order polynomial, a smooth term with one degree of freedom (DF), with 2 DFs, and 3 DFs. The smoother used in the smooth terms was a spline smoother.



Figure 11-4 Two generalized additive models fitting, separately the dependence of the *Fagus* frequency on the altitude and the annual precipitation amount

The stepwise selection selected a smooth term with 3 degrees of freedom for the model with the altitude as a predictor, while a more simple model (a smooth term with 1 degree of freedom) was selected for the other model. There we see a similar shape for the dependency of the *Fagus* frequency upon the annual precipitation, but for the altitude, the model unveils another, probably more realistic picture. There is an asymptric unimodal response curve, with the frequency of beech increasing promptly from the approximate altitude of 400 meters up to 700 meters (an

⁺ This statistics is called an *Akaike Information Criterion (AIC)* and is described for example in Hastie & Tibshirani (1990).

optimum?) and then decreasing again with the same speed until 1000 meters a.s.l. Nevertheless, the decrease of the *Fagus* frequency is much slower above this altitude, gradually declining until 2500 meters.

Now, we can check how the generalized additive model looks if we use both explanatory variables at the same time. We should realize that even when the a GAM models the effects of the two predictors independently, in an additive way, the character of their contribution into the model will differ from that they display when they are used separately. Also the model complexity selection must be done again. Surprisingly enough, the stepwise selection selects a more complex model this time.⁺ The smooth term for the altitude uses 4 degrees of freedom, the smooth term for the precipitation takes 3 DFs.



Figure 11-5 Generalized additive model of dependence of the *Fagus* frequency on both the altitude and the annual precipitation amount

The Figure 11-5 represents the most common type of representing the generalized linear models: it plots the smooth functions s_j for the individual predictors, as described in the section 11.3. The vertical axis shows, therefore a contribution of a particular smooth term to the additive predictor in the generalized additive model. The estimated confidence regions for the fitted transformation curves are also indicated. The smooth term for the altitude looks quite similar to that for the separate GAM for the altitude shown earlier, but the smooth term for precipitation indicates a loss of the decreasing segment for the low precipitation region.

At this place, we might review how comparable regression models would look the generalized linear model, generalized additive model, and the loess smoother. All these are in Figure 11-6 where both the 3-D surface plots and the contour plots are shown.

⁺ This might be due to a newly "recovered" patterns related to a possible interaction of the two predictors or it might be a deficiency of the AIC statistics used to select a parsimonious model.



Figure 11-6 Comparison of three response surfaces modeling frequency of *Fagus* using altitude and annual precipitation amount as predictors and using (from left to right, looking sideways) GLM, GAM, and loess smoother

The loess model does not retain the additivity of the two predictors' effects, so it is more flexible at modelling their interaction, yet there is a price to pay for this. First, we cannot evaluate the two effects independently: the shape of response curve along, say, the altitudinal gradient depends on the values of the precipitation. Second, to fit such more flexible model usually needs more data points (more degrees of freedom to be available). Finally, we will check what the regression trees methodology has to offer when modelling response of the *Fagus* frequency to the mesoclimatic factors. This time, we add another predictor to the two considered so far. The *DegreeDays* variable provides a measure of how warm a particular site is.

As we might recall from the section 11.6, we first let the algorithm grow as large a regression tree as possible. The result is displayed in the Figure 11-7. The height of the branches is proportional to the decrease of the model' deviance or (expressed in a different way), to the difference between the values of the response for the two groups resulting from the particular split. We can clearly see that the difference in altitude (with the threshold value being somewhere near the 1000 meters a.s.l.) is the most decisive one. The *Fagus* prefers the lower altitudes (the top-left branch) where the frequency is generally higher, except the sites which are extremely warm (DegreeDays value over approx. 19000).



Figure 11-7 The newly grown regression tree with an overly complicated contents

Before we continue with the interpretation, we should work-out how complicated tree is justified by its performance for predicting new values. This can be done using the cross-validation procedure and the Figure 11-8 shows its results.



Figure 11-8 Results of the tree cross-validation

The regression tree model residual deviance is plotted against the tree complexity expressed by two different characteristics on the upper and lower horizontal axis. There seems to be a minimum of the prediction errors for the tree with 10 terminal branches (10 "leaves"), but for easier viewing, we have fitted, at last, even simpler tree, displayed in the Figure 11-9.



Figure 11-9 The final regression tree

Note that while the result of fitting a regression tree is not as quantitative as from the other presented models, it has other advantages. For example, it allows to easily pin-point the interactions between the predictors, as well as the effects of predictors with different spatial resolution etc.

12. References

- Borcard D., Legendre P. & Drapeau P. (1992): Partialling out the spatial component of ecological variation. Ecology, 73: 1045 1055
- Cleveland W.S. & Devlin S.J. (1988): Locally-weighted regression: an approach to regression analysis by local fitting. J. Am. Statist. Assoc., 83: 597 610
- Grime J.P., Hodgson J.G. & Hunt R. (1988): Comparative plant ecology. Unwin Hyman, London.
- Hastie T.J. & Tibshirani R.J. (1990): Generalized Additive Models. Chapman and Hall, London. 335 pp.
- Hill M.O. & Gauch H.G. (1980): Detrended correspondence analysis, an improved ordination technique. Vegetatio, 42: 47 58
- Knox R.G. (1989): Effects of detrending and rescaling on correspondence analysis: solution stability and accuracy. Vegetatio, 83: 129 136
- Kovář P. & Lepš J. (1986): Ruderal communities of the railway station Ceska Trebova (Eastern Bohemia, Czechoslovakia) - remarks on the application of classical and numerical methods of classification. - Preslia, 58: 141 - 163
- Lepš J. (1999): Nutrient status, disturbance and competition: an experimental test of relationship in a wet meadow. Journal of Vegetation Science, in press.
- McCullagh P. & Nelder J.A. (1989): Generalised Linear Models. Second Edition. - Chapman and Hall, London. 511 pp.
- Pyšek P. & Lepš J. (1991): Response of a weed community to nitrogen fertilization: a multivariate analysis. Journal of Vegetation Science, 2: 237 244
- Sokal R.R. & Rohlf F.J. (1995): Biometry. 3rd edition, W.H. Freeman, New York, USA. 887 pp.
- Šmilauer P. (1992): CanoDraw User's Guide v. 3.0 Microcomputer Power, Ithaca, USA. 118 pp
- Špačková, I., Kotorová, I. & Lepš, J. (1998): Sensitivity of seedling recruitment to moss, litter and dominant removal in an oligotrophic wet meadow. - Folia Geobot,. 33: 17 - 30
- Ter Braak C.J.F. (1994): Canonical community ordination. Part I: Basic theory and linear methods. Ecoscience, 1: 127 140
- Ter Braak C.J.F. & Looman C.W.N. (1986): Weighted averaging, logistic regression and the Gaussian response model. Vegetatio, 65: 3 11
- Ter Braak C.J.F. & Prentice I.C. (1988): A theory of gradient analysis. Advances in Ecological Research, 18: 93 138
- Ter Braak C.J.F. & Šmilauer P. (1998): CANOCO Reference Manual and User's Guide to Canoco for Windows. Microcomputer Power, Ithaca, USA. 352 pp
- Van der Maarel E. (1979): Transformation of cover-abundance values in phytosociology and its effect on community similarity. Vegetatio, 38: 97 114
- Von Ende C.N. (1993): Repeated measures analysis: growth and other timedependent measures. In: Scheiner S.M. & Gurevitch J. [eds]: Design and analysis of ecological experiments, pp. 113 - 117. Chapman and Hall, New York, NY, USA.