Educational Tutorial of Biodiversity Data Processing
I. Biodiversity in Environmental Research
Biodiversity in contemporary biology and environmental sciences

Biodiversity is one of the leading concepts in biology, that is defined at many levels ... 

- Genes
- Individuals
- Populations
- Species
- Communities
- Ecosystems

... and very suitable for bioindication

- Effects made in past
- Effects of chronic exposure
- Impact of low dose exposures
- Impact on relationships in communities

Biodiversity bears very useful and lasting information, even when „short-term“ methods fail (biotests, chemical analyses)
Biodiversity as end-point in environmental studies

Biodiversity is one of the most complex measures /“integrating endpoint“/
### Biological communities as typical assessment endpoint at ecosystem level

<table>
<thead>
<tr>
<th><strong>VALUES &amp; BENEFITS</strong></th>
<th><strong>DRAWBACKS</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>Long-term „memory“</td>
<td>Time-consuming measurement</td>
</tr>
<tr>
<td>Interpretation at ecosystem level</td>
<td>Problem with representative sampling</td>
</tr>
<tr>
<td>Relation to functionality and stability of ecosystems</td>
<td>Dependent on experts (taxonomists, …)</td>
</tr>
<tr>
<td>End-point with evolutionary background</td>
<td>Long-term representative data are required for benchmarking</td>
</tr>
</tbody>
</table>

Diversity = variability: numerous masking effects
Biological communities as typical assessment endpoint at ecosystem level

**DRAWBACKS**
- Time-consuming measurement
- Problem with representative sampling
- Dependent on experts (taxonomists, …)
- Long-term representative data are required for benchmarking
  - Diversity = variability: numerous masking effects

**CONSEQUENCES IN BIOMONITORING**
- Trend to simplify design of studies or to reduce number of examined species
- Searching for indicator species of components of communities, susceptible to stress
Biodiversity as end-point in environmental studies

Also objective stratification of communities strongly decreases sample size!!
How to quantify and communicate complex assessment endpoint?

- As aesthetic nature
- As parametrically standardized end-point

Species rank plot

No. of individuals

Cumulative species-abundance profile

Species rank

Log (no. of individuals)
Complex endpoints typically offers redundant list of possible parameters

Species-abundance profile as standard output

- Number of individuals
- Species rank

Multivariate analysis
- Niche-oriented modelling

Model level
- Species – abundance profile

Analytic level
- Diversity indices
- Stochastic indices

Descriptive level
- Indicative species
- Species richness
- Dominance
Questions in analysis of biodiversity data

- Species abundance profile
- Dominance
- Communities structure
- Species interaction
- Biodiversity and environment
- Time and spatial changes in biodiversity
- Number of species
Methods of biodiversity analysis

- Diversity indices
- Species abundance models
- Multivariate analyses
Diversity indices: advantages and disadvantages

• The biodiversity indices can be considered as analogy to descriptive statistics

• The whole biological community is aggregated in one number which represents number of species and/or dominance of species in community;

• The confidence intervals can be obtained for this type of computation and tests for indices comparison are available;

• Advantages:
  • measure of diversity in one number

• Disadvantages:
  • reduction of species individuality
  • sometimes unclear interpretation of indices (the same value of index can originated in very different communities)
Species abundance models: advantages and disadvantages

• Can be considered as **analogy to distribution fitting** in statistics

• The **quantitative structure** of biological community (species abundance profile) results from the **underlying ecological processes**
  • the model profile with possible ecological interpretation can be fitted to profile of observed communities

• **Advantages:**
  • whole abundance profile of community is included in the analysis
  • species are represented only by the ranked abundance profile (i.e. the hypothesis of ecological process influencing quantitative structure of community is tested)

• **Disadvantages:**
  • unfinished methodology of fitting of expected and observed abundance profiles
  • slack of information on testing of ecological hypothesis connected to model profiles
Multivariate analysis of communities: advantages and disadvantages

• Biodiversity data can be subject of many types of cluster, ordination, regression and classification multivariate techniques.

• These methods search for **multivariate patterns in huge datasets** of communities and can answer questions:
  • relation of species to environment,
  • spatial relationships
  • species interactions

• **Advantages:**
  • comprehensive results aggregating all aspects of the data
  • reveal hidden interactions and links among variables

• **Disadvantages:**
  • data demanding and rather complicated
  • require expertise both in their statistical methodology and biological communities under study otherwise can lead to incorrect conclusions and interpretations
II. Tutorial Data and Tasks
Tutorial data

- Benthic macroinvertebrates data from the biomonitoring network of the Czech Republic
  - For training purposes: selection of Ephemeroptera community on reduced set of localities
  - Both community and environmental data
Tutorial tasks

• What is the diversity of biological communities in our dataset?

• What is the best approach for biodiversity measurement?

• Is there relationship between the diversity and the environment?

• What is the multivariate structure of analysed communities?

• Spatial variability of diversity data?
III. How to Sample and Measure Biodiversity
What is biodiversity?

- Place on Earth ➔ there are some organisms i.e. biodiversity ➔ description, explanation and what does it mean?
- Dual concept of biodiversity:
  1. Number of taxa (quality)
  2. Relative abundance of taxa (quantity)
Biodiversity definition

• Many definitions of biodiversity
  • The simplest one:
    The number of species (in fact species richness (Fiedler and Jain, 1992)

• Complex definition:

Biodiversity is an attribute of an area and specifically refers to the variety within and among living organisms, assemblages of living organisms, biotic communities, and biotic processes, whether naturally occurring or modified by humans. Biodiversity can be measured in terms of genetic diversity and the identity and number of different types of species, assemblages of species, biotic communities and biotic processes, and the amount (e.g., abundance, biomass, cover, rate) and structure of each. It can be observed and measured at any spatial scale ranging from microsites and habitat patches to the entire biosphere (DeLong, 1996)
Individuals, population and community

- Theoretical minimal unit in ecology is individual organism
Types of biodiversity

- Many possible views of diversity
  - Taxonomical diversity
  - Genetic diversity
  - Ecological/functional diversity
  - Physiological/biochemical diversity
Quality measures of biodiversity

- We have to define minimal qualitative unit of biodiversity

- The utilization of different stratification of community provide us with the information from the different points of view: good strategy

- For some organisms only limited number of approaches is available (lack of information etc.)
Quantity measures of biodiversity

- Many approaches for measurement of quantitative part of biodiversity

  - Some quantity measures are suitable only for some organisms or in some situations

  - Quantity of organisms reflect their utilization of the environment
Biodiversity sampling

- The same theoretical background as any statistical sampling
  - Random and representative sampling
  - Correct sampling design and reproducible sampling procedure
IV. Storage and Visualization of Diversity Data
Good practice of data storage

- Correct storage of data is the necessary prerequisite of their further statistical processing
- Tabular form of data is crucial for the computer processing of data
- The most suitable storage is database table
  - There is only one type of information in one column (taxa, abundance, etc.)
  - The row defines the most detailed level of data (locality or sample or fish etc.)
- The data in this form can be utilized in many spreadsheet and database programs

<table>
<thead>
<tr>
<th>Taxon</th>
<th>Abundance</th>
<th>Lokalita</th>
<th>etc.</th>
</tr>
</thead>
</table>
Visualization of community data

- Various types of visualization – different information on community
- Simple optical comparison of communities and/or species abundance models
Plotting number of species I

- Number of species is plotted against their abundance
Plotting number of species II

- Number of species is plotted against their abundance classes
- These classes are often based on log-scale and solve the problem of outliers visualisation
Rank abundance plot

- The species are ordered according to their abundance.
- The most common and rare species are well shown.
K- dominance plot

- Cumulative abundance of species is plotted against log rank of species
- Often used for the optical comparison of several samples
Cumulative number of species

- Cumulative number of species is plotted against abundance logarithm
- Connected with Q statistic diversity index
Software

- There is no universal software for biodiversity analysis
- Many types of software can be adopted

- Universal statistical packages:
  - Statistica, SPSS, SAS, Matlab
  - R project – vegan plugin
- Spreadsheet software
  - MS Excel, OpenOffice.org
- Database
- Specialized software
  - Canoco, PAST

- And many others ….
V. Comparing Species Richness – Rarefaction
Sample size and the number of species

- The number of species have non-linear dependence on number of individuals in the sample.

- How to compare species richness in samples of different size?

RAREFACTION
Rarefaction

• This method deals with the problem of comparison of species richness in samples of different size
• It standardize both samples to sample size of the smallest sample

It is the method for estimation of number of species in random selection of n individuals from the sample

Both point estimate and its confidence interval can be computed
Rarefaction: formula

\[
E(S_n) = \sum_{i=1}^{s} \left[ 1 - \frac{N - N_i}{N \choose n} \right]
\]

• Where \( E(S_n) \) is the estimated number of species in sample with \( n \) individuals, \( N \) is total number of individuals, \( N_i \) is the number of individuals of species \( i \) and \( n \) is number of individuals for which the estimate is computed.

• Variability of the estimate can be also computed:

\[
\text{var}(\hat{S}_n) = \left( \frac{N}{n} \right)^{-1} \sum_{i=1}^{s} \left( \frac{N - N_i}{n} \right) \left[ 1 - \frac{N - N_i}{N \choose n} \right] + 2 \sum_{i=1}^{s-1} \sum_{j=i+1}^{s} \left( \frac{N - N_i - N_j}{n} \right) - \left( \frac{N - N_i}{n} \right) - \left( \frac{N - N_j}{n} \right)
\]
Rarefaction: prerequisites

- Communities compared by rarefaction should have similar taxonomic composition.
- Equal sampling methods for compared samples
- The rarefaction curve could not be extrapolated beyond the sum of abundance in the biggest sample
- Individuals should be randomly distributed in the community
  - In the reality the distribution is often clustered and the rarefaction method overestimates under these conditions

\[
\sigma^2 < \mu \quad \sigma^2 > \mu \quad \sigma^2 = \mu
\]

Uniform \quad Clustered \quad Random
Software

- PAST software

- Diversity software by Institute of Biostatistics and Analyses
VI. Diversity indices
Diversity indices: higher diversity – what does it mean?

- Diversity can be measured by recording the number of species, by describing their relative abundances or by using a measure which combines the two components.
Species diversity measures

- Three main categories of species diversity measures:
  1. Species richness indices
  2. Species abundance models
  3. Indices based on the proportional abundances of species

Indices

Number of species

Dominance

Number of species and their eveness
Diversity indices

S, ‘species richness’  
McIntosh, 1967

\[ R_1 = \frac{(S-1)}{\ln N} \]  
Margalef, 1958

\[ R_2 = \frac{S}{\sqrt{N}} \]  
Menhinick, 1964

\[ H' = -\sum_i (p_i \ln p_i) \]  
Shannon & Weaver, 1949

\[ H_{adj} \]  
Hutcheson, 1970

\[ N_1 = e^{H'} \]  
Simpson, 1949

\[ N_2 = \frac{1}{\sum_i p_i^2} \]  
Hill, 1973

\[ E_1 = \frac{H'}{\ln S} = \frac{H'}{H'_{max}} \]  
Pielou, 1969

\[ E_2 = \frac{N_1}{S} \]  
Sheldon, 1969

\[ E_3 = \frac{(N_1-1)}{(S-1)} \]  
Heip, 1974

\[ E_4 = \frac{N_2}{N_1} \]  
Hill, 1973

\[ E_5 = \frac{(N_2-1)}{(N_1-1)} \]  
Alatalo, 1981

\[ D = \frac{(N-\sqrt{\sum_i n_i^2})}{(N-\sqrt{N})} \]  
McIntosh, 1967

\[ DE = \frac{(N-\sqrt{\sum_i n_i^2})}{[N-(N/\sqrt{S})]} \]  
Pielou, 1969
Diversity indices

\[ \text{PIE} = 1 - \sum_i p_i^2 \quad \text{Hurlbert, 1971} \]
\[ \text{NMS} = \sum_i (R_i p_i - 1) \quad \text{Fager, 1972} \]
\[ Q = (0.5n_{Q25} + \sum n_r + 0.5n_{Q75})/\ln(Q75/Q25) \]
\[ \text{HB} = (\ln N! - \sum_i \ln n_i!)N \quad \text{Brillouin, 1956} \]
\[ \text{HBe} = \text{HB}/\text{HB}_{\text{max}} \quad \text{Pielou, 1969} \]
\[ \text{G} = E5 \cdot (\arcsin E5/90) \quad \text{Molinari, 1989} \]
\[ \text{G} = (E5)^3 \quad \text{Molinari, 1989} \]
\[ 1/d = N/n_{\text{max}} \quad \text{Berger & Parker, 1970} \]
\[ R100 = \sum_i \left\{ 1 - \left[ \frac{(N-n_i)}{n} \right]/\left( \begin{array}{c} N \cr n \end{array} \right) \right\} \quad \text{Hurlbert, 1971} \]
Species richness indices

- Number of species (**species richness**) $S$
- Number of species per specified number of individuals or biomass
  it is not always possible to ensure equal sample sizes (number of species increases with sample size and sampling effort) => RAREFACTION - method for calculating the number of species expected in each sample if all samples were of a standard size (for example 1000 individuals)
- Species density – number of species per specified collection area

  - **Margalef index** $R_1 = (S-1)/\ln N$
  - **Menhinick index** $R_2 = S/\sqrt{N}$

  Number of species is weighted by number of individuals.

- **Disadvantage of species richness indices:** shape of community (ratios of abundances of taxa; species abundance curve) is not considered.
Species richness

- Species richness increases with sample size.

Relationship between number of species and area for flowering plants in England.
**Shannon index**

The Shannon index is given by the formula:

\[ H' = -\sum p_i \ln p_i \]

where \( p_i = \frac{n_i}{N} \),

- \( n_i \) is the abundance of the \( i \)-th species,
- \( N \) is the total number of individuals.

**Species** | **Abundance** | **Shannon index of diversity** | **\( p_i \)** | **\( \ln(p_i) \)** | **\( p_i \ln(p_i) \)**
---|---|---|---|---|---
Baetis alpinus | 736 | | 0.9472 | -0.0542 | -0.0513
Rhithrogena semicolorata | 28 | | 0.0360 | -3.3232 | -0.1198
Epeorus sylvicola | 8 | | 0.0103 | -4.5760 | -0.0471
Baetis rhodani | 4 | | 0.0051 | -5.2691 | -0.0271
Ephemera mucronata | 1 | | 0.0013 | -6.6554 | -0.0086

**Total number of individuals** | **777** | **Shannon index of diversity (using natural logarithm)** | **0.2539**

Shannon index is the most popular diversity index among ecologists.
# Shannon index

The Shannon index is a measure of biodiversity that takes into account both the number of species and the relative abundances of those species. The formula for the Shannon index is:

\[ H' = - \sum p_i \ln p_i \]

where \( p_i \) is the proportion of the total population that belongs to species \( i \), and \( \ln \) is the natural logarithm function.

\[ p_i = \frac{n_i}{N} \]

where \( n_i \) is the number of individuals in species \( i \) and \( N \) is the total number of individuals.

## Table of Species and Shannon Indices

<table>
<thead>
<tr>
<th>Species</th>
<th>Abundance</th>
<th>Shannon index of diversity</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>( p_i \log_2(p_i) )</td>
</tr>
<tr>
<td>Baetis alpinus</td>
<td>736</td>
<td>-0.0741</td>
</tr>
<tr>
<td>Rhithrogena semicolorata</td>
<td>28</td>
<td>-0.1728</td>
</tr>
<tr>
<td>Epeorus sylvicola</td>
<td>8</td>
<td>-0.0680</td>
</tr>
<tr>
<td>Baetis rhodani</td>
<td>4</td>
<td>-0.0391</td>
</tr>
<tr>
<td>Ephemera munronata</td>
<td>1</td>
<td>-0.0124</td>
</tr>
<tr>
<td>Total number of individuals</td>
<td>777</td>
<td></td>
</tr>
</tbody>
</table>

Shannon index of diversity:

| Shannon index of diversity | 0.3663 | 0.2539 | 0.1103 |

**Base of logarithm**

- \( \log_2 \)
- \( \ln \)
- \( \log \)

**Shannon diversity**

- 0.3663
- 0.2539
- 0.1103
Shannon index

Shannon index

\[ H' = -\sum p_i \ln p_i \]

\[ p_i = \frac{n_i}{N} \]

Base of logarithm \hspace{1cm} Shannon diversity

\[ f(x) \]

\[ x \]

\[ \log_2 (x) \]

\[ \ln (x) \]

\[ \log_{10} (x) \]
Berger – Parker index

Berger-Parker index

\[ d = \frac{N_{\text{max}}}{N} \]

### Species Abundance

<table>
<thead>
<tr>
<th>Species</th>
<th>Abundance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baetis alpinus</td>
<td>736</td>
</tr>
<tr>
<td>Rhithrogena semicolorata</td>
<td>28</td>
</tr>
<tr>
<td>Epeorus sylvicola</td>
<td>8</td>
</tr>
<tr>
<td>Baetis rhodani</td>
<td>4</td>
</tr>
<tr>
<td>Ephemerella mucronata</td>
<td>1</td>
</tr>
</tbody>
</table>

**Total number of individuals**: 777

**Berger-Parker index**: 0.9472

\( N_{\text{max}} \) … number of ind. in the most dominant species

\( N \) … total number of individuals

1 – d is used instead of d
VII. Species Abundance Models
What is biodiversity?

• Place on Earth ➔ there are some organisms i.e. biodiversity ➔ description, explanation and what does it mean?

• Dual concept of biodiversity:
  I. Number of taxa (quality)
  II. Relative abundance of taxa (quantity)
Species abundance curves
Species abundance models

• Models of theoretical species abundance curve
• Species abundance model should reflect **ecological prerequisites** of the model
• Comparison of real abundance curve and model should give us information on **ecological processes which build this community**
Species abundance models

- Some models can be members of several groups of models (for example niche oriented models with mathematical definition) or reach the same results by different ways
- The most important is the ecological meaning of the model

Mathematical models
- big communities
- mathematically defined distributions – deterministic models
- simple testing

Biological models

Niche oriented
- small communities
- stochastic models
- problems with testing
- exact mathematical definition is rare

Others
- models of population dynamics etc.
Prerequisites of the models

There is some “space” on the locality – niche

- Niche is distributed among species - abundance of taxa reflect this distribution
- The partitioning of niche is based on ecological properties of organisms and every model reflect different ecological processes which lead to given abundance distribution
- Of course, the models are only simplification of the reality which can be much more complicated
Mathematical models

- Deterministic mathematical definition
- Geometric series, logarithmic series, log-normal distribution, broken stick
- The dominance decrease in this order
- Some of these models have also stochastic (probabilistic) version
Niche oriented models

- Some of them called also Tokeshi models
- Various ecological algorithms of niche partitioning
- The computation is based on computer simulation
Niche oriented models: comprehensive summary

Simultaneous niche division

Successive niche division

- DD (dominance decay)
  - selection of the biggest part

- MF (MacArthur fraction = broken stick)
  - probability of selection depends on size

- RF (random fraction)
  - random selection
  - random division

- DP (dominance preemption)
  - geometric series
  - selection of the smallest part
  - species occupies more than half of this part

CM (composite model) = DD, MF, RF, DP for the most dominant species + RS for the rest

RS (random assortment) random division of niche in dynamic communities where abundance of species is independent on each other
Computer assisted estimation of niche oriented models as proposed by Tokeshi (1993)

RA = Random Assortment
RF = Random Fraction
MF = MacArthur Fraction
GS = Geometric Series
DP = Dominance Preemption
DD = Dominance Decay

The computation of expected model and its confidence interval is based on computer simulation for given number of species.

Several methods for testing of fit between observed and predicted model - multivariate distance, Monte carlo test etc.

Which one fits to real data?

Conclusion: the data fits one of niche oriented models.
Software

- Mathematical models are included in biodiversity softwares:
  - PAST
  - Biodiversity professional
  - Genstat

- Niche oriented models
  - Models with both stochastic and deterministic definition are included in biodiversity software
  - Software developed by Institute of Biostatistics and Analyses
VIII. Multivariate Analysis of Biodiversity Data: Overview
Univariate and multivariate statistics

- Multivariate analysis uses approaches of univariate statistics and follows its prerequisites.
- It brings new quality in revealing hidden relationships in multivariate space.
Multivariate analysis

- **Multivariate methods:**
  The name multivariate - data consist of objects (sites), each of them is characterized by different parameters (species)
  each of these parameters we can consider as one dimension of the object (site)

```
<table>
<thead>
<tr>
<th>DATA MATRIX</th>
</tr>
</thead>
<tbody>
<tr>
<td>site 1</td>
</tr>
<tr>
<td>species 1</td>
</tr>
<tr>
<td>species 2</td>
</tr>
<tr>
<td>species 3</td>
</tr>
<tr>
<td>site 2</td>
</tr>
<tr>
<td>site 3</td>
</tr>
<tr>
<td>site 4</td>
</tr>
<tr>
<td>site 5</td>
</tr>
<tr>
<td>site 6</td>
</tr>
</tbody>
</table>
```

Values of species (presence/absence; abundance; dominance) for each site

Ordination and cluster analysis are the only available techniques when one has no measured environmental data.
Input matrices of multivariate analysis

**NxP MATRIX**

Parameter values of cases

**ASSOCIATION MATRIX**

Correlation, covariance, distance, similarity metric

Computation of similarity/distance metric
## Basic typology of multivariate statistics in ecology

### CLUSTER ANALYSIS

- Clusters of objects on the base of their similarities
- Identification of existing groups of objects (clusters)

### ORDINATION METHODS

- Simplification of multivariate space into reduced number of dimensions
- New dimensions with higher amount of system variability than the original variables
Basic typology of multivariate statistics in ecology

**CLUSTER ANALYSIS**
- Classification of sites (localities) or species
- Searching for groups in data

**ORDINATION**
- Arrange sites and species along an underlying trend in data
- Searching for trends in data (dimensions explaining variability)

![Diagram illustrating cluster analysis and ordination](image-url)
Basic typology of multivariate statistics in ecology

**CLUSTER ANALYSIS**

- classifies sites, species or variables
- identifies groups in raw data

**ORDINATION**

- orders samples to emphasize underlying trends or patterns in the data
IX. Similarity of Biological Communities
**Taxa and their abundance: multivariate description of community**

- The list of taxa can be considered as the list of community parameters i.e. its multivariate description.
- The communities can be compared on the base of their distance in multivariate space.
- Theoretically any distance/similarity metric can be adopted for the comparison of communities.
Euclidean distance

- The basic metric for comparison in multivariate space is the Euclidean distance
- The Euclidean distance is based on Pythagorean theorem
- The method is sensitive to range of input parameters (standardisation is often adopted) and double zero problem

\[ D_1(x_1, x_2) = \sqrt{\sum_{j=1}^{p} (y_{1j} - y_{2j})^2} \]
Double zero problem !!!

- For biodiversity data, the presence of species in both communities (11) has not the same meaning as absence in both communities (00).
- The co-occurrence of taxon in both communities means similar environmental conditions (suitable for given taxon).
- Absence of taxon in both communities says nothing about their similarity – the reason for the absence can be very variable.
- Specialised similarity coefficient should be adopted for biodiversity data.
Similarity coefficient

- Many similarity coefficients based on presence/absence of species or their abundance are used in ecology

**Binary similarity coefficients**

<table>
<thead>
<tr>
<th>Community 1</th>
<th>Community 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>a b</td>
</tr>
<tr>
<td>0</td>
<td>c d</td>
</tr>
</tbody>
</table>

a, b, c, d = number of cases when different combinations of 1/0 occur

\[ a + b + c + d = p \]

**Symmetric binary coefficients** - 1-1 and 0-0 are treated equally

**Asymmetric binary coefficients** - 1-1 and 0-0 are treated unequally

- robust to double zero problem
- Jaccard coefficient

\[ S_7(x_1, x_2) = \frac{a}{a + b + c} \]
Quantitative similarity coefficients insensitive to double zero problem

- Sørensen quantitative coefficient, where $a_N$ and $b_N$ are total numbers of individuals in community A and B, $j_N$ is sum of abundances of less abundant species for species present in both communities

$$C_N = \frac{2j_N}{(a_N + b_N)}$$

- Morisita-Horn index, where $a_N$ is total number of individuals in community A and $a_{ni}$ is number of individuals of species $i$ in community A (the same for community B)

$$C_{mH} = \frac{2\sum (a_{ni}b_{ni})}{(da + db).a_N.b_N}$$

$$da = \frac{\sum a_{ni}^2}{a_N^2}$$

Lots of information on similarity measures are given in

X. Clustering of Biodiversity Data
Cluster analysis

- Hierarchical techniques
  - Agglomerative clustering
    - Single-linkage clustering
    - Average-linkage clustering
    - Complete-linkage clustering
    - ... (more methods)
  - Divisive clustering
    - Monothetic method
    - Polythetic method
    - Association analysis
    - Two way indicator species analysis
- Non-hierarchical techniques
  - K-means clustering
Hierarchical agglomerative clustering

- starts from single objects, which are agglomerated into larger clusters
- require a similarity or dissimilarity matrix (site by site) to start from
- for presence/absence data as well as for quantitative data many similarity indices are available
- all agglomerative methods are based on fusion of single entities (sites) or clusters into larger groups

\[ \text{distance in single linkage} \]
\[ \text{distance in complete linkage} \]

between cluster distance in other methods:
- distance between the centroid average distance

...
Hierarchical agglomerative clustering

REAL DATA
- from 6 localities, each locality monitored in 3 periods
- data matrix: 18 sites x 63 planktonic species; values = dominance scale

Dendrograms representative of single-linkage, average, and complete-linkage clustering. The single-linkage example shows chaining.
Hierarchical divisive clustering

- **Hierarchical techniques**
  - **Divisive clustering**
    - **Monothetic method**
      - Divisions are made on the basis of one attribute
      - First of all, the species that is maximally associated to the other species is selected; the groups are divided according to presence/absence of this species
    
    - **Association analysis**

- **Polythetic method**
  - Two way indicator species analysis
    - Group divisions are based on all species
    - Works also with abundance data
    - It is developed a qualitative equivalent of species abundance, the so-called pseudo-species

A binary key for identifying types of salt-marsh habitat (Iverney-Cook, Proctor 1966)

- I
- II
- III
- IV
- V
- VI
- VII
- VIII
Hierarchical divisive clustering

- starts with all objects as a group
- the group is divided into two smaller groups, ...

**Hierarchical techniques**

- divisive clustering

**Monothetic method**

- association analysis

- provides a simple binary key which can be used in the field for classifying further samples

- only for presence/absence data

- obtained groups - less homogenous than groups resulting from polythetic methods

- final classification - not very robust

**Polythetic method**

- two way indicator species analysis

- obtained groups are more homogenous than groups resulting from monothetic methods

- it does not provide a simple key suitable for field application, which would allow new samples to be assigned to their appropriate class
Hierarchical clustering

**Hierarchical techniques**

- **Agglomerative clustering**
  - Clustering is intuitive and superficially simple => it is the most popular classification method
  - Result is summarized in dendrogram
  - There is no “correct” joining rule
  - The results change dramatically if one selects
    - different joining rules
    - different distance measure
  - Agglomerative clustering is inefficient for very large data sets

- **Divisive clustering**
  - Simple interpretation of results
  - Divisive technique is more efficient for very large data sets than agglomerative clustering
  - Monothetic method is not very robust
  - Polythetic method does not provide a simple key suitable for field application, which would allow new samples to be assigned to their appropriate class
Non-hierarchical clustering

- the groups are not contained by larger groups, nor do they contain smaller groups
- divide objects into specified number of groups
- $K$-means clustering works with Euclidean distances

Non-hierarchical methods might be more appropriate than clustering:
- in cases with huge datasets
- in cases with no hierarchical structure in the data

The number of groups $K$ must be specified in advance by the user:

$K$-means clustering works with Euclidean distances $\Rightarrow$ problems when Euclidean distance is not „the best“ metric
XI. Ordination of Biodiversity Data and Definition of Environmental Gradients
Introduction

CLUSTER ANALYSIS
- Classification of sites (localities) or species
- Searching for groups in data

ORDINATION
- Arrange sites and species along an underlying trend in data

Factor axes
Ordination

- Ordination and cluster analysis can be used without environmental data.
- Explanatory variables in ordination are theoretical (underlying) variables = environmental gradients.

Each sample includes values of many species.

**Response = dependent variables**

species data
Ordination analysis: types of data

Biodiversity data:

quantitative data:
- number of individuals of defined species
- percentage cover
- biomass

semiquantitative data
- Braun-Blanquet scale

qualitative data
- presence / absence
Ordination methods, gradient analysis

• We use the term **gradient analysis** for any method arranging species data along environmental gradient – this gradient can be measured or can be hypothetical.

• Gradient analysis deals with relationship of community structure to (known or unknown) environmental gradients.

**Indirect gradient analysis**

• Axes constructed on species data.

**Direct gradient analysis**

• Ordination combined with regression – ordination axes are *constrained* or *canonical* – linearly dependent on measured variables.
Species response on environmental gradient

Two types of species response on environmental (known or unknown) gradient

- **linear** – the most simple estimation (in the case the sites from short gradient are available, linear approximation works very good)
- **unimodal** – species has an optimum on the environmental gradient (linear approximation does not work in case we have a long gradient)

![Diagram of species response on environmental gradient with linear and unimodal approximations]
Correspondence analysis

Correspondence analysis – a tool for analysis of relationships among rows and columns of contingency table => two categorical variables.

Correspondence analysis and biodiversity data

- indirect gradient analysis
- based on unimodal response – estimates species optimum at underlying theoretical gradient
Correspondence analysis (CA): reciprocal averaging or eigenanalysis

<table>
<thead>
<tr>
<th>Species</th>
<th>Site1</th>
<th>Site2</th>
<th>Site3</th>
<th>WA1</th>
<th>WA2</th>
<th>WA3</th>
<th>WA4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rhitrogena</td>
<td>0</td>
<td>0</td>
<td>3</td>
<td>13.000</td>
<td>10.000</td>
<td>10.000</td>
<td>10.000</td>
</tr>
<tr>
<td>Alainites</td>
<td>5</td>
<td>2</td>
<td>1</td>
<td>4.625</td>
<td>1.363</td>
<td>1.312</td>
<td>1.310</td>
</tr>
<tr>
<td>Baetis</td>
<td>6</td>
<td>2</td>
<td>0</td>
<td>3.250</td>
<td>0.113</td>
<td>0.062</td>
<td>0.060</td>
</tr>
<tr>
<td>Epeorus</td>
<td>8</td>
<td>1</td>
<td>0</td>
<td>2.556</td>
<td>0.050</td>
<td>0.028</td>
<td>0.027</td>
</tr>
</tbody>
</table>

Initial value: 2, 7, 13

WA1 scores: 3.319, 3.661, 10.906
WA1resc.: 0.000, 0.450, 10.000
WA2 scores: 0.415, 0.600, 7.841
WA2resc.: 0.000, 0.249, 10.000
WA3 scores: 0.377, 0.555, 7.828
WA3resc.: 0.000, 0.240, 10.000
WA4 scores: 0.375, 0.553, 7.827
WA4resc.: 0.000, 0.239, 10.000
Correspondence analysis (CA): reciprocal averaging or eigenanalysis

<table>
<thead>
<tr>
<th></th>
<th>Site1</th>
<th>Site2</th>
<th>Site3</th>
<th>WA4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rhitrogena</td>
<td>0</td>
<td>0</td>
<td>3</td>
<td>10.000</td>
</tr>
<tr>
<td>Alainites</td>
<td>5</td>
<td>2</td>
<td>1</td>
<td>1.310</td>
</tr>
<tr>
<td>Baetis</td>
<td>6</td>
<td>2</td>
<td>0</td>
<td>0.060</td>
</tr>
<tr>
<td>Epeorus</td>
<td>8</td>
<td>1</td>
<td>0</td>
<td>0.027</td>
</tr>
</tbody>
</table>

Correspondence analysis (CA): reciprocal averaging or eigenanalysis
Correspondence analysis: results

- **Ordination diagram:** ordination axes are orthogonal (linearly independent)
- Species and site scores (rows and columns)
- Eigenvalues, eigenvectors

![Ordination diagram]

**Eigenvalue** represents the information (variability) explained by defined axis.

High score: species with low frequency

Usually we use and interpret only 2-3 **ordination axes**.
Correspondence analysis (CA)

- CA is related to a unimodal response model; any species occurs in a limited range of values of each of the latent variables.
- CA is recommended for reduced-spaced ordination of species data tables containing a good number of zeros.

REAL DATA

- mayflies
- data matrix: 30 sites x 40 mayfly species
- values = logarithmically transformed abundances

The first two ordination axes of correspondence analysis explain 24.7% of species variance.

Only species with best fit shown (sites not shown).
Correspondence analysis: „arch effect“

- CA is based on unimodal response model
- Strong unimodal response can lead to „arch effect“ in ordination diagram
- Detrended form of CA copes with „arch effect“

REAL DATA
- terrestrial snails
- data matrix: 42 sites x 33 molluscan species
  values = dominance scale
Correspondence analysis: „arch effect“

„arch effect“, „horse shoe effect“

DCA

Arbitrarily assign Site Scores

Assign Species Scores as WA of Site Scores

Assign new Site Scores as WA of Species Scores

Detrend Site Scores

Any change in scores?

YES

End

NO
Detrended correspondence analysis (DCA)

- DCA is related to a unimodal response model
- DCA removes the "arch effect" by various methods

REAL DATA
- terrestrial snails
- data matrix: 42 sites x 33 molluscan species
  values = dominance scale
Software

- Canoco for Windows (ter Braak & Šmilauer 2004)
- SYN-TAX 2000 (Podani 1997)
- Statistica (StatSoft, Inc. 2005)
- PAST (Hammer, Harper, Ryan 2001)
XII. GIS in Analysis of Diversity Data
Geographic Information System (GIS)

• GIS is specialized, database-driven computer information system.
• The database contains observations of spatially distributed features, activities or events, which can be defined in space as points, lines or areas.
• It allows users to capture, store, display, manipulate and analyze geographically referenced data.
• We can use GIS to discover and demonstrate spatial relationships, making GIS a valuable tool to assess the biodiversity.
• GIS can be a powerful tool for displaying and analyzing biodiversity data during the planning, sampling and its evaluation.
Geographic Information System (GIS)

- GIS technology is increasingly being used in the environmental sciences
- Ideal tool for observation and connection of the data concerning distributions for each taxon and information on the taxonomy, ecology, biology and ecodistribution
- It allows to integrate the diversity data with many affecting data layers (geography, climate, geology, pedology, inhabitation…)
Geographic Information System (GIS)

- Three functions of GIS are important for biodiversity modelling:
  - terrain analysis
  - data integration
  - data visualization

- Terrain analysis can be used to identify micro, meso and macro terrain indices.
- Data integration can be used to determine the environmental characteristics of known habitats of species.
- Data visualisation uses maps, graphs and statistics to make the enormous amount of data that can be derived on a species' habitat easy to understanding.
Terrain Analysis

• Terrain analysis commonly refers to the identification of elevation, slope and aspect of a site.

• However, it can also be used to calculate the curvature of the surface, distinguish upslope or downslope areas, define catchment boundaries, determine runoff, and identify morphological landform types, such as terraces and gullies.

• Terrain, along with climate, soil and landcover, is an important determinant of species' habitat.

• Terrain is also an important abiotic determinant of the soil properties and moisture content of a site and is therefore a variable used in soil landscape modelling.
Data Integration

- GIS has the ability to integrate environmental character information with known locations of species or ecosystems.
- We need a range of GIS data sets that can be used to derive environmental characteristics of each plot.
Data Visualization

- By applying the functionality of GIS, information on the location of species can be ascertained using a variety of visualisation tools.
- The use of GIS to map data is well known, as a map is a basic spatial analysis tool.
- However, the development of GIS has lead to functions such as statistical analysis and charting also being available as visualisation tools.
XIII. Spatial Analysis of Diversity Data
Spatial Analysis

• Processes operating in space produce patterns

• Is the data randomly distributed in space or is there a regular pattern to how the data are distributed in space?

• Spatial Analysis:
  • Identifying and describing the pattern
  • Identifying and understanding the process
Types of Spatial Distribution

Random

Regular

Clustered
Spatial Autocorrelation

• “everything is related to everything else, but near things are more related than distant things” Waldo Tobler

• Spatial autocorrelation measures the level of interdependence between the variables, and the nature and strength of that interdependence.

• Classic spatial autocorrelation statistics - Moran’s $I$ and Geary’s $C$

• It may be classified as either positive or negative
Spatial Autocorrelation

- **spatial autocorrelation** - If there is any systematic pattern in the spatial distribution of a variable

- Positive spatial autocorrelation - spatial objects that are located more closely have more similar attributes; similar objects are closely located

- Negative spatial autocorrelation - spatial objects that are located more closely have more different attributes

- Random patterns exhibit *no spatial autocorrelation*
Spatial Autocorrelation

Negative

Random

Positive
Spatial Correlogram

- Spatial correlogram – autocorrelations values are plotted against the distance classes among sites
- can be computed for single variables (Moran’s I or Geary’s C) or for multivariate data (Mantel correlogram)
Variogram

- Variogram - a function which summarises the strength of association between responses as a function of distance, and possibly direction

- evaluates the spatial variation of a Surface

- The variogram function is the average (square) difference of the surface value $f(x)$ between two points of distance $h$
Variogram Models

- Exponential
- Spherical
- Gaussian

\( h \): distance
XIV. Predictive modelling of spatial distribution of species
PHDMs - potential habitat distribution maps

- PHDMs – modelling potential distributions of species or communities

  1) the probability of occurrence
  2) the most probable abundance
  3) the predicted occurrence based on non-probabilistic metrics
  4) the most probable entity (vegetation maps)
PHDMs - potential habitat distribution maps

Guisan A., 2000
Analytical methods

- Ordination methods (PCA, RDA, CA, CCA, ENFA)
- Regression (GLM, GAM)
- Neural network
- Classification trees
- Discriminant analysis
Relationship

Simple ordination of matrix $X$
PCA, CA, ENFA, distances

Ordination of $y$ (single axis) under constraint of $X$
Regression, Classification

Ordination of $Y$ under constraint of $X$
RDA and CCA
Predictive modelling of potential distribution of plant associations in the Czech Republic

- data were taken from the Czech National Phytosociological Database at the Department of Botany

- 36 associations – with 4700 phytosociological plots

- abiotic environmental variables - elevation, average temperature (June, January, annual), annual precipitation and soil acidification

- geographical information system ArcGIS
  - 37 685 squares of 2.21 km²
The data set

- training x testing
- outliers of predictive variables were excluded
- presence/absence - comparing the phytosociological plots assigned to the given association with plots assigned to other associations
Models

• variable with binary response
  ordination method - Canonical Correspondence Analysis (CCA)
  regression techniques - Generalized Linear Models (GLM)
  Generalized Additive Models (GAM)

• categorical variable
  neural network - Multilayer Perceptron (MLP),
  Radial Basis Function (RBF)
  classification trees - CART, CHAID
Classification methods for association of *Aphano arvensis-Matricarietum chamomillae*
Results

• **GLM and GAM** - simply implemented in the GIS  
  - suitable for small numbers of phytosystematic plots  
  - giving probability of occurrence  
  - presumption of binomial distribution

• **CART and CHAID** - unsuitable for small numbers of plots  
  - giving just presence/absence results  
  - better predictive power in cases of non-linear relationships and non-binomial distributions

• **Neural networks** - non-linear relationships and non-binomial distributions  
  - did not provide successful results  
  - giving probability of occurrence  
  - unsuitable for small numbers of plots

• **CCA** - non-binomial distributions  
  - giving centroid distances
Comparison of predictions for ten associations

<table>
<thead>
<tr>
<th>Associations</th>
<th>GLM accuracy</th>
<th>CART accuracy</th>
<th>CHAID accuracy</th>
<th>MLP accuracy</th>
<th>RBF accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Setario viridis-Fumarietum</td>
<td>87,1</td>
<td>76,50</td>
<td>74,19</td>
<td>65,03</td>
<td>64,38</td>
</tr>
<tr>
<td>Euphorbio exiguae-Melandrietum noctiflori</td>
<td>85,4</td>
<td>77,50</td>
<td>78,25</td>
<td>64,39</td>
<td>64,15</td>
</tr>
<tr>
<td>Vaccinio-Callunetum vulgaris</td>
<td>84,1</td>
<td>79,59</td>
<td>76,53</td>
<td>64,81</td>
<td>66,05</td>
</tr>
<tr>
<td>Angelico sylvestris-Cirsietum palustris</td>
<td>82,3</td>
<td>66,67</td>
<td>50,00</td>
<td>61,56</td>
<td>62,10</td>
</tr>
<tr>
<td>Chaerophyllo hirsuti-Filipenduletum ulmariae</td>
<td>82,2</td>
<td>80,98</td>
<td>50,00</td>
<td>64,41</td>
<td>62,71</td>
</tr>
<tr>
<td>Echinochloo crus-gali-Setarietum pumilae</td>
<td>79,6</td>
<td>54,28</td>
<td>62,85</td>
<td>59,59</td>
<td>60,96</td>
</tr>
<tr>
<td>Filipendulo ulmariae-Geranietum palustris</td>
<td>64,8</td>
<td>62,30</td>
<td>67,20</td>
<td>36,30</td>
<td>53,08</td>
</tr>
<tr>
<td>Aphano arvensis-Matricarietum chamomillae</td>
<td>62,4</td>
<td>76,60</td>
<td>68,80</td>
<td>50,00</td>
<td>33,73</td>
</tr>
<tr>
<td>Ranunculo bulbosi-Arrhenatheretum elatioris</td>
<td>57,6</td>
<td>75,40</td>
<td>59,80</td>
<td>45,80</td>
<td>50,38</td>
</tr>
<tr>
<td>Poo-Trisetetum flavescentis</td>
<td>56,9</td>
<td>65,00</td>
<td>66,10</td>
<td>41,10</td>
<td>47,85</td>
</tr>
</tbody>
</table>
Potential distribution maps of *Aphano arvensis* – *Matricarietum chamomillae* in the Czech Republic
Classification tree CART

Classification tree CHAID

Regression technique GLM

Regression technique GAM
XV. Conclusions
Conclusions

• Biodiversity analysis is suitable and useful in environmental studies

• Set of statistical methods supported by expert knowledge should be applied

• Combination of software should be used because there is no complex software package for biodiversity analysis
THANK YOU FOR YOUR ATTENTION !!!