**Introduction and Overview of Major Numerical Methods**

John Birks

Quantitative Methods in Palaeoecology and Palaeoclimatology

PAGES Valdivia October 2010

---

**WHAT IS PALAEOECOLOGY?**

Palaeoecology is, in theory, the ecology of the past and is a combination of biology and geology.

In practice, it is largely concerned with the reconstruction of past communities, landscapes, environments, and ecosystems.

It is difficult to study the ecology of organisms in the past and hence deduce organism – environment relationships in the past. Often the only record of the past environment is the fossil record. Cannot use the fossil record to reconstruct the past environment, and then use the past environment to explain changes in the fossil record!

---

**Contents**

- What is palaeoecology?
- What are palaeoecological data?
- Why attempt quantification in palaeoecology?
- What are the main approaches to quantification in palaeoecology?
- What are the aims of the course?
- What is the level of the course?
- What are the major numerical techniques in quantitative palaeoecology?
- How to transform palaeoecological data?
- What are the basics behind the major techniques used in quantitative palaeoecology?

---

There are several approaches to palaeoecology

1. Descriptive – basic description, common
2. Narrative – ‘story telling’, frequent
3. Analytical – rigorous hypothesis testing, rare

---

1. Qualitative - common
2. Quantitative – increasing

---

1. Descriptive - common
2. Deductive – rare, but increasing
3. Experimental – very rare
Why Study Palaeoecology?

1. Present-day ecology benefits from historical perspective
   “Palaeoecology can provide the only record of complete *in situ* successions. The framework of classical succession theory (probably the most well known and widely discussed notion of ecology) rests largely upon the inferences from separated areas in different stages of a single hypothetical process (much like inferring phylogeny from the comparative analogy of modern forms). Palaeoecology can provide direct evidence to supplement ecological theory.”
   S.J. Gould, 1976
2. Past analogue for future
3. Intellectual challenge and desire to understand our past
4. Reconstruction of past environment important to evaluate extent of natural variability
5. ‘Coaxing history to conduct experiments’ (Deevey, 1969)
6. Fun!

Philosophy of Palaeoecology

1. Descriptive historical science, depends on inductive reasoning.
2. Uniformitarianism “present is key to the past”.
4. Simplicity “Ockham’s razor”.
5. Sound taxonomy essential.
7. Data frequently quantitative and multivariate.

WHAT ARE PALAEOECOLOGICAL DATA?

Presence/absence or, more commonly, counts of fossil remains in sediments (lake muds, peats, marine sediments, etc).

**Fossils** - pollen diatoms chironomids cladocera radiolaria testate amoebae mollusca ostracods plant macrofossils foraminifera chrysophyte cysts - biochemical markers (e.g. pigments, lipids, DNA)

**Sediments** - geochemistry grain size physical properties composition magnetics stable isotopes (C,N,O)
Data are usually quantitative and multivariate (many variables (e.g. 30-300 taxa), many samples (50-300)).
Quantitative data usually expressed as percentages of some sum (e.g. total pollen).
Data may contain many zero values (taxa absent in many samples).
Closed, compositional data, containing many zero values, strong inter-relationships between variables.
If not percentages, data are presence/absence, categorical classes (e.g. <5, 5-10, 10-25, >25 individuals), or ‘absolute’ values (e.g. pollen grains cm$^{-2}$ year$^{-1}$).
Samples usually in known stratigraphical order (time sequence).
Some types of data may be modern 'surface' samples (e.g. top 1 cm of lake mud) and associated modern environmental data. Such data form 'training sets' or 'calibration data-sets'.

Palaeoecological data are thus usually
1. **stratigraphical sequences** at one point in space or samples from one point in time but geographically dispersed
2. **percentage** data
3. contain many zero values

**WHY ATTEMPT QUANTIFICATION IN PALAEOECOLOGY?**
1. Data are very time consuming (and expensive) to collect.
2. Data are quantitative counts. Why spend time on counting if the quantitative aspect of the data is then ignored?
3. Data are complex, multivariate, and often stratigraphically ordered. Methods to help summarise, describe, characterise, and interpret data are needed (Lectures 2, 3 and 4).
4. Quantitative environmental reconstructions (e.g. lake-water pH, mean July temperature) important in much environmental science (e.g. to validate model hindcasts or back-predictions) (Lecture 5).
5. Often easier to test hypotheses using numerical methods (Lecture 6).
Reasons for Quantifying Palaeoecology

1: Data simplification and data reduction
   “signal from noise”
2: Detect features that might otherwise escape attention.
3: Hypothesis generation, prediction, and testing.
4: Data exploration as aid to further data collection.
5: Communication of results of complex data.
   Ease of display of complex data.
6: Aids communication and forces us to be explicit.
   “The more orthodox amongst us should at least reflect that many of the
   same imperfections are implicit in our own cerebrations and welcome
   the exposure which numbers bring to the muddle which words may obscure”.
   D Walker (1972)
7: Tackle problems not otherwise soluble. Hopefully better science.
8: Fun!

Model Building in Palaeoecology

Model building approach
Cause of sudden and dramatic extinction of large mammals in North America 10-12,000 years ago at end of Pleistocene.
One hypothesis - arrival and expansion of humans into the previously uninhabited North American continent, resulting in overkill and extinction.
Model - arrival of humans 12,000 years ago across Bering Land Bridge. Start model with 100 humans at Edmonton, Alberta. Population doubles every 30 years. Wave of 300,000 humans reaching Gulf of Mexico in 300 years, populated area of 780 x 10^6 ha.
Population could easily kill a biomass of 42 x 10^9 kg corresponding to an animal density of modern African plains.
Model predicts mammal extinction in 300 years, then human population crash to new, low population density.

WHAT ARE THE MAIN APPROACHES TO QUANTIFICATION IN PALAEOECOLOGY?

1. Model building
   explanatory
   statistical
2. Hypothesis generation ‘exploratory data analysis’ (EDA)
   detective work
3. Hypothesis testing ‘confirmatory data analysis’ (CDA)
   CDA and EDA – different aims, philosophies, methods
   “We need both exploratory and confirmatory”
   J.W. Tukey (1980)

A hypothetical model for the spread of man and the overkill of large mammals in North America. Upon arrival the population of hunters reached a critical density, and then moved southwards in a quarter-circle front. One thousand miles south of Edmonton, the front is beginning to sweep past radiocarbon-dated Palaeoindian mammoth kill sites, which will be overrun in less than 2000 years. By the time the front has moved nearly 2000 miles to the Gulf of Mexico, the herds of North America will have been hunted to extinction.
(After Mosimann and Martin, 1975.)
A well-designed palaeoecological study combines both

1) Two-phase study
- Initial phase is exploratory, perhaps involving subjectively located plots or previous data to generate hypotheses.
- Second phase is confirmatory, collection of new data from defined sampling scheme, planned data analysis.

2) Split-sampling
- Large data set (>100 objects), randomly split into two (75/25) – exploratory set and confirmatory set.
- Generate hypotheses from exploratory set (allow data fishing); test hypotheses with confirmatory set.
- Rarely done in palaeoecology.

WHAT ARE THE AIMS OF THE COURSE?

Provide introductory understanding to the most appropriate methods for the numerical analysis of complex palaeoecological data. Recent maturation of methods.
Provide introduction to what these methods do and do not do.
Provide some guidance as to when and when not to use particular methods.
Provide an outline of major assumptions, limitations, strengths, and weaknesses of different methods.
Indicate to you when to seek expert advice.
Encourage numerical thinking (ideas, reasons, potentialities behind the techniques). Not so concerned here with numerical arithmetic (the numerical manipulations involved).

Data diving with cross-validation: an investigation of broad-scale gradients in Swedish weed communities
ERIK HALLGREN, MICHAEL W. PALMER and PER MILBERG

Flow chart for the sequence of analyses. Solid lines represent the flow of data and dashed lines the flow of analysis.

Teaching of statistics in ecology
At its best, statistical analysis sharpens thinking about data, reveals new patterns, prompts creative thinking, and stimulates productive discussions in multidisciplinary research groups. For many scientists, these positive possibilities of statistics are over-shadowed by negatives; abstruse assumptions, emphasis of things one can’t do, and convoluted logic based on hypothesis rejection. One colleague’s reaction to this Special Feature (on statistical analysis of ecosystem studies) was that “statistics is the scientific equivalent of a trip to the dentist.”

This view is probably widespread. It leads to insufficient awareness of the fact that statistics, like ecology, is a vital, evolving discipline with ever-changing capabilities.

At the end of the semester, could my students fully understand all of the statistical methods used in a typical issue of Ecology? Probably not, but they did have the foundation to consider the methods if authors clearly described their approach. Statistics can still mislead students, but students are less apt to see all statistics as lies and more apt to constructively criticise questionable methods. They can dissect any approach by applying the conceptual terms used throughout the semester. Students leave the course believing that statistics does, after all, have relevance, and that it is more accessible than they believed at the beginning of the semester.

J.S. Clark 1994
A warning!

"Truths which can be proved can also be known by faith. The proofs are difficult and can only be understood by the learned; but faith is necessary also to the young, and to those who, from practical preoccupations, have not the leisure to learn. For them, revelation suffices."

_— Bertrand Russell 1946_  
The History of Western Philosophy

"It cannot be too strongly emphasised that a long mathematical argument can be fully understood on first reading only when it is very elementary indeed, relative to the reader’s mathematical knowledge. If one wants only the gist of it, he may read such material once only, but otherwise he may expect to read it at least once again. Serious reading of mathematics is best done sitting bolt upright on a hard chair at a desk. Pencil and paper are indispensable."

_L. Savage 1972_  
The Foundations of Statistics.

BUT:

"A journey of a thousand miles begins with a single step"  
_Lao Tsu_

---

**WHAT IS THE LEVEL OF THE COURSE?**

- Approach from practical palaeoecological, biological, and geological viewpoint, not statistical theory viewpoint.
- Assume no detailed background in matrix algebra, eigenanalysis, or statistical theory.
- Emphasis on techniques that are palaeoecologically realistic and useful and that are computationally feasible.

---

**WHAT ARE THE MAJOR NUMERICAL TECHNIQUES IN PALAEOECOLOGY?**

1. Exploratory data analysis
   1a. Numerical summaries - means, medians, standard deviations, ranges
   1b. Graphical approaches - box-and-whisker plots, scatter plots, stratigraphical diagrams
   1c. Multivariate data analysis - classification, ordination (including discriminant analysis)
Numerical Techniques in Palaeoecology (cont.)

2. Confirmatory data analysis or hypothesis testing
3. Statistical modelling (regression analysis)
4. Quantitative environmental reconstruction (calibration = inverse regression)
5. Time-series analysis

1. Exploratory Data Analysis

1a. Summary Statistics

(A) Measures of location

(1) Arithmetic mean
\[ \bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i \]

(2) Weighted mean
\[ \bar{x} = \frac{\sum_{i=1}^{n} x_i w_i}{\sum_{i=1}^{n} w_i} \]

(3) Mode 'most frequent' value

(4) Median 'middle values' Robust statistic

(5) Trimmed mean 1 or 2 extreme observations at both tails deleted

(6) Geometric mean
\[ \log GM = \frac{1}{n} \sum_{i=1}^{n} (\log x_i) \]
\[ GM = \exp \left( \frac{1}{n} \sum_{i=1}^{n} \log x_i \right) \]
\[ = \text{antilog} \left( \frac{1}{n} \sum_{i=1}^{n} (\log x_i) \right) \]

(B) Measures of dispersion

\[ B \] smaller scatter than \[ A \]

Precision

Accuracy

(1) Range
\[ A = 0.37, \quad B = 0.07 \]

(2) Interquartile range 'percentiles'
\[ Q_1, Q_2, Q_3 \]

(3) Mean absolute deviation
\[ \frac{1}{n} \sum_{i=1}^{n} |x_i - \bar{x}| \]

Mean absolute difference
\[ \frac{1}{n} \sum_{i=1}^{n} |x_i - \bar{x}| \]

Ignore negative signs
\[ \frac{1}{n} \sum_{i=1}^{n} |x_i - \bar{x}| \]

Mean absolute difference
\[ \frac{1}{n} \sum_{i=1}^{n} |x_i - \bar{x}| \]

Ignore negative signs
\[ \frac{1}{n} \sum_{i=1}^{n} |x_i - \bar{x}| \]

10/n = 2.5

MAJOR USES OF NUMERICAL METHODS IN PALAEOECOLOGY

1. Data collection and assessment
   - Identification
   - Error estimation
2. Data summarisation – summarise major patterns
   - Single data set
   - Two or more stratigraphical sequences
   - Two or more geographical data sets
3. Data analysis – estimate particular numerical characteristics
   - Sequence splitting
   - Rate-of-change analysis
   - Time-series analysis
   - Environmental reconstructions
4. Data interpretation
   - Vegetation reconstruction
   - Causative or ’forcing’ factors
(B) Measures of dispersion (cont.)

(4) Variance and standard deviation
\[ s^2 = \frac{1}{n-1} \sum (x - \bar{x})^2 \] Variance = mean of squares of deviation from mean
\[ SD(s) = \sqrt{s^2} \] Root mean square value

(5) Coefficient of variation
\[ CV = \frac{s}{\bar{x}} \times 100 \] Relative standard deviation
Percentage relative SD (independent of units)

(6) Standard error of mean
\[ SEM = \frac{s}{\sqrt{n}} \]

1b. Graphical Approaches

(A) Graphical display of univariate data

Box-and-whisker plots – box plots

Box plots for samples of more than ten wing lengths of adult male winged blackbirds taken in winter at 12 localities in the southern United States, and in order of generally increasing latitude. From James et al. (1984a). Box plots give the median, the range, and upper and lower quartiles of the data.
4.15 ESTIMATION OF CORRELATION

Our estimation of correlation is affected by the area of the data rectangle divided by the area of the scale-line rectangle. In this example, the ratio is close to 1.

4.16 ESTIMATION OF CORRELATION

The data of Figure 4.15 are graphed again and the area of the data rectangle divided by the area of the scale-line rectangle is much smaller. The amount of correlation now appears greater.

- **Fig 2.15**
  - Triangular arrangement of all pairwise scatter plots for four variables.
  - Variables describe the length and width of sepals and petals for 150 iris plants, comprising 3 species of 50 plants.

- **Fig 2.21**
  - Three-dimensional perspective view for the first three variables of the iris data (fig 2.15). Plants of the three species are coded A, B, and C.

**FOURIER PLOTS**

Andrews (1972)

Plot multivariate data into a function.

\[ f_x(t) = x_1/\sqrt{2} + x_2\sin(t) + x_3\cos(t) + x_4\sin(2t) + x_5\cos(2t) \ldots \]

where data are \([x_1, x_2, x_3, x_4, x_5, \ldots, x_n]\)

Plot over range \([-\pi \leq t \leq \pi]\)

Each object is a curve. Function preserves distances between objects. Similar objects will be plotted close together.

**MULTI-PLOT**
Other types of graphical display of multivariate data involve some dimension reduction methods (e.g. ordination or clustering techniques), namely multivariate data analysis.

1c. Multivariate Data Analysis

**EUROPEAN FOOD**

(From A Survey of Europe Today, The Reader’s Digest Association Ltd.) Percentage of all households with various foods in house at time of questionnaire, Foods by countries.

<table>
<thead>
<tr>
<th>Food Type</th>
<th>NL</th>
<th>GB</th>
<th>P</th>
<th>A</th>
<th>CH</th>
<th>S</th>
<th>DK</th>
<th>N</th>
<th>SF</th>
<th>E</th>
<th>IRL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Coffee beans</td>
<td>91</td>
<td>85</td>
<td>94</td>
<td>83</td>
<td>85</td>
<td>85</td>
<td>95</td>
<td>95</td>
<td>98</td>
<td>98</td>
<td>99</td>
</tr>
<tr>
<td>Instant coffee</td>
<td>81</td>
<td>80</td>
<td>87</td>
<td>83</td>
<td>76</td>
<td>76</td>
<td>90</td>
<td>90</td>
<td>92</td>
<td>92</td>
<td>93</td>
</tr>
<tr>
<td>Sugar</td>
<td>90</td>
<td>88</td>
<td>92</td>
<td>90</td>
<td>92</td>
<td>92</td>
<td>95</td>
<td>95</td>
<td>97</td>
<td>97</td>
<td>98</td>
</tr>
<tr>
<td>Tea</td>
<td>88</td>
<td>85</td>
<td>88</td>
<td>86</td>
<td>88</td>
<td>88</td>
<td>91</td>
<td>91</td>
<td>93</td>
<td>93</td>
<td>94</td>
</tr>
<tr>
<td>Sugarless sugar</td>
<td>19</td>
<td>17</td>
<td>19</td>
<td>17</td>
<td>19</td>
<td>19</td>
<td>22</td>
<td>22</td>
<td>25</td>
<td>25</td>
<td>26</td>
</tr>
<tr>
<td>Biscuits</td>
<td>57</td>
<td>55</td>
<td>59</td>
<td>57</td>
<td>59</td>
<td>59</td>
<td>70</td>
<td>70</td>
<td>72</td>
<td>72</td>
<td>73</td>
</tr>
<tr>
<td>Soup (packages)</td>
<td>51</td>
<td>41</td>
<td>43</td>
<td>41</td>
<td>43</td>
<td>43</td>
<td>55</td>
<td>55</td>
<td>57</td>
<td>57</td>
<td>58</td>
</tr>
<tr>
<td>Soup (tinned)</td>
<td>19</td>
<td>17</td>
<td>17</td>
<td>17</td>
<td>17</td>
<td>17</td>
<td>25</td>
<td>25</td>
<td>26</td>
<td>26</td>
<td>27</td>
</tr>
<tr>
<td>Frozen fish</td>
<td>27</td>
<td>4</td>
<td>4</td>
<td>14</td>
<td>14</td>
<td>14</td>
<td>20</td>
<td>20</td>
<td>22</td>
<td>22</td>
<td>23</td>
</tr>
<tr>
<td>Frozen vegetables</td>
<td>21</td>
<td>2</td>
<td>2</td>
<td>14</td>
<td>14</td>
<td>14</td>
<td>17</td>
<td>17</td>
<td>19</td>
<td>19</td>
<td>20</td>
</tr>
<tr>
<td>Fresh apples</td>
<td>81</td>
<td>67</td>
<td>87</td>
<td>83</td>
<td>76</td>
<td>76</td>
<td>85</td>
<td>85</td>
<td>85</td>
<td>85</td>
<td>86</td>
</tr>
<tr>
<td>Fresh oranges</td>
<td>75</td>
<td>71</td>
<td>79</td>
<td>76</td>
<td>78</td>
<td>78</td>
<td>82</td>
<td>82</td>
<td>84</td>
<td>84</td>
<td>85</td>
</tr>
<tr>
<td>Tinned fruit</td>
<td>44</td>
<td>9</td>
<td>9</td>
<td>40</td>
<td>40</td>
<td>40</td>
<td>61</td>
<td>61</td>
<td>64</td>
<td>64</td>
<td>65</td>
</tr>
<tr>
<td>Jam</td>
<td>71</td>
<td>46</td>
<td>45</td>
<td>81</td>
<td>57</td>
<td>57</td>
<td>91</td>
<td>91</td>
<td>92</td>
<td>92</td>
<td>93</td>
</tr>
<tr>
<td>Garlic clove</td>
<td>22</td>
<td>80</td>
<td>88</td>
<td>16</td>
<td>29</td>
<td>29</td>
<td>91</td>
<td>91</td>
<td>94</td>
<td>94</td>
<td>95</td>
</tr>
<tr>
<td>Butter</td>
<td>91</td>
<td>66</td>
<td>94</td>
<td>31</td>
<td>84</td>
<td>84</td>
<td>97</td>
<td>97</td>
<td>98</td>
<td>98</td>
<td>99</td>
</tr>
<tr>
<td>Margarine</td>
<td>85</td>
<td>24</td>
<td>47</td>
<td>97</td>
<td>80</td>
<td>80</td>
<td>94</td>
<td>94</td>
<td>97</td>
<td>97</td>
<td>98</td>
</tr>
<tr>
<td>Olive, corn oil</td>
<td>74</td>
<td>94</td>
<td>36</td>
<td>13</td>
<td>83</td>
<td>84</td>
<td>57</td>
<td>57</td>
<td>59</td>
<td>59</td>
<td>60</td>
</tr>
<tr>
<td>Yoghurt</td>
<td>30</td>
<td>5</td>
<td>5</td>
<td>57</td>
<td>53</td>
<td>53</td>
<td>20</td>
<td>20</td>
<td>22</td>
<td>22</td>
<td>23</td>
</tr>
<tr>
<td>Crispbread</td>
<td>26</td>
<td>18</td>
<td>3</td>
<td>15</td>
<td>15</td>
<td>15</td>
<td>24</td>
<td>24</td>
<td>28</td>
<td>28</td>
<td>29</td>
</tr>
</tbody>
</table>

Country
Clustering

Dendrogram showing the results of minimum variance agglomerative cluster analysis of the 16 European countries for the 20 food variables listed in the table.

Key:
Countries: A Austria, B Belgium, CH Switzerland, D West Germany, E Spain, F France, GB Great Britain, I Italy, IRL Ireland, L Luxembourg, N Norway, NL Holland, P Portugal, S Sweden, SF Finland

Ordination

Correspondence analysis of percentages of households in 16 European countries having each of 20 types of food.

Geometric models

**Polien data - 2 pollen types x 15 samples**

Depths are in centimetres, and the units for pollen frequencies may be either in grains counted or percentages.

<table>
<thead>
<tr>
<th>Samples</th>
<th>Variables</th>
</tr>
</thead>
<tbody>
<tr>
<td>Depth</td>
<td>Type A</td>
</tr>
<tr>
<td>1</td>
<td>20</td>
</tr>
<tr>
<td>2</td>
<td>40</td>
</tr>
<tr>
<td>3</td>
<td>60</td>
</tr>
<tr>
<td>4</td>
<td>80</td>
</tr>
<tr>
<td>5</td>
<td>100</td>
</tr>
<tr>
<td>6</td>
<td>120</td>
</tr>
<tr>
<td>7</td>
<td>140</td>
</tr>
<tr>
<td>8</td>
<td>160</td>
</tr>
</tbody>
</table>

Adam (1970)
2. Hypothesis Testing or Confirmatory Data Analysis

Hypothesis of interest may be 'human impact on the landscape caused major changes in the lake-water nutrient status'. Called $H_1$ – alternative hypothesis.

Require 'response' variables ($Y$) e.g. lake-water total $P$ reconstructed from fossil diatoms.

Require 'predictor' or 'explanatory' variables ($X$) e.g. terrestrial pollen of unambiguous indicators of human impact (e.g. cereal pollen).

Need to quantify the predictive power of $X$ to explain variation in $Y$

$$Y = f(X)$$

e.g. $Y = b_0 + b_1X$ (linear regression)
Null hypothesis ($H_0$) is the opposite of our hypothesis ($H_1$), namely that human impact had no effect on the lake-water nutrient status; 

i.e. $b_1 = 0$ in $Y = b_0 + b_1X$ ($H_0$)  
$b_1 > 0$ in $Y = b_0 + b_1X$ ($H_1$) 

Can do a regression-type analysis of $Y$ in relation to $X$, estimate $b_1$. How to evaluate statistical significance when data are non-normal and samples are not random? Use so-called randomisation or Monte Carlo permutation tests (Lectures 3 & 6).

### 3. Statistical Modelling or Regression Analysis

Regression model  
$Y = b_0 + b_1X$  

[Inverse regression (= calibration) $X = a_0 + a_1Y$]  

Types of regression depend on numbers of variables in $Y$ and $X$  
- $Y = 1$, $X = 1$: simple linear or non-linear regression  
- $Y = 1$, $X > 1$: linear or non-linear multiple regression  
- $Y > 1$, $X > 1$: linear or non-linear multivariate regression  

($Y =$ response variable(s)  
$X =$ predictor or explanatory variable(s))

Lectures 3 & 5

### 4. Calibration (=Inverse Regression) Quantitative Environmental Reconstruction

$X_m = g Y_m + \text{error}$  
where $X_m =$ modern environment (e.g. July temperature)  
$Y_m =$ modern biological data (e.g. diatom %)  
$g =$ modern ‘transfer function’

$X_i = g Y_i$  
where $X_i =$ past environmental variable  
$Y_i =$ fossil biological data

Lecture 5

### 5. Time-Series Analysis

Values of one or more variables recorded over a long period of time as in a stratigraphical sequence. Values may vary with time. 

Variations may be long-term trends, short-term fluctuations, cyclical variation, and irregular or ‘random’ variation. 

Time-series analysis looks at correlation structure within a variable in relation to time, between variables in relation to time, trends within a variable, and periodicities or cycles within and between variables. 

Lecture 6
HOW TO TRANSFORM PALAEOECOLOGICAL DATA?

Percentage data – square-root transformation helps to stabilise the variances and maximises the ‘signal’ to ‘noise’ ratio.

Absolute data – log transformations (log(y+1)) help to stabilise the variances and may maximise the ‘signal’ to ‘noise’ ratio. Often also very effective with percentage data.

Stratigraphical data are in a fixed order. Need numerical methods that take account of this ordering (constrained classifications, constrained ordinations, restricted or constrained Monte Carlo permutation tests, time-series analysis).

Basis of much quantitative palaeoecology is not only the stratigraphical ordering but also age chronology of the samples.

Transformation of depth to age key stage.

WHAT ARE THE BASICS BEHIND THE MAJOR TECHNIQUES?

1. Multivariate data analysis
   - Classification (Lecture 4)
   - Ordination (Lectures 2 & 4)
   - Constrained ordination (Lectures 3, 4, 5, & 6)

2. Confirmatory data analysis (Lectures 3 & 6)

3. Statistical modelling (Lectures 3, 5, & 6)

4. Quantitative environmental reconstruction (Lecture 5)

5. Time-series analysis (Lecture 6)

1. MULTIVARIATE DATA ANALYSIS

   Clustering and Partitioning – Two Major Types used in Palaeoecology

   1. Agglomerative Hierarchical Cluster Analysis

      i. Calculate matrix of proximity or dissimilarity coefficients between all pairs of n samples (\( \frac{1}{2}n(n-1) \))

      ii. Clustering of objects into groups using stated criterion – ‘clustering’ or sorting strategy

      iii. Graphical display of results

      iv. Check for distortion

i. Proximity or Distance or Dissimilarity Measures

   Quantitative Data

   \[
   d_{ij} = \sqrt{(x_{ij} - x_{j})^2 + (x_{i} - x_{j})^2} \quad \text{Euclidean distance}
   \]

   Manhattan or city-block metric

   \[
   d_{ij} = \sum_{k=1}^{n} |x_{ik} - x_{jk}| \quad \text{dominated by large values}
   \]

   Bray & Curtis (percentage similarity)

   \[
   d_{ij} = \frac{\sum_{k=1}^{n} |x_{ik} - x_{jk}|}{\sum_{k=1}^{n} (x_{ik} + x_{jk})} \quad \text{sensitive to extreme values}
   \]

   relates minima to average values and represents the relative influence of abundant and uncommon variables
Percentage Data (e.g. pollen, diatoms)

Standardised Euclidean distance - gives all variables 'equal' weight, increases noise in data
Euclidean distance - dominated by large values, rare variables almost no influence
Chord distance (= Euclidean distance of square-root transformed data; Hellinger distance)

Transformations
Normalise samples - 'equal' weight
Normalise variables - 'equal' weight, rare species inflated
No transformation - quantity dominated
Double transformation - equalise both, compromise

Simple Distance Matrix

<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>9</td>
<td>4</td>
</tr>
<tr>
<td>5</td>
<td>8</td>
<td>5</td>
</tr>
</tbody>
</table>

\[ d_{ij} = \sqrt{\sum_{k=1}^{m}(x_{ik} - x_{jk})^2} \]

Objects

ii. Clustering Strategy using Single-Link Criterion

Find objects with smallest \( d_j = d_{12} = 2 \)

Calculate distances between this group (1 and 2) and other objects

\[
\begin{align*}
\ell_{12} & = \min (d_{13}, d_{23}) = d_{23} = 5 \\
\ell_{14} & = \min (d_{14}, d_{24}) = d_{24} = 9 \\
\ell_{15} & = \min (d_{15}, d_{25}) = d_{25} = 8
\end{align*}
\]

Find objects with smallest \( d_j = d_{15} = 3 \)

Calculate distances between (1, 2), 3, and (4, 5)

\[
\begin{align*}
D & = 2+3 = 5 \\
D & = 2+3 = 5 \\
D & = 4+5 = 9
\end{align*}
\]

Find object with smallest \( d_j = d_{4,5} = 4 \)

Fuse object 3 with group (4 + 5)

Now fuse (1, 2) with (3, 4, 5) at distance 5

Single-link (nearest neighbour) - fusion depends on distance between closest pairs of objects, produces 'chaining'

Complete-link (furthest neighbour) - fusion depends on distance between furthest pairs of objects

Median - fusion depends on distance between \( K \) and mid-point (median) of line \( IJ \)

Centroid - 'weighted' because \( I = J(1 \text{ compared with } 4) \)

'Unweighted' as the size of \( J \) is taken into account
Also:

Unweighted group-average distance between $K$ and $(i,j)$ is average of all distances from objects in $i$ and $j$ to $K_i$, i.e. $\frac{\sum d}{5}$

Weighted group-average distance between $K$ and $(i,j)$ is average of distance between $K$ and $i$ (i.e. $\sum d/K$) and between $i$ and $K$ i.e. $\frac{d_K + \sum d_{iK}}{4}$

Minimum variance, sum-of-squares
Orci 1967 J. Ecology 55, 193-206

Ward’s method

$Q_i Q_j Q_k$ within-group variance

Fuse $j$ with $K$ to give $(i,j)$ if and only if $Q_{ij} - (Q_i + Q_j) \times Q_{ik} - (Q_i + Q_k)$
or $Q_K - (Q_i + Q_j)$

i.e. only fuse $j$ and $K$ if neither will combine better and make lower sum-of-squares with some other group.

### iii. Graphical display

Dendrogram 'Tree Diagram'

![Dendrogram](image)

### iv. Tests for Distortion

Cophenetic correlations. The similarity matrix $S$ contains the original similarity values between the OTUs (in this example it is a dissimilarity matrix $D$ of taxonomic distances). The UPGMA phenogram derived from it is shown, and from the phenogram the cophenetic distances are obtained to give the matrix $C$. The cophenetic correlation coefficient $r_C$ is the correlation between corresponding pairs from $C$ and $D$, and is 0.9911.

![Cophenetic Correlation](image)
Which Cluster Method to Use?

**SINGLE LINK**

General Behaviour of Different Methods

- **Single-link**: Often results in chaining
- **Complete-link**: Intense clustering
- **Group-average (weighted)**: Tends to join clusters with small variances
- **Group-average (unweighted)**: Intermediate between single and complete link
- **Median**: Can result in reversals
- **Centroid**: Can result in reversals
- **Minimum variance**: Often forms clusters of equal size

**General Experience**

Minimum variance is usually most useful but tends to produce clusters of fairly equal size, followed by group average. Single-link is least useful.

---

2. **TWINSPAN – Two-Way Indicator Species Analysis**

Mark Hill (1979)

**Differential variables** characterise groups, i.e. variables common on one side of dichotomy. Involves **qualitative** (+/−) concept, have to analyse numerical data as **PSEUDO-VARIABLES** (conjoint coding).

- **Species A**
  - 1-9%
  - → SPECIES A1
  - 5-10%
  - → SPECIES A2
  - 10-25%
  - → SPECIES A3

↑ cut level

**Basic idea** is to construct **hierarchical classification** by successive division.

Ordinate samples by correspondence analysis, divide at middle → group to left **negative**; group to right **positive**. Now **refine classification** using variables with maximum indicator value, so-called iterative character weighting and do a second ordination that gives a greater weight to the ‘preferentials’, namely species on one or other side of dichotomy.

Identify number of indicators that differ most in frequency of occurrence between two groups. Those associated with positive side +1 score, negative side -1. If variable 3 times more frequent on one side than other, variable is good indicator. Samples now reordered on basis of indicator scores. Refine second time to take account of other variables. Repeat on 2 groups to give 4, 8, 16 and so on until group reaches below minimum size.
**Variables classified** in much the same way. Variables classified using sample weights based on sample classification. Classified on basis of fidelity - how confined variables are to particular sample groups. Ratio of mean occurrence of variable in samples in group to mean occurrence of variable in samples not in group. Variables are ordered on basis of degree of fidelity within group, and then print out structured two-way table.

Concepts of INDICATOR SPECIES

DIFFERENTIALS and PREFERENTIALS

FIDELITY


Very robust - considers overall data structure

**TWINS PAN, T WINGRP, T WINDEND, WINTWINS**

**Pseudo-species Concept**

Each species can be represented by several pseudo-species, depending on the species abundance. A pseudo-species is present if the species value equals or exceeds the relevant user-defined cut-level.

<table>
<thead>
<tr>
<th>Original data</th>
<th>Sample 1</th>
<th>Sample 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cirsium palustre</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>Filipendula ulmaria</td>
<td>6</td>
<td>0</td>
</tr>
<tr>
<td>Junco effusus</td>
<td>15</td>
<td>25</td>
</tr>
</tbody>
</table>

Cut levels 1, 5, and 20 (user-defined)

<table>
<thead>
<tr>
<th>Pseudo-species</th>
<th>Cirsium palustre 1</th>
<th>0</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Filipendula ulmaria 1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>Junco effusus 1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>Junco effusus 2</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>Junco effusus 3</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

Thus quantitative data are transformed into categorical nominal (1/0) variables.

**Extensions to TWINS PAN**

Basic ordering of objects derived from correspondence analysis axis one. Axis is bisected and objects assigned to positive or negative groups at each stage. Can also use:

1. First PRINCIPAL COMPONENTS ANALYSIS axis

**ORBACLAN** C.W.N. Looiman

Ideal for TWINS PAN style classification of environmental data, e.g. chemistry data in different units, standardise to zero mean and unit variance, use PCA axis in **ORBACLAN** (cannot use standardised data in correspondence analysis, as negative values not possible).

2. First CANONICAL CORRESPONDENCE ANALYSIS axis.

**COINS PAN** T.J. Carleton et al. (1996) J. Vegetation Science 7: 125-130

First CCA axis is axis that is a linear combination of external environmental variables that maximises dispersion (spread) of species scores on axis, i.e. use a combination of biological and environmental data for basis of divisions. **COINS PAN** is a constrained TWINS PAN - ideal for stratigraphically ordered palaeoecological data if use sample order as environmental variable.
Ordination – Two Major Types

Indirect gradient analysis (Lecture 2)
Direct gradient analysis (Lecture 3)

Aims of Indirect Gradient Analysis

1. Summarise multivariate data in a convenient low-dimensional geometric way. **Dimension-reduction technique.**
2. Uncover the fundamental underlying structure of data. Assume that there is underlying **LATENT** structure. Occurrences of all species are determined by a few unknown environmental variables, LATENT VARIABLES, according to a simple response model. In ordination trying to recover and identify that **underlying structure.**

Underlying Response Models

A straight line displays the linear relation between the abundance value (y) of a species and an environmental variable (x), fitted to artificial data (●). (a = intercept; b = slope or regression coefficient).

A Gaussian curve displays a unimodal relation between the abundance value (y) of a species and an environmental variable (x), t = optimum or mode; t = tolerance; c = maximum = exp(a)).
Indirect gradient analysis can be viewed as being like regression analysis but with the MAJOR difference that in ordination the explanatory variables are not known environmental variables but theoretical "latent" variables.

Constructed so that they "best" explain the species data.

As in regression, each species is a response variable but in contrast to regression, consider all response variables simultaneously.

Principal Components Analysis

Estimation of fitting straight line and planes by least-squares regression

Fit a predictor variable to all the species in data by a series of least-squares regression of \( E_y = b_0 + b_1 x_1 + b_2 x_2 + \epsilon \),

we obtain for each regression the RESIDUAL SUM OF SQUARES, the sum of squared vertical distances between observed and fitted line.

Total of the separate residual sum of squares for all species, total residual SS, is a measure of how badly the predictor explains the data of all species.

What is the best fit that is theoretically possible with straight-line regression?

\[ y = b_0 + b_1 x_1 + \epsilon \]

or, if we have centred the data (subtracted the mean)

\[ y = b_1 x_1 + \epsilon \]

Defines an ORDINATION problem – construct the single hypothetical variable (latent variable) that gives the best fit to the data according to our linear equation.

PCA is the ordination technique that constructs the theoretical variable that mini-mises the total residual SS after fitting straight lines or planes to the data.

Principal Components Analysis

\[ y = b_0 + b_1 x_1 \]

\( y = b_1 x_1 \) (if centred data)

Three dimensional view of a plane fitted by least-squares regression of responses (●) on two explanatory variables PCA axis 1 and PCA axis 2. The residuals, i.e. the vertical distances between the responses and the fitted plane are shown. Least squares regression determines the plane by minimisation of the sum of these squared distances.

PCA-ordination diagram of the Dune Meadow Data in covariance biplot scaling with species represented by arrows. The b scale applies to species, the x scale to sites. Species not represented in the diagram lie close to the origin (0,0).

Total sum-of-squares (variance) = 1598 = SUM OF EIGENVALUES

Axis 1 or eigenvalue 1 = 471 = 29 %, 51% 1598

Axis 2 or eigenvalue 2 = 344 = 22 %

Each axis – vector of species slopes or scores (b) EIGENVECTORS

[Regression \( E_y = b_0 + b_1 x_1 \) = \( b_1 x_1 \) (if centred data)]
Correspondence Analysis (CA)

Invented independently numerous times:

1. **Correspondence Analysis**: Weighted Principal Components with Chi-squared metric.

2. **Optimal or Dual Scaling**: Find site and species scores so that (i) all species occurring in one site are as similar as possible, but (ii) species at different sites are as different as possible, and (iii) sites are dispersed as widely as possible relative to species scores.

3. **Reciprocal Averaging**: species scores are weighted averages of site scores, and simultaneously, site scores are weighted averages of species scores.

Like PCA in finding underlying latent variables but under the assumption of a unimodal response model.

How Many PCA Axes to Retain for Interpretation?


Scree plot.

Broken-stick.

Total variance (=\(\Sigma\)) divided randomly amongst the axes, eigenvalues follow a broken stick distribution.

\[ b_k = \sum_{i=0}^{k-1} \frac{1}{p} \]

\( p = \) number of variables (= mo) \( b_k = \) size of eigenvalue

e.g. 6 eigenvalues

\% variance – 40.8, 24.2, 15.8, 10.7, 6.1, 2.8

Artificial example of unimodal response curves of five species (A-E) with respect to standardized variables, showing different degrees of separation of the species curves.

a: moisture; b: First axis of CA; C: First axis of CA folded in this middle and the response curves of the species lowered by a factor of about 2. Sites are shown as dots at \( y = 1 \) if Species D is present and at \( y = 0 \) if Species D is absent.
Detrended Correspondence Analysis (DCA)

Aim to correct three 'artefacts' or 'faults' in CA:

1. **Detrending** to remove 'spurious' curvature in the ordination of strong single gradients
2. **Rescaling** to correct shrinking at the ends of ordination axes resulting in packing of sites at gradient ends
3. **Downweighting** to reduce the influence of rare species

Implemented originally in DECORANA and now in CANOCO

---

**CA Joint Plot**

CA ordination diagram of the Dune Meadow Data in Hill's scaling. The first axis is horizontal and the second axis vertical; the sites are represented by crosses. ($\lambda_2 = 0.26, \lambda_1 = 0.17$)

---

**Adding 'Unknown' Samples to PCA or CA**

**Passive samples**

$$X_i = \sum_{k=1}^{m} y_{ik} b_k$$

A PCA biplot showing the scores of the first and second components of the modern pollen spectra, the vectors of the pollen taxa, and the means and standard deviations of the five pollen zones from the Lateral Pond fossil site (zone 1 is the oldest); o represents the projection of the origin.

---

**CANOCO**

---

**Artificial Axis**

**Artemisia**

**Cyperaceae**

**Forest-tundra**

**Grassland**

**Grasstundra**

**Grasstundra plus shrub**

**Grasstundra plus shrub plus forest**

**Grasstundra plus shrub plus forest plus forest**

**Pine forest**

**Pine forest plus shrub**

**Pine forest plus shrub plus forest**

**Pine forest plus shrub plus forest plus forest**

**Pinus**

---

**ORDINATION**

**1st component**

**2nd component**

**3rd component**

**4th component**

**5th component**

---

**Ordination by CA of the two-way Petrie matrix in the table above.**

- 2: Arch effect in the ordination diagram (Hill's scaling; sites labelled as in table above; species shown). 3: One-dimensional DCA ordination, obtained by non-linearly rescaling the first CA axis. The sites would not show variation on the second axis of DCA.
Artificial example of unimodal response curves of five species (A-E) with respect to standardized variables, showing different degrees of separation of the species curves.

a: Moisture.
b: First axis of CA.
c: First axis of CA folded in this middle and the response curves of the species lowered by a factor of about 2.

Sites are shown as dots at y = 1 if species D is present and y = 0 if Species D is absent.

\[ \hat{\omega}_k = \frac{\sum y_k x_k}{\sum y_k} \quad \text{Species optima or score } \hat{\omega}_k \]

\[ \hat{x}_k = \frac{\sum y_k \hat{\omega}_k}{\sum y_k} \quad \text{Sample score} \]

Non-Linear Rescaling in DCA

Assume a species-packing model, variance of optima of species at a site (‘within-site variance’) is an estimate of average response curve breadth (‘tolerance’) of those species. Because of edge effect, species curves are narrower at edges of axes than in centre and within-site variance is correspondingly smaller in sites near the ends.

Rescale by equalising within-site variance at all points along axis by dividing into small segments, expand those with small within-site variance and contract those with large within-site variance.

Site scores then calculated as WA of species scores and standardised so that within-site variance is 1.

Length of axis is range of site scores in ‘standard deviation’ units.

Measure of total compositional change. Useful estimate in palaeoecology.

Detrending by Segments

Method of detrending by segments (simplified). The crosses indicate site scores before detrending; the dots are site scores after detrending. The dots are obtained by subtracting, within each of the five segments, the mean of the trial scores of the second axis (after Hill & Gauch, 1980).

PCA or CA/DCA?

PCA – linear response model
CA/DCA – unimodal response model

How to know which to use?

Gradient lengths important.
If short, good statistical reasons to use LINEAR methods.
If long, linear methods become less effective, UNIMODAL methods become more effective.
Range 1.5–3.0 standard deviations both are effective.

In practice:
Do a DCA first and establish gradient length.
If less than 2 SD, responses are monotonic. Use PCA.
If more than 2 SD, use CA or DCA.

When to use CA or DCA more difficult.
Ideally use CA (fewer assumptions) but if arch is present, use DCA.
Hypothetical diagram of the occurrence of species A-J over an environmental gradient. The length of the gradient is expressed in standard deviation units (SD units). Broken lines (A', C', H', J') describe fitted occurrences of species A, C, H and J respectively. If sampling takes place over a gradient range $<1.5$ SD, this means the occurrences of most species are best described by a linear model (A' and C'). If sampling takes place over a gradient range $>3$ SD, occurrences of most species are best described by an unimodal model (H' and J').

**Direct Gradient Analysis or Constrained (= Canonical) Ordination**

Lecture 3 and 5

Outline of ordination techniques. DCA (detrended correspondence analysis) was applied for the determination of the length of the gradient (LG). LG is important for choosing between ordination based on a linear or on a unimodal response model. In cases where LG $<3$, ordination based on linear response models is considered to be the most appropriate. PCA (principal component analysis) visualizes variation in species data in relation to best fitting theoretical variables. Environmental variables explaining this visualized variation are deduced afterwards, hence, indirectly, RDA (redundancy analysis) visualizes variation in species data directly in relation to quantified environmental variables. Before analysis, covariables may be introduced in RDA to compensate for systematic differences in experimental units. After RDA, a permutation test can be used to examine the significance of effects.

**Canonical Ordination Techniques**

Ordination and regression in one technique

Search for a weighted sum of environmental variables that fits the species best, i.e. that gives the maximum regression sum of squares

Ordination diagram
1) patterns of variation in the species data
2) main relationships between species and each environmental variable

Redundancy analysis = constrained or canonical PCA
Canonical correspondence analysis (CCA) = constrained CA
(Detrended CCA) = constrained DCA

Aves constrained to be linear combinations of environmental variables.

In effect PCA or CA with one extra step:

Do a multiple regression of site scores on the environmental variables and take as new site scores the fitted values of this regression.

**Multivariate regression** of Y on X.

Major use in analysing modern calibration data sets (assemblages in surface samples and associated modern environmental data)
**Primary Data in Gradient Analysis**

<table>
<thead>
<tr>
<th>Variable</th>
<th>Abundances or +/- variables</th>
<th>Values</th>
<th>Predictor or explanatory variables</th>
</tr>
</thead>
<tbody>
<tr>
<td>Species</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Canonical or Constrained Correspondence Analysis (CCA)**

Ordinary correspondence analysis gives:
1. **Site scores** which may be regarded as reflecting the underlying gradients.
2. **Species scores** which may be regarded as the location of species optima in the space spanned by site scores.

Canonical or constrained correspondence analysis gives in addition:
3. **Environmental scores** which define the gradient space.

These optimise the interpretability of the results.

CCA selects **linear combination of environmental variables that maximises dispersion of species scores**.

---

**Basic Terms**

- **Eigenvalue** = Maximalised dispersion of species scores along axis. In CCA usually smaller than in CA. If not, constraints are not useful.
- **Canonical coefficients** = 'best' weights or parameters of final regression.
- **Multiple correlation of regression** = Species-environment correlation. Correlation between site scores that are linear combinations of the environmental variables and site scores that are WA of species scores. Multiple correlation from the regression. Can be high even with poor models. Use with care!
- **Species scores** = WA optima of site scores, approximations to Gaussian optima along individual environmental gradients.
- **Site scores** = Linear combinations of environmental variables ('fitted values' of regression) (1).

Can also be calculated as weighted averages of species scores that are themselves WA of site scores (2).

1. LC scores are predicted or fitted values of multiple regression with constraining predictor variables ‘constraints’.
2. WA scores are weighted averages of species scores.

Generally always use (1) unless all predictor variables are 1/0 variables.

---

**Canonical Correspondence Analysis**

Canonical correspondence analysis: canonical coefficients (100 x c) and intra-set correlations (100 x r) of environmental variables with the first two axes of CCA for the Dune Meadow Data. The environmental variables were standardised first to make the canonical coefficients of different environmental variables comparable. The class SF of the nominal variable 'type of management' was used as a reference class in the analysis.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Coefficients</th>
<th>Correlations</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Axes 1</td>
<td>Axes 2</td>
</tr>
<tr>
<td>A1</td>
<td>9</td>
<td>-37</td>
</tr>
<tr>
<td>Moisture</td>
<td>71</td>
<td>-29</td>
</tr>
<tr>
<td>Use</td>
<td>25</td>
<td>5</td>
</tr>
<tr>
<td>Manure</td>
<td>-7</td>
<td>-27</td>
</tr>
<tr>
<td>SF</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>BF</td>
<td>-9</td>
<td>16</td>
</tr>
<tr>
<td>HF</td>
<td>18</td>
<td>19</td>
</tr>
<tr>
<td>NM</td>
<td>20</td>
<td>92</td>
</tr>
</tbody>
</table>

**CANOCO**

**R**
CCA of the Dune Meadow Data. a: Ordination diagram with environmental variables represented by arrows, the c scale applies to environmental variables, the u scale to species and sites, the types of management are also shown by closed squares at the centroids of the meadows of the corresponding types of management.

b: Inferred ranking of the species along the variable amount of manure, based on the biplot interpretation of Part a of this figure.

Outline of ordination techniques presented here. DCA (detrended correspondence analysis) was applied for the determination of the length of the gradient (LG). LG is important for choosing between ordination based on a linear or on an unimodal response model. Correspondence analysis (CA) is not considered any further because in “microcosm experiment discussed here LG < or = 1.5 D units. LG < 3 D units are considered to be typical in experimental ecotoxicology. In cases where LG < 3, ordination based on linear response models is considered to be most appropriate. PCA (principal component analysis) visualizes variation in species data directly in relation to quantified environmental variables. Environmental variables explaining this visualised variation are deduced afterwards, hence, indirectly. RDA (redundancy analysis) visualises variation in species data directly in relation to quantified environmental variables. Before analysis, covariables may be introduced in RDA to compensate for systematic differences in experimental units. After RDA, a permutation test can be used to examine the significance of effects.

Passive “fossil” samples added into CCA of modern data

Redundancy Analysis – Constrained PCA

Short (< 2SD) compositional gradients
Linear or monotonic responses

Reduced-rank regression
PCA of y with respect to x
Two-block mode C PLS
PCA of instrumental variables Rao (1964)

PCA - best hypothetical latent variable is the one that gives the smallest total residual sum of squares
RDA - selects linear combination of environmental variables that gives smallest total residual sum of squares

ter Braak (1994) Ecocience 1, 127–140 Canonical community ordination
Part I: Basic theory and linear methods
Partial Constrained Ordinations (Partial CCA, RDA, etc)

- e.g. pollution effects
- seasonal effects → COVARIABLES

Eliminate (partial out) effect of covariables. Relate residual variation to pollution variables.

Replace environmental variables by their residuals obtained by regressing each pollution variable on the covariables.

Statistical Testing of Constrained Ordination Results

Statistical significance of species-environmental relationships. Monte Carlo permutation tests.

Randomly permute the environmental data, relate to species data 'random data set'.
Calculate eigenvalue and sum of all canonical eigenvalues (trace). Repeat many times (99).

If species react to the environmental variables, observed test statistic ($t_i$ or trace) for observed data should be larger than most (e.g. 95%) of test statistics calculated from random data. If observed value is in top 5% highest values, conclude species are significantly related to the environmental variables.

Partial CCA

Ordination diagram of a partial CCA of diatom species A in dyes with 3 explanatory variables:
24 variables of interest (arrows) and 2 covariables (chlorides concentration and season).
The diagram is symmetrical scaled (23) and shows selected species and standardized variables and, instead of individual dyes, centroids (of dye clusters). The variables of interest shown are: BOD = biological oxygen demand, Ca = calcium, Fe = ferrous compounds, N = nitrates/nitrites, O2 = oxygen, P = ortho phosphate, Si = silicon compounds, WIDTH = dye width, and soil types (CLAY, PEAT). All variables except BOD, WIDTH, CLAY and PEAT were transformed to logarithms because of their skew distribution. The diatoms shown are: Ach_th = Achnanthes hungarica; Ach_m = A. minutissima; Apl_h = Amphora hustedtii; Apl_v = A. vanderplanki; Coc_pa = Cocconeis placentula; Dun_k = Dunaliella koliae; Euc_l = Eucampia littoralis; Euc_p = E. pectinata; Gom_s = Gomphonema autumnale; Gom_p = G. peltatum; Hap_j = Haplochromis jorgerdii; Hap_ac = H. acceptabilis; Ne_n = Nitzschia naviculata; Ne_k = N. kotikovii; Ne_b = N. balci; Ne_a = N. aesticum; Ne_s = N. seminulum; Ne_s = N. seminulum; Nillep = Nilopsis lepadiformis; Nil_le = N. lepadiformis; Nil_b = N. brunnescens; Nil_d = N. dissipata; Nil_p = N. peltata; Tha_c = Thalassiothrix corvus.
(Adapted from H. Sorensen in prep.)

Natural variation due to sampling season and due to gradient from fresh to brackish water partialled out by partial CCA.

Variation due to pollution could now be assumed.
Partitioning Variance

Regression → total SS = regression SS + residual SS

Borcard et al. (1992) Ecology 73, 1045–1055

Variance decomposition into 4 components using (partial) CCA or RDA

Discriminant Analysis

Discriminant analysis - a form of constrained or direct gradient analysis where the constraints are a priori group membership.
- a statistician's concept of classification, analysis of classes and assignment of unknowns to classes

Discriminant Analysis

1. Taxonomy – species discrimination
2. Pollen analysis – pollen grain separation
3. Morphometrics – sexual dimorphism
4. Geology – distinguishing rock samples

Discriminant function – **linear combination of variables** \( x_1 \) and \( x_2 \),

\[
z = b_1 x_1 + b_2 x_2
\]

where \( b_1 \) and \( b_2 \) are weights attached to each variable that determine the relative contributions of the variable.

**Geometrically** – line that passes through where group ellipsoids cut each other \( L \), then draw a line perpendicular to \( L \), \( M \) that passes through the origin, \( O \) Project ellipses onto the perpendicular to give two univariate distributions \( S_1 \) and \( S_2 \) on discriminant function \( M \).
Can position the means of group A and of group B on the discriminant function
\[
F_a = \lambda_1 x_1 + \lambda_2 x_2 = -783.63 \times 0.340 + 75.62 \times 1.210 = -378.81
\]
\[
= -346.64
\]

**Solve from:**

\[
\lambda = S_a^{-1}(\bar{x} - \bar{y}) = S_a^{-1} D
\]

**Identification of Unknown Objects**

Assumption that probability of unknown object belonging to either group only is equal. Presupposes no other possible groups it could come from.

Closeness rather than either/or identification.

If unknown, \( u \) has position on discriminant function:

\[
R_a = \lambda_1 u_1 + \lambda_2 u_2
\]

then:

\[
x^2 = (\bar{y} - \mu_a)^T S^{-1} (\bar{y} - \mu_a)
\]

\[
x^2 = (\bar{y} - \mu_b)^T S^{-1} (\bar{y} - \mu_b)
\]

\( m \) degrees of freedom


*Picea glauca* (white spruce) pollen

*Picea mariana* (black spruce) pollen
Quantitative characters of *Picea* pollen (variables $x_1 - x_5$). The means (vertical line), ± 1 standard deviation (open box), and range (horizontal line) are shown for the reference populations of the three species.

Fig. 7. Manifestion of fossil *Picea* pollen of Holocene age from two levels at Paradise Lake, Labrador. The *Picea* pollen curve is shown, along with selected radiocarbon dates. The positions of the fossil grains on the discriminant function for the samples examined are presented in comparison with the position of modern *P. glauca* and *P. mariana* pollen on the discriminant function.
2. Confirmatory Data Analysis

Constrained ordination techniques (CCA, RDA) and associated Monte Carlo permutation tests.

In reality multivariate regression of \( Y \) (response variables) on \( X \) (predictor or explanatory variables), possibly with covariables ('nuisance' variables) \( Z \).

Lectures 3 & 6

3. Statistical Modelling or Regression Analysis

Explore relationships between variables and their environment
+/- or abundances for species (responses)

Individual species, one or more environmental variable (predictors)
Species abundance or presence/absence - response variable \( Y \)
Environmental variables - explanatory or predictor variables \( X \)

Aims

1. To describe response variable as a function of one or more explanatory variables. This RESPONSE FUNCTION usually cannot be chosen so that the function will predict responses without error. **Try to make these errors as small as possible and to average them to zero.**

2. To predict the response variable under some new value of an explanatory variable. The value predicted by the response function is the **expected response, the response with the error averaged out.**

Main Uses

1. Estimate ecological parameters for species, e.g. optimum, amplitude (tolerance) - ESTIMATION AND DESCRIPTION. (Lecture 5)

2. Assess which explanatory variables contribute most to a species response and which explanatory variables appear to be unimportant. Statistical testing - MODELLING.

3. Predict species responses (+/-, abundance) from sites with observed values of explanatory variables - PREDICTION.

4. Predict environmental variables from species data - CALIBRATION. (Lecture 5)
Response Model

Systematic part - regression equation
Error part - statistical distribution of error

\[ Y = \beta_0 + \beta_1 x + \varepsilon \]

\( \beta_0, \beta_1 \) fixed but unknown coefficients
\( \beta_0 = \text{intercept} \)
\( \beta_1 = \text{slope} \)

**Error part** is distribution of \( \varepsilon \), the random variation of the observed response around the expected response.

**Aim is to estimate systematic part** from data while taking account of error part of model.

In fitting a straight line, systematic part simply estimated by estimating \( \beta_0 \) and \( \beta_1 \).

Least squares estimation - error part assumed to be normally distributed.

---

**Quantitative Response Variable, Quantitative Explanatory or Predictor Variable**

Straight line fitted by least-squares regression of log-transformed relative cover on mean water-table. The vertical bar on the far right has length equal to twice the sample standard deviation \( \sigma_y \); the other two smaller vertical bars are twice the length of the residual standard deviation \( \sigma_{\epsilon} \). The dashed line is a parabola fitted to the same data (*).

**Error part** - responses independent and normally distributed around expected values \( \hat{y} \).

---

**Quantitative Response Variable, Quantitative Explanatory Variable**

**Does expected response depend on water table?**

\[ F = 27.56 >> 4.4 \quad \text{(critical value 5%) \quad df (1, 18)} \]

\( F = \text{MS regression} \quad \text{df} = \text{parameters - 1}, \text{MS residual} \quad n - \text{parameters} \)

**Does slope \( \beta_1 = 0? \)**

\[ t \text{ of } \beta_1 \left( \frac{\beta_1 - 0}{\sigma_{\beta_1}} \right) = -5.25 > \text{absolute value of critical value of two-tailed } t \text{-test at } 5\% \]

\[ t_{0.025,18} = 2.10 \]

\( \beta_1 \) not equal to 0 \[ \text{[exactly equivalent to } F \text{ test } \left( \frac{\beta_1}{\sigma_{\beta_1}} \right)^2 = F \] \)

**Construct 95% confidence interval for \( \beta_1 \)**

\[ \text{estimate} \times t_{0.025,18} \times \text{se} = -0.052 / -0.022 \]

Does not include 0 \( \Rightarrow \) is unlikely value for \( \beta_1 \)

**Check assumptions of response model**

Plot residuals against \( x \) and \( \hat{y} \).
Could we fit a curve to these data better than a straight line?

**Parabola**

\[
EY = b_0 + b_1 x + b_2 x^2
\]

Straight line fitted by least-squares regression of log-transformed relative cover on mean water table. The vertical bar on the far right has a length equal to twice the sample standard deviation \(s\); the other two smaller vertical bars are twice the length of the residual standard deviation \(s_e\).

The dashed line is a parabola fitted to the same data (+).

Polynomial regression

---

**Many Explanatory Variables, All Quantitative**

Response variable expressed as a function of two or more explanatory variables. Not the same as separate analyses because of correlations between explanatory variables and interaction effects.

**MULTIPLE LEAST-SQUARES REGRESSION**

**Planes**

\[
EY = b_0 + b_1 x_1 + b_2 x_2
\]

**explanatory variables**

- \(b_0\) – expected response when \(x_1\) and \(x_2\) = 0
- \(b_1\) – rate of change in expected response along \(x_1\) axis
- \(b_2\) – rate of change in expected response along \(x_2\) axis

\(b_1\) measures change of \(EY\) with \(x_1\) for a fixed value of \(x_2\)

\(b_2\) measures change of \(EY\) with \(x_2\) for a fixed value of \(x_1\)

---

**Parabola fitted by least-squares regression: parameter estimates and ANOVA table for the transformed relative cover of above figure.**

<table>
<thead>
<tr>
<th>Term</th>
<th>Parameter</th>
<th>Estimate</th>
<th>s.e.</th>
<th>t</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>(b_0)</td>
<td>3.988</td>
<td>0.819</td>
<td>4.88</td>
</tr>
<tr>
<td>Water-table</td>
<td>(b_1)</td>
<td>-0.0187</td>
<td>0.0317</td>
<td>-0.59</td>
</tr>
<tr>
<td>(Water-table)²</td>
<td>(b_2)</td>
<td>-0.000169</td>
<td>0.000284</td>
<td>-0.59</td>
</tr>
</tbody>
</table>

\(t\) not different from 0

<table>
<thead>
<tr>
<th>ANOVA table</th>
</tr>
</thead>
<tbody>
<tr>
<td>d.f.</td>
</tr>
<tr>
<td>Regression</td>
</tr>
<tr>
<td>Residual</td>
</tr>
<tr>
<td>Total</td>
</tr>
</tbody>
</table>

\(R^2_{adj} = 0.57\)

\(R^2_{adj} = 0.58\) for linear model

---

A straight line displays the linear relationship between the abundance value \(y\) of a species and an environmental variable \(x\), fitted to artificial data (+); \(a = \) intercept; \(b = \) slope or regression coefficient.

A plane displays the linear relationship between the abundance value \(y\) of a species and two environmental variables \(x_1\) and \(x_2\), fitted to artificial data (+).
Three-dimensional view of a plane fitted by least-squares regression of responses \( y \) on two explanatory variables \( x_1 \) and \( x_2 \). The residuals, i.e. the vertical distances between the responses and the fitted plane, are shown. Least-squares regression determines the plane by minimization of the sum of these squared vertical distances.

Estimates of \( b_0 \), \( b_1 \), \( b_2 \) and standard errors and \( t \) (estimate / se)

<table>
<thead>
<tr>
<th>ANOVA total SS, residual SS, regression SS</th>
</tr>
</thead>
<tbody>
<tr>
<td>( R^2 = 1 )</td>
</tr>
<tr>
<td>( \hat{y} = b_0 + b_1 x_1 + b_2 x_2 )</td>
</tr>
</tbody>
</table>

**MULTICOLLINEARITY**

Selection of explanatory variables: Forward selection Backward selection 'Best-set' selection Stepwise

**Regression Analysis Summary**

**GENERAL LINEAR MODEL**

Response variable: \( Y = \hat{Y} + e \)

where \( \hat{Y} \) is the expected value of \( Y \) for particular values of the predictors and \( e \) is the variability ("error") of the true values around the expected values \( \hat{Y} \).

The expected value of the response variable is a function of the predictor variables

\[ EY = f(X_1, \ldots, X_m) \]

\( EY \) = systematic component, \( e \) = stochastic or error component.

Simple linear regression

\[ EY = f(X) = \beta_0 + \beta_1 X \]

Polynomial regression

\[ EY = \beta_0 + \beta_1 X + \beta_2 X^2 \]

Null model

\[ EY = \beta_0 \]

What to do with response variables that are presence/absence or proportions?

What to do with predictor variables that do not follow normal error distribution?

Need GENERALISED LINEAR MODELS
Generalised Linear Models (GLM)

Primary aim - provide a mathematical expression for use in the description, interpretation, prediction, involving the reconstruction of relationship between variables

\[ y = a + bx \]

Want to find linear combinations of predictor (= explanatory or independent) \( x \) variables which best predict the response variable \( y \).

\[ y = a + bx + \varepsilon \]

Systematic component

Error component

Influences estimates of \( a \) and \( b \)

Five steps:
1. Identification of response \( (y) \) and predictor \( (x) \) variables.
2. Identification of model equation.
3. Choice of appropriate error function for response variable.
5. Appropriate model evaluation procedures.

Advantages of GLM

1. **Error function** can follow several distributions, not just normal distribution.

   Errors may be:
   - strongly skewed
   - kurtotic
   - strictly bounded (0/1, proportions, %)
   - cannot lead to negative fitted values (counts)

2. Linear combination of the \( x \) variables, **LINEAR PREDICTOR** \( \eta \) (eta) may be used to predict \( y \) through a non-linear intermediary function, so-called **LINK FUNCTION**. Use of a non-linear link function allows the model to use response and predictor variables that are measured on different scales by effectively mapping the linear predictor onto the scale of the response variable.

3. Common framework for regression and **ANOVA**.

4. Can handle many problems that look non-linear.

5. Not necessary to transform data since the regression is transformed through the link function.

Structure of Generalised Linear Model

(1) **ERROR FUNCTION**

<table>
<thead>
<tr>
<th>Distribution</th>
<th>Probability</th>
<th>Range of ( y )</th>
<th>Variance function</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gaussian</td>
<td>( \rightarrow \rightarrow ) ( \rightarrow \rightarrow )</td>
<td>( \rightarrow \rightarrow )</td>
<td>( 1 )</td>
</tr>
<tr>
<td>Poisson</td>
<td>( 0 (1) ) ( \rightarrow \rightarrow )</td>
<td>( 0 \rightarrow \rightarrow )</td>
<td>( \mu )</td>
</tr>
<tr>
<td>Binomial</td>
<td>( 0 (1) ) ( n ) ( \mu (1 - \mu/n) )</td>
<td>( 0 \rightarrow \rightarrow )</td>
<td>( \mu^2 )</td>
</tr>
<tr>
<td>Gamma</td>
<td>( 0 \rightarrow \rightarrow ) ( \rightarrow \rightarrow )</td>
<td>( 0 \rightarrow \rightarrow )</td>
<td>( \mu^2 )</td>
</tr>
<tr>
<td>Inverse Gaussian</td>
<td>( 0 \rightarrow \rightarrow ) ( \rightarrow \rightarrow )</td>
<td>( 0 \rightarrow \rightarrow )</td>
<td>( \mu^3 )</td>
</tr>
</tbody>
</table>

Choice depends on range of \( y \) and on the proportional relationship between variance and expected value \( \eta \).
(2) LINEAR PREDICTOR

\[ \eta = \sum_{j=1}^{m} \chi_{j} \beta_{j} \]

unknown parameters

LINEAR STRUCTURE

predictor variables

To determine fit of a given model, linear predictor is needed for each value of response variable and then compares predicted value with a transformed value of \( y \); the transformation to be applied specified by LINK FUNCTION. The fitted value is computed by applying the inverse of the link function to get back to the original scale of measurement of \( y \).

- Log-link - Fitted values are anti-log of linear predictor
- Reciprocal link - Fitted values are reciprocal of linear predictor

(3) LINK FUNCTION

Link function relates the mean value of \( y \) to its linear predictor \( \eta \).

\[ \eta = g(\mu) \]

where \( g(\cdot) \) is link function and \( \mu \) are fitted values of \( y \).

\[ y = \text{predictable component} + \text{error component} \]

\[ y = \mu + \epsilon \]

Linear predictor is sum of terms for each of the parameters and value of \( \eta \) is obtained by transforming value of \( y \) by link function and obtaining predicted value of \( y \) as inverse link function.

\[ \mu = g^{-1}(\eta) \]

Can combine link function and linear predictor to form basic or core equation of GLM.

\[ y = g^{-1}(\eta) + \epsilon \]

OR

\[ \eta = g(y) \]

Type of analysis

- Probit
- Logit
- GLIM

Examples of some error distributions and link functions

<table>
<thead>
<tr>
<th>Error Distribution</th>
<th>Link Function</th>
<th>Type of Analysis</th>
</tr>
</thead>
<tbody>
<tr>
<td>Normal</td>
<td>Identity</td>
<td>Continuous</td>
</tr>
<tr>
<td>Binomial</td>
<td>Logit</td>
<td>Proportional</td>
</tr>
<tr>
<td>Poisson</td>
<td>Log</td>
<td>Continuous</td>
</tr>
</tbody>
</table>

Some common link functions in generalised linear models

<table>
<thead>
<tr>
<th>Link Function</th>
<th>Definition</th>
<th>Range of fitted value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Identity</td>
<td>( \eta = \mu )</td>
<td>0 to ( \infty )</td>
</tr>
<tr>
<td>Log</td>
<td>( \eta = \log \mu )</td>
<td>0 to ( \infty )</td>
</tr>
<tr>
<td>Probit</td>
<td>( \eta = \Phi^{-1}(\mu) )</td>
<td>0 to 1</td>
</tr>
</tbody>
</table>

SOURCE: "Applied Regression and Other Multivariate Methods" by John Fox

Types of GLM Analysis

Examples of some error distributions and link functions

<table>
<thead>
<tr>
<th>Error Distribution</th>
<th>Link Function</th>
<th>Type of Analysis</th>
</tr>
</thead>
<tbody>
<tr>
<td>Normal</td>
<td>Identity</td>
<td>Continuous</td>
</tr>
<tr>
<td>Binomial</td>
<td>Logit</td>
<td>Proportional</td>
</tr>
<tr>
<td>Poisson</td>
<td>Log</td>
<td>Continuous</td>
</tr>
</tbody>
</table>

Some common link functions in generalised linear models

<table>
<thead>
<tr>
<th>Link Function</th>
<th>Definition</th>
<th>Range of fitted value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Identity</td>
<td>( \eta = \mu )</td>
<td>0 to ( \infty )</td>
</tr>
<tr>
<td>Log</td>
<td>( \eta = \log \mu )</td>
<td>0 to ( \infty )</td>
</tr>
<tr>
<td>Probit</td>
<td>( \eta = \Phi^{-1}(\mu) )</td>
<td>0 to 1</td>
</tr>
</tbody>
</table>

SOURCE: "Applied Regression and Other Multivariate Methods" by John Fox

R
Generalised Linear Models – A Summary

Mathematical extensions of linear models that do not force data into unnatural scales. Thereby allow for non-linearity and non-constant variance structures in the data.

Based on an assumed relationship (link function) between the mean of the response variable and the linear combination of the predictor variables.

Data can be assumed to be from several families of probability distributions – normal, binomial, Poisson, gamma, etc – which better fit the non-normal error structures of most real-life data.

More flexible and better suited for analysing real-life data than ‘conventional’ regression techniques.

Gaussian Logit Model

Gaussian logit model

Gaussian response function: GLM estimation

\[ \mu = h \exp \left( \frac{x - u}{2\sigma^2} \right) \]

\[ \log \mu = b_0 + b_1 x + b_2 x^2 \]

- Gaussian response function can be written as a generalized linear model (which is easy to fit)
  - Linear predictor: explanatory variables x and x^2
  - Link function log (or logit)
  - Error Poisson (or Binomial)

- The original Gaussian response parameters can be found by

\[ u = -b_2/2b_1 \] OPTIMUM

\[ t = \sqrt{-1/2b_1} \] TOLERANCE

\[ h = \exp(b_1 - b_2^2 / 4b_1) \] HEIGHT
Results of fitting **Gaussian logit, linear logit and null models** to the SWAP 167-lake training set and lake-water pH

**225 taxa**

<table>
<thead>
<tr>
<th>Model</th>
<th>No. of taxa</th>
</tr>
</thead>
<tbody>
<tr>
<td>Non-converging</td>
<td>1</td>
</tr>
<tr>
<td>Gaussian unimodal curves with maxima (b2 &lt; 0)</td>
<td>88</td>
</tr>
<tr>
<td>Linear logit sigmoidal curves</td>
<td>78</td>
</tr>
<tr>
<td>Gaussian unimodal curves with minima (b2 &gt; 0)</td>
<td>5</td>
</tr>
<tr>
<td>No pattern</td>
<td>53</td>
</tr>
<tr>
<td>Significant Gaussian logit model</td>
<td>88</td>
</tr>
<tr>
<td>Significant linear logit model</td>
<td>78</td>
</tr>
<tr>
<td>Non-significant fit to pH</td>
<td>58</td>
</tr>
</tbody>
</table>

**Assessing Assumptions of Regression Models**

Regression diagnostics

1. relationship between Y and X is linear, perhaps after transformation
2. variance of random error is constant for all observations
3. errors are normally distributed
4. errors for n observations are independently distributed

Assumption (2) required to justify choosing estimates of \( b \) parameters so as to minimise residual SS and needed in tests of \( t \) and \( F \) values. Clearly in minimising SS residuals, essential that no residuals should be larger than others.

Assumption (3) needed to justify significance tests and confidence intervals.

**RESIDUAL PLOTS**

Plot \((Y - \hat{Y})\) against \( \hat{Y} \) or \( X \)

**Example of GLM Model Criticism**

- **Residual plots** from the multiple regression of gene frequencies on environmental variables for *Euphydryas editha*.
  - (a) standardised residuals plotted against Y values from the regression equation,
  - (b) standardised residuals against \( X_1 \)
  - (c) standardised residuals against \( X_2 \)
  - (d) standardised residuals against \( X_3 \)
  - (e) standardised residuals against \( X_4 \) and
  - (f) normal probability plot.

- Normal probability plot—plot ordered standardised residuals against expected values assuming standard normal distribution.

If \((Y - \hat{Y})\) is standard residual for \( Z \) expected value is value for standardised normal distribution that exceeds proportion \((1 - \phi(k))\) / \( (n + 1) \)

**Standardised residual** \( \frac{Y - \hat{Y}}{\sqrt{\text{MSE}}} \)

**Residuals**

- *P. vivipara* and *E. editha*.
- 110 lakes data.
- Frustulia rhomboides

**Poisson error, offset total**

- **Fitted value**

**Normal deviate**

- **Anscombe residual**

- **Gaussian residual**

- **Linear residual**

- **Log residual**

- **Corrected residual**

- **Standardised residual**

**Fitted value**

- 5.5
- 6
- 6.5

**Normal deviate**

- 0.06
- 0.12
- 0.14
- 0.16

**Anscombe residual**

- 12
- 8

**Poisson error, offset total**

- 110 lakes data
Locally Weighted Regression

W. S. Cleveland

Locally weighted or regression scatterplot smoothing

May be unreasonable to expect a single functional relationship between $Y$ and $X$ throughout range of $X$.

(Running averages for time-series – smooth by average of $y_{i-1}$, $y_{i}$, $y_{i+1}$ or add weights to $y_{i}$)

LOESS - more general

1. Decide how ‘smooth’ the fitted relationship should be.
2. Each observation given a weight depending on distance to observation $x_i$ for adjacent points considered.
3. Fit simple linear regression for adjacent points using weighted least squares.
4. Calculate residuals (difference between observed and fitted $y$).
5. Estimate robustness weights based on residuals, so that well-fitted points have high weight.
6. Repeat LOESS procedure but with new weights based on robustness weights and distance weights.
7. Repeat for different degree of smoothness to find ‘optimal’ smoother.

R

An air pollutant, ozone, is graphed against wind speed from the graph we can see that ozone tends to decrease as wind speed increases, but judging whether the pattern is linear or nonlinear is difficult.

Loess, a method for smoothing data, is used to compute a curve summarizing the dependence of ozone on wind speed. With the curve superposed, we can now see that the dependence of ozone on wind speed is nonlinear.

---

Generalised Additive Models (GAM)

Semi-parametric extension of generalised linear models GLM:

GLM

$g(E(y)) = \alpha + \beta x$

predictor variables

link function

modelled abundance of response variable $y$

$E(y) = \alpha + \beta \mathbf{x}$

regression coefficients or model parameters

e.g. Ordinary least-squares regression - identity link, normal error distribution

Logit (p) = $\log \left( \frac{p}{1-p} \right) = \alpha + \beta x_1 + \beta x_2 + \beta x_3$

Required a priori statistical model, e.g. Gaussian logit model, b-response model, etc.

What if the response is bimodal, is badly skewed, or is more complex than a priori model?

GLM may not be flexible enough to approximate the true response adequately. GLM are model-driven.
GAM

\[ g(E_y) = \alpha + f(x) = \alpha + \sum_{j=1}^{n} f_j(x_j) \]

predictor variables

intercept or constant

modelling abundance of response variable \( y \)

unspecified smoothing functions estimated from data using smoothers to give maximum explanatory power

\( f \) are unspecified smoothing functions estimated from the data using techniques developed for smoothing scatter plots, e.g. loess, cubic splines.

Data determine shape of response curve rather than being limited by the shapes available in parametric GLM. Can detect bimodality and extreme skewness.

Regression surface \( g(E_y) \) for taxon \( y \) is expressed as a sum of the functions for each variable \( x_j \) so each has an additive effect, hence GAM.

**GAM are data-driven**, the resulting fitted values do not come from an *a priori* model. Still some statistical framework with link functions and error specification

Need to specify the type of smoother and their complexity in terms of their degrees of freedom.
**Generalised Additive Models – A Summary**

GAMs are semi-parametric extensions of GLMs.

Only underlying assumptions are that the functions are additive and that the components are smooth. Like GLM, uses a link function to establish a relationship between the mean of the response variable and a ‘smoothed’ function of the predictor variable(s).

Strength is ability to deal with highly non-linear and non-monotonic relationships between the response variable and the set of predictor variables.

Data-driven rather than model-driven (as in GLM). Data determine the nature of the relationship between response and predictor variables.

Can handle non-linear data structures. Very useful exploratory tool.

---

**4. Quantitative Environmental Reconstruction**

Builds on statistical modelling

\[ Y = f(X) \]

where \( Y \) is biological variable and \( X \) is environmental variable

Reconstruction is two-step process:

1. \( Y_m = f(X_m) \) Regression

   where \( m \) refers to modern assemblages

2. \( X_p = f_m(Y_p) \) Calibration

   where \( p \) refers to past or fossil assemblages

   Lecture 5

---

**5. Time-series Analysis**

Set of specialised techniques for analysing the behaviour of one or more variables over time and for investigating the relationships of two or more variables over time.

   Lecture 6

---

**A Continuum of Regression Models**

Simple Linear Regression \( \geq \) Multiple Linear Regression > GLM > GAM

SLR and MLR - most restrictive in terms of assumptions but are most used (and misused!)

GLM - fairly general but still model-based

GAM - most general as data-based
Course Powerpoints

In some classes, some of the slides are rather technical.
They are included for the sake of completion to the topic under discussion.
They are for reference only and are marked REF