Hypothesis Testing and Time-Series Analysis

John Birks

Quantitative Methods in Palaeoecology and Palaeoclimatology

PAGES Valdivia October 2010

CONTENTS

Randomisation tests

Simple introductory example

Mandible lengths of male and female jackals in Natural History Museum

Male	120	107	110	116	114	111	113	117	114	112	mm
Female	110	111	107	108	110	105	107	106	111	111	mm

Is there any evidence of difference in mean lengths for two sexes? Male mean larger than female mean.

Null hypothesis (H_o) – no difference in mean lengths for two sexes, any difference is purely due to chance. If H_o consistent with data, no reason to reject this in favour of alternative hypothesis that males have a larger mean that females.

Classical hypothesis testing – t-test for comparison of 2 means

Group 1	n_1 objects	Group 2	n_2
	$\overline{x}_1$ mean		$\overline{x}_2$
	s_1		s_2

Assume that values for group 1 are random sample from a normal distribution with μ_1 mean and standard deviation σ, and mean μ_2 and standard deviation σ

H_0 $\mu_1 = \mu_2$ H_1 $\mu_1 > \mu_2$

Test null hypothesis with estimate of common within-group s.d.

$S = \sqrt{[\{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2\}/(n_1 + n_2 - 2)]}$

$T = (\bar{x}_1 - \bar{x}_2)/(S\sqrt{(1/n_1 + 1/n_2)})$

If H_0 true, T will be a random value from t-distribution with $n_1 + n_2 - 2$ d.f.

Jackal data

$\bar{x}_1 = 113.4$mm $\quad s_1 = 3.72$mm $\quad s = 3.08$

$\bar{x}_2 = 108.6$mm $\quad s_2 = 2.27$mm

$\therefore T = 3.484 \quad 18$ d.f.

Probability of a value this large is 0.0013 if null hypothesis is true.

$\therefore$ Sample result is nearly significant at 0.1% level. Strong evidence against null hypothesis. Support for alternative hypothesis.

Assumptions of T-test

1. Random sampling of individuals from the populations of interest
2. Equal population standard deviations for males and females
3. Normal distributions within groups

Alternative Approach

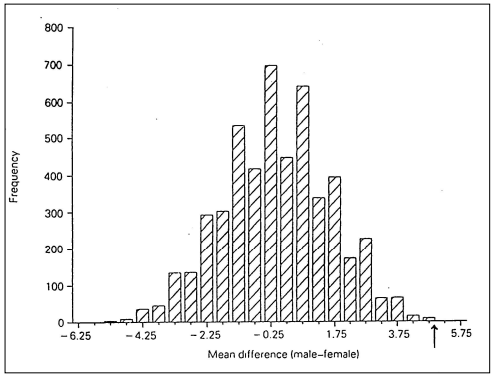
If there is no difference between the two sexes, then the length distribution in the two groups will just be a typical result of allocating 20 lengths at random into 2 groups each of size 10. Compare observed difference with distribution of differences found with random allocation.

TEST:

1. Find mean scores for male and female and difference D_0 for observed data.
2. Randomly allocate 10 lengths to male group, remaining 10 to female. Calculate D_1.
3. Repeat many times (n e.g. 999 times) to find an empirical distribution of D that occurs by random allocation. RANDOMISATION DISTRIBUTION.
4. If D_0 looks like a 'typical' value from this randomisation distribution, conclude that allocation of lengths to males and females is essentially random and thus there is no difference in length values. If D_0 unusually large, say in top 5% tail of randomisation distribution, observed data unlikely to have arisen if null hypothesis is true. Conclude alternative model is more plausible.

   If D_0 in top 1% tail, significant at 1% level
   If D_0 in top 0.1% tail, significant at 0.1% level

The distribution of the differences observed between the mean for males and the mean for females when 20 measurements of mandible lengths are randomly allocated, 10 to each sex. **4999 randomisations.**



$\overline{x}_1$ = 113.4mm $\overline{x}_2$ = 108.6mm D_0 = 4.8mm

Only nine were 4.8 or more, including D_0.

Six were 4.8 2 > 4.8

Significance level = $\frac{9}{5000}$. = 0.0018 = 0.18%

(*cf*. t-test 0.0013 0.13%)

$^{20}C_{10}$ = 184,756. 5000 only 2.7% of all possibilities.

Three main advantages

1. Valid even without random samples.
2. Easy to take account of particular features of data.
3. Can use 'non-standard' test statistics.

   Tell us if a certain pattern could or could not be caused/arisen by chance. Completely specific to data set.

Randomisation tests and Monte Carlo permutation tests

If all data arrangements are equally likely, RANDOMISATION TEST with random sampling of randomisation distribution. Otherwise, MONTE CARLO PERMUTATION TEST.

Validity depends on validity of permutation types for particular data-type – time-series stratigraphical data, spatial grids, repeated measurements (BACI). All require particular types of permutations.

pH changes in Round Loch of Glenhead

Monte Carlo permutation tests

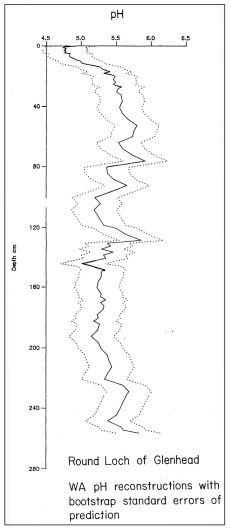
ROUND LOCH OF GLENHEAD

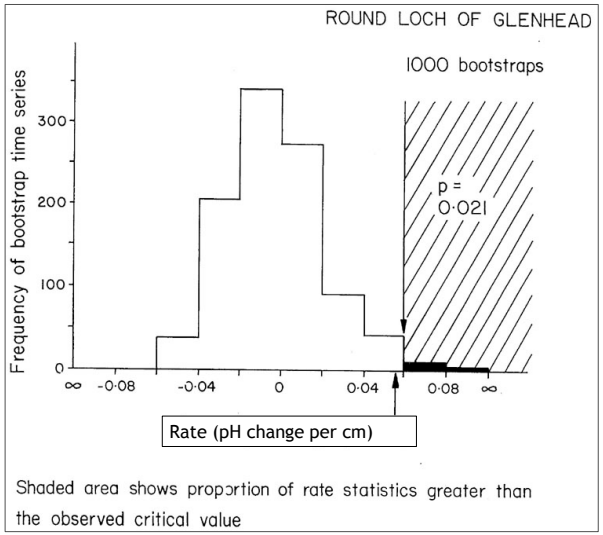
pH change 1874-1931 (17.3-7.3cm) very marked.

Is it any different from other pH fluctuations over last 10,000 years?

Null hypothesis – no different from rates of pH change in pre-acidification times.

Randomly resample with replacement 1,000 times to create temporally ordered data of same thickness as the interval of interest – time-duration or elapsed-time test. As time series contains unequal depth intervals between pH estimates, not possible for each bootstrapped time series to contain exactly 10cm. Instead samples are added in time series until depth interval equals or exceeds 10cm.





Statistical methods for testing competing causal hypotheses

Response variable(s)	Y	e.g. lake-water pH, sediment LOI, tree pollen stratigraphy
Predictor variable(s)	X	e.g. charcoal, age, land-use indicators, climate

Also covariables

Basic statistical model:

Y = BX

Y	X	Method
1	1	Simple linear regression
1	>1	Multiple linear regression, principal components regression, partial least squares (PLS)
>1	≥1	Redundancy analysis (= constrained PCA, reduced-rank regression, PCA of y with respect to x, etc.)

Statistical testing by Monte Carlo permutation tests to derive empirical statistical distributions

Variance partitioning or decomposition to evaluate different hypotheses.

Assessing Impacts of Laacher See Volcanic Ash on Terrestrial and Aquatic Ecosystems

A.F. Lotter & H.J.B. Birks (1993) J. Quat. Sci. 8, 263 - 276

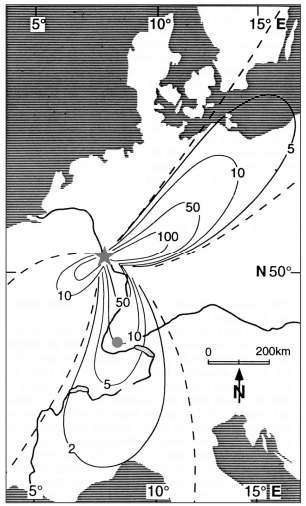
11000 BP

? Any impact on terrestrial and aquatic systems

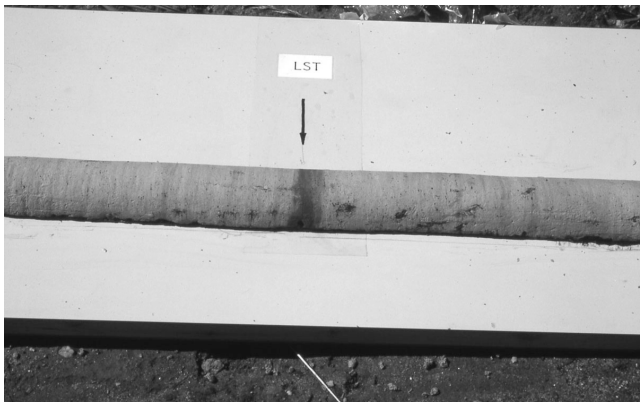
Also:

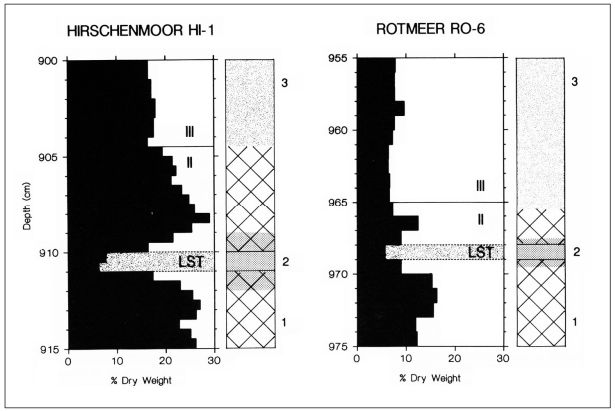
H.J.B. Birks & A.F. Lotter (1994) J. Paleolimnology 11, 313 - 922

A F Lotter *et al.* (1995) J. Paleolimnology 14, 23 - 47

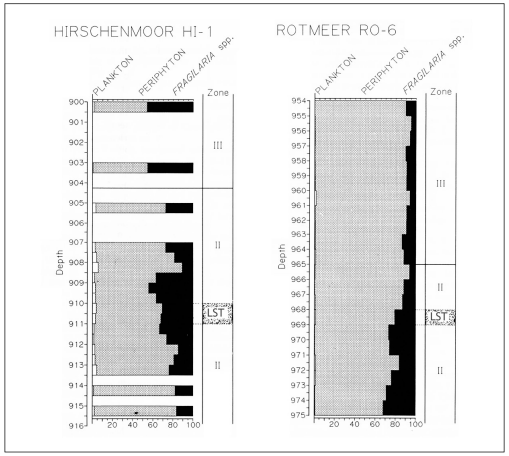


Map showing the location of Laacher See (red star), as well as the location of the sites investigated (blue circle). Numbers indicate the amount of Laacher See Tephra deposition in millimetres (modified from van den Bogaard, 1983).

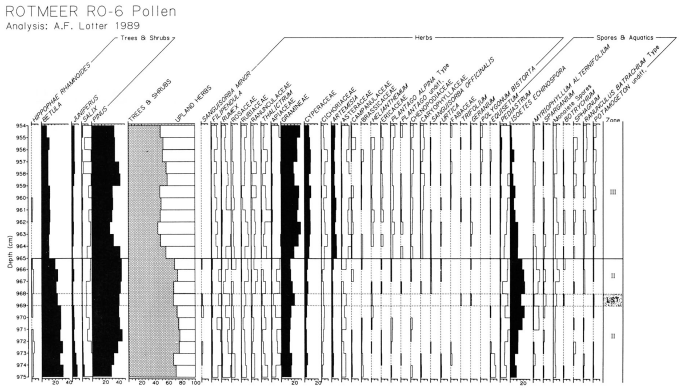


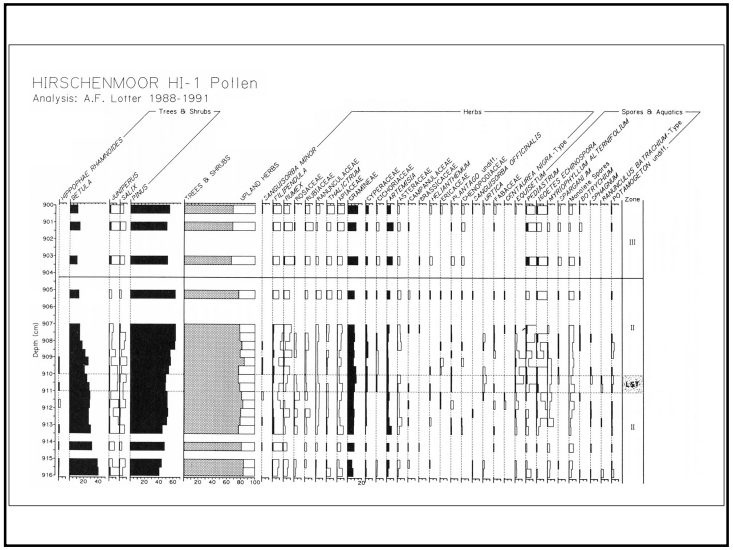


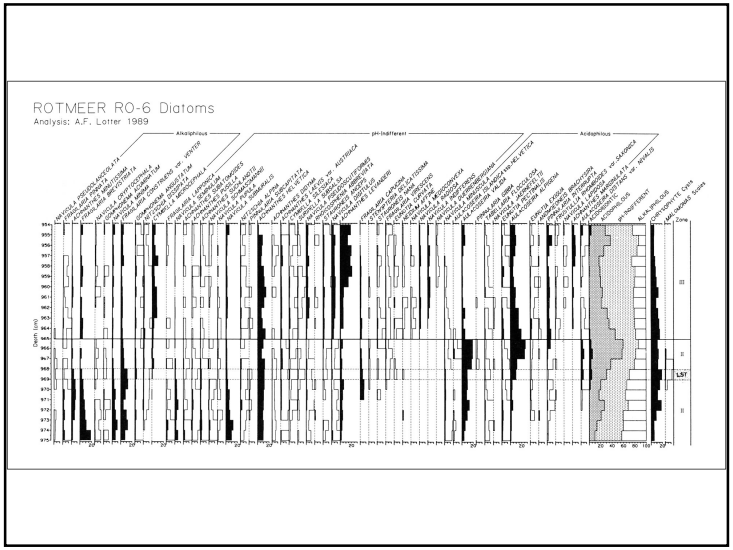
Loss-on-ignition of cores Hirschenmoor HI-1 and Rotmeer RO-6. The line marks the transition from the Allerød (II) to the Younger Dryas (III) biozone. LST = Laacher See Tephra.

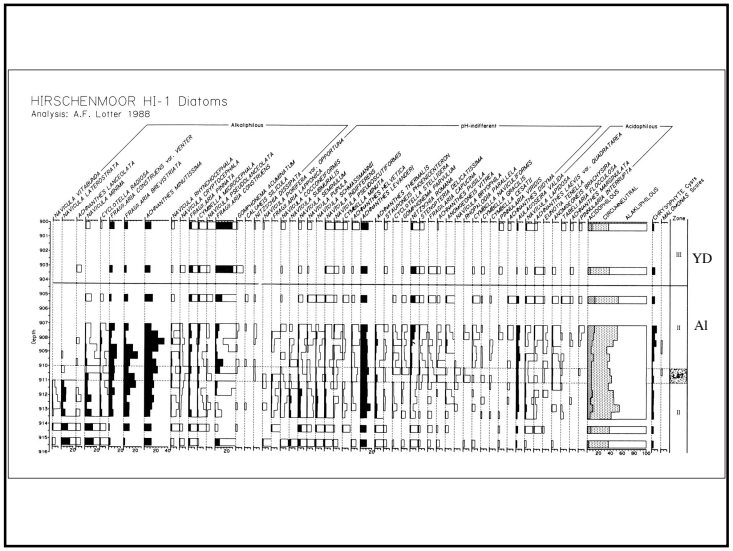


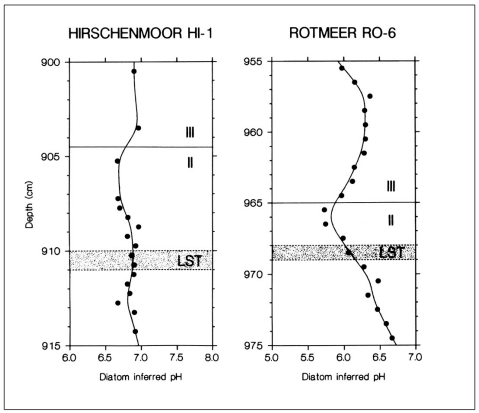
Diatoms in cores HI-1 and RO-6 grouped according to life-forms.
LST = Laacher See Tephra.











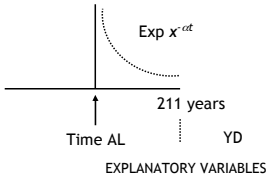
Diatom-inferred pH values for cores HI-1 and RO-6. The interpolation is based on distance-weighted least-squares (tension = 0.01). The line marks the transition from the Allerød (II) to the Younger Dryas (III) biozone. LST = Laacher See Tephra.

Data

Terrestrial pollen and spores (9, 31 taxa)
Aquatic pollen and spores (6, 8 taxa)
Diatoms (42,54 taxa)

RESPONSE VARIABLES
% data

Biozone	(Allerød, Allerød/Younger Dryas, Younger Dryas)	+/-
Lithology	(gyttja, clay/gyttja)	+/-
Depth	("age")	Continuous
Ash	Exponential decay process	Continuous



α = 0.5
x = 100
t = time

NUMERICAL ANALYSIS

(Partial) redundancy analysis

Restricted (stratigraphical) Monte Carlo permutation tests

Variance partitioning

Log-ratio centring because of % data

The biostratigraphical data sets used in the (partial) redundancy analyses

(SD = standard deviation units)

HIRSCHENMOOR CORE HI-1

	Terrestrial pollen	Aquatic pollen and spores	Diatoms
Number of samples	16	16	16
Number of taxa	9	6	42
Gradient length (SD)	0.48	0.84	1.44

ROTMEER CORE RO-6

	Terrestrial pollen	Aquatic pollen and spores	Diatoms
Number of samples	21	21	21
Number of taxa	31	8	54
Gradient length (SD)	0.74	0.71	1.68

RESULTS OF (PARTIAL) RESUNDANCY ANALYSIS OF THE BIOSTRATIGRAPHICAL DATA SETS AT ROTMEER (RO-6) AND HIRSCHENMOOR (HI-1) UNDER DIFFERENT MODELS OF EXPLANATORY VARIABLES AND COVARIABLES.

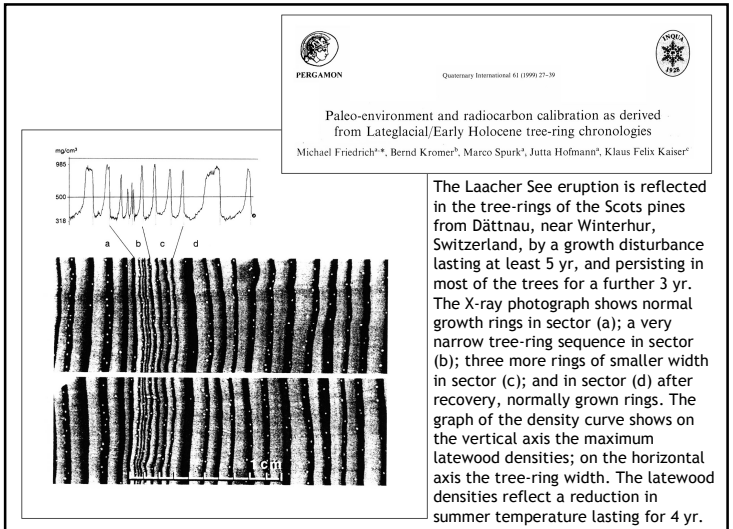
Entries are significance levels as assessed by restricted Monte Carlo permutation tests (n = 99)

			Data Set			
Site	Explanatory variables	Covariables	Terrestrial pollen	Aquatic pollen & spores	Diatoms	
RO-6	Depth + biozone + ash + lithology	-	**0.01**[a]	**0.01**[a]	**0.01**[a]	
HI-1	Depth + biozone + ash + lithology	-	**0.01**[a]	0.10	**0.01**[a]	
RO-6	Ash	Depth + biozone	0.09[ns]	0.48[ns]	0.16[ns]	Unique ash effect (no lithology)
HI-1	Ash	Depth + biozone	0.28[ns]	0.13[ns]	**0.01**[a]	
RO-6	Ash + lithology	Depth + biozone	-	0.88[ns]	0.17[ns]	Unique ash + lithology effect
HI-1	Ash + lithology	Depth + biozone	-	0.10[ns]	**0.01**[a]	
RO-6	Ash	Depth + biozone + lithology	-	0.53[ns]	0.08[ns]	Unique ash effect (lithology considered)
HI-1	Ash	Depth + biozone + lithology	-	0.10[ns]	0.19[ns]	
RO-6	Ash + lithology + ash*lithology	Depth + biozone	-	0.25[ns]	**0.03**[b]	Unique ash + lithology + (ash*lithology) interaction effect
HI-1	Ash + lithology + ash*lithology	Depth + biozone	-	0.12[ns]	**0.05**[b]	

[a] $p \leq 0.01$ [b] $0.01 < p \leq 0.05$

Results of partitioning the variance in the biostratigraphical data sets at Rotmeer (RO-6) and Hirschenmoor (HI-1) under the most appropriate model of explanatory variables. Entries are sum of squares (= variance) expressed as proportions of the total variance in each data set. The models used are as follows: terrestrial pollen, depth + biozone + ash; aquatic pollen and spores, depth + biozone + ash + lithology; diatoms, depth + biozone + ash + lithology + ash*lithology

	Pollen		Aquatic pollen and spores		Diatoms	
Source	RO-6	HI-1	RO-6	HI-1	RO-6	HI-1
Unexplained variance	0.35	0.64	0.56	0.59	0.27	0.46
Temporal and climatic change independent of any LST effects	0.57	0.30	0.17	0.11	0.12	0.27
LST effects independent of temporal and climatic change	0.03	0.06	0.06	0.31	0.15	0.20
Depth- and biozone-structured LST effects	0.05	0.00	0.21	−0.01	0.46	0.07
Total variance	1.00	1.00	1.00	1.00	1.00	1.00



PERGAMON

Quaternary International 61 (1999) 27–39

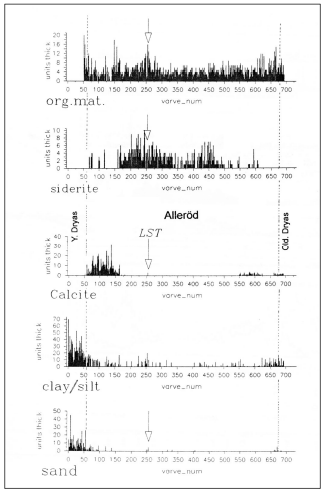
Paleo-environment and radiocarbon calibration as derived from Lateglacial/Early Holocene tree-ring chronologies

Michael Friedrich[a,*], Bernd Kromer[b], Marco Spurk[a], Jutta Hofmann[a], Klaus Felix Kaiser[c]

The Laacher See eruption is reflected in the tree-rings of the Scots pines from Dättnau, near Winterhur, Switzerland, by a growth disturbance lasting at least 5 yr, and persisting in most of the trees for a further 3 yr. The X-ray photograph shows normal growth rings in sector (a); a very narrow tree-ring sequence in sector (b); three more rings of smaller width in sector (c); and in sector (d) after recovery, normally grown rings. The graph of the density curve shows on the vertical axis the maximum latewood densities; on the horizontal axis the tree-ring width. The latewood densities reflect a reduction in summer temperature lasting for 4 yr.

Hämelsee - annually laminated sediments

Effects on lake sediments lasted no more than 20 years.



8 winter layers contain clay and silt.

Assessing Potential External 'Drivers' on an Aquatic Ecosystem

Bradshaw *et al*. 2005 The Holocene 15: 1152-1162

Dalland Sø, a small (15 ha), shallow (2.6 m) lowland eutrophic lake on the island of Funen, Denmark.

Catchment (153 ha) today

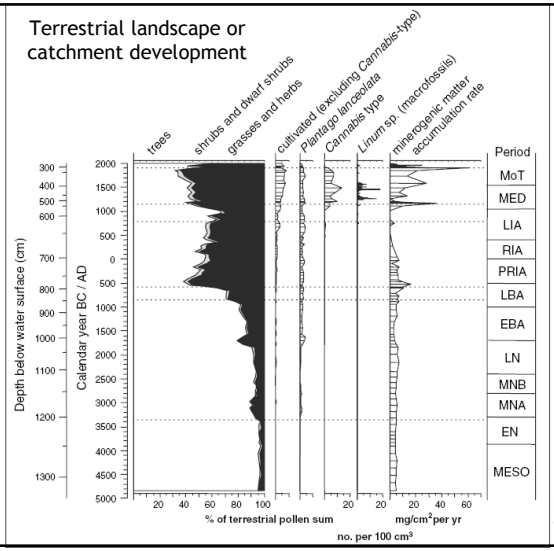
agriculture	77 ha
built-up areas	41 ha
woodland	32 ha
wetlands	3 ha

Nutrient rich – total P 65-120 $\mu g\ l^{-1}$

Multi-proxy study to assess role of potential external 'drivers' or forcing functions on changes in the lake ecosystem in last 7000 yrs.

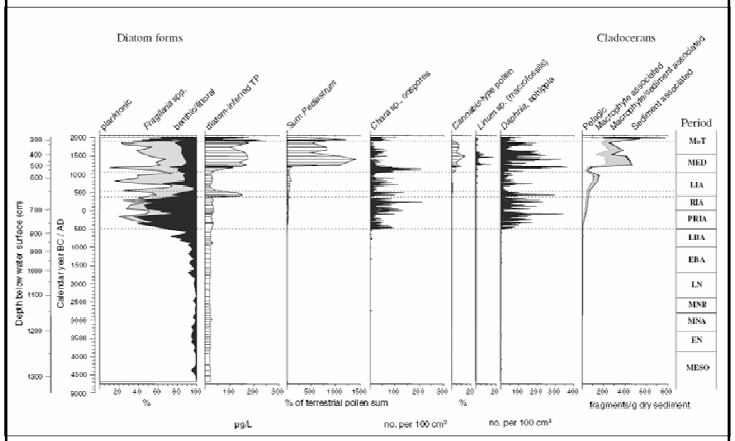
Data:	No. of samples	Transformation
Sediment loss-on-ignition %	560	None
Sediment dry mass accumulation rate	560	Log (x + 1)
Sediment minerogenic matter accumulation rate	560	Log (x + 1)
Plant macrofossil concentrations	280	Log (x + 1)
Pollen %	90	None
Diatoms %	118	None
Diatom inferred total P	118	None
Biogenic silica	84	Not used
Pediastrum %	90	None
Zooplankton	31	Not used

Terrestrial landscape or catchment development



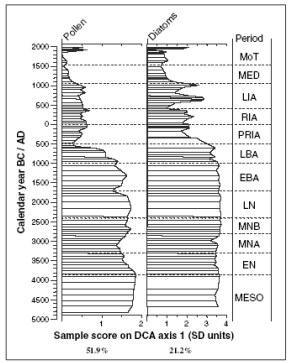
Bradshaw *et al*. 2005

Aquatic ecosystem development



Bradshaw *et al*. 2005

DCA of pollen and diatom data separately to summarise major underlying trends in both data sets

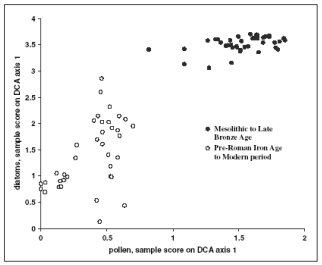


Pollen – high scores for trees, low scores for light-demanding herbs and crops

Diatom - high scores mainly planktonic and large benthic types, low scores for *Fragilaria* spp. and eutrophic spp. (e.g. *Cyclostephanos dubius*)

Bradshaw *et al*. 2005

Major contrast between samples before and after Late Bronze Age forest clearances



Prior to clearance, lake experienced few impacts.

After the clearance, lake heavily impacted.

Bradshaw *et al*. 2005

Canonical correspondence analysis

Response variables

Diatom taxa

Predictor variables

Pollen taxa, LOI, dry mass and minerogenic accumulation rates, plant macrofossils, *Pediastrum*

Covariable

Age

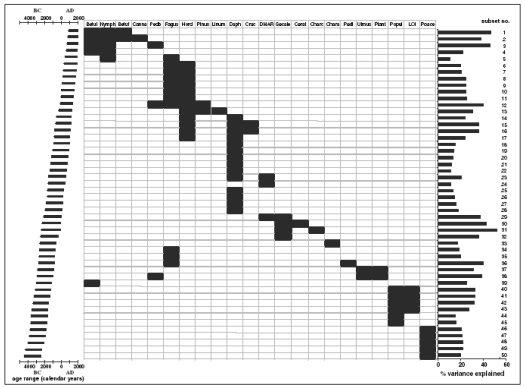
69 matching samples

Partial CCA with age partialled out as a covariable. Makes interpretation of effects of predictors easier by removing temporal trends and temporal autocorrelation

Partial CCA all variables

18.4% of variation in diatom data explained by Poaceae pollen, Cannabis-type pollen, and *Daphnia* ephippia.

As different external factors may be important at different times, divided data into 50 overlapping data sets – sample 1-20, 2-21, 3-22, etc.



Bradshaw *et al*. 2005

CCA of 50 subsets from bottom to top and % variance explained

1. 4520-1840 BC Poaceae is sole predictor variable (20-22% of diatom variance)
2. 3760-1310 BC LOI and *Populus* pollen (16-33%)
3. 3050-600 BC *Betula, Ulmus, Populus, Fagus, Plantago*, etc. (17-40%)

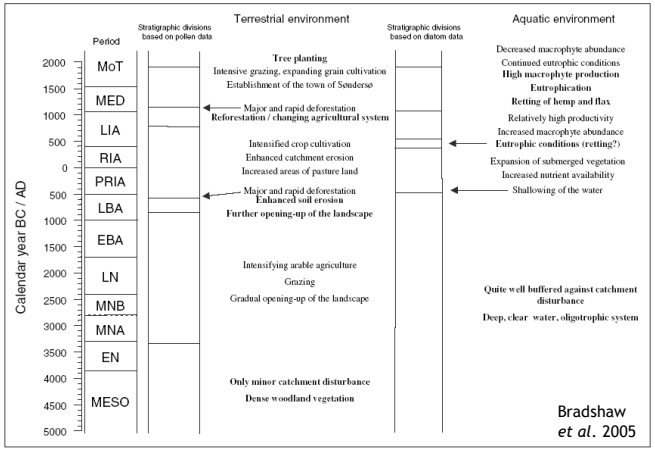
i.e. in these early periods, diatom change influenced to some degree by external catchment processes and terrestrial vegetation change.

4. 2570 BC – 1260 AD Erosion indicators (charcoal, dry mass accumulation), retting indicator *Linum* capsules, *Daphnia* ephippia, *Secale* and *Hordeum* pollen (11-52%)

   i.e. changing water depth and external factors

5. 160 BC – 1900 AD *Hordeum*, *Fagus*, Cannabis pollen, *Pediastrum boryanum*, *Nymphaea* seeds (22-47%)

   i.e. nutrient enrichment as a result of retting hemp, also changes in water depth and water clarity



Strong link between inferred catchment change and within-lake development. Timing and magnitude are not always perfectly matched, e.g. transition to Mediæval Period

Impact to Quaternary Palaeoecology of Hypothesis Testing

Descriptive phase -	patterns are detected, described and classified
Narrative phase -	plausible, inductively-based explanations, generalisations, or reconstructions are proposed for observed patterns
Analytical phase -	falsifiable or testable hypotheses are proposed, evaluated, tested and rejected

Why is there so little analytical hypothesis-testing in palaeoecology?

MONTE CARLO PERMUTATION TESTS are valid without random samples, can be developed to take account of the properties of the data of interest, can use "non-standard" test statistics, and are completely specific to the data-set at hand. Ideal for palaeoecology.

Time-Series Analysis
Introduction

'Time-series analysis' – series of techniques for analysing the behaviour of a variable over time and for investigating the relationship of two or more variables over time.

Time-series – values of one or more variables recorded, usually at a regular interval, over a long period of time.

For example, y_{1k}, y_{2k}, ..., y_{nk} could denote the pollen-accumulation rates of taxon *k* at *n* different times.

The observed values and fluctuations in such series may be comprised of

(1) long-term trend

(2) short-term variation

(3) cyclical variation

(4) phases of values well above or below long-term means or trend

(5) irregular or random variation

Such time-series data usually require special methods for analysis because of the presence of auto-correlation (serial correlation) between individual observations.

Auto-correlation

The internal correlation of observations in a time series, usually expressed as a function of the time lag between observations.

The auto-correlation at lag k, $\gamma(k)$ is

$$\gamma(k) = \frac{E(y_t - \mu)(y_{t+k} - \mu)}{E(y_t - \mu)^2}$$

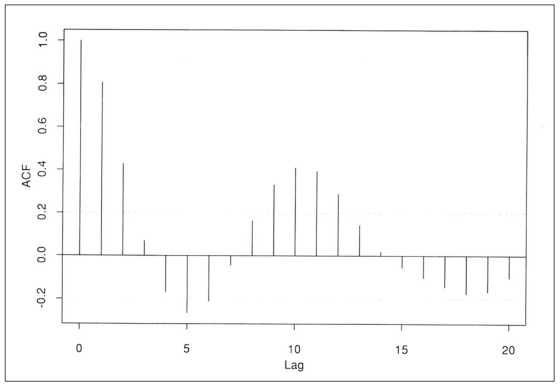
where yt, t = 0, ±1, ±2, ... represent the values of the series, μ is the mean of the series, and E denotes the expected value.

The corresponding sample statistic is

$$\hat{\gamma}(k) = \frac{\sum_{t=1}^{n-k}(y_t - \bar{y})(y_{t+k} - \bar{y})}{\sum_{t=1}^{n}(y_t - \bar{y})^2}$$

where $\bar{y}$ is the mean of the series of the observed values, $y_1, y_2, y_3, \ldots, y_n$

A plot of values of the auto-correlation values against the lag is auto-correlation function or auto-correlogram.



Main Domains of Time-Series Analysis

Two main domains of analysing time series

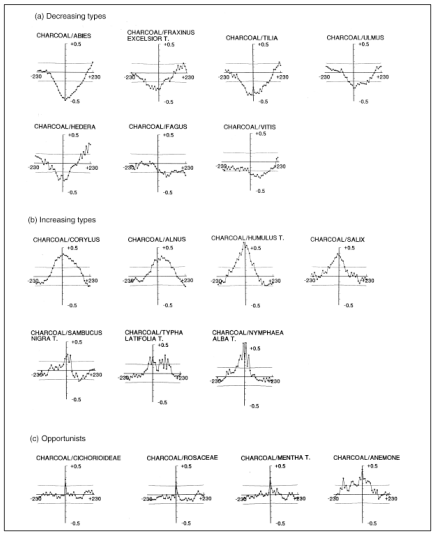
1. Time domain - autocorrelation coefficient

   at lag k is the correlation coefficient between observations in the same sequence which are k time intervals apart

   is a measure of similarity of observations separated by k time intervals

   auto-correlation plot can help in assessing behaviour of the variable over time

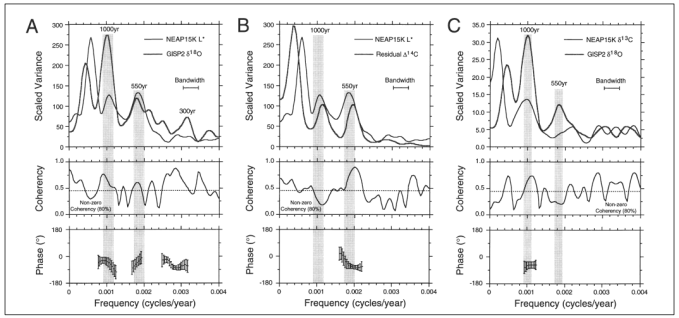
   can also compare two different variables by correlation coefficient between two variables, cross-correlation coefficient and associated cross-correlogram



Tinner *et al.* (1999)

2. Frequency domain - power spectrum of a time series gives an indication of the different frequencies of variation which account for most of the variability in the time series.

can help detect periodicities within the time series.



Chapman & Shackleton (2000)

These assumptions rarely, if ever, fulfilled with palaeoecological data.

Two approaches:

1. Use special procedures for unevenly spaced time series e.g. Lomb-Scargle-Fourier transformation for unevenly spaced data in combination with the Welch-Overlapped-Segment-Averaging procedure prior to frequency-domain or spectral analysis.

   SPECTRUM Shultz & Stattegger (1997) Computers & Geosciences 23: 929-945

2. Randomisation and permutation tests

   RT B.F.J. Manly (1997) Randomisation, bootstrap and Monte Carlo methods in biology. Chapman & Hall

Basic Assumptions of Time-Series Analysis

1. Equal time intervals between observations.
2. Data are stationary, namely that the time series contains no trends in means or variances. Normally detrended prior to analysis.
3. Data are normally distributed about the series mean. This assumption is necessary to enable statistical testing of the significance of correlations and power spectra. Can still do general analyses in the time and frequency domains without such assumptions in an 'exploratory' mode.

Randomisation Tests

Time series is a set of time-ordered observations, each of which has an associated observation time (e.g. age). Because of this ordering, observations are inherently not interchangeable unless the series is 'random', namely that all the observations are independent values from the same distribution.

In principle therefore can only test a series for time structure against the null hypothesis (Ho) that there is no structure at all – tests for randomness or tests for independence.

With randomisation tests, the significance of a test statistic can be determined by comparing it with the distribution obtained by randomly re-ordering observations. With *n* observations, there are *n*! possible orderings. A full randomisation distribution can be determined for *n* up to 8 (8! = 40,320).

Straightforward to estimate randomisation distribution by sampling it repeatedly.

Given the nature of time-series data, only justification for randomisation testing is the belief that the mechanism generating the data may be such as to make any observed value equally likely to have occurred at any time or position in the series.

Can use randomisation tests or tests of randomness to test for the absence (H_0) of

(1) serial correlation (auto-correlation)

(2) trend

(3) periodicity

Randomisation Tests for Serial Correlation (Auto-Correlation)

Can calculate observed serial correlation r_k for time series of n observations

$$r_k = \frac{\sum_{t=1}^{n-k}(y_i - \bar{y})(y_{i+k} - \bar{y})/(n-k)}{\sum_{t=1}^{n}(y_i - y)^2/n}$$

For a random series, r_k will approximate a value from a normal distribution with mean $-1/(n-1)$ and variance $1/n$ when n is large.

Tests for significance can therefore be constructed by comparing the statistics

$$z_k = \{r_k + 1/(n-1)\} \Big/ \sqrt{(1/n)}$$

for k = 1, 2, ... with percentage points of the standard normal distribution

If the serial correlations are tested at the same time for a range of values of k there is a high probability of declaring some of them significant by chance alone. Need to select a strict significance level by applying Bonferroni inequality which proposes that if k serial correlations should all be tested using the $(100\alpha/k)\%$ level of significance, there is then a probability of α' or less of declaring any of them significant by chance (α set by convention at 0.05).

Alternatively, use a randomisation test where the y values are randomly ordered a large number of times and compare the observed r_k with the randomisation distributions of r_k. From these randomisation distributions, the minimum significance level observed for the serial correlations is found from each randomised data set. The significance level for testing individual serial correlations is the minimum significance level that is exceeded for 95% of all the randomised data sets.

A simple alternative approach to randomness in a time series is a Markov process of the form

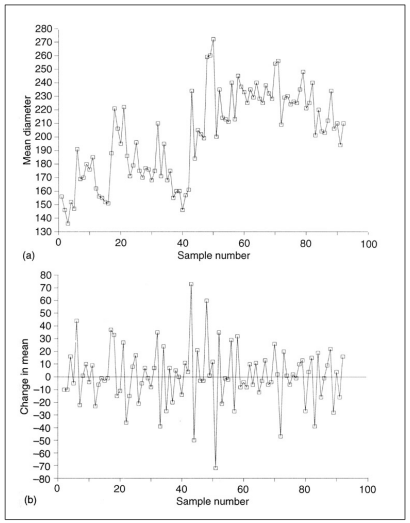
$$y_i = \tau y_{i-1} + \varepsilon_i$$

where τ is a constant and ε values are independent random variables with mean zero and constant variance.

For this alternative, the von Neumann ratio (v) is

$$v = \sum_{i=2}^{n}(y_i - y_{i-1})^2 \Big/ \sum_{i=1}^{n}(y_i - \bar{y})^2$$

a suitable statistics with a mean of 2 and variance of $4(n-2)/(n^2-1)$ for a random series. With randomisation, can compare observed v with the randomisation distribution obtained when the y values are randomly permuted. No assumptions of normality are made.



Evolutionary trends in a Cretaceous foraminifer

Clear that there is positive correlation between close samples. Not surprising because of the continuity of the fossil record. Are the differences in the mean between successive samples random?

Manly (1997)

Observed serial correlations for the mean difference series and their percentage significance levels for two-sided tests using 4999 randomisations.

	k									
	1	2	3	4	5	6	7	8	9	10
r_k	-0.44	0.10	-0.1	0.07	-0.14	-0.00	0.13	-0.18	0.10	-0.17
Sig. %	0.02	30.88	32.16	51.76	19.64	99.48	21.40	8.92	35.64	11.48

Evidence for a non-zero r_k for the first auto-correlation only (%<5%).

von Neumann ratio v = 2.86. Greater than any of the 4999 values obtained by randomisation; significance is 0.02%.

Other tests with Bonferroni inequality also indicate that only r_1 gives clear evidence of non-randomness.

Randomisation Tests for Trend

Trend in a time series is usually thought of as a broad-term tendency to move in a certain direction. Tests for trends should therefore be sensitive to this type of effect as distinct from being sensitive to a similarity between values that are close in time. Note that high positive serial correlations may produce series that have the appearance of trending in one direction for a long time period.

Various tests for trend involving the observed test statistics being compared with the randomisation distributions based on the randomisation of *y* values.

1. Regression of the time-series values *y* against the observation times. Compare observed β values in the model

   $$Y = \alpha + \beta X + \varepsilon$$

   and establish if β (slope) is significantly different from zero (no trend).

2. Runs above and below the median.

   Replace each value in time series by 1 if it is greater than or equal to the median and by 0 if it is below the median.

   Number of runs of same value is determined, and compared with the distribution expected with the 1 and 0s randomised.

   e.g. 1 2 3 4 5 6 7 9 8 median = 5

   $\therefore$ 0 0 0 0 1 1 1 1 1

   There are only two runs. Compare with randomisation distribution based on repeated randomisation of the 0 and 1 values.

3. Sign test

Test statistic is number of positive signs for the differences

$y_2 - y_1, y_3 - y_2, \ldots, y_n - y_{n-1}$

If there are m differences after zeroes have been eliminated, then the distribution of the number of positive differences has mean $\mu = m/2$ and variance $\sigma^2 = m/12$ under null hypothesis of randomness.

Significantly low number of positive differences indicates significant downward trend, significantly high number indicates upward trend.

4. Runs up and down test

Also based on differences between successive terms in the time series. Test statistic is the observed number of 'runs' of positive or negative differences.

Series 1 2 5 4 3 6 7 9 10 the signs of the differences are

\- - + + - - - -

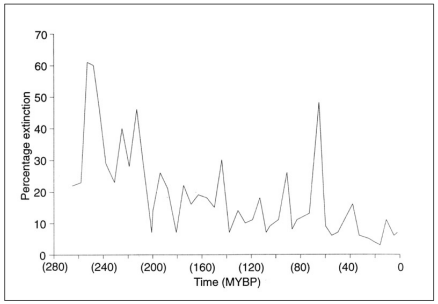
Three runs present

Null hypothesis of randomness

$\mu = (2m + 1)/3$ and $\sigma^2 = (16m - 13/90)$

where m is the number of differences.

Significantly small number of runs suggests trends, significantly high number suggests rapid oscillations.



Extinction rates for marine genera from late Permian to today (263 million years).

Manly (1997)

Observed test statistic		Randomisation p
Regression	-0.115	0.0002 *
Runs & medians	16	0.008 *
Sign test	23	1.0
Runs up & down	28	0.36

Two tests suggest trend, both looking at overall time-series. Other tests look at fine-scale patterns only.

Randomisation Tests for Periodicity

Attractive alternative to randomness in time series is some form of periodicity e.g. 11 year cycles and sunspot numbers.

Conventional approach is to model time series as a sum of sine and cosine terms at different frequencies and test to see if there is significant variance associated with these frequencies.

The model for n observations takes observation at time i to be

$$yi = A(0) + \sum_{k=1}^{m-1}\{A(k)\cos(w_k i) + B(k)\sin(w_k i)\} + A(m)\cos(w_m i)$$

where $w_k = 2\pi k/n$

The $B(m)$ term is absent because $\sin(w_m i) = \sin(\pi i)$ is always zero.

There are n unknown coefficients A(0), A(1), ..., A(m), B(1), ..., B(m-1) on the right-hand side. The n equations for the different values of y give n linear equations with n unknowns for these coefficients. Solving them gives

$$A(0) = \bar{y}$$

$$A(k) = (2/n)\sum_{i=1}^{n} y_i \cos(w_k i)$$

$$B(k) = (2/n)\sum_{i=1}^{n} y_i \sin(w_k i) \qquad \text{for } k = 1, 2, \ldots, m\text{-}1, \text{ and}$$

$$A(m) = (1/n)\sum_{i=1}^{n} y_i (-1)^i$$

If $S^2(k) = A^2(k) + B^2(k)$

then $$n\left\{\sum_{k=1}^{m-i} S^2(k)/2 + A(m)^2\right\} = \sum_{i=1}^{n}(y_i - \bar{y})^2$$

In other words we have partitioned the total sum of squares about the mean of the time series into m-1 components, representing variation associated with the frequencies

$w_1 = 2\pi/n$, $w_2 = 4\pi/n$, ..., $w_{m-1} = 2\pi(m-1)/n$ and $A(m)^2$

which represents the variation associated with a frequency of π

A plot of $nS^2(k)$ against w_k gives a periodogram. Also a plot of $nS^2(k)$ against cycle length is also called a periodogram.

Randomisation tests for peaks in the periodogram can be based on the $S^2(k)$ values

$$p(k) = \begin{cases} S^2(k)\Big/\sum_{i=1}^{n}(y_i - \bar{y})^2 & k < m \\ A(m)^2\Big/\sum_{i=1}^{n}(y_i - \bar{y})^2 & k = m \end{cases}$$

As the $p(k)$ values with $\Sigma p(k) = 1$, estimate the proportions of the variation in the series that are associated with different frequencies. High $p(k)$ values indicate important frequencies.

Significance levels can be determined by comparing each $p(k)$ to the distribution found for this statistic from randomising the order of the time-series values. The $p(k)$ values are equivalent statistics to the $S^2(k)$ and $A(m)^2$ because the total sum of squares of the y values remains constant for all randomisations.

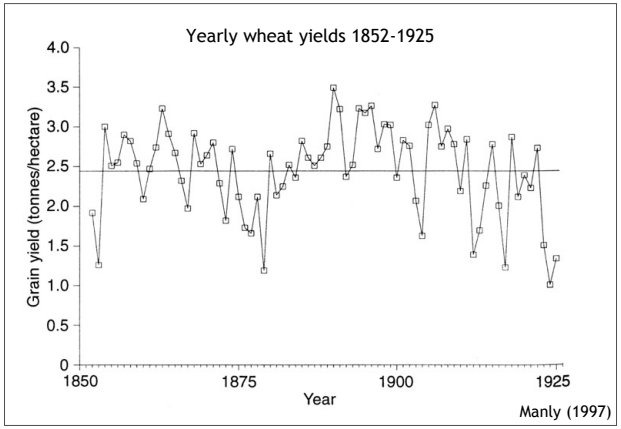
Another test for the null hypothesis of randomness against the alternative hypothesis of at least one periodic component is Bartlett's Kolmogorov-Smirnov test for overall randomness.

$D = \max(D^+, D^-)$	Overall deviation
$D^+ = \max\{j/(m\text{-}1) - u_j\}$	Maximum number of u values that fall below expectation
$D^- = \max\{u_j - (j\text{-}1)/(m\text{-}1)\}$	Above expectation

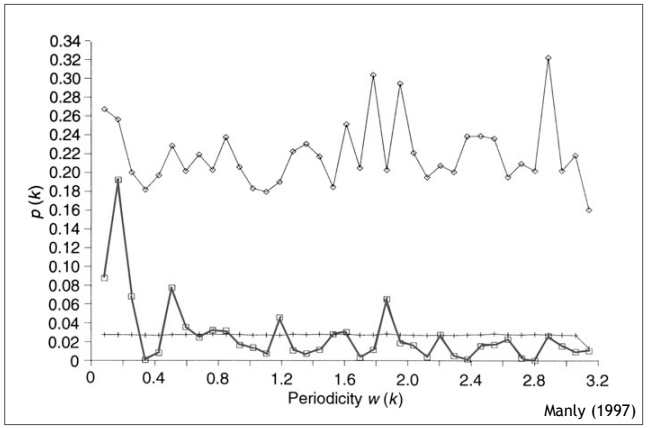
where $$uj = \sum_{k=1}^{j} S^2(k)\Big/\sum_{k=1}^{m-1} S^2(k)$$

Randomisation test – compare observed D with randomisation distribution when the time series is randomised.

A significantly high D value indicates that at least one periodic component is present.



D = 0.3139, p = 0.12% level in randomisations. Series not random; go on to look for individual periodicities.



□ = real data ◊ = randomisation maximum + = randomisation mean

k	w(k)	Cycle length	p(k)	Sig. level (%)
1	0.085	74.0	0.0877	3.48 ←
2	0.170	37.0	0.1923	0.08 ←
3	0.255	24.7	0.0678	7.92
4	0.340	18.5	0.0007	97.68
5	0.425	14.8	0.0082	73.86
6	0.509	12.3	0.0772	6.12
7	0.594	10.6	0.0355	27.46
8	0.679	9.3	0.0245	41.36
9	0.764	8.2	0.0320	31.26
10	0.849	7.4	0.0318	32.64
11	0.934	6.7	0.0165	56.06
12	1.019	6.2	0.0137	60.82
13	1.104	5.7	0.0074	77.04
14	1.189	5.3	0.0455	18.38
15	1.274	4.9	0.0107	69.20
16	1.359	4.6	0.0074	75.78
17	1.443	4.4	0.0114	67.20
18	1.528	4.1	0.0279	37.36
19	1.613	3.9	0.0301	34.74
20	1.698	3.7	0.0032	89.32
21	1.783	3.5	0.0114	66.62
22	1.868	3.4	0.0653	9.96
23	1.953	3.2	0.0185	51.44
24	2.038	3.1	0.0162	55.90
25	2.123	3.0	0.0039	86.84
26	2.208	2.8	0.0276	36.82
27	2.293	2.7	0.0050	83.62
28	2.377	2.6	0.0016	94.46
29	2.462	2.6	0.0154	58.42
30	2.547	2.5	0.0169	56.60
31	2.632	2.4	0.0225	44.84
32	2.717	2.3	0.0024	92.10
33	2.802	2.2	0.0002	99.04
34	2.887	2.2	0.0262	39.28
35	2.972	2.1	0.0155	57.48
36	3.057	2.1	0.0093	71.68
37	3.142	2.0	0.0104	38.52

Manly (1997)

Only two potentially significant periodicities, 74 and 37 years (3.48%, 0.08%). Allowing for multiple tests and Bonferroni inequality, $\alpha' = (5/37)\% = 0.14\%$. Thus only real evidence is for the 37-year cycle (p = 0.08%).

Interpretation needs common sense! There are only two 37-year cycles in the data. How to interpret the patterns?

Null hypothesis of complete randomness rejected in favour of periodicity. But a time series with positive auto-correlation can easily give the observed patterns.

Need to test hypotheses of randomness first, then auto-correlation, then trend, and then periodicity. What is the logical order?

40% of variation is explained by a linear regression of yield with rainfall. The 37-year cycle in the wheat yield may be a response to rainfall cycles.

Irregularly Spaced Time-Series

Common in palaeoecology

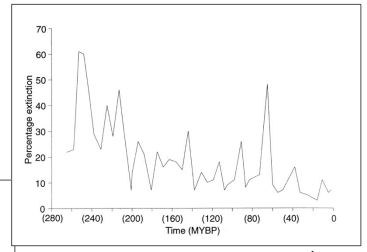
Very difficult to resolve, especially for tests for periodicity, even when using randomisation tests.

Not sensible to perform randomisation tests for periodicity when the series is non-random because it contains a trend, unless the trend cannot affect the test statistics. Detrending usually necessary but time-series often not very large.

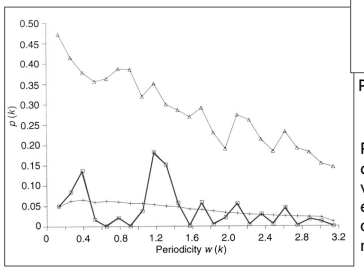
Test statistics must be chosen in such a way that when a significant result is obtained by chance, there is no bias towards indicating any particular cycle length as being important.

Marine genera extinction rates revisited.

Raup & Sepkoski (1984) proposed a periodicity in extinctions of about 29 million years (periodicity $w(9) = 1.18$ corresponding to cycle length of 29 million years). Kolmogorov-Smirnov $D = 0.307$.



Plot of extinction rates against time (10^6 yr BP)



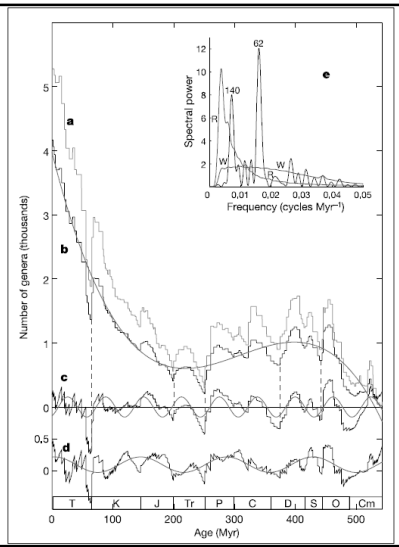
Periodogram for the marine genera extinction data, with the mean and maximum $p(k)$ values determined by simulating 1000 sets of extinction data with a trend similar to that observed. □ = real data, △ = simulation maximum, + = simulation mean.

Manly (1997)

But if randomisation tests applied, the $w(9)$ peak lies within the range of randomisations. This peak equalled or exceeded by 2.7% of the randomisations. $w(10)$ equalled or exceeded by 4.2%. ?statistically significant.

Bonferroni inequality suggests $\alpha = (5/24)\% = 0.2\%$ in order to have a probability of 0.05 or less in declaring $p(k)$ value significant. Values of 2.7% and 4.2% not significant. Observed Kolmogorov-Smirnov D of 0.307 exceeded by 28.6% of the randomised values. Not statistically significant.

Perhaps no periodicity in extinctions after all.



Cycles in fossil diversity revisited yet again!

Rohde & Muller 2005 Nature 434; 208-210

36,380 marine genera over last 542 million years

(a) number of known genera (36,380)

(b) as in (a) but with single occurrences and poorly dated genera removed, leaving 17,797 genera. Fitted line is third-order polynomial

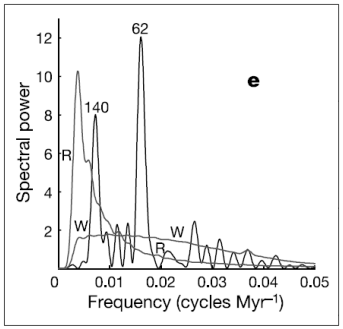
(c) subtracting fitted trend from (b), 'residuals' and a 64 M yr cycle added

(d) subtract (detrend) the 62 M yr cycle from data in (b) and add a 140 M yr sine curve.

(e) spectral analysis of residuals (c), strong 62 ± 3 M yr peak and a 140 ± 15 M yr cycle

Are the 62 ± 3 M yr and 140 ± 15 M yr cycles statistically significant?

Assume that all diversity changes are random walks. Simulate this with random permutations of the steps in (b). 30,000 simulations detrended (third-order polynomial) and their spectral power estimated (R)



Also broke detrended data into 20 groups, scrambled their order (preserves short-term correlations but randomises placement of major events). 30,000 simulations, spectral power (W).

Statistical probabilities of peaks

	At this frequency		Anywhere	
	R	W	R	W
62 M yr	$<5 \times 10^{-5}$	3.6×10^{-4}	<0.0013	0.010
140 M yr	0.12	0.0056	0.71	0.13

62 M yr peak highly significant

140 M yr peak may entirely be a random process

Possible causes of 62 M yr cycle – geological hypotheses

1. $\delta^{18}O$ (climate) – strong 135 M yr cycle
2. Volcanism – minor feature at 62 M yr cycle
3. $\delta^{13}C$ (biomass proxy) – no 62 M yr cycle or 140 M yr cycle
4. Sea-level change – peaks at 62 M yr and 140 M yr but low statistical significance
5. Impact craters – no significant 62 M yr or 140 M yr cycles
6. Geological formations – no significant 62 M yr or 140 M yr cycles

But 62 M yr cycle is strong in the data!

Now to science fiction!

1. Periodic passage of solar system through molecular clouds or Galactic arms could periodically perturb the Oort cloud and cause variations in the rates of comet impacts on Earth.
2. Mantle plumes reaching Earth show cycles and could cause periodic volcanism.
3. Sun currently oscillates up and down across the Galactic plane every 52-74 M yr.
4. Solar cycles.
5. Earth orbital oscillations.
6. Companion stars to the Sun could trigger periodic comet showers.
7. 'Planet X' is a proposed large planet that perturbs the Kuiper belt and could yield periodic comet showers on the right time scales. No evidence!
8. Maybe the 62 M yr cycle is caused by a biological pendulum that swings so slowly that we cannot detect its underlying mechanisms.

"It is often said that the best discoveries in science are those that raise more questions than they answer, and that is certainly the case here"

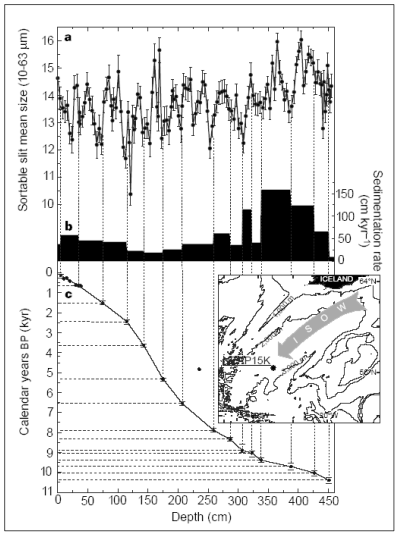
"The 62 million year wave is too big to ignore - why has it not been seen before?"

Kirchner & Weil (2005)

Limiting factors in palaeoecology are the irregular sampling in time, the quality of datings and the resulting age-depth models, and the number of samples (ideally 100-300).

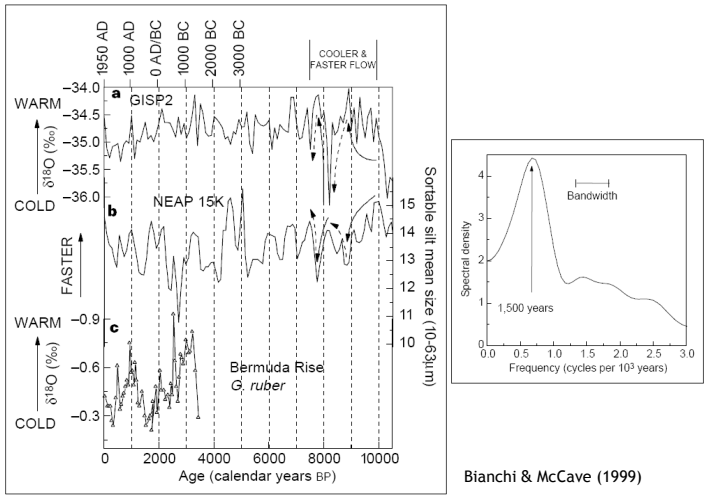
Unless sediments are annually laminated, age-depth models for the Holocene may have 1 standard deviation uncertainties of 60-120 years, or 2 standard deviation uncertainties of 120-240 years.

Must be cautious to many published spectral analyses in Holocene palaeoecology.



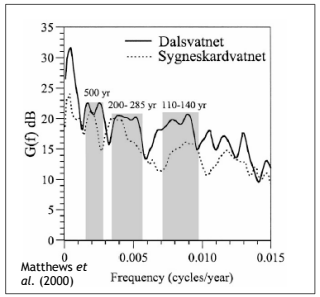
Sortable silt mean size record from North Atlantic marine core NEAP-I5K (about 15 AMS ^{14}C dates).

Bianchi & McCave (1999)



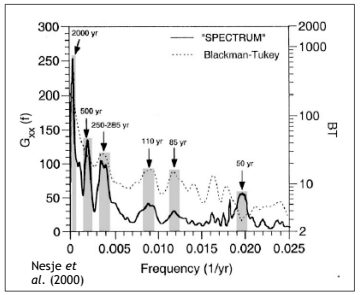
Bianchi & McCave (1999)

Loss-on-ignition data – Matthews *et al.* (2000), Nesje *et al.* (2000).



17 AMS ^{14}C dates

Rejected 13!

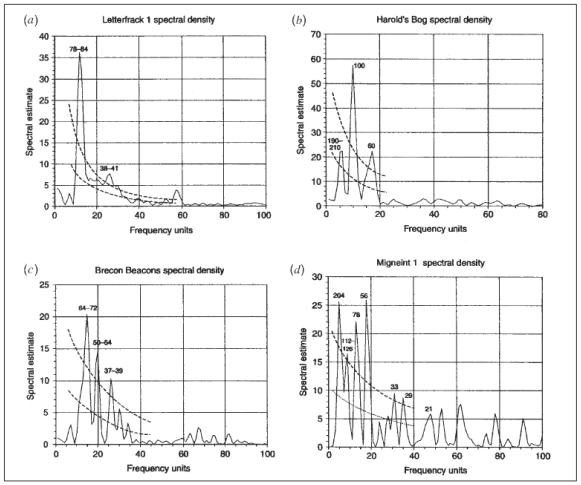


16 AMS ^{14}C dates

Rejected 2

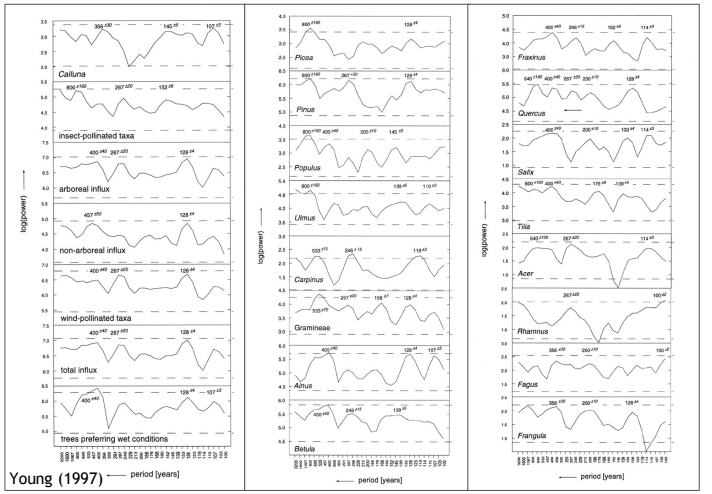
Suggesting cycles of 110-140 years, 85 years, 50 years, etc. How good is the underlying chronology?

Peat humification – each sequence 3 or 4 AMS ^{14}C dates only. Chambers & Blackford (2001).



Chambers & Blackford (2001)

Pollen influx data Lake Gosciaz, Poland (annually laminated sediments). Young (1997).



Young (1997)

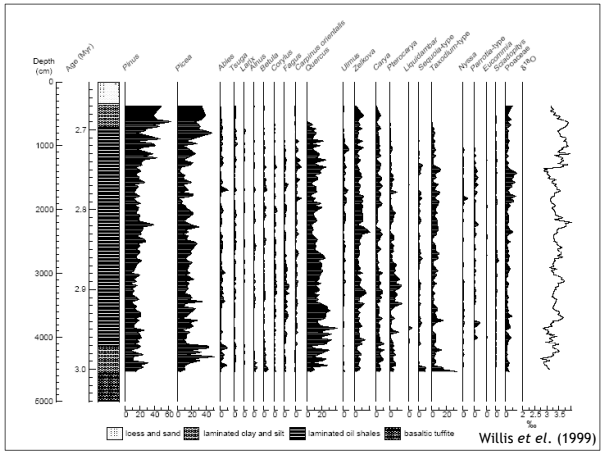
PSEUDOSIGNIFICANT POWERS (*)
AND THE MOST IMPORTANT OTHER PEAKS IN POWER SPECTRA

taxon/var.	period								
Picea	800 *						128		
Pinus	800				267		128		
Populus	800 *		400			200			
Ulmus	800 *								
Carpinus		533 *		246 *				119 *	
Gramineae		533 *			267		128		
Alnus			400 *				128		
Betula			400	246					
Fraxinus			400	246					
Quercus			400		267	200	128		
Salix			400			200			
Tilia			400						
Acer					267				
Rhamnus					267				100 *
Fagus						200			100
Frangula						200	128		
Calluna							128		
occurrences (*)	4 (3)	2 (2)	7 (1)	3 (1)	5	5	7	1 (1)	2 (1)
insect pollin.	800			246					
arboreal			400		267		128		
non-arbor.			400				128		
wind pollin.			400		267		128		
total influx			400		267		128		
wet prefer.			400 *				128		

Young (1997)

Late Pliocene pollen data – Willis *et al.* (1999).

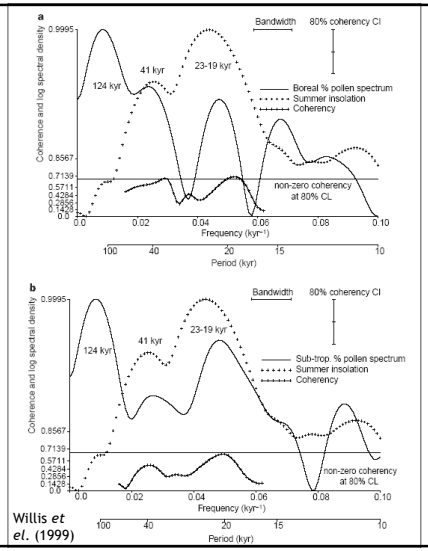
Laminated sequence 2.6 – 3.0 million years in Hungary.



Willis *et el.* (1999)

Spectral and cross-spectral (coherency) analysis.

Boreal pollen, sub-tropical pollen, summer insolation



Willis *et el.* (1999)

Boreal pollen and sub-tropical pollen linked to 23-19 k yr (precession) and 41 k yr (obliquity).

Out of phase – sub-tropical pollen increases in response to increased insolation whilst boreal pollen increases in response to decreasing insolation.

Good coherence at 23-19 k yr and 41 k yr.

Also strong low frequency component at 124 k yr. Not present in summer insolation time-series. Similar 124 k yr periods in dust content in marine cores, possibly reflecting continental aridity.

SiZer and SiNos - Smoothing Procedures in Palaeoecology

Combination of hypothesis-testing and time-series analysis

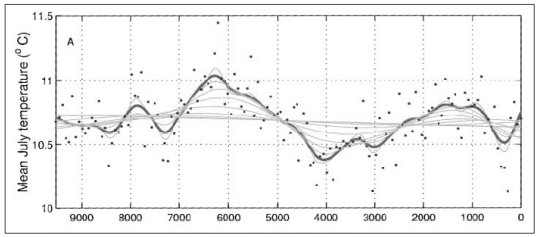
SiZer = Significant Zero crossings of Derivatives

Chaudhuri & Marron 1999 J. Amer. Stat. Assoc. 94: 807-823

Godtliebsen *et al*. 2003 Geophysical Research Letters 30(12) doi: 10.1029/2003GL017229

Holmström & Erästö 2002 Computational Stats. & Data Analysis 41: 289-309

SiZer approach uses a whole family of smooth curves, each providing information about the underlying curve at different levels of detail. Features detected typically depend on the level of detail for which the time series is considered. Obvious example – recent global warming over last few decades barely detectable if viewed over long (10 000 yr) time scale.



Different smoothers based on various smoothing window sizes

Main line is 'optimal' smooth based on Ruppert *et al*. (1995) criterion

Korhola *et al*. (2000)

Based on confidence intervals of the derivatives of the smooths, SiZer enables an assessment of which observed features are statistically significant, i.e. what is 'signal' and what may be 'noise'.

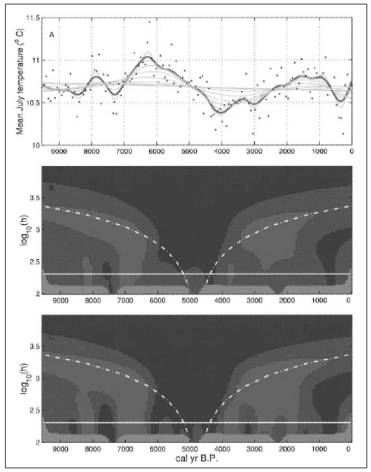
Shown as colour SiZer maps that are a function of location and scale.

Amount of smoothing is controlled by parameter *h* and, for each value of $\log_{10}(h)$, the effective smoothing window is described by a horizontal space between the two dash-dotted curves.

The white line corresponds to the 'optimal' smoother used based on Ruppert *et al*. (1995) criterion.

Red indicates that the curve is significantly increasing; blue that the curve is significantly decreasing; purple indicates no conclusions about the slope can be made; and the grey areas indicate that the data are too sparse at that smoothing level for any conclusions to be made about significance.

Here 80% level of significance (α = 0.2) is used



Family plot

Space below dashed white curve illustrates the effective size of smoothing window used

Global significance analysis – dependence of nearby smoothed values are taken into account

Local significance analysis – analysis is independent of each instant in time

Korhola *et al*. (2000)

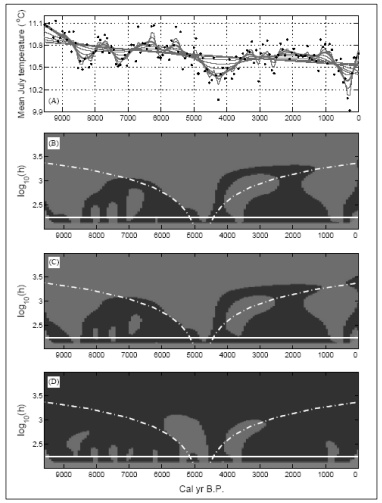
SiZer consists of four steps

1. Fix a level of smoothing by specifying a parameter $h>0$. This is the window width (or span) used in the local averaging or regression procedure.
2. Using window width h, turn the time-series estimates into a smooth derivative function.
3. Construct a confidence band around the derivative function and use it to make inferences about trends in the (unknown) values of the time-series at the time scale corresponding to h.
4. Repeat steps 1 – 3 for different values of h to make inferences about significant trends in the time series at various time scales.

Construction of confidence band around the derivative function can be done in several ways for palaeoecological reconstructions

1. treat the training set X_m as random and the fossil set X_f as fixed
2. treat the training set X_m as fixed and X_f as random
3. treat both X_m and X_f as random

Can derive Gaussian confidence bands from statistical theory or by bootstrapping



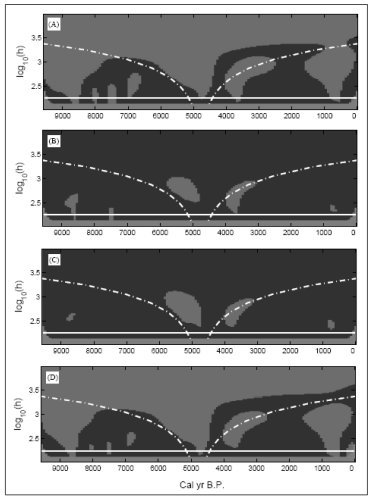
Family plot

X_m fixed X_f random
Gaussian bands
$\alpha = 0.2$

X_m fixed X_f random
Bootstrap bands
$\alpha = 0.2$

X_f fixed X_m random
Bootstrap bands
$\alpha = 0.2$

Holmström & Erästö (2002)



X_m fixed
$\alpha = 0.05$

X_m and X_f random
Gaussian bands
$\alpha = 0.2$

X_m and X_f random
Bootstrap bands
$\alpha = 0.2$

X_m fixed
Bootstrap bands
$\alpha = 0.2$

Holmström & Erästö (2002)

Results depend on significance value chosen (α = 0.05 or 0.2) and on how the confidence bands are estimated

Inferences are simultaneous only in time and not in the level of smoothing.

At the level of smoothing shown by the white line, the probability that the true time-series record exhibits the indicated features shown as red (warming) and blue (cooling) is 1 – α.

Using a fixed larger smoothing parameter value for the same map, one can infer a long-term change with the same confidence.

Cannot claim to have the same confidence in both of these statements simultaneously.

To do this, need to incorporate the bootstrap confidence intervals simultaneously in the smoothing parameter (3D). Little difference between 2C (X_m fixed, X_f random, α = 0.2) and 3D (X_m fixed, α = 0.2, confidence intervals in the smoothing).

REF

Technical details of SiZer

1. Non-parametric regression

   $y_i = m(x_i) + \varepsilon_{ij} \qquad i = 1,\ldots,n$ where $m(x)$ is the target curve.

   Assume x_i are equally spaced and independent (a big assumption!), m is smooth, and ε_i are independent with mean = 0 and variance = σ_1^2

2. At each location, a local linear kernel estimator is used to produce smooths of the observed signal. In this parameter h, the band-width, controls the degree of smoothness in the estimate of $\hat{m}_h$

   In detail, at point x_j, $\hat{m}_h(x_j)$ equals the fit $\hat{\alpha}_0$ where $(\hat{\alpha}_0, \hat{\alpha}_1)$ minimises

   $$\sum_{i=1}^{n} \left(y_i - \alpha_0 - \alpha_1(x_i - x_j)\right)^2 K_h(x_i - x_j) \qquad (1)$$

   $K_h(\cdot) = \left(\frac{1}{h}\right) K\left(\frac{\cdot}{h}\right)$ where K is a kernel function chosen as a unimodal probability density function that is symmetric around zero.

REF

3. For each scale (bandwidth in the kernel estimator) and location of the signal, a test is performed to see if the smooth has a derivative significantly different from zero.

   Test if $\alpha_i \neq 0$ in (1) for each (x,h) location.

4. Try to detect significant features at different scales. What is significant at one scale may not be significant at another scale.

SiNos – Significant Non-stationarities

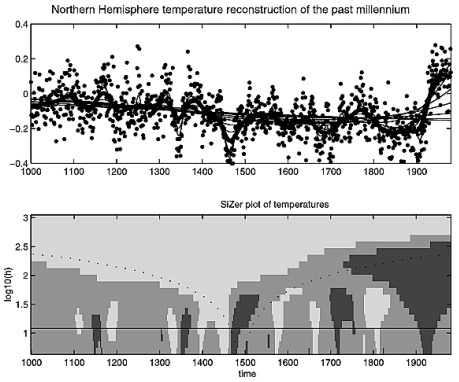
Godtliebsen *et al*. 2003 Geophysical Research Letter 30 (12)

Handle time series where there is stochastic DEPENDENCE between different data points.

Looks simultaneously for significant changes in the mean, variance, and first-lag auto-correlation of the observed time series when the null hypothesis suggests that the process is stationary.

Change in mean is claimed if the means of the window widths to the right and left of the location are significantly different. Similar tests of variances and first-lag auto-correlation.

Northern Hemisphere temperature data for past millennium



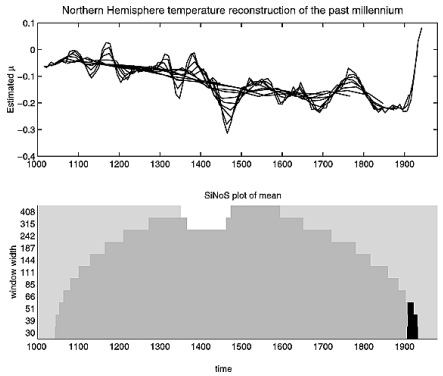
SiZer

Black = positive

Light grey = negative

Dark grey = too few data

Godtliebsen *et al*. (2003)



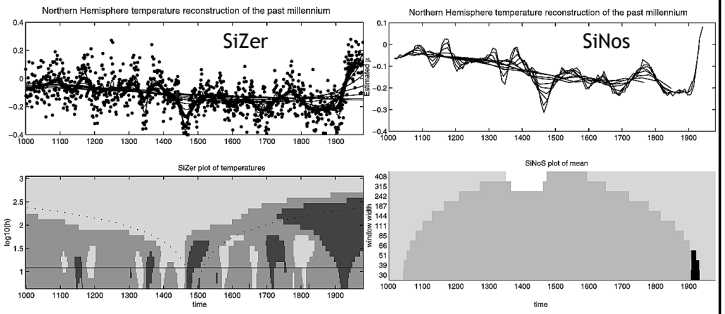
Godtliebsen *et al*. (2003)

SiNos analysis of the mean for Northern Hemisphere temperatures

SiNos - decreases in temperature are significant at scales of 300 years or more

\- recent increase detected as significant at scales of 30-50 years.

SiZer - because the time-series data are not independent, can easily suggest spurious details as significant at small sizes.



Godtliebsen *et al*. (2003)

SiZer typically detects too many features for dependent data.

SiZer superior to SiNos for independent data.

SiNos can detect other types of stationarities (e.g. changes in the first-lag auto-correlation) in a time-series that SiZer cannot do.

Both SiZer and SiNos are useful tools that help to detect 'signal' in palaeoecological time-series and to test if particular changes are statistically significant or not.

Conclusions

1. Palaeoecological data are rarely really suitable for time-series analysis because of uneven sampling and dating uncertainties.
2. Use of randomisation tests eases some of the restrictive assumptions of time-series analysis.
3. Can test for randomness in time series.
4. Can test for auto-correlation in time series but should allow for multiple testing (Bonferroni inequality).
5. Can test for trends in time series. Piece-wise regression potentially valuable as no reason for one trend only.

6. Can test for periodicity by modelling a time series as a sum of sine and cosine functions corresponding to cycles of different lengths. The variances associated with different cycle lengths can be tested directly by randomisation, allowing for multiple testing.
7. Deriving reliable and robust estimates of statistical significance of time-series test statistics is the key.
8. Data quality, especially the number of observations and the dating quality, is major limiting factor in palaeoecological time-series analysis.
9. Tests for auto-correlation and trends in time series much less 'data demanding' but they still need good dating control, as does rate of change analysis.
10. Approach time-series analysis with caution!
11. Sizer and SiNos approach valuable in detecting potentially significant features in time series.

Major Uses of Numerical Methods in Palaeoecology

1. Data collection and assessment

Identification	Lecture 1	Discriminant analysis
Error estimation	Lectures 1,4	Summary statistics, LOWESS, RMSEP, bootstrapping, etc

2. Data summarisation

Single data sets	Lectures 2, 3, 4	Ordinations, zonation
2+ stratigraphical sequences	Lectures 2, 3, 4	Ordinations, sequence slotting
2+ stratigraphical & spatial data sets	Lectures 2, 3, 4	Ordinations, mapping

Major Uses of Numerical Methods in Palaeoecology

3. Data analysis

Sequence-splitting	Lecture 4
Rate-of-change analysis	Lecture 4
Time-series analysis	Lecture 6
Environmental reconstructions	Lecture 5

4. Data interpretation

Vegetation reconstruction	Lecture 4	Modern analogue techniques
Causative or 'forcing' factors	Lectures 3, 6	Canonical ordinations, randomisation and permutation tests

General Conclusions about Quantitative Palaeoecology

1. Powerful tool for summarising stratigraphical, time-ordered multivariate data – numerical zonation, ordination (e.g. PCA, CA). Repeatable methods.
2. Methods for quantitative reconstruction of past environment (e.g. lake-water pH, total P, dissolved organic carbon, summer temperature, etc) with sample specific error estimates.
3. Permutation tests provide a means of statistically testing specific palaeoecological hypotheses even though palaeoecological data are time-ordered and hence not independent in a statistical sense, are closed percentage data and thus do not follow any simple normal distribution, are highly multivariate, and are from 'undesigned' experiments.
4. Useful tools for aiding critical identification of fossils (e.g. linear discriminant analysis).
5. Powerful tools for developing age-depth models in calibrated years. Time is the basis for almost all palaeoecology. Essential stage in any study.
6. Time-series analysis potentially very attractive but in practice very data demanding in terms of both data quantity and data quality.

   Quantitative palaeoecology provides a powerful means of 'coaxing history to conduct experiments'.
7. But remember the social scientists and the statisticians!