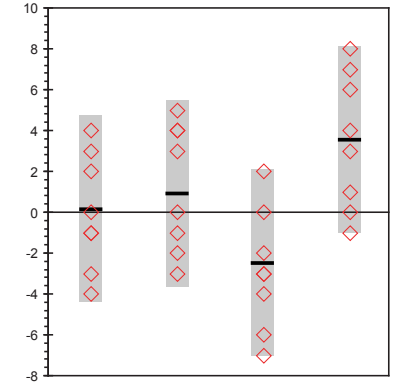
Data analysis and Geostatistics - lecture VII

Analysis of time-series data

Analysis of variance - ANOVA

The analytical data for the four marble units:

	unit 1	unit 2	unit 3	unit 4
	-3	+3	-3	+4
	+3	-1	-6	+7
	-1	-2	-2	-1
	-1	+4	+2	+1
	+4	0	-3	+6
	-4	-3	-4	+3
	+2	+5	0	0
	0	+4	-7	+8
mean	0	1.25	-2.88	3.5
s^2	8	9.6	8.7	11.1
n	8	8	8	8
SS	56	67.5	60.9	78



difference between needs to exceed difference within

ANOVA - Analysis of variance

Input the data into PAST with two factors: unit and geologist

	sum of squares	degrees of freedom	variance	F-ratio	F-crit
between geol	3200	2	1600	4	5.14
between units	600	3	200	0.5	4.76
within/residual	2400	6	400		
total	6200	11			

From this it is clear that the variance between units is larger than the within variance, but this is not true for the variance between geologists

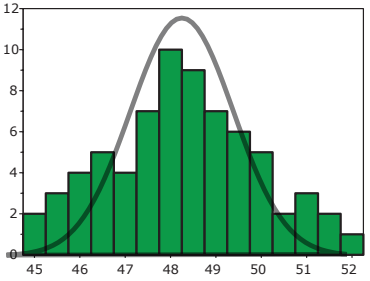
However, at α = 5%, **neither** exceeds the critical probability: all are the same

Testing of “goodness-of-fit”

comparison of curves: predicted and observed values

the cumulative discrepancy between the predicted and observed values is a measure of the goodness-of-fit

if this exceeds a critical value: can reject the fit that we are testing



this is the Chi-squared (X^2) test:

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

with O_i = observed value of i
and E_i = predicted value of i

Time series analysis

Time is a critical variable in geology and a whole subfield of geostatistics is devoted to it: time series analysis

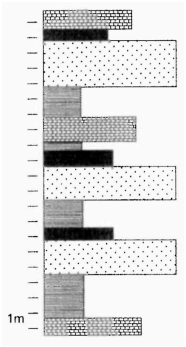
aims: detect trends and systematics with time for process identification and to predict the future

time is only rarely absolute, in most cases we have only qualitative information on time (strat sequence, growth zoning, younger-older)

Time series analysis - Markov chain

Systematics in the lithology changes for a log (time is qualitative)

strat log



transition matrix

from \ to	sh	s	c	l	total
sh	0	3	0	1	4
s	0	0	3	0	3
c	2	0	0	1	3
l	2	0	0	0	2
total	4	3	3	2	12

transition prob matrix

from \ to	sh	s	c	l	total
sh	0	3/4	0	1/4	1
s	0	0	1	0	1
c	2/3	0	0	1/3	1
l	1	0	0	0	1
total					

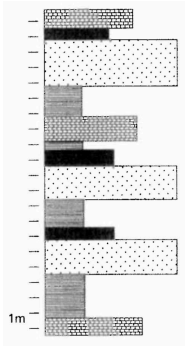
find systematic chains: sh —> s —> c —> sh
l —> sh —> s —> c —> l

are these sequences significant or pure chance ?

Time series analysis - Markov chain

Systematics in the lithology changes for a log (time is qualitative)

strat log



transition matrix

from \ to	sh	s	c	l	total
sh	4	3	0	1	8
s	0	4	3	0	7
c	2	0	0	1	3
l	2	0	0	0	2
total	8	7	3	2	20

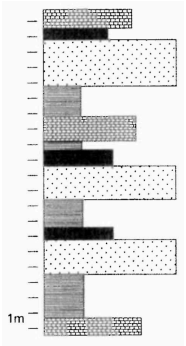
transition prob matrix

from \ to	sh	s	c	l	total
sh	1/2	3/8	0	1/8	1
s	0	4/7	3/7	0	1
c	2/3	0	0	1/3	1
l	1	0	0	0	1
total					

Time series analysis - Markov chain

Systematics in the lithology changes for a log (time is qualitative)

strat log



transition matrix

from \ to	sh	s	c	l	total
sh	4	3	0	1	8
s	0	4	3	0	7
c	2	0	0	1	3
l	2	0	0	0	2
total	8	7	3	2	20

transition prob matrix

from \ to	sh	s	c	l	total
sh	1/2	3/8	0	1/8	1
s	0	4/7	3/7	0	1
c	2/3	0	0	1/3	1
l	1	0	0	0	1
total					

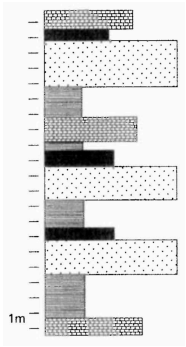
random transition prob matrix

from \ to	sh	s	c	l	total
sh	8/20	7/20	3/20	2/20	1
s	8/20	7/20	3/20	2/20	1
c	8/20	7/20	3/20	2/20	1
l	8/20	7/20	3/20	2/20	1
total					

Time series analysis - Markov chain

Systematics in the lithology changes for a log (time is qualitative)

strat log



transition matrix

from \ to	sh	s	c	l	total
sh	4	3	0	1	8
s	0	4	3	0	7
c	2	0	0	1	3
l	2	0	0	0	2
total	8	7	3	2	20

random transition matrix

from \ to	sh	s	c	l	total
sh	3.2	2.8	1.2	0.8	8
s	2.8	2.45	1.05	0.7	7
c	1.2	1.05	0.45	0.3	3
l	0.8	0.7	0.3	0.2	2
total	8	7	3	2	20

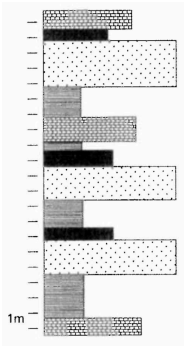
random transition prob matrix

from \ to	sh	s	c	l	total
sh	8/20	7/20	3/20	2/20	1
s	8/20	7/20	3/20	2/20	1
c	8/20	7/20	3/20	2/20	1
l	8/20	7/20	3/20	2/20	1
total					

Time series analysis - Markov chain

Systematics in the lithology changes for a log (time is qualitative)

strat log



transition matrix

from \ to	sh	s	c	l	total
sh	4	3	0	1	8
s	0	4	3	0	7
c	2	0	0	1	3
l	2	0	0	0	2
total	8	7	3	2	20

observed

random transition matrix

from \ to	sh	s	c	l	total
sh	3.2	2.8	1.2	0.8	8
s	2.8	2.45	1.05	0.7	7
c	1.2	1.05	0.45	0.3	3
l	0.8	0.7	0.3	0.2	2
total	8	7	3	2	20

expected

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

X^2_{calc} = 19.23

df = (c -1)(r-1) = 9

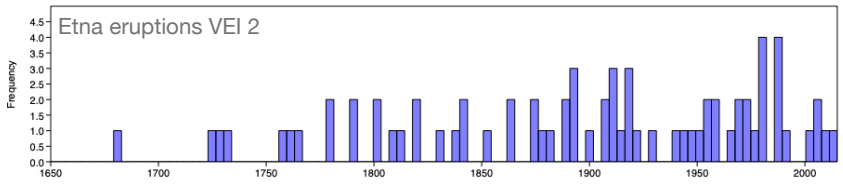
$X^2_{0.05:9}$ = 16.92 : calc exceed critical : not random

Time series analysis - randomness of

same number of eruptions through time

however, non uniform spacing in time?

The past is the key to the future: but only if the past was non-random !



periode 1890 - 2015: 44 eruptions

split this up in 25 year time periods:

	obs	exp
1890 - 1915:	11	8.8
1915 - 1940:	5	8.8
1940 - 1965:	8	8.8
1965 - 1990:	14	8.8
1990 - 2015:	6	8.8

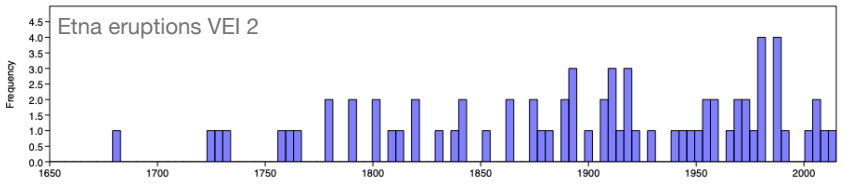
X^2_{calc} = 6.23

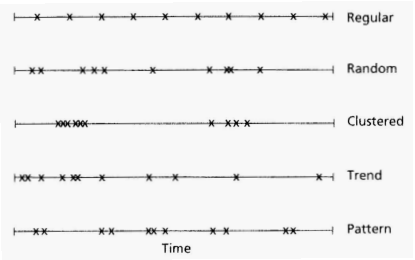
$X^2_{0.05:4}$ = 9.488

d.f. = class - 1 = 5 - 1 = 4

Time series analysis - randomness of events

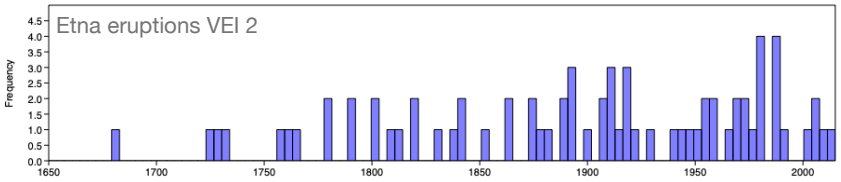
The past is the key to the future: but only if the past was non-random !





Time series analysis - randomness of events

The past is the key to the future: but only if the past was non-random !



periode 1890 - 2015: 44 eruptions, 43 intervals

time between eruptions:		obs
	<2:	19
	2-4:	14
	5-7:	5
	8-10:	5

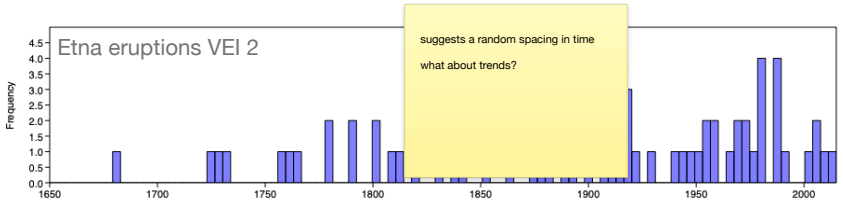
we will calculate the expected random occurrence from the Poisson distribution (2.3.7.3):

$E_j = T \cdot e^{(-n/T)} \cdot (n/T)^j / j!$

where n = total no. events = 44, T = no. intervals = 43

Time series analysis - randomness of events

The past is the key to the future: but only if the past was non-random !



periode 1890 - 2015: 44 eruptions, 43 intervals

time between eruptions:		obs	exp	exp
	<2:	19	15.46	15.46
	2-4:	14	15.81	15.81
	5-7:	5	8.09	8.09
	8-10:	5	2.76	3.61
	11-13:		0.71	
	14-16:		0.14	

$X^2_{calc} = 2.74$

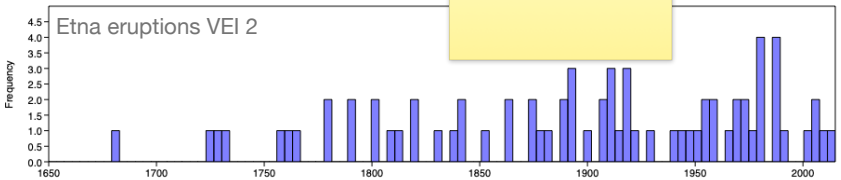
d.f. = class - 1 = 3

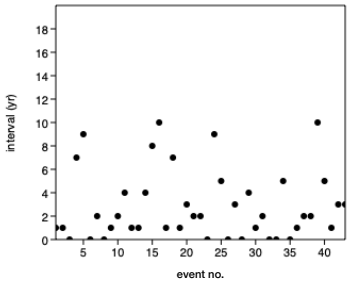
$X^2_{0.05:3} = 7.815$

Time series analysis - randomness of events

The past is the key to the future: but c… non-random !

the occurrence of an eruption has no info on occurrence of another





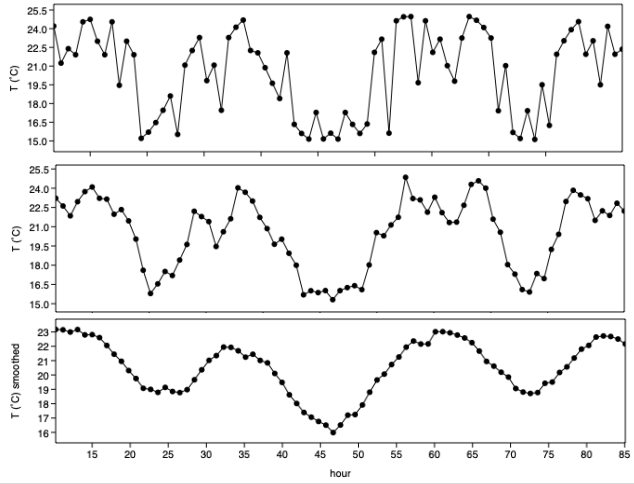
event number does not have a normal distribution: robust test

Spearman r analysis: H_o = no trend
H_A = trend

r' = 0.07, p_{uncorr} = 0.65

Time series analysis - systematics with time

Time series data consists of noise + signal: smoothing allows for noise to be reduced by assuming its frequency to be different

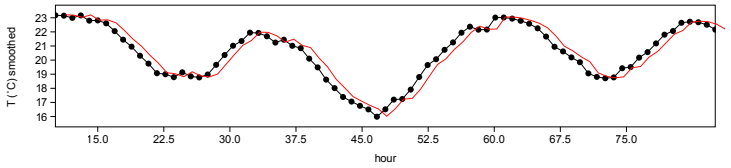


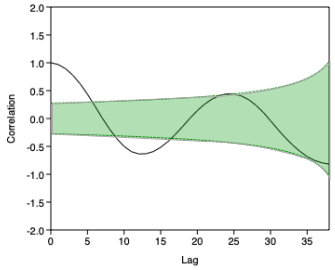
smoothing:
3 point average

smoothing:
weighted
polynomial

Time series analysis - systematics with time

Assessing periodicity in your time series: auto-correlation

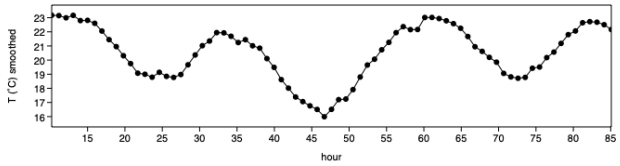




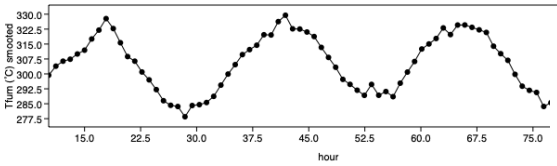
95% confidence window

Time series analysis - systematics with time

Analysing multiple variables against time: cross-correlation



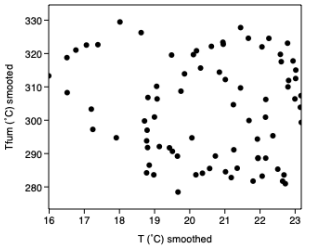
air temperature



fumarole temperature

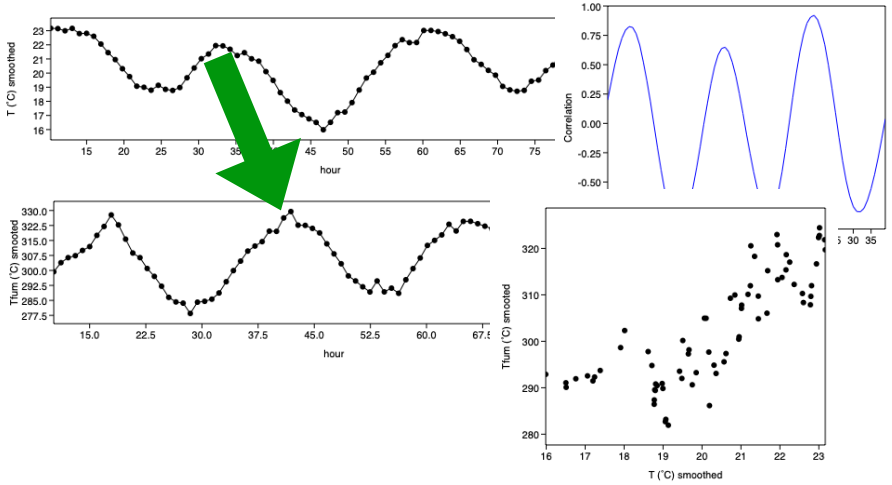
$r = -0.09$, $p_{uncorr} = 0.41$

$r' = -0.07$, $p_{uncorr} = 0.57$



Time series analysis - systematics with time

Analysing multiple variables against time: cross-correlogram

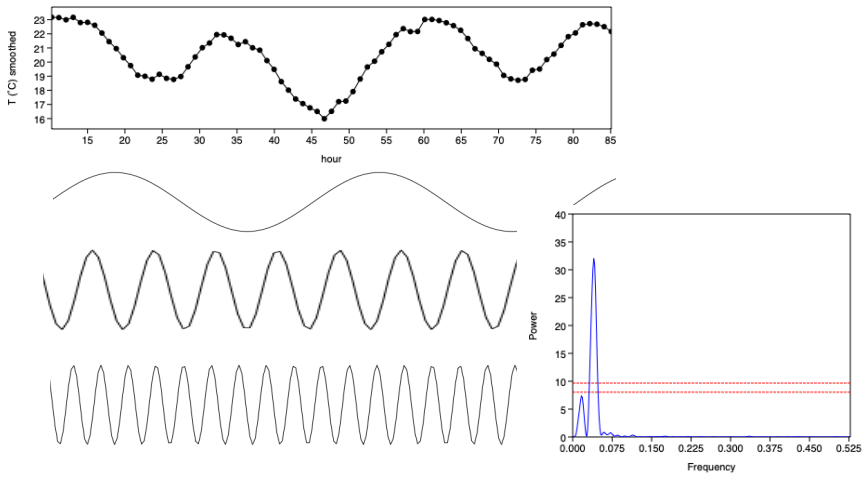


Time series analysis - systematics with time

Analysing multiple variables against time: the need for interpolation

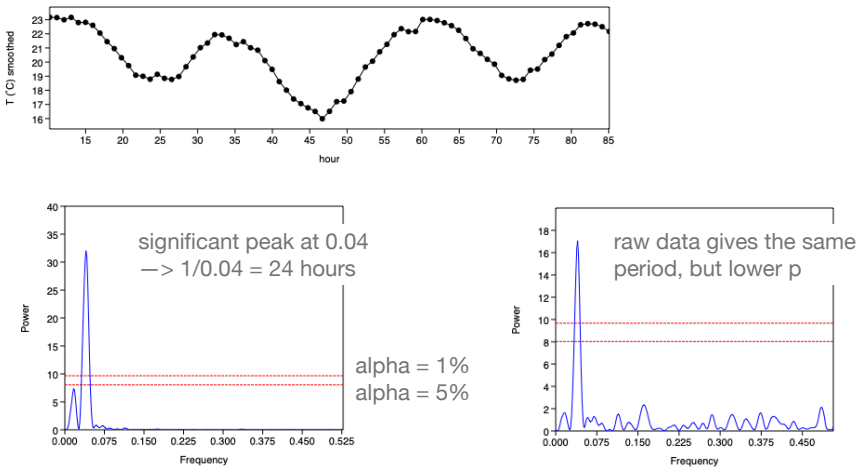
Time series analysis - periodicity

Analysing variables against time: periodogram

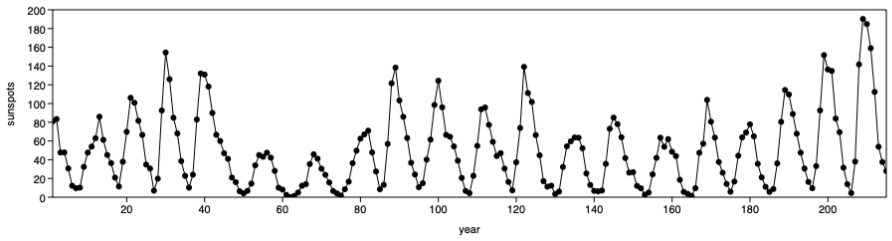


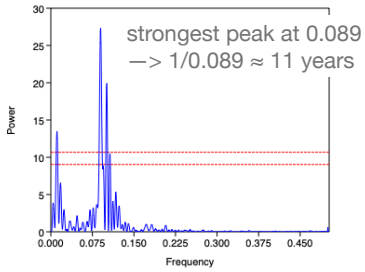
Time series analysis - periodicity

Analysing variables against time: periodogram



Time series analysis - periodicity





This plot is essentially a histogram of the sum of squares for each frequency:

the amount of variance in the time series explained by each frequency

Course logistics

Switch to multivariate statistics after Spring break

- Regression analysis
- Discriminant Function Analysis
- Principal Component Analysis, Factor Analysis & Partial Least Squares
- Cluster analysis (hierarchical and partitioning methods)
- Geostatistics, kriging and semi-variance

Midterm: Tuesday the 15th of March 9:30 to 11:00

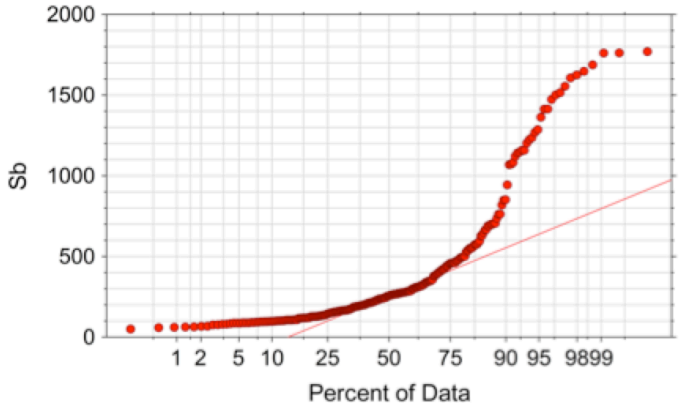
Example midterm

The midterm covers everything up to Spring break

- Generally consists of 3 questions
- Focuses on theory, but statistical tests are commonly included that require some simple calculations
- Closed book exam, but I provide any probability distribution tables and equations you may need
- Bring a hand calculator (no phone)
- The exam is 1.5 hours

Example midterm

QUESTION 1. (35) The figure below shows the cumulative frequency distribution of a data set. What will the histogram for this data set look like and sketch it. Mark the mean, median, mode, standard deviation, interquartile range, and 90% percentile field on your sketch. Which, if any, of these parameters will be meaningful for this data set?



Example midterm

QUESTION 2. (30) Correlation coefficients

(15) What is the purpose of correlation analysis? Provide a geological example where you would use this type of analysis.

(15) Does the significance of a correlation coefficient go up or down when you increase the number of samples? Illustrate your answer using the t-distribution.

QUESTION 3. (35) You just won a bidding war on eBay for a bag of scrap gold for a fantastic price. Your friends are however very skeptical and convinced that you have bought a bag of painted lead, so you decide to test the density of the material. You do 8 measurements and obtain (in g/cm^3):

22.1	18.7	17.8	23.2
20.4	25.5	22.4	27.0

According to wikipedia, the density of pure gold is 19.3 g/cm^3.

(10) What test do you use to determine whether the material is gold ?

(5) What is the H_0 and H_A of your test ?

(15) Set your confidence level to 1% and conduct the test. What do you conclude ?

(5) Your friends are unhappy with your choice of confidence level. At what level does your conclusion change ?