

Data Screening and Transformations

➤ *Objectives:*

Discuss Steps for Analysis: Data Screening, Data Manipulation

Go over the principles of data exploration

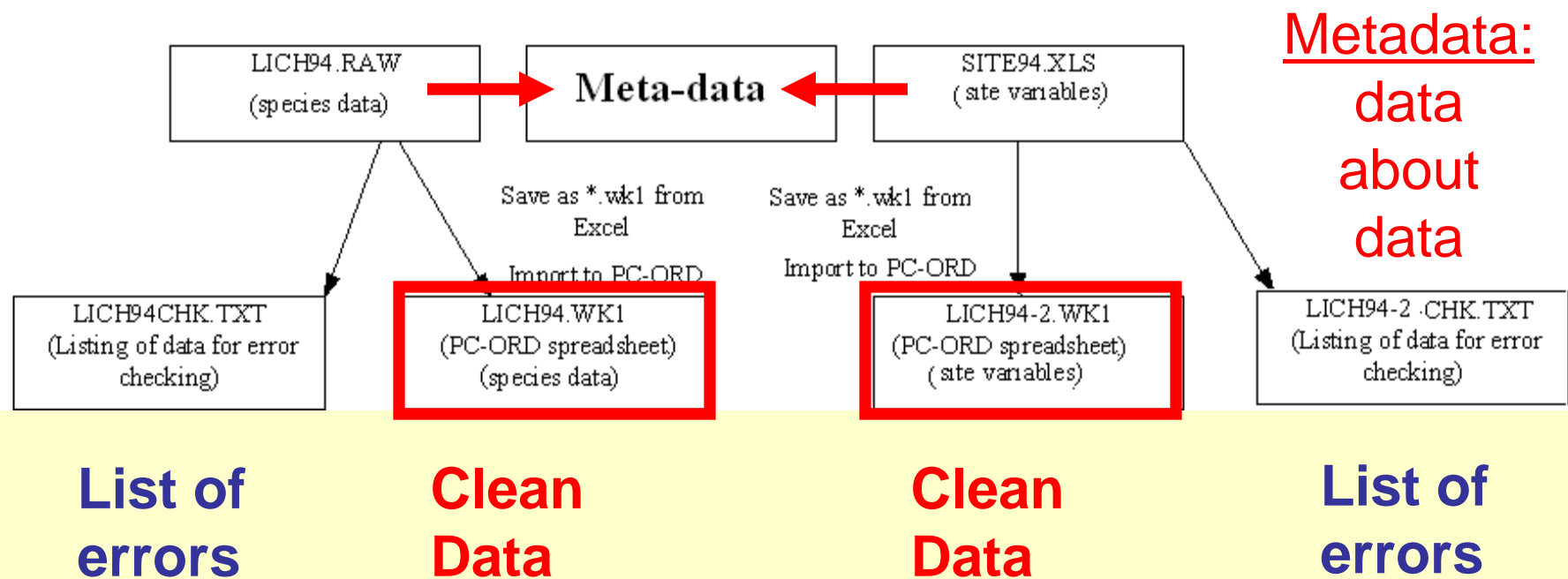
➤ *Learning Outcomes:*

Be ready to plan your analysis: Develop Metadata and Analysis Log

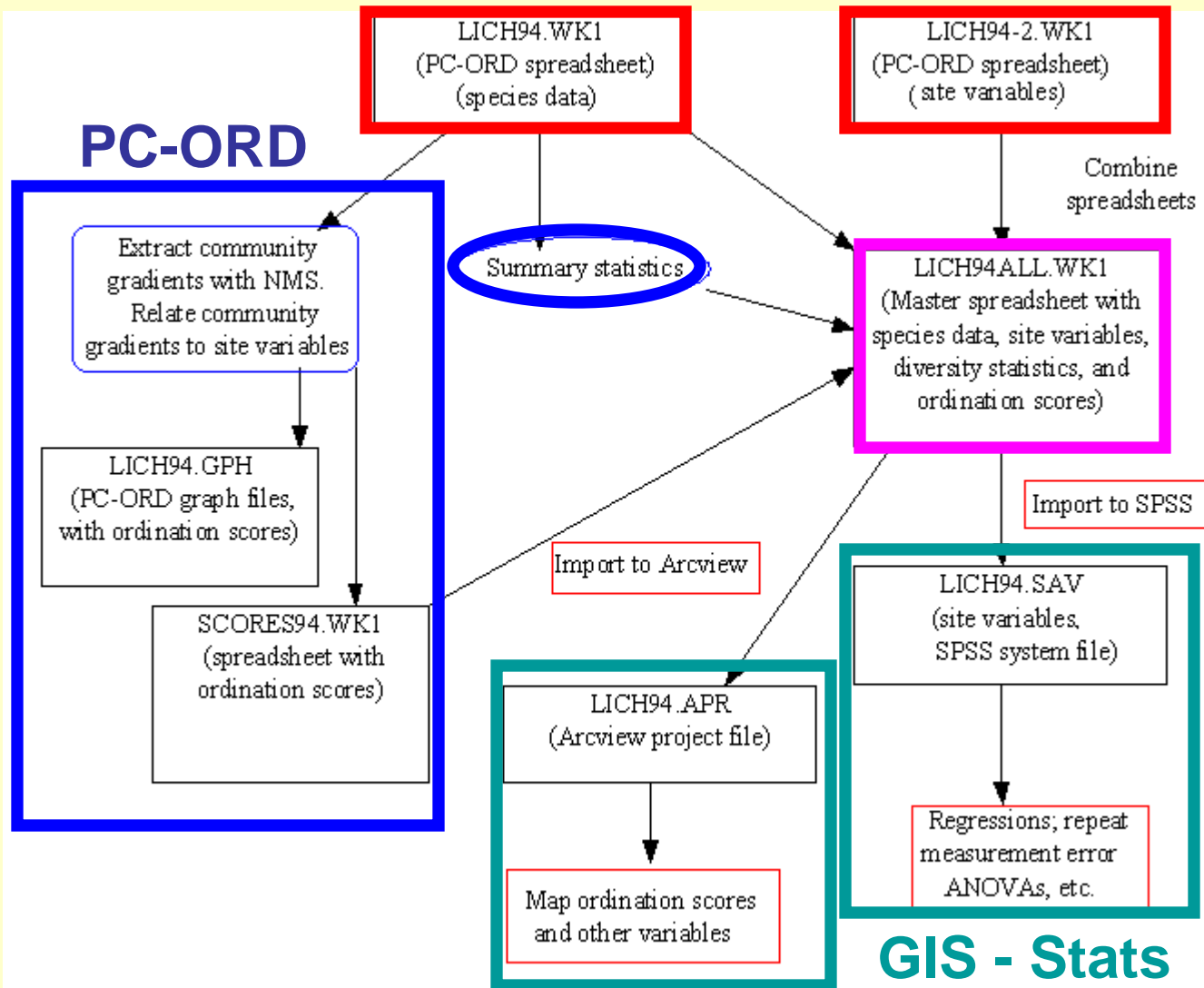
Be able to screen and manipulate your data with PC-ORD

Data Exploration – Documenting Flow

- Flow diagram: sequence of changes / analysis
- Analysis log: input, output, results
- Save all input and output files – and data edits



Data Exploration – Documenting Flow



- File Names
- File Contents
- Connections
- Links to other software
- Products:
 - figures
 - tables
 - results

Data Exploration – Documenting Flow

- Screening:
 - Are column / rows means and ranges reasonable?
 - Are the sample sizes correct?
 - Are there missing data / outliers?
- Cleaning:
 - Fix typos
 - Erase / Correct incomplete data
 - Check effects of corrections
- Transformations:
 - Look up assumptions of test
 - Check data distributions
 - Make transformations (re-check)

Data Exploration – Data Screening

Main - PCA1M.WK1					
96	stands				
5	variables				
	Q	Q	Q	Q	Q
	time	MEI	PDO	up_36	up_39
1997jan	1997	0.23	-8	-21	
1997feb	1997.083	-0.482	0.28	-8	8
1997mar	1997.167	-0.174	0.65	88	65
1997apr	1997.25	0.468	1.05	49	50
1997may	1997.333	1.1	1.83	126	45
1997jun	1997.417	2.316	2.76	-88	-86
1997jul	1997.5	2.648	2.35	75	145
1997aug	1997.583	2.876	2.79	19	105
1997sep	1997.667	2.847	2.19	-27	-58
1997oct	1997.75	2.217	1.61	29	18

Metadata

- 96 samples
- 5 variables
- Data type?
- Explanation



➤ Show Current Profile

% zeros, data ranges, skewness

Data Exploration – Current Profile

% zeros:
species data

Lowest / highest value:

typos (errors)

Skewness:
non-normality

$$-1 < SK < 1$$

Outliers:
(in SD units)

$$2 \text{ SD} \rightarrow 96\%$$

Main matrix: PCA1M.WK1

Main matrix

% zeros	1.0
Average distance - Rela.Eucl.	0.05615
Lowest nonzero value	-162.000
Highest value	2004.917

	Rows	Columns
Contents:	96 stands	5 variable
Skewness		
Average	2.2	0.0
Maximum	2.2	0.9
Minimum	2.1	-1.9
CV of totals, %	5.9	215.5

Potential Outliers

Distance measure: Rela.Eucl.

SD-Item	SD-Item
4.1-1999jun	0.0-
3.6-2002jul	0.0-
2.9-1999may	0.0-
2.6-2003jan	0.0-

Data Exploration – Summary I

➤ Data Summary:

Summary Ordination Graph Groups

Compact-Format Data Summary

Row And Column Summary

Outlier Analysis

Species-area Curves

Species Lists

Write Distance Matrix

		Mean	SD	Range		Diversity				
Summary of:		5 variable			N = 96 stands					
Num.	Name	Mean	Stand.Dev.	Sum	Minimum	Maximum	S	E	H	D'
1	time	2000.958	2.321	192092.0000	1997.000	2004.917	96	1.000	4.564	0.9896
2	MEI	0.302	1.033	28.9850	-1.194	2.876	95	0.951	4.329	0.9847
3	PDO	0.043	1.371	4.0800	-8.000	2.790	96	0.996	4.545	0.9893
4	up_36	22.448	57.692	2155.0000	-111.000	200.000	94	0.983	4.467	0.9877
5	up_39	37.302	75.555	3581.0000	-162.000	240.000	94	0.988	4.487	0.9881
AVERAGES:		412.2	27.59	0.3957E+05	343.0	490.1	95.0	0.983	4.479	0.9879

S = Richness = number of non-zero elements in row

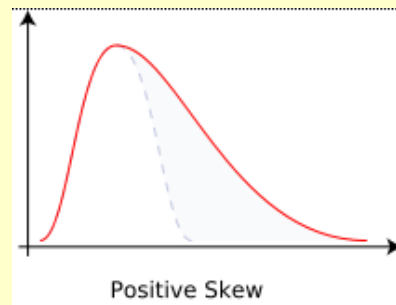
E = Evenness = $H / \ln(\text{Richness})$

H = Diversity = $-\sum (P_i \cdot \ln(P_i))$ = Shannon`s diversity index

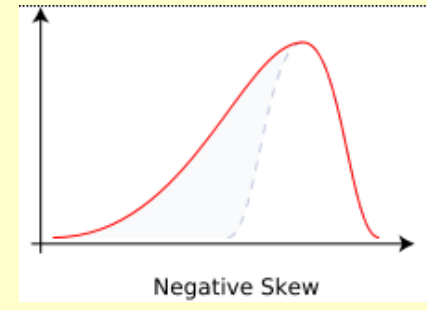
D = Simpson`s diversity index for infinite population = $1 - \sum (P_i \cdot P_i)$

Data Exploration – Summary II

➤ Skewness:



0



Steps to Fix Skewness:

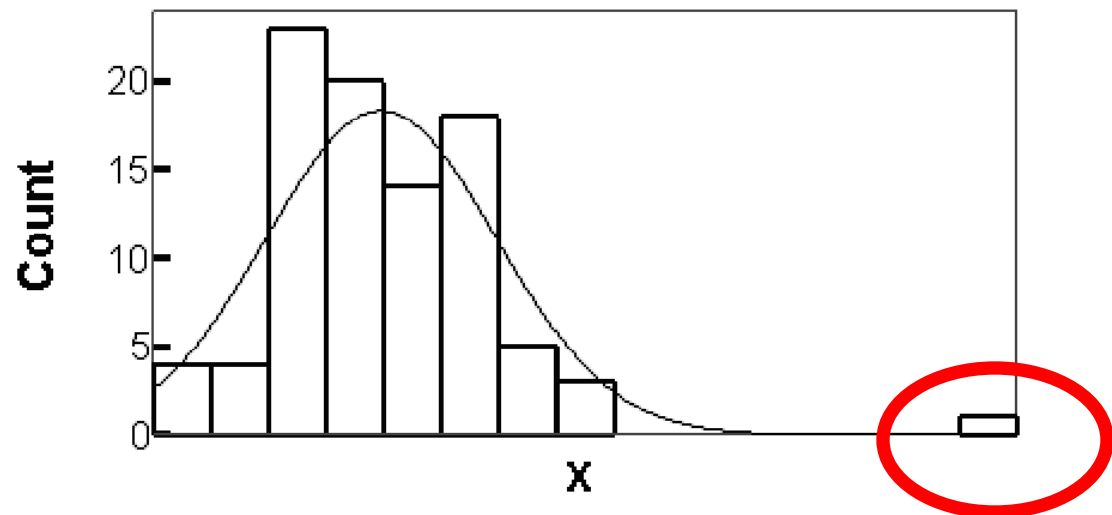
Taking the log or square root works for data with moderate skewness

	Skewness	Kurtosis
1 time	0.000	-1.103
2 MEI	0.945	0.456
3 PDO	0.221	-0.140
4 up_36	0.609	1.040
5 up_39	0.246	0.424
Averages:	0.404	0.136

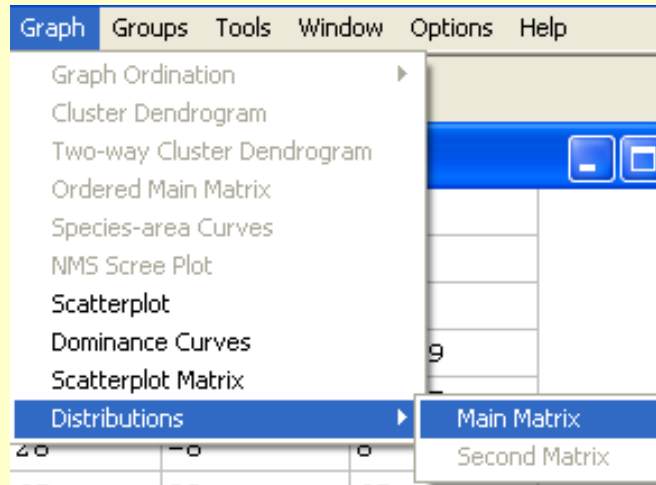
Data Exploration – 1-D Outliers

Type	Detect →	Describe →
Univariate	Histograms, boxplots, normal probability plots, etc.	Note which variables and sample units are involved.

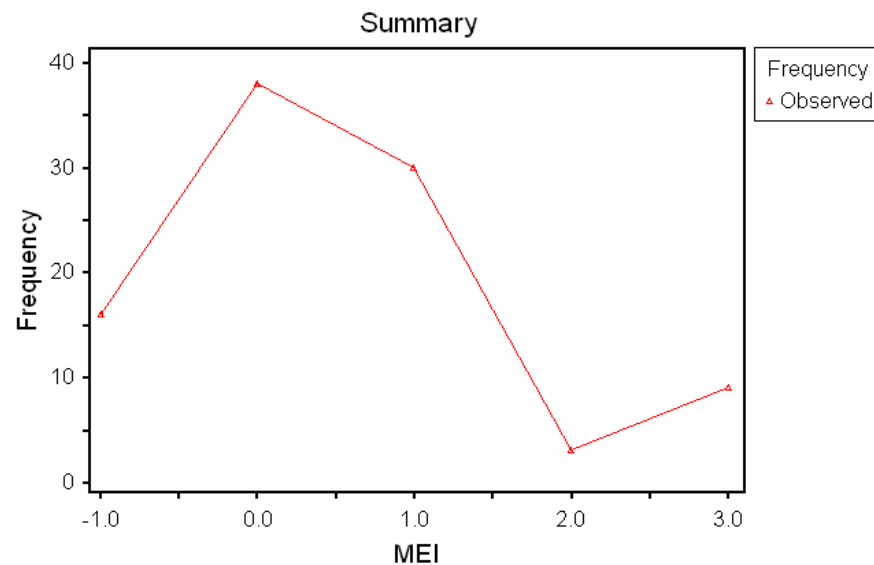
Frequency distribution of a univariate outlier falling 5.5 standard deviations above mean



Data Exploration – 1-D Outliers



- Describe the distribution:
In graph and tabular form

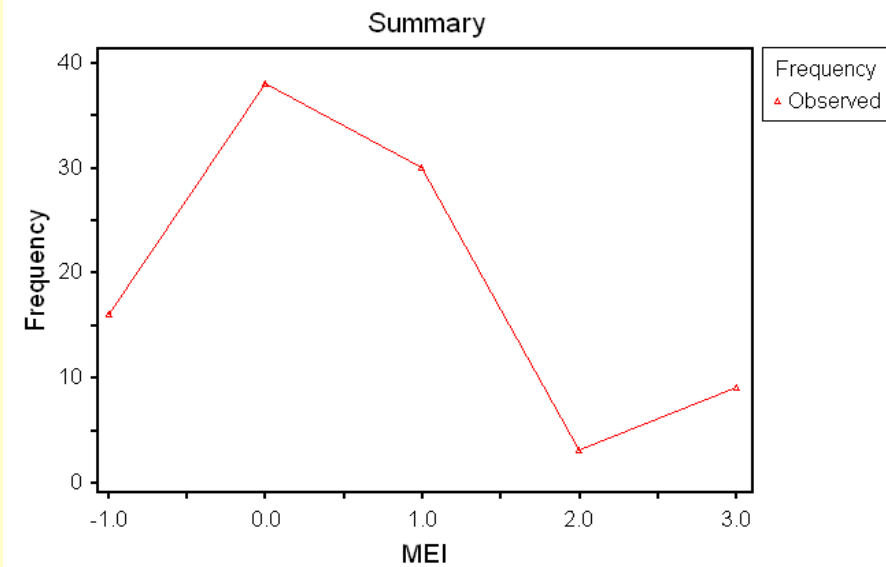


Result - RESULT.TXT

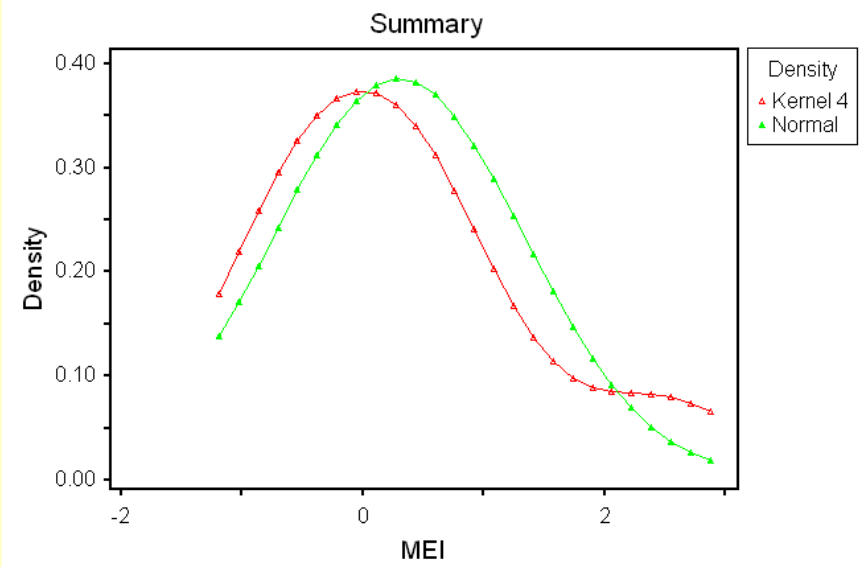
Basic descriptive statistics

```
0.2946      = mean
1.074       = variance
1.036       = standard deviation
0.9446      = skewness
0.9300E-01  = median
-1.194      = minimum
0.1000E-02  = smallest positive value
2.876       = maximum
4.070       = range
```

Data Exploration – 1-D Outliers



Discrete Distribution



Continuous Distribution

DISTRIBUTIONS for variable: MEI

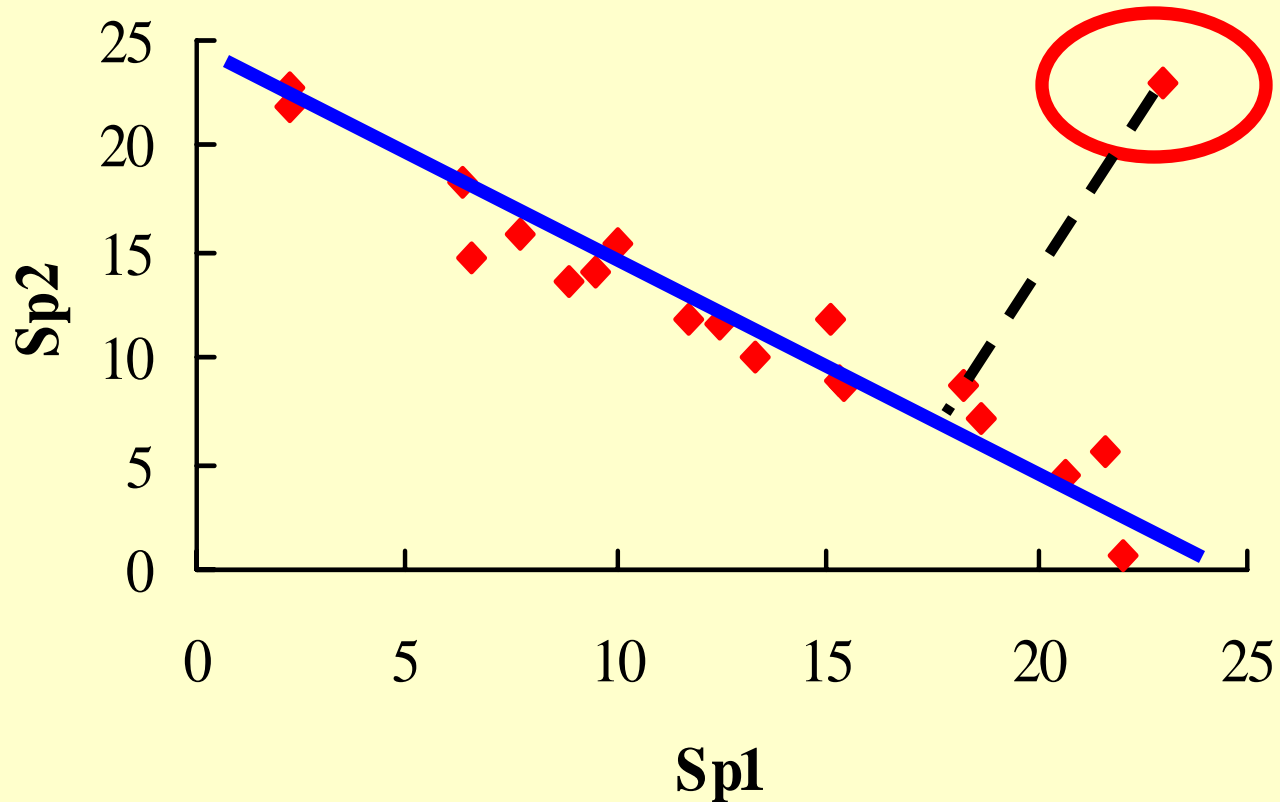
Values are density estimates from various models.

-99.9999 = not enough data or incompatible data.

Test for significance: off PC-ORD

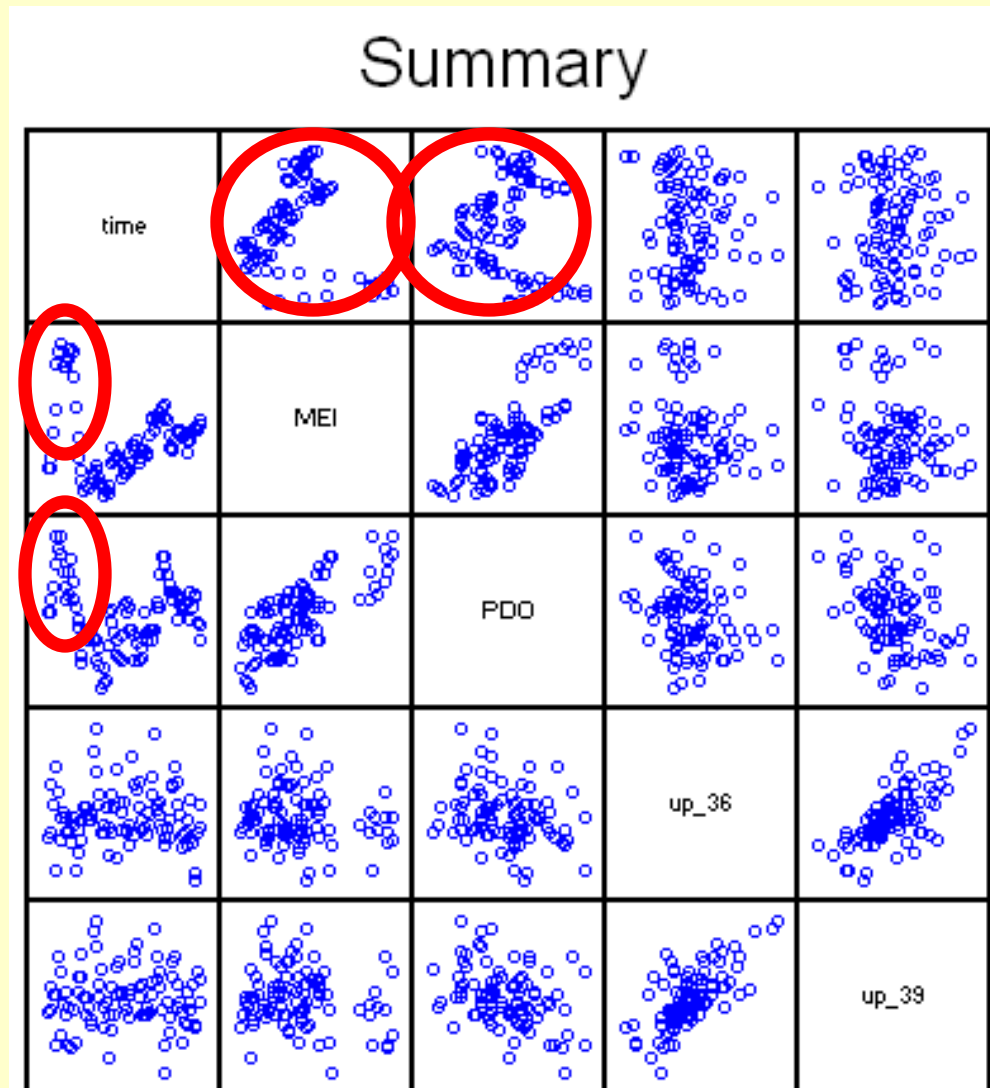
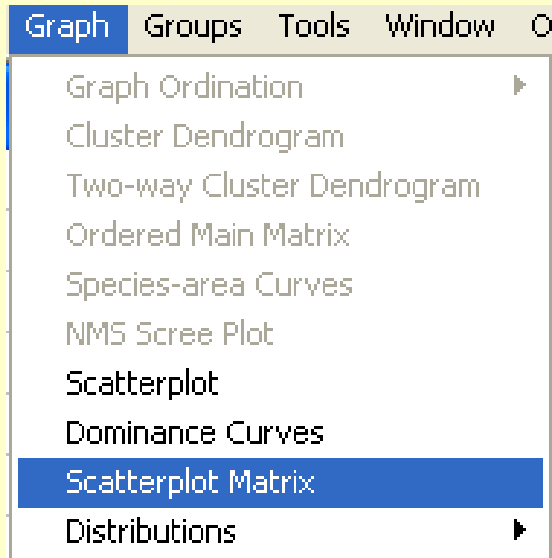
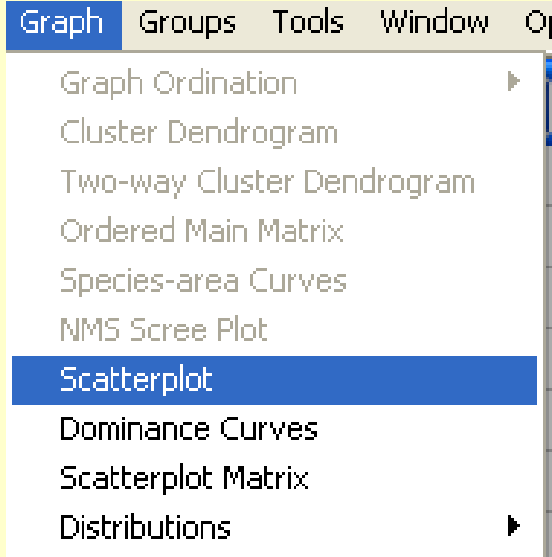
Step	MEI	Distribution type					
		Kernel 1	Kernel 2	Kernel 3	Kernel 4	Normal	Lognorm
0	-1.194	0.1687	0.1711	0.1827	0.1773	0.1372	-100.0
1	-1.031	0.2539	0.2235	0.2106	0.2179	0.1698	-100.0

Data Exploration – 2-D Outliers



A bivariate outlier that is not a univariate outlier for either of the two variables Sp1 and Sp2

Data Exploration – 2-D Outliers



Data Exploration – 2-D Outliers

Type	Detect →	Describe →
Multivariate	<p>A. Compute Mahalanobis distance (distance from a sample unit to the group of remaining sample units). Use a very conservative probability, e.g., $p < 0.001$. Use a chi-square table with degrees of freedom equal to the number of variables. If data are grouped, seek outliers in each group.</p> <p>OR</p> <p>B. Calculate average distance, using your choice of distance measure, from each sample unit to every other sample unit. Examine SUs that fall greater than some number of standard deviations (say 2 or 3) above the mean for average distance.</p> <p>In PC-ORD you can do this by selecting <i>Summary Outlier Analysis</i>.</p>	<p>If there are only a few outlying sample units (cases), examine each case. If there are many outliers, create a dummy grouping variable where outlier cases are given a 1 and others get dummy=0. Use the dummy variable as the grouping variable in discriminant analysis or as a dependent variable in multiple regression.</p>

Data Exploration – Outliers

Summary Ordination Graph Groups

- Compact-Format Data Summary
- Row And Column Summary
- Outlier Analysis**
- Species-area Curves
- Species Lists
- Write Distance Matrix

Outlier Analysis [X]

Outlier Analysis

Rows: stands

Columns: variables

Distance Measure

Sorensen (Bray-Curtis)

Relative Sorensen

Jaccard

Euclidean (Pythagorean)

Relative Euclidean

Correlation

Chi-squared

Squared Euclidean

Outlier

Graph frequency distribution

Write list of outliers

Write ranked list

All the above

Cutoff = s.d.'s to flag outliers

OK Cancel Help

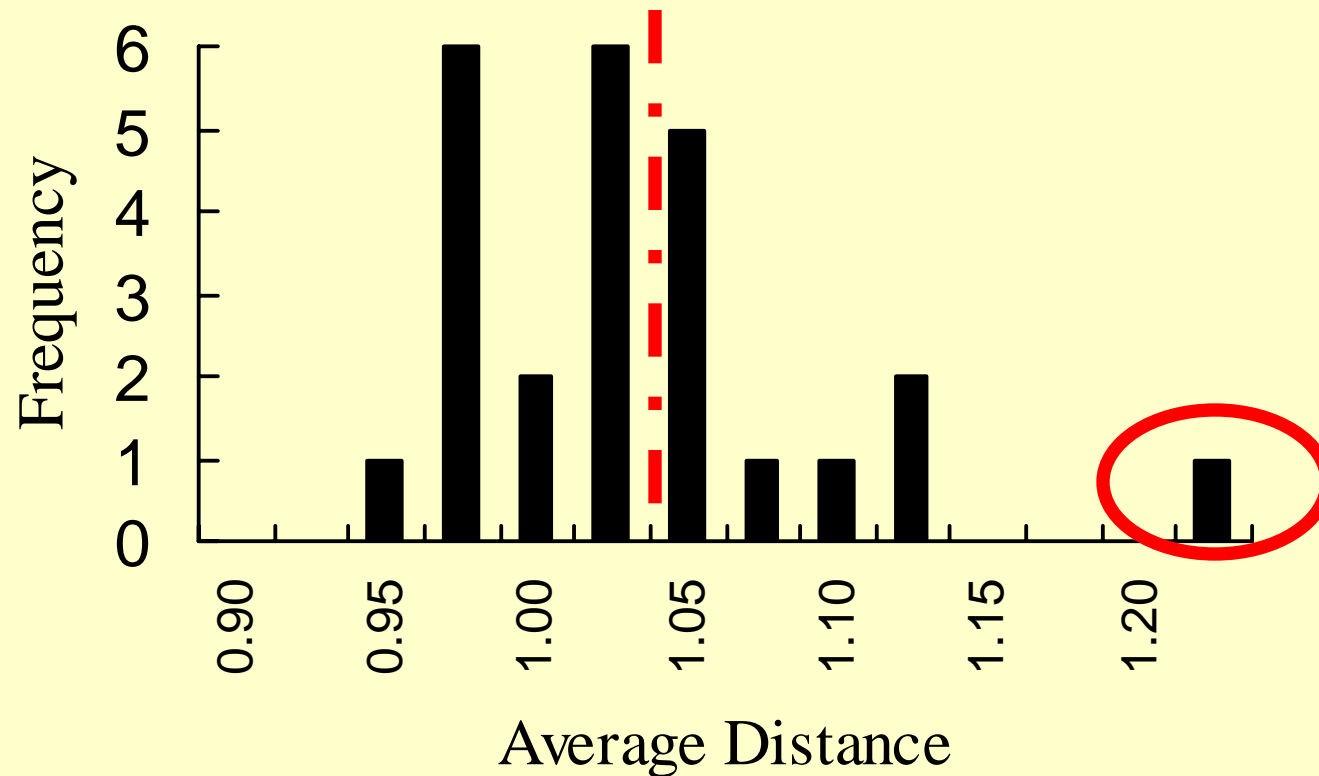
Distance*	Frequency	(each "X" represents one entity)
261.99130	X	1999jun
248.87854	X	2002jul
235.76579		
222.65305	X	1999may
209.54030	X	2003jan
196.42755		
183.31480	X	
170.20206	XXXX	
157.08932	XX	
143.97658	XXXXXXXX	
130.86383	XXXXX	
117.75108	XXXXX	
104.63834	XXXXXXXXXXXXXXXX	
91.52559	XXXXXXXXXXXXXXXX	
78.41284	XX	

* Distances at left are lower end of that bin's range.

Summary

RANK	ENTITY NAME	AVERAGE DISTANCE	STANDARD DEVIATIONS
1	1999jun	275.10403	4.10900
2	2002jul	255.51050	3.61313
3	1999may	229.00174	2.94226
4	2003jan	214.17082	2.56692

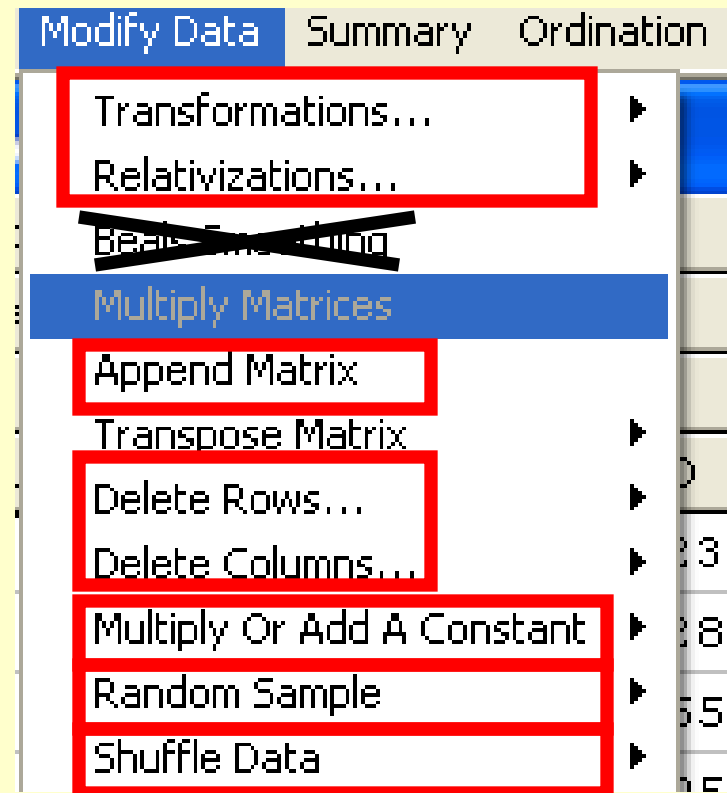
Data Exploration – Outliers



Frequency distribution of average relative Euclidean distances to a sample unit, given a sample size of 25. The sample marked with the red circle is 3.2 SD units above the mean of the average distances

Data Manipulations

- You can manipulate data directly in PC-ORD
 - Modify / Append Data
 - Delete Columns / Rows
 - Multiply / Add Constant
 - Randomly Sample
 - Shuffle Data



Note: Beals smoothing is Experimental – **DO NOT USE**

Data Transformations

- What are the two reasons for data transformations?
 - Statistical:
 - Meet assumptions (normality, linearity, variances,...)
 - Express variables in the same units (km, km/hr):
 - Ecological:
 - Make distance measures work better
 - Reduce influence of total quantity (sample totals)
 - Deal with importance of “rare” / “common” species
 - Identify informative species

Data Transformations - Nomenclature

- **Monotonic:** Element values are changed, but ranks stay the same (e.g., change unit from km to m)

- **Relativization:** Adjusts matrix elements by one column / row standard (e.g., total, maximum)

Note: Not all transformations are reasonable or feasible with all types of data (e.g., negative, P/A)

Data Transformation

Modify Data Summary Ordination Graph Groups Tools Window Options Help

Transformations... Main Matrix Power Transformation
 Relativizations... Second Matrix Logarithmic Transformation
 Beals Smoothing Presence-absence
 Multiply Matrices Arcsine Transformation
 Append Matrix Arcsine Squareroot
 Transpose Matrix
 Delete Rows...
 Delete Columns...
 Multiply Or Add A Constant
 Random Sample
 Shuffle Data

	Q	
0	up_36	up_39
23	-0.13888	-0.27743
28	-0.13888	0.105688
55	1.52768	0.858719
75	0.8506402	0.660553

Data Transformation

- Monotonic transformations retain ranks, but change values

	Reasonable and acceptable domain of x	Range of $f(x)$	
MONOTONIC TRANSFORMATIONS			
x^0 (power)	all	0 or 1 only	P / A
$x^{1/2}$ (power)	nonnegative	nonnegative	
$\log(x)$	positive	all	
$(2/\pi) \cdot \arcsin(x)$	$0 \leq x \leq 1$	0 to 1 inclusive	
$(2/\pi) \cdot \arcsin(x^{1/2})$	$0 \leq x \leq 1$	0 to 1 inclusive	

(x)  **f(x)**

Power exponents: $1/2$ (square root), 2 (squared), 3 (cubed)

Note: 0 used to recode data as Presence / Absence (0 / 1)

Data Transformations – Example I

Logarithmic transformation $f(x) = \ln(x)$ OR $\log(x)$

TRANSFORMATION	Reasonable and acceptable domain of x	Range of $f(x)$
$\log(x)$	positive	all

- This transformation is useful when:
 - high degree of variation within attributes (e.g., Chl Conc.)
 - high degree of variation among attributes within a sample
 - helps if there are large outliers and lots of zeros
- Note: to log-transform data containing zeros, a small number should be added to all data points.
 - With count data, add one, so that: $f(x) = \log(0+1) = 0$
 - With density data, add constant smaller than smallest possible sample, so that: $f(x) = \log(0+0.001) = -3$

Data Transformations – Example II

Arcsine / Arcsine-squareroot transformation

TRANSFORMATION	Reasonable and acceptable domain of x	Range of $f(x)$
$(2/\pi) \cdot \arcsin(x)$	$0 \leq x \leq 1$	0 to 1 inclusive
$(2/\pi) \cdot \arcsin(x^{1/2})$	$0 \leq x \leq 1$	0 to 1 inclusive

➤ This transformation is useful when:

- normalizing proportion data (e.g., Percent Cover)

➤ Note: data must range between zero and one, inclusive. If they are not, you should **relativize** (general relativization or relativization by maximum) before selecting this option.

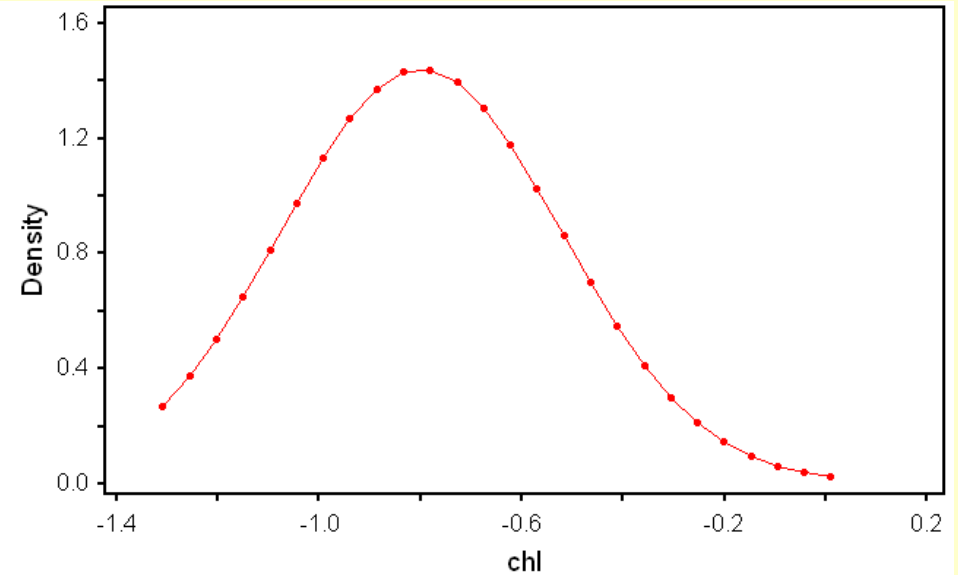
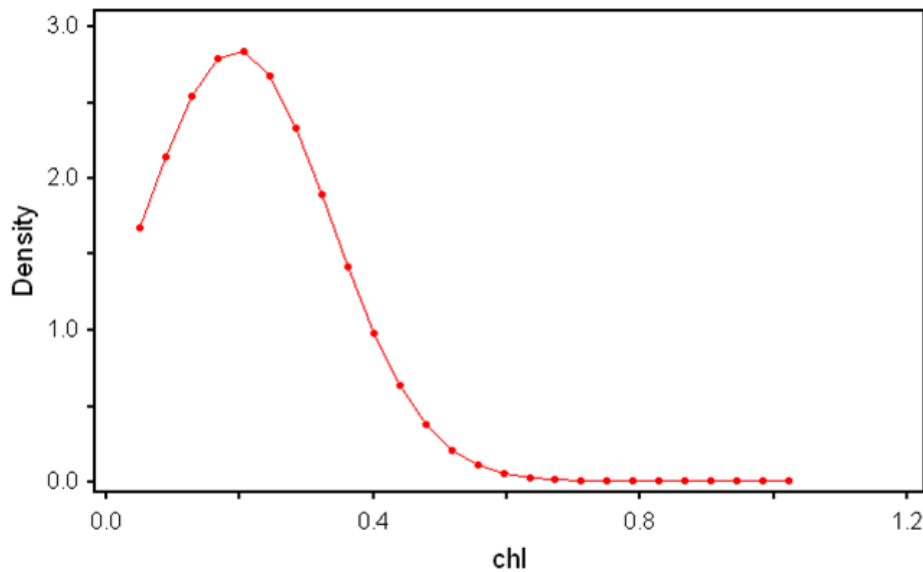
The constant $2 / \pi$ scales the result of $\arcsin(x)$ [in radians] to range from 0 to 1, assuming that $0 < x < 1$.

Data Transformation – How to...



Note:

Need to accept TEMP file



Data Relativization

- Relativization re-scales data using some criterion / standard.
- When its done by columns (e.g., species), variation across plots is retained, but variation across species is standardized.
- Two approaches:

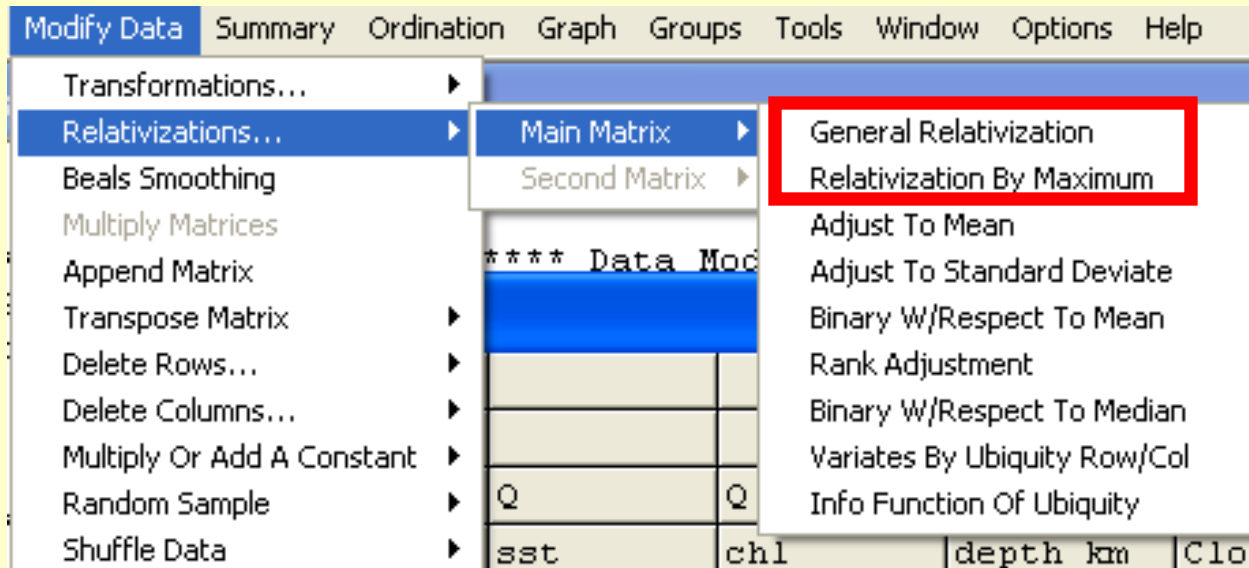
General Relativization: (by totals or sums) makes area under each species distribution response curve = 1.

(input: $x \geq 0$; output: from 0 to 1)

Relativization by Maximum: (by max for column or row) equalizes the heights of the peaks along the gradient

(input: $x \geq 0$; output: from 0 to 1)

Data Relativization



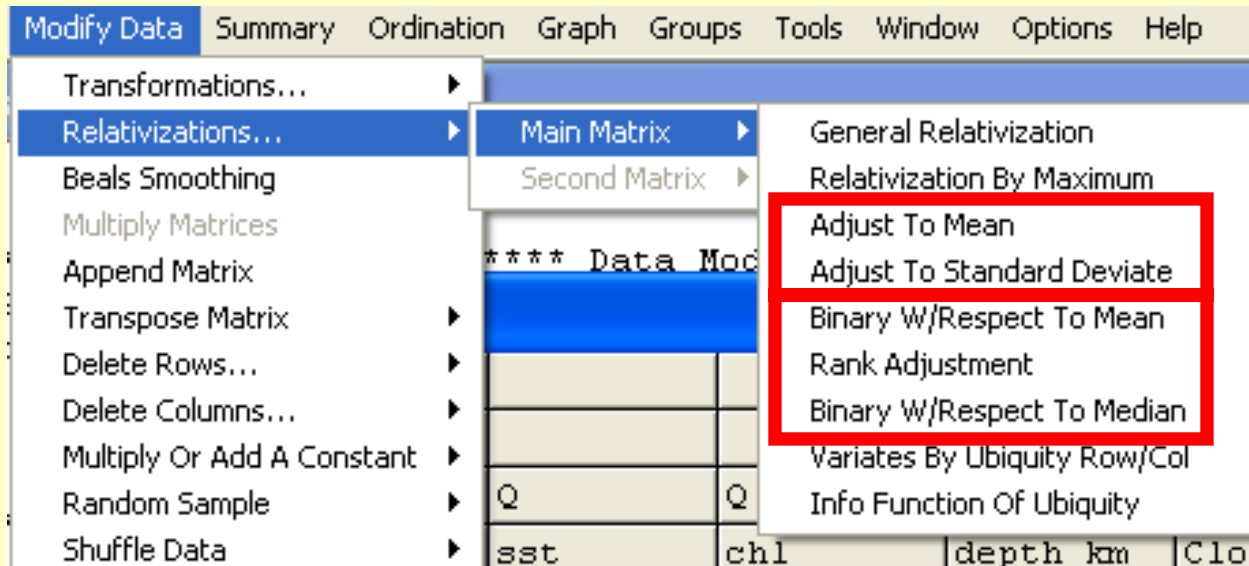
- **General Relativization:** (by totals or sums) makes area under each species distribution response curve = 1.

(input: $x \geq 0$; output: from 0 to 1)

- **Relativization by Maximum:** (by max for column or row) equalizes the heights of the peaks along the gradient

(input: $x \geq 0$; output: from 0 to 1)

Data Relativization



- **Deviations:** Value – Mean
- **Z scores:** $(\text{Value} - \text{Mean}) / \text{SD}$
- **Binary response:** Above (1) / Below (0)
- **Ranks:** Assigns ranks
(e.g., 0, 0, 6, 9 would receive the ranks 1.5, 1.5, 3, 4)

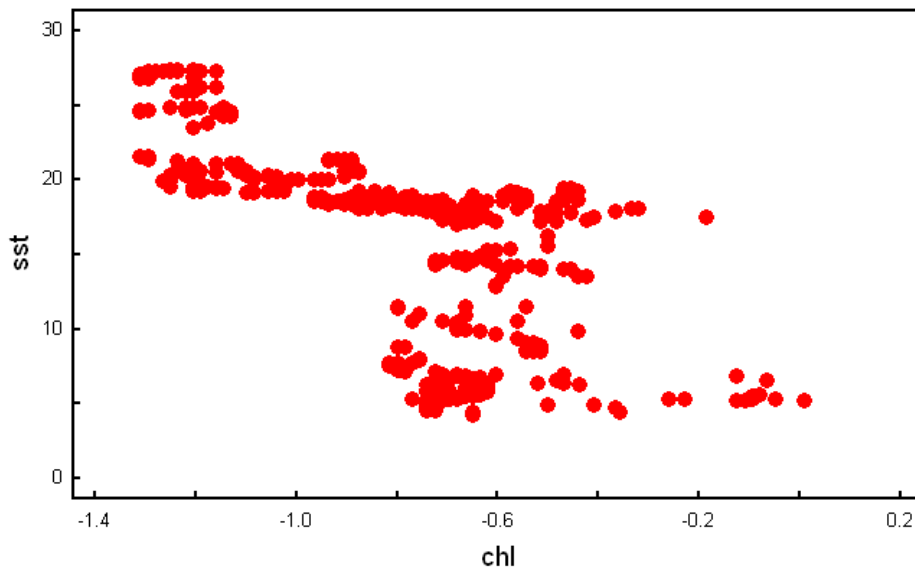
Data Relativization – How to...



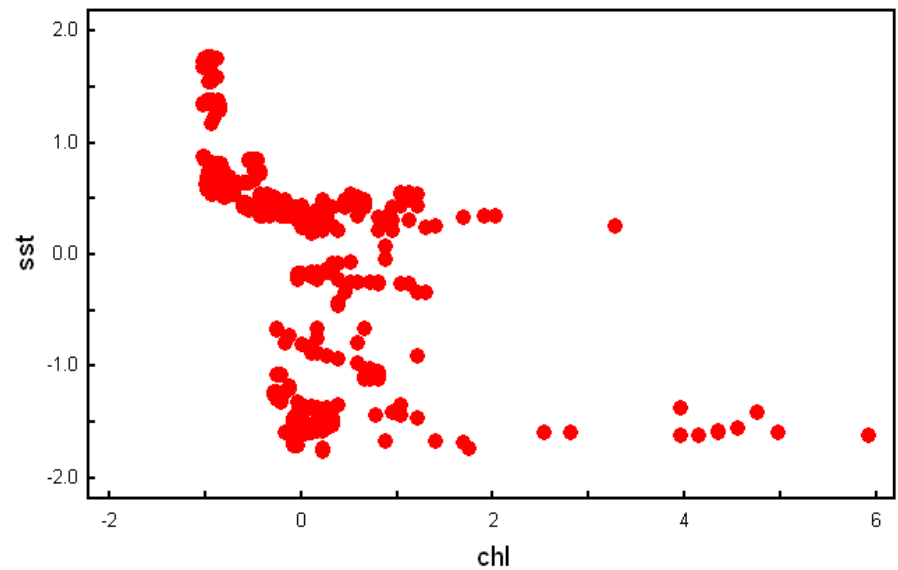
Note:

Need to accept TEMP file

SST vs Chl



Standardized Anomalies (SST vs Chl)



Data Exploration - Summary

- Create naming convention for your files (metadata record)
(DATE_AREA_SP_suffix)
9710_Oahu_WTSH_raw
- Create a data flow archive in your analysis notebook
- Check assumptions of statistical tests / approaches
(PCA: normality of data, linear relationships)
- Visually inspect your data: 1-D, 2-D, many-D.
 - Look for missing data and outliers in individual datasets
 - Inspect relationships between variables (pairs, multiple)

Data Manipulation - Summary

- Add missing data and fix typos
- Ensure variables expressed in the same units (km / m)
- Select the number and identify of species
(Rare species that occur in a single sample contribute virtually no information, but add noise)
- Look for and deal with outliers: (Remove OR Transform)
- Deal with confounding factors, such as the different magnitude of environmental variables (e.g., depth in m or km) and the proportional representation of different species

(Relativize your data)

CHAPTER 9

Data Transformations

Tables, Figures, and Equations

From: McCune, B. & J. B. Grace. 2002. *Analysis of Ecological Communities*. MjM Software Design, Gleneden Beach, Oregon <http://www.pcord.com>

A general procedure for data adjustments

Species data

Table 9.3. Suggested procedure for data adjustments of species data matrices.

Action to be considered	Criteria
1. Calculate descriptive statistics. <u>Repeat this</u> after each step below. (In PC-ORD run <i>Row & column summary</i>) Beta diversity (community data sets) Average skewness of columns Coefficient of variation (CV, %) CV of row totals CV of column totals	Always
2. Delete rare species (< 5% of sample units)	Usually applied to community data sets, unless contrary to study goals

Species data, cont.

3. Monotonic transformation (if applied to species, then usually applied uniformly to all of them, so that all are scaled the same)

A. Average skewness of columns (species)

B. Data range over how many orders of magnitude?
(Count and biomass data often are extreme.)

C. Beta diversity. (Consider presence/absence transformation for community data when β is high.)

Species data, cont.

3. Monotonic transformation (if applied to species, then usually applied uniformly to all of them, so that all are scaled the same)

A. Average skewness of columns (species)

B. Data range over how many orders of magnitude? (Count and biomass data often are extreme.)

C. Beta diversity. (Consider presence/absence transformation for community data when β is high.)

4. Row or column relativizations

What is the question?

Are units for all variables the same?

Is relativization built into the subsequent analysis?

CV of row totals

CV of column totals

What distance measure do you intend to use?

Note: regardless of your decision to relativize or not, you should state your decision and justify it briefly on biological grounds.

Species data, cont.

5. Check for outliers based on the average distance of each point from all other points. Calculate standard deviation of these average distances. Describe outliers and take steps to reduce influence, if necessary

standard
deviation

< 2

2 - 2.3

2.3 - 3

>3

degree of
problem

no problem

weak outlier

moderate outlier

strong outlier

Environmental data

Table 9.4. Suggested procedure for data adjustments of quantitative variables in environmental data matrices.

Action to be considered	Criteria
1. Calculate descriptive statistics for quantitative variables. <u>Repeat this</u> after each step below. (In PC-ORD run <i>Row & column summary</i>) Skewness and range for each variable (column)	Always
2. Monotonic transformation (applied to individual variables, depending on need)	Consider log or square root transformation for variables with skewness > 1 or ranging over several orders of magnitude. Consider arcsine squareroot transformation for proportion data.

Environmental data

Table 9.4. Suggested procedure for data adjustments of quantitative variables in environmental data matrices.

Action to be considered	Criteria
1. Calculate descriptive statistics for quantitative variables. <u>Repeat this</u> after each step below. (In PC-ORD run <i>Row & column summary</i>) Skewness and range for each variable (column)	Always
2. Monotonic transformation (applied to individual variables, depending on need)	Consider log or square root transformation for variables with skewness > 1 or ranging over several orders of magnitude. Consider arcsine squareroot transformation for proportion data.
3. Column relativizations	Consider column relativization (by norm or standard deviates) if environmental variables are to be used in a distance-based analysis that does not automatically relativize the variables (for example, using MRPP to answer the question: do groups of sample units defined by species differ in environmental space?). Column relativization is not necessary for analyses that use the variables one at a time (e.g., ordination overlays) or for analyses with built-in standardization (e.g., PCA of a correlation matrix).

Environmental data

Table 9.4. Suggested procedure for data adjustments of quantitative variables in environmental data matrices.

Action to be considered	Criteria
<p>1. Calculate descriptive statistics for quantitative variables. <u>Repeat this</u> after each step below. (In PC-ORD run <i>Row & column summary</i>)</p> <p style="padding-left: 40px;">Skewness and range for each variable (column)</p>	Always
<p>2. Monotonic transformation (applied to individual variables, depending on need)</p>	<p>Consider log or square root transformation for variables with skewness > 1 or ranging over several orders of magnitude.</p> <p>Consider arcsine squareroot transformation for proportion data.</p>
<p>3. Column relativizations</p>	<p>Consider column relativization (by norm or standard deviates) if environmental variables are to be used in a distance-based analysis that does not automatically relativize the variables (for example, using MRPP to answer the question: do groups of sample units defined by species differ in environmental space?). Column relativization is not necessary for analyses that use the variables one at a time (e.g., ordination overlays) or for analyses with built-in standardization (e.g., PCA of a correlation matrix).</p>
<p>4. Check for univariate outliers and take corrective steps if necessary.</p>	<p>Examine scatterplots or frequency distributions or relativize by standard deviates (“z scores”) and check for high absolute values.</p>