

PCA Examination Key

This exam is worth 10 points (two homeworks).

Just like in the homeworks, make sure you explain what you are doing and how you are getting the answers. This way, I can give you partial credit for incomplete answers.

In particular, explicitly state what PC-ORD command you used to obtain the various figures / results.

You will turn in a ppt file with your images and text inserted into the body of the presentation. To copy text from PC-ORD screen, use “CONTROL + Print Screen”

When answering the questions, back up your responses with figures / tables / numbers. An image / table is worth 1000 words!!!

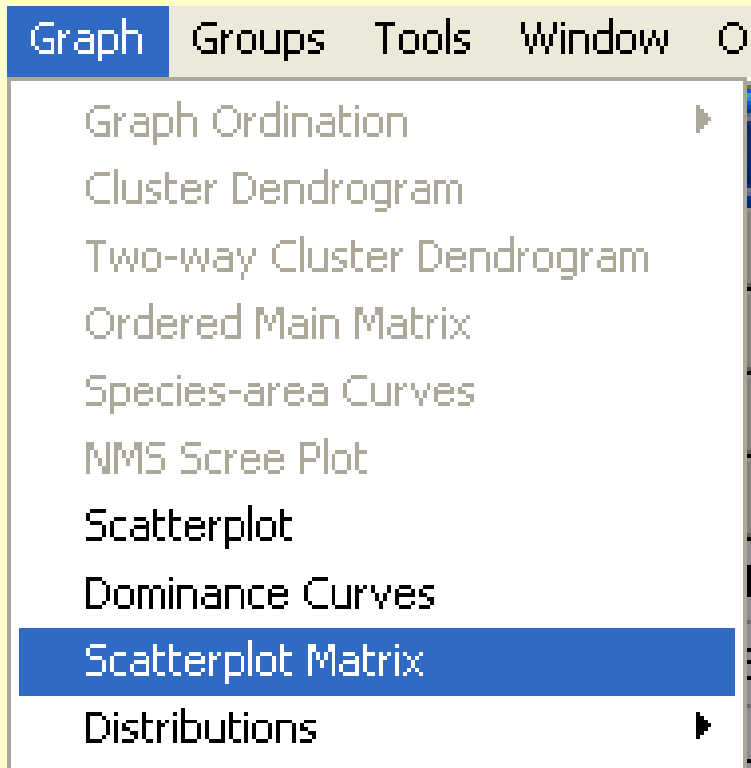
Dataset

- Data file: PCA1M.wk1 (main matrix)

```
96 samples
5 variables
      Q      Q      Q      Q      Q
      time  MEI  PDO  up_36  up_39
1997jan  1997 -0.47  0.23  -21  -0.27
```

- 96 samples and 5 variables
- Samples are monthly values (Jan. 97 - Dec. 04)
- Variables:
 - Time: decimal year
 - MEI: El Niño Multivariate Index (positive: warm, negative: cold)
 - PDO: Pacific Decadal Oscillation (positive: warm, negative: cold)
 - Up36: upwelling at 36 N (positive: upwelling, negative: downwelling)
 - Up39: upwelling at 39 N (positive: upwelling, negative: downwelling)

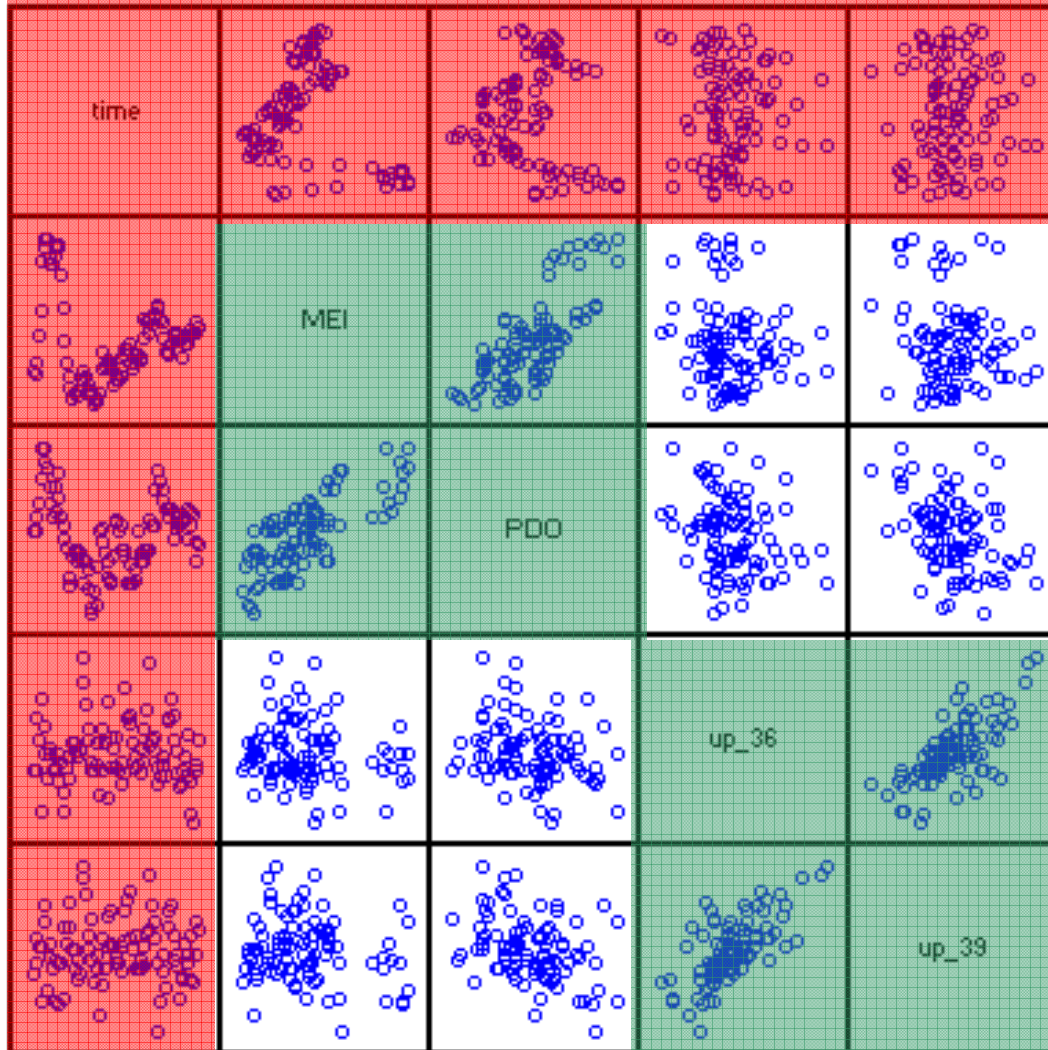
Data Exploration



Use scatterplot matrix to make a plot of all possible pair-wise combinations of the 5 environmental variables

Data Exploration – Correlograms

correlations



➤ Time Trends ?

➤ Regional Indices
(PDO / MEI)

➤ Local Indices
(up36 / up39)

Data Exploration – Advisor

Main matrix

% zeros	0.8
Average distance - Rela.Eucl.	0.05615
Lowest nonzero value	-162.000
Highest value	2004.917

Contents:	Rows	Columns
	96 samples	5 variable
Skewness		
Average	2.2	0.4
Maximum	2.2	0.9
Minimum	2.1	0.0
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-

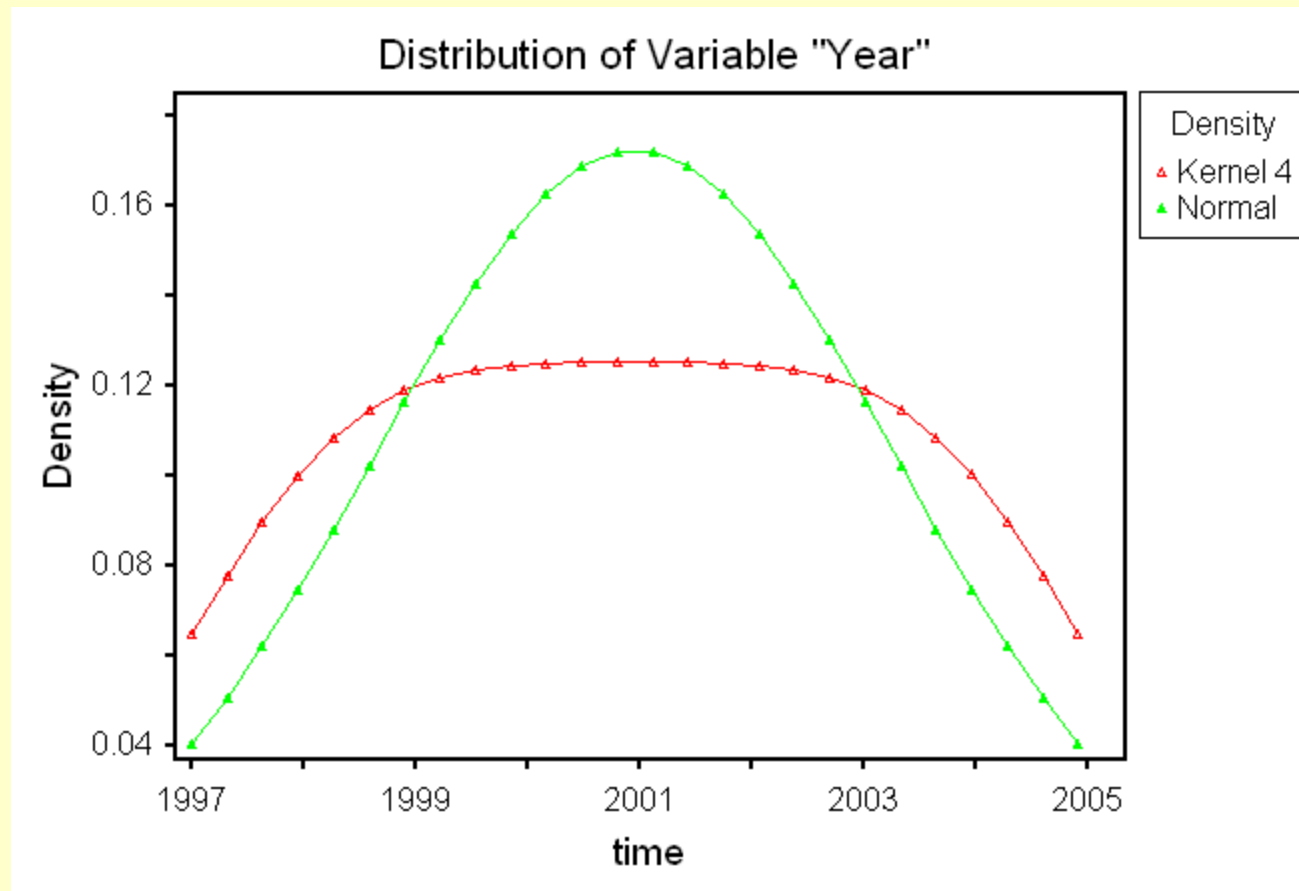
➤ Rows -
Skewed

➤ Columns -
Not Skewed

➤ Outliers: Samples

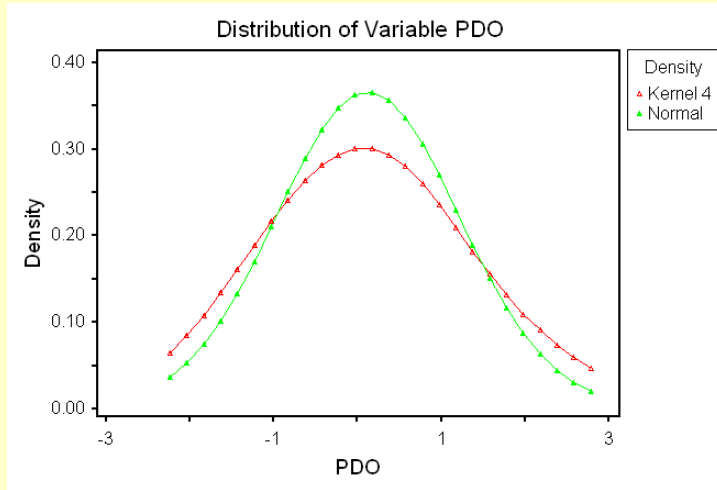
**Look out for these
in the plot results**

Data Exploration – Variable Year

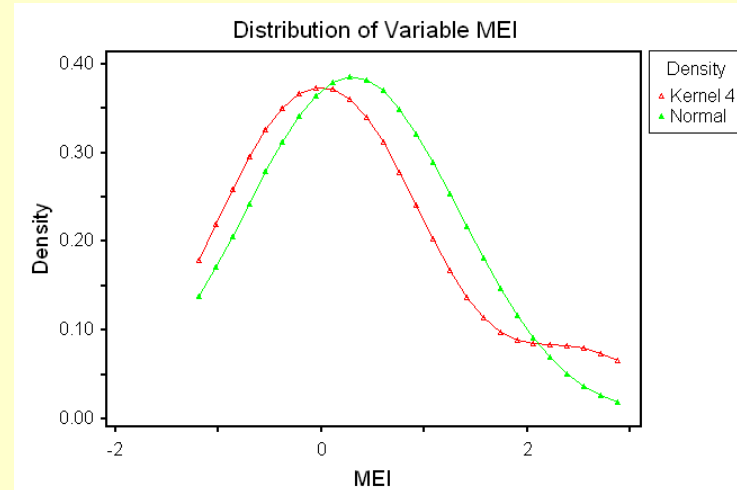


- 0.2330 E-07 = skewness

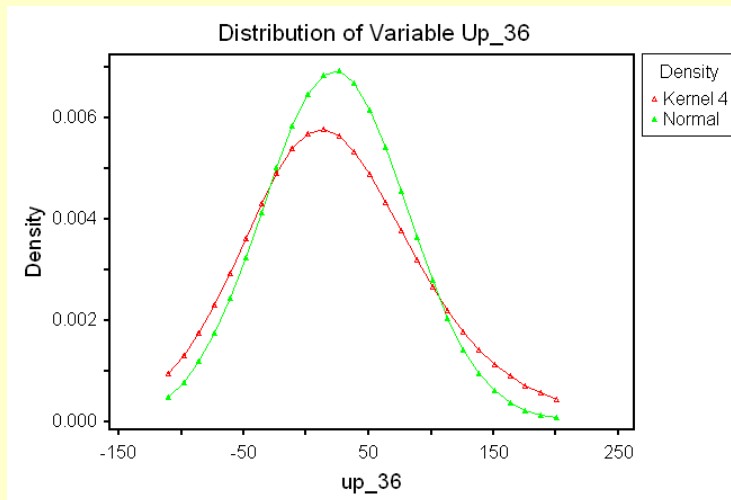
Data Exploration – Skewness



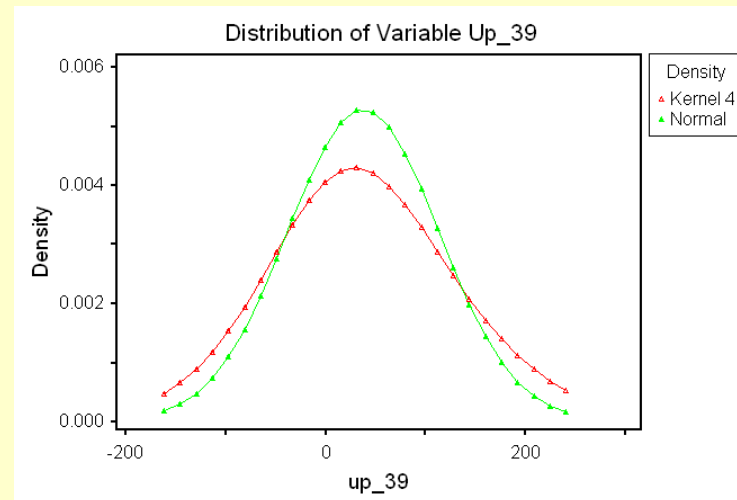
0.22 = skewness



0.94 = skewness



0.60 = skewness



0.24 = skewness

Statistical Results – With Year

VARIANCE EXTRACTED, FIRST 5 AXES

AXIS	Eigenvalue	% of Variance	Cum.% of Var.	Broken-stick Eigenvalue
1	740731.375	85.936	85.936	393628.812
2	120521.281	13.982	99.918	221236.641
3	511.527	0.059	99.977	135040.547
4	172.044	0.020	99.997	77576.484
5	24.721	0.003	100.000	34478.437

RANDOMIZATION RESULTS

99 = number of randomizations

Axis	Eigenvalue from	Eigenvalues from randomizations			
	real data	Minimum	Average	Maximum	p *
1	0.74073E+06	0.54505E+06	0.55140E+06	0.57766E+06	0.010000
2	0.12052E+06	0.28359E+06	0.30985E+06	0.31621E+06	1.000000
3	511.53	446.67	502.24	519.74	0.230000
4	172.04	98.667	114.72	132.27	0.010000
5	24.721	78.014	92.908	101.46	1.000000

* p-value for an axis is $(n+1)/(N+1)$, where n is the number of randomizations with an eigenvalue for that axis that is equal to or larger than the observed eigenvalue for that axis. N is the total number of randomizations.

Statistical Results – With Year

➤ Important Axes:

Eigenvalue: 1,2,3,4 Broken-stick: 1 P-values: 1,4

➤ Interpretation: **Loadings > 0.5 highlighted (arbitrary)**

FIRST 5 EIGENVECTORS, scaled to unit length.
These can be used as coordinates in a distance-based biplot,
where the distances among objects approximate their Euclidean
distances.

	Eigenvector				
variable	1	2	3	4	5
time	0.0000	0.0046	0.9970	0.0702	0.0310
MEI	-0.0019	-0.0001	-0.0714	0.6997	0.7108
PDO	-0.0041	-0.0030	-0.0281	0.7109	-0.7027
up_36	0.5617	-0.8273	0.0038	0.0009	0.0009
up_39	0.8273	0.5617	-0.0028	0.0046	-0.0025

up36 / up39

Together

up36 / up39

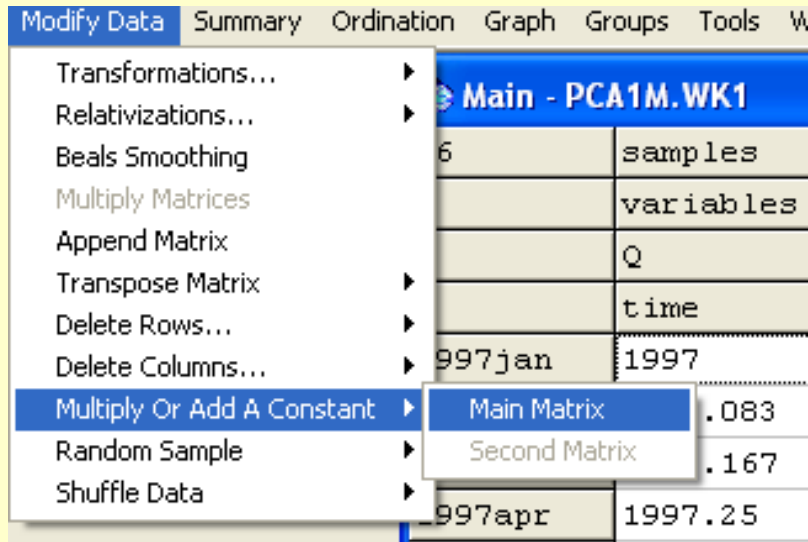
Opposite

Time

MEI / PDO

Together

Data Transformation – Time since Start



➤ Transformation:

- Subtract 1970 (first year sample)
- Recode as “Time Since Start”

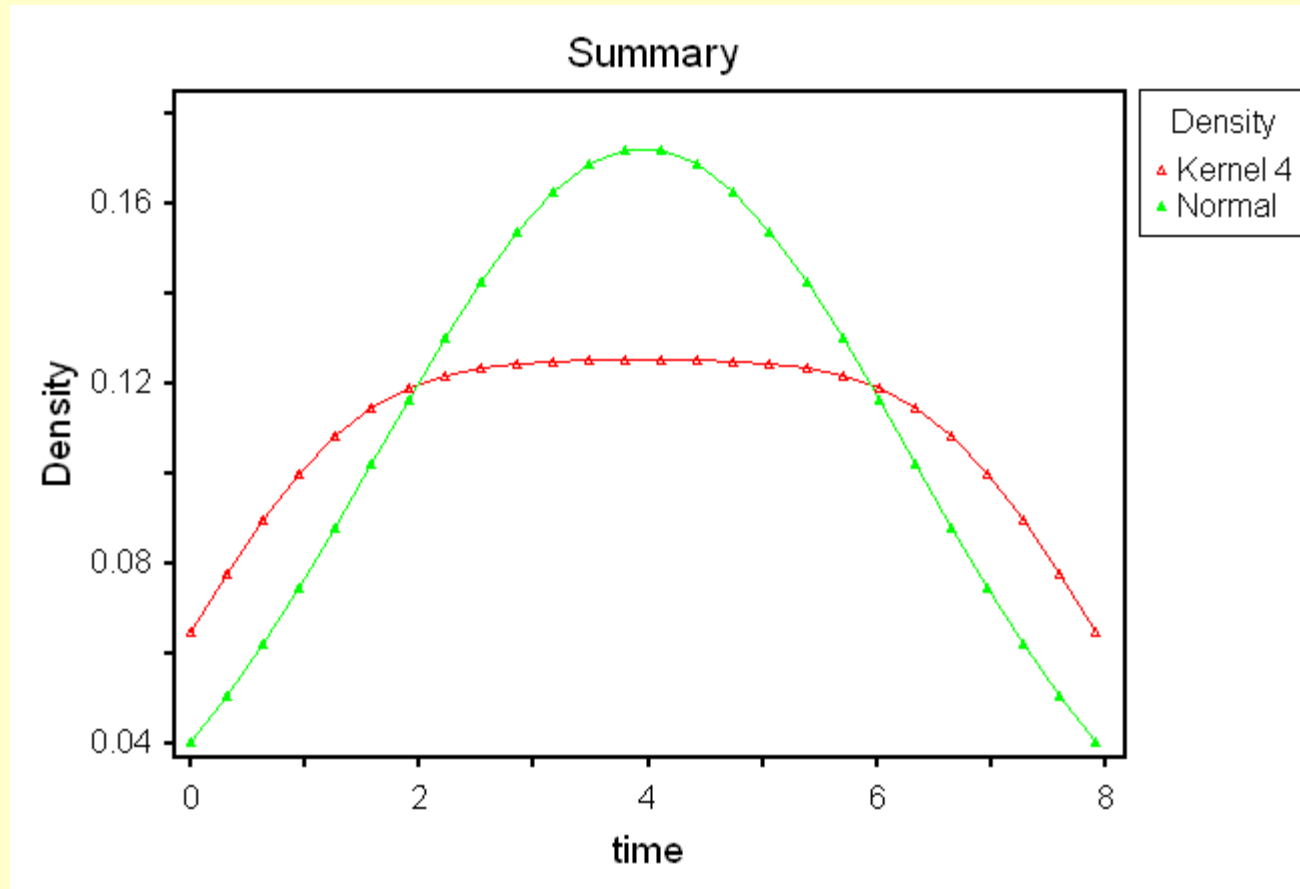
➤ Similar skewness

➤ No more outliers

```
-----
                                Main matrix
-----
% zeros                               1.0
Average distance - Rela.Eucl.        1.14712
Lowest nonzero value                  -162.000
Highest value                          240.000
-----
Contents:
Skewness
  Average                               0.5
  Maximum                               2.2
  Minimum                              -2.2
CV of totals, %                       191.9
-----
Potential Outliers
Distance measure:  Rela.Eucl.
                  SD-Item      SD-Item
None found.
-----
```

	Rows	Columns
Contents:	96 samples	5 variable
Skewness		
Average	0.5	0.4
Maximum	2.2	0.9
Minimum	-2.2	0.0
CV of totals, %	191.9	128.4

Data Exploration – Skewness



- 0.2330 E-07 = skewness

Statistical Results – With Time

VARIANCE EXTRACTED, FIRST 5 AXES

AXIS	Eigenvalue	% of Variance	Cum.% of Var.	Broken-stick Eigenvalue
1	740731.375	85.936	85.936	393628.812
2	120521.281	13.982	99.918	221236.641
3	511.527	0.059	99.977	135040.547
4	172.044	0.020	99.997	77576.484
5	24.721	0.003	100.000	34478.437

RANDOMIZATION RESULTS

99 = number of randomizations

Axis	Eigenvalue from	Eigenvalues from randomizations			
	real data	Minimum	Average	Maximum	p *
1	0.74073E+06	0.54505E+06	0.55140E+06	0.57766E+06	0.010000
2	0.12052E+06	0.28359E+06	0.30985E+06	0.31621E+06	1.000000
3	511.53	446.67	502.24	519.74	0.230000
4	172.04	98.667	114.72	132.27	0.010000
5	24.721	78.014	92.908	101.46	1.000000

* p-value for an axis is $(n+1)/(N+1)$, where n is the number of randomizations with an eigenvalue for that axis that is equal to or larger than the observed eigenvalue for that axis. N is the total number of randomizations.

Statistical Results – With Time

➤ Important Axes:

Eigenvalue: 1,2,3,4 Broken-stick: 1 P-values: 1,4

➤ Interpretation: **Loadings > 0.5 highlighted (arbitrary)**

FIRST 5 EIGENVECTORS, scaled to unit length.
 These can be used as coordinates in a distance-based biplot,
 where the distances among objects approximate their Euclidean
 distances.

	Eigenvector				
variable	1	2	3	4	5
time	0.0000	0.0046	0.9970	0.0702	0.0310
MEI	-0.0019	-0.0001	-0.0714	0.6997	0.7108
PDO	-0.0041	-0.0030	-0.0281	0.7109	-0.7027
up_36	0.5617	-0.8273	0.0038	0.0009	0.0009
up_39	0.8273	0.5617	-0.0028	0.0046	-0.0025

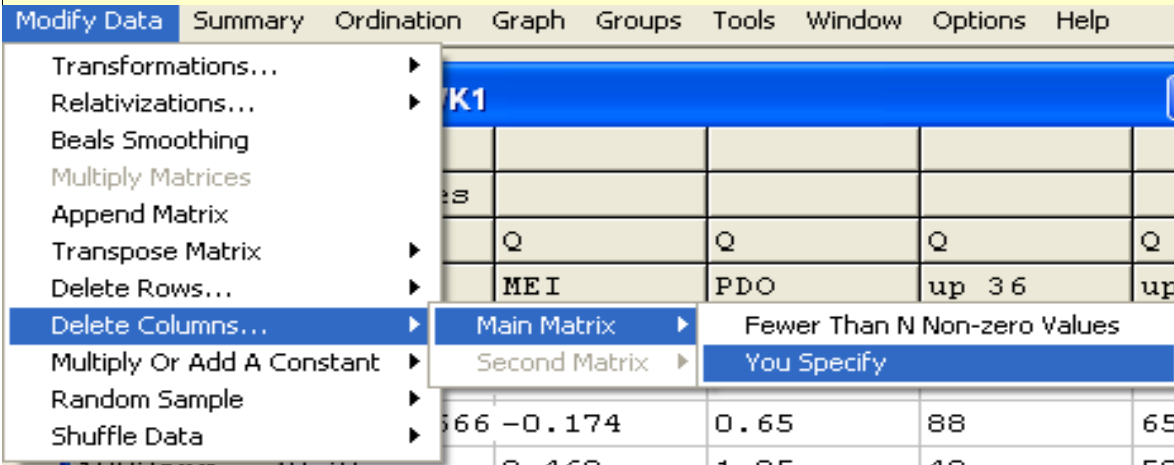
up36 / up39
Together

up36 / up39
Opposite

Time

MEI / PDO
Together

Data Transformation – Remove Time



➤ Remove Column:

- No Time

➤ Less Skewness
(for rows)

➤ Still No Outliers

```

-----
                                Main matrix
-----
% zeros                               1.0
Average distance - Rela.Eucl.         1.14816
Lowest nonzero value                   -162.000
Highest value                           240.000
-----
Contents:                               Rows          Columns
                                           96 samples    4 variable
Skewness
  Average                               0.3           0.5
  Maximum                               2.0           0.9
  Minimum                               -2.0          0.2
CV of totals, %                         204.6         120.6
-----
Potential Outliers
  Distance measure:   Rela.Eucl.
                     SD-Item      SD-Item
None found.
-----
    
```

Statistical Results – Remove Time

VARIANCE EXTRACTED, FIRST 4 AXES

AXIS	Eigenvalue	% of Variance	Cum.% of Var.	Broken-stick Eigenvalue
1	740731.375	85.987	85.987	448671.375
2	120518.727	13.990	99.977	233309.109
3	173.720	0.020	99.997	125627.984
4	25.184	0.003	100.000	53840.566

RANDOMIZATION RESULTS

99 = number of randomizations

Axis	Eigenvalue from	Eigenvalues from randomizations			p *
	real data	Minimum	Average	Maximum	
1	0.74073E+06	0.54505E+06	0.55205E+06	0.59287E+06	0.010000
2	0.12052E+06	0.26837E+06	0.30919E+06	0.31619E+06	1.000000
3	173.72	103.67	116.26	130.66	0.010000
4	25.184	79.994	94.932	101.95	1.000000

* p-value for an axis is $(n+1)/(N+1)$, where n is the number of randomizations with an eigenvalue for that axis that is equal to or larger than the observed eigenvalue for that axis. N is the total number of randomizations.

Statistical Results – Without Time

➤ Important Axes:

Eigenvalue: 1,2,3 Broken-stick: 1 P-values: 1,3

➤ Interpretation: **Loadings > 0.5 highlighted (arbitrary)**

FIRST 4 EIGENVECTORS, scaled to unit length.

These can be used as coordinates in a distance-based biplot, where the distances among objects approximate their Euclidean distances.

	Eigenvector			
variable	1	2	3	4
MEI	-0.0019	-0.0001	-0.7066	0.7077
PDO	-0.0041	-0.0030	-0.7076	-0.7066
up_36	0.5617	-0.8274	-0.0006	0.0008
up_39	0.8273	0.5617	-0.0047	-0.0025

up36 / up39

up36 / up39

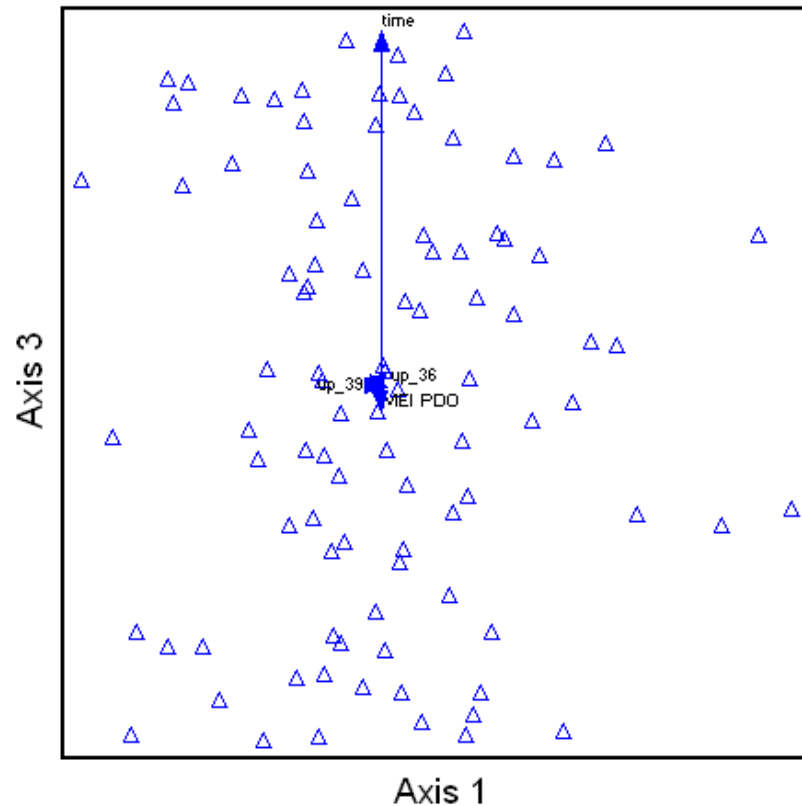
MEI / PDO

Together

Opposite

Together

Data Exploration – With Time



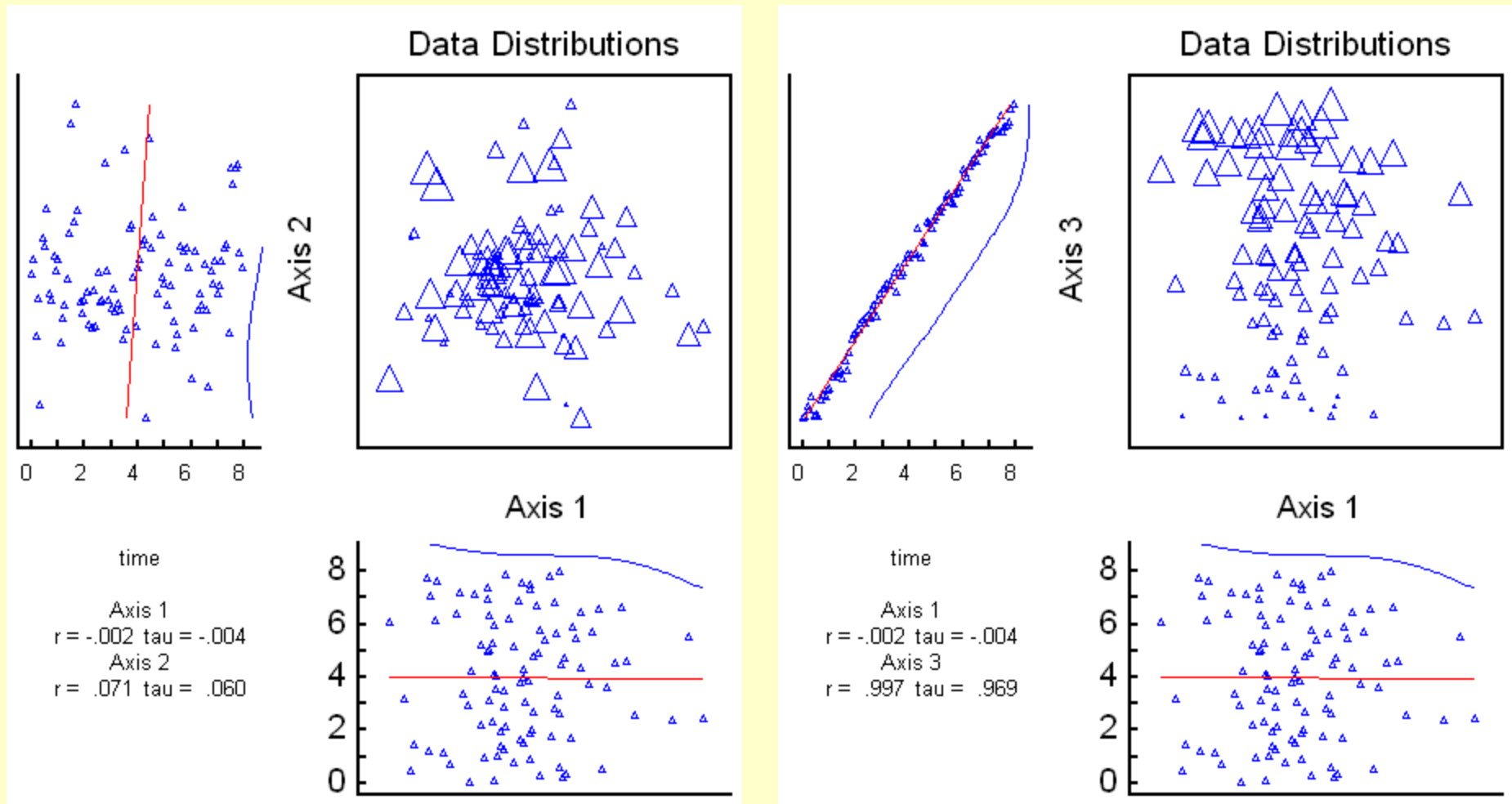
variable	Eigenvector	
	1	3
time	0.0000	0.9970
MEI	-0.0019	-0.0714
PDO	-0.0041	-0.0281
up_36	0.5617	0.0038
up_39	0.8273	-0.0028

Increment and cumulative R-squared were adjusted for any lack of orthogonality of axes.

Axis pair	r	Orthogonality,% = 100(1-r ²)
1 vs 2	0.000	100.0
1 vs 3	0.000	100.0
2 vs 3	0.000	100.0

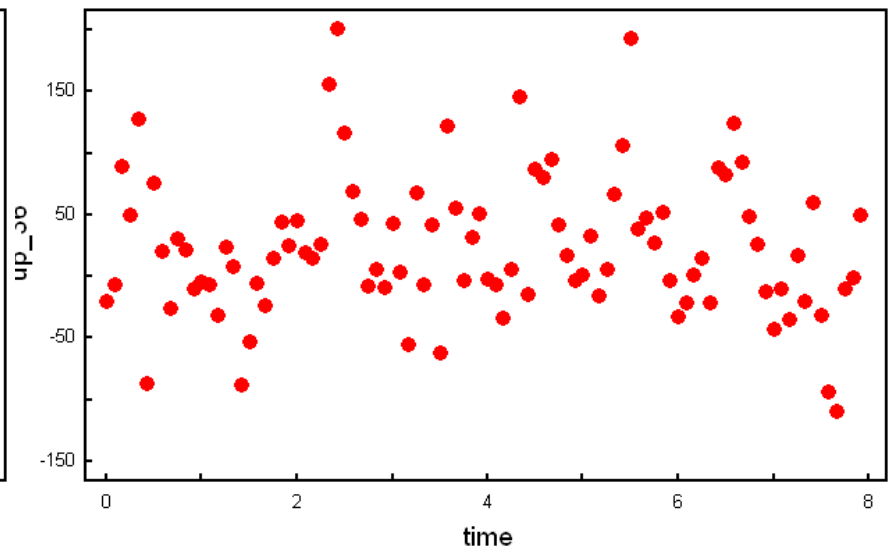
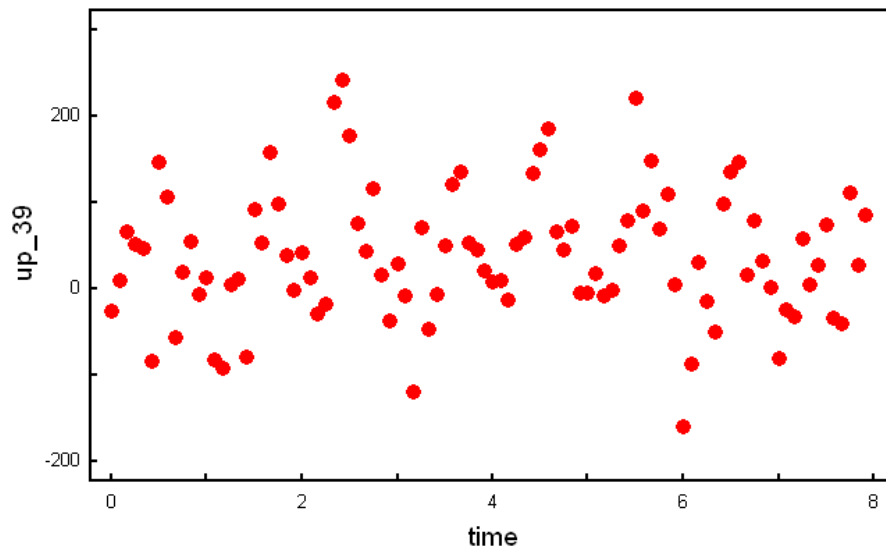
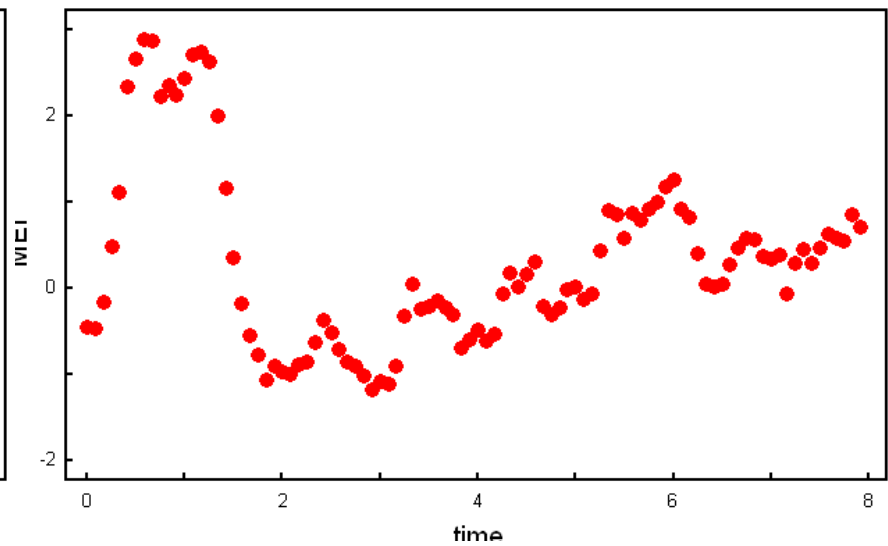
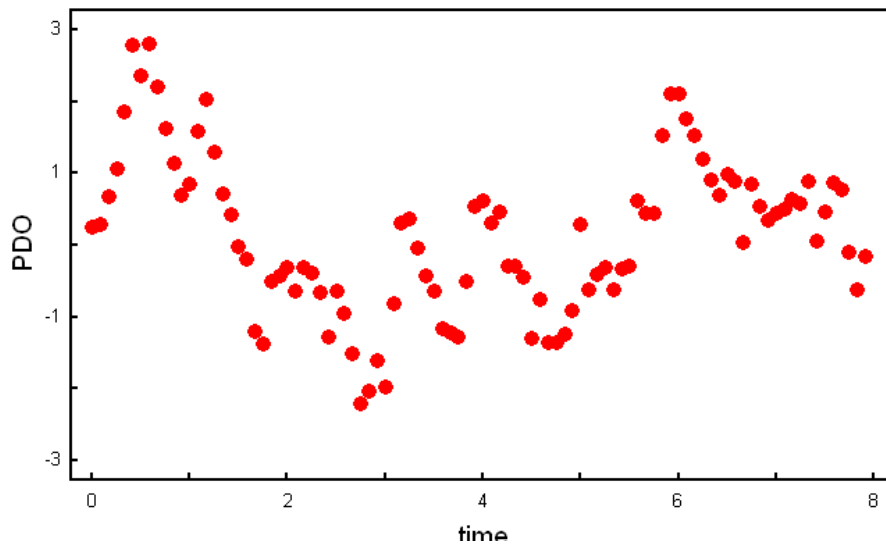
➤ **Independent
(orthogonal) variables**

Data Exploration – Time

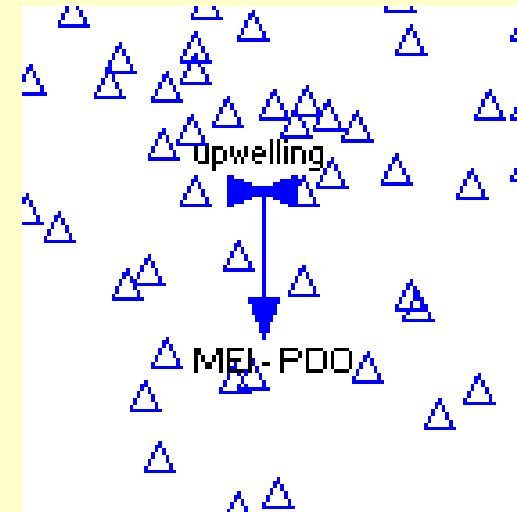
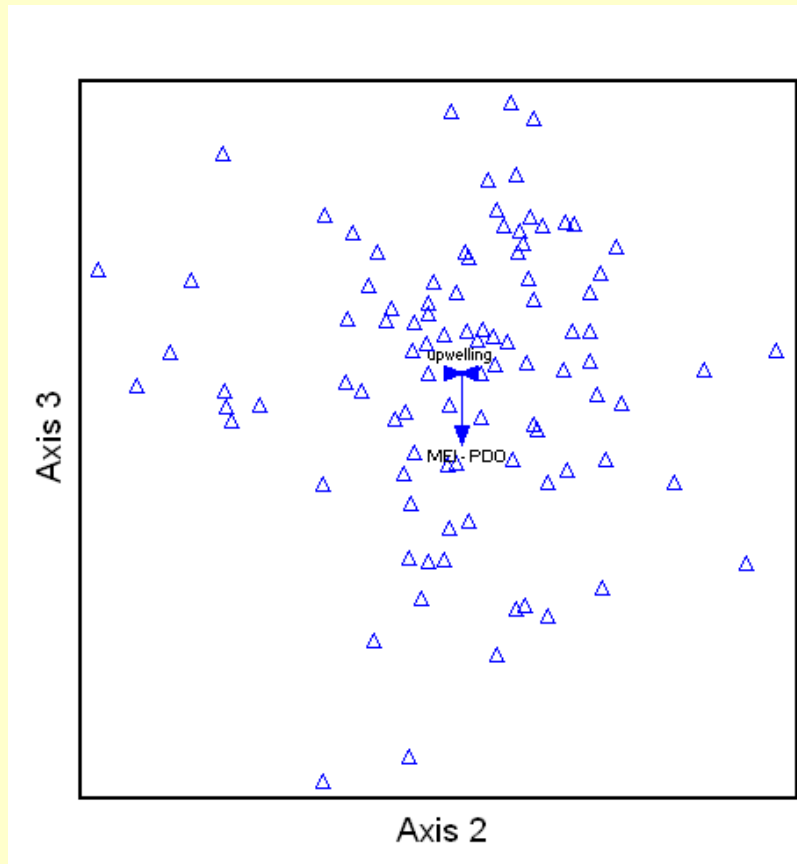


- No correlation with axis 1 or 2
- Positive correlation with axis 3

Data Exploration – Time



Data Exploration – Without Time



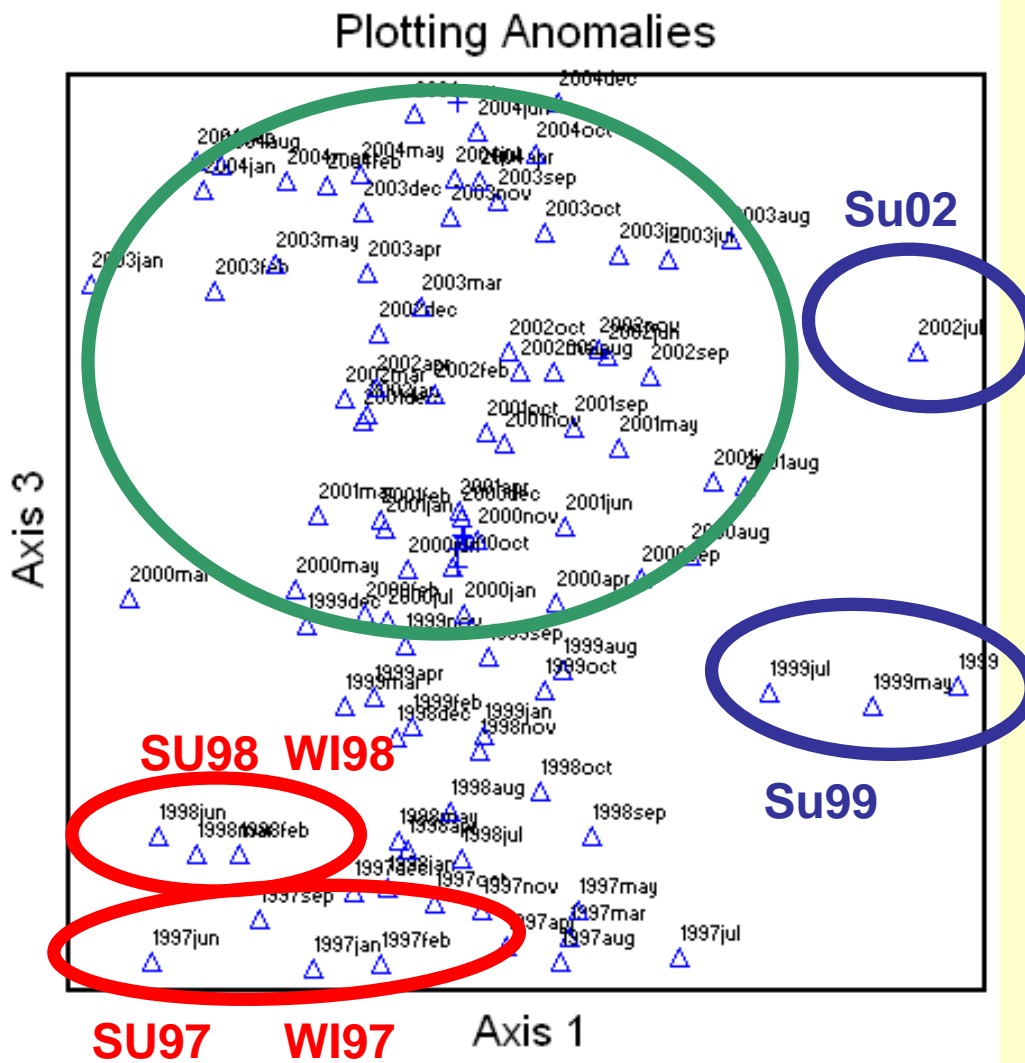
variable	Eigenvector	
	2	3
MEI	-0.0001	-0.7066
PDO	-0.0030	-0.7076
up_36	-0.8274	-0.0006
up_39	0.5617	-0.0047

Increment and cumulative R-squared were adjusted for any lack of orthogonality of axes.

Axis pair	r	Orthogonality,% = 100(1-r ²)
1 vs 2	0.036	99.9
1 vs 3	0.000	100.0
2 vs 3	0.000	100.0

➤ **Independent (orthogonal) variables**

Data Exploration – With Time



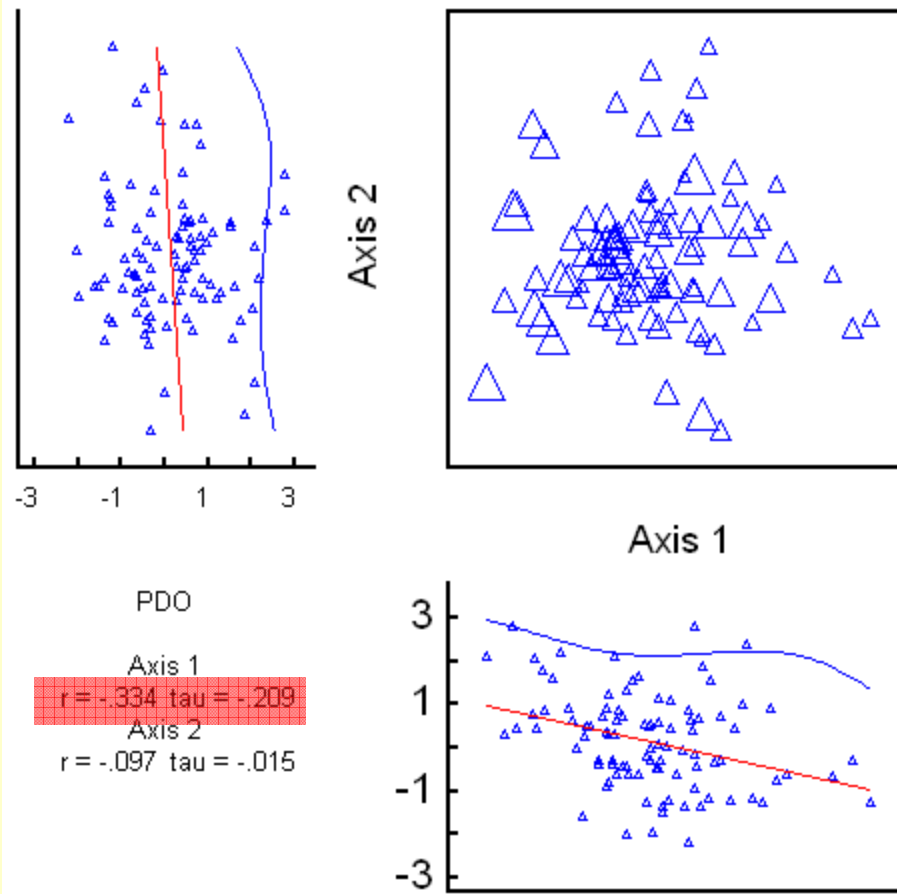
➤ Axis 1:

- Big: More Upwelling
- Small: Less Upwelling

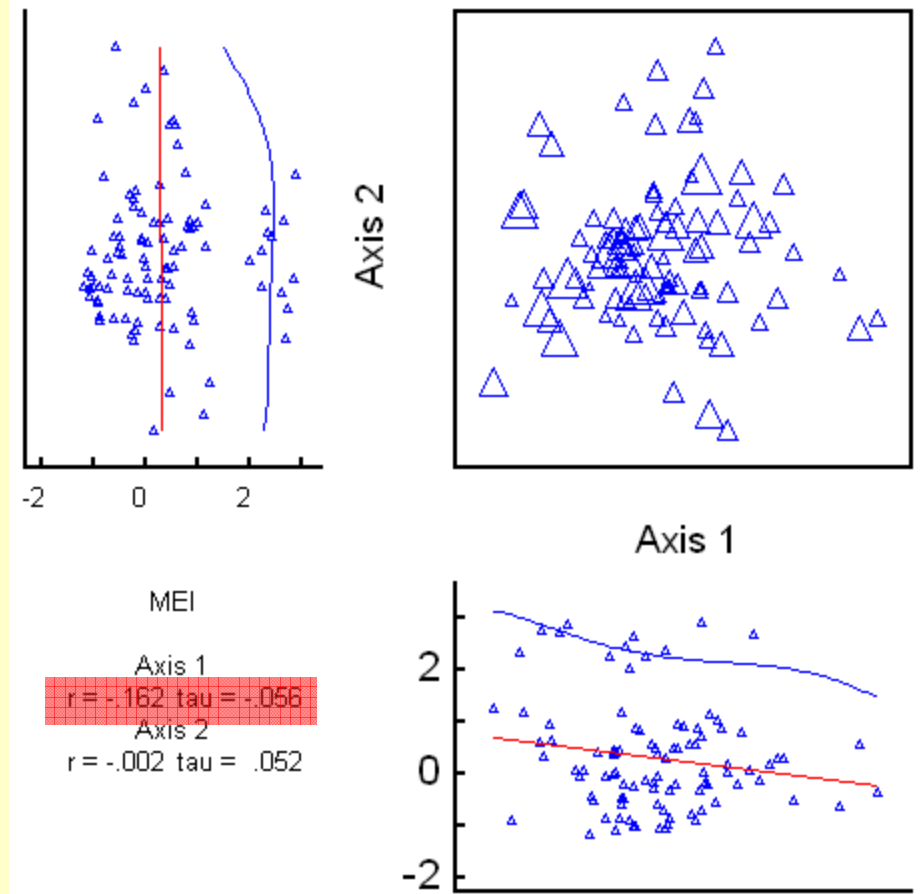
➤ Axis 3:

- Small: Warm
- Big: Cool

Data Exploration – Without Time

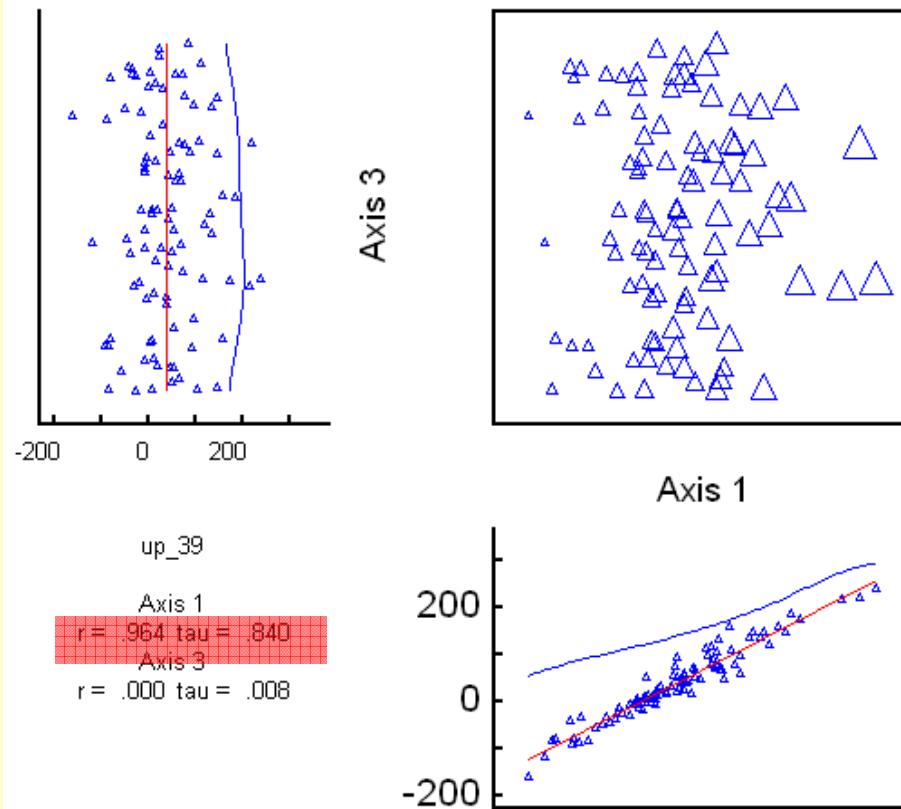


PDO – axis 1

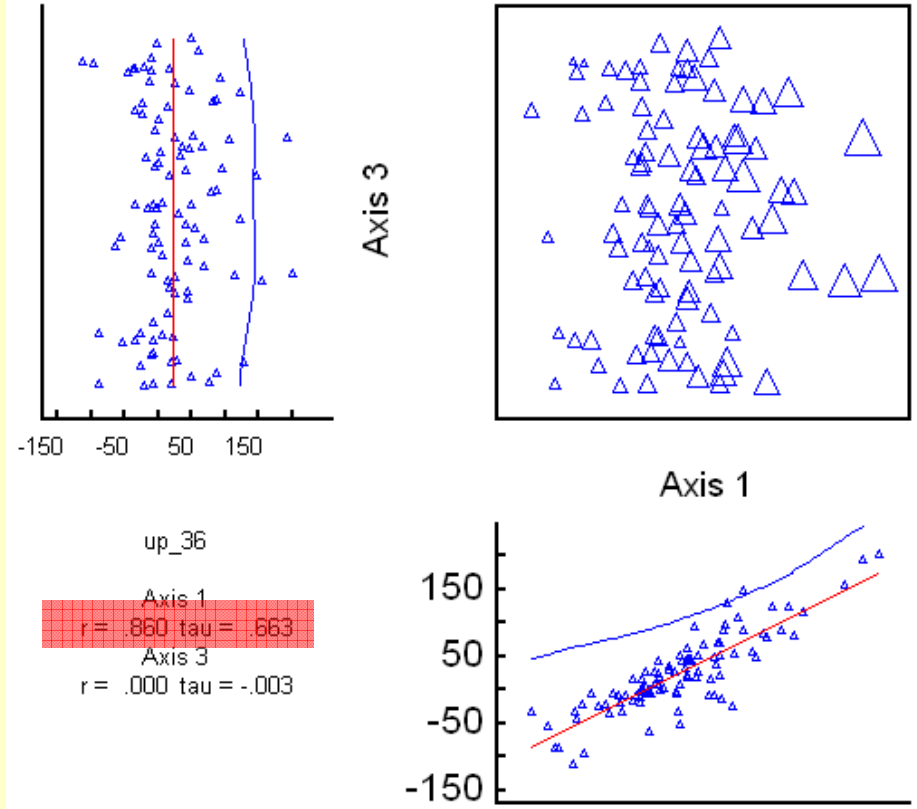


MEI – axis 1

Data Exploration – Without Time



Upwelling 39 – axis 1



Upwelling 36 – axis 1

Conclusions

Number of eigenvalues = Number of variables

Eigenvalues loadings did not change

- even after transforming YEAR data

Broken-stick results did not vary: YEAR / TIME

Randomization results did vary: YEAR / TIME

Removing time (linear variable)

- one less eigenvalue
- highlighted upwelling / PDO / MEI influence