Multivariate Statistics (**MARS6300**) - Homework 3 **Name:** _______KEY________________

Distributed: Wednesday, February 26 , 2018 **Due:** Friday, March 2, 2018

Instructions: Copy and paste your answers below and turn in a word file and two excel files by the end of due day via email to khyrenba@gmail.com. Please use email title “MARS 6300 hw#3” and label all files with a suffix including your name (e.g., MARS6300_hw3_hyrenbach). Unlabeled emails / files will be penalized 10% of points.

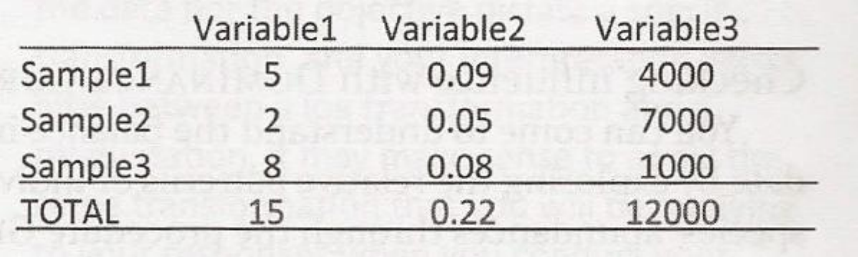
You are free to use any reference materials of your choice. While you are encouraged to work together, make sure you turn your own assignment. This homework is worth 5 points.

The objectives of this homework are:
A) To review and practice data relativizations.
B) To investigate the effects of relativizations on PCA.
C) To perform and interpret PCA analyses.
D) To critically evaluate the PCA literature.
E) To evaluate the reporting of PCA analyses.

To complete this homework, you will need:

- Instruction file: “MARS6300_hw3.doc” (open with word file) – turn in
- “upwell.xls” data file: (open with excel) – do not turn in

1) **Relativizations:**
You have three samples with three variable measurements.
Use the column summary tool in PC_ORD and calculate column totals shown below:



	Variable1	Variable2	Variable3
Sample1	5	0.09	4000
Sample2	2	0.05	7000
Sample3	8	0.08	1000
TOTAL	15	0.22	12000

The column totals were found to be the same values as pasted in the table here.

Because you want each variable to have an equivalent weight in the analysis, you will relativize the data on the basis of the columns.

First, relativize the data by each column maxima – fill in the cells below:

	Variable 1	**Variable 2**	**Variable 3**
Sample 1	0.625	1	0.5714286
Sample 2	0.25	0.5555555	1
Sample 3	1	0.8888888	0.1428571
TOTAL	1.875	2.4444443	1.714286

Did this relativization work well? Why / why not? – Explain:
This relativization did not seem to work well as it is not a monotonic relativization. Each variable DOES NOT have an equivalent weight in the analysis. Before the relativization, the total of the Variable 3 column was by far the largest value of the 3 column totals. After the relativization, it has become the smallest value. And the total of the column of Variable 2 has become the largest value, which was initially the smallest. The same problem occurs when comparing values along rows (Samples 1-3) – the monotonic rank of individual values along rows changes. Additionally, now that Variable 2 has the largest total sum, it may have the largest influence on any patterns, even though it initially had the smallest values before relativization, by far.

This type of relativization should be used when one of the values within the column represents the potential maximum abundance for the column. In this example, this may not have been the case, which is why this relativization did not work.

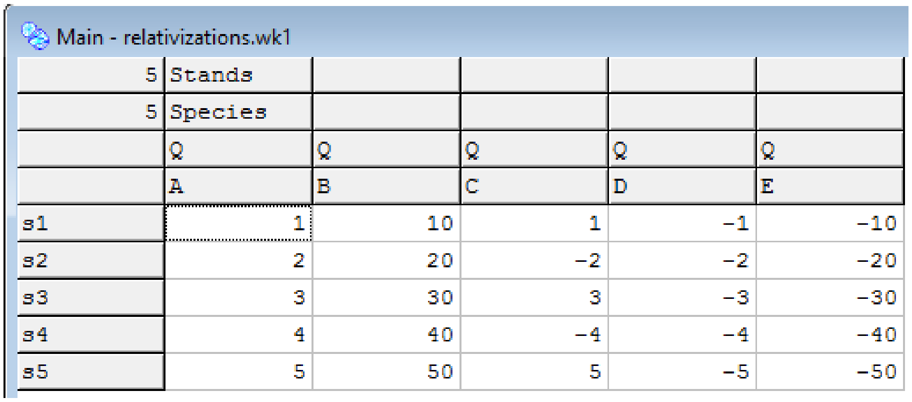
Next, relativize the data by each column totals (in PC-ORD, this is the general relativization approach, with p = 1) – fill in the cells below:

	Variable 1	**Variable 2**	**Variable 3**
Sample 1	0.3333333	0.4090909	0.3333333
Sample 2	0.1333333	0.2272727	0.5833333
Sample 3	0.5333334	0.3636363	0.0833333
TOTAL	1.0000000	1.0000000	1.0000000

Did this relativization work well? Why / why not? – Explain:
This type of relativization is used if you wish to focus on the relative contribution of a response compared to the total abundance within a sample unit. Basically, this method gives every sample the same weight. The influence of variables with high total abundance is reduced as the observations have become proportional to the intra-response total abundance. The key is that the variation in abundances across sample units is maintained, but the influence of rare species (Variable 2) is enhanced, and the influence of common species (Variable 3) is reduced. Now that all 3 columns have equivalent weight in analysis, **this relativization did seem to work well**!

C) Load the file "relativizations.wk1" into PC-ORD and perform the following relativizations by columns. For each instance, report two pieces of information:

This is the raw data:



Main - relativizations.wk1

5	Stands				
5	Species				
	Q	Q	Q	Q	Q
	A	B	C	D	E
s1	1	10	1	-1	-10
s2	2	20	-2	-2	-20
s3	3	30	3	-3	-30
s4	4	40	-4	-4	-40
s5	5	50	5	-5	-50

i) paste resulting matrix. (Note: you can copy and paste the data from the matrix – select all and copy – or you can get a screen shot. If you do the latter, please crop image before pasting)

ii) summarize the rows / columns using PC-ORD tools and report the sums of columns / rows. (Note: copy table from "results.txt" file and paste)

- Relativize by adjusting by the mean: What did PC-ORD do?

i) Resulting Matrix:

5	Stands				
5	Species				
	Q	Q	Q	Q	Q
	A	B	C	D	E
s1	-2	-20	0.4	2	20
s2	-1	-10	-2.6	1	10
s3	0	0	2.4	0	0
s4	1	10	-4.6	-1	-10
s5	2	20	4.4	-2	-20

In order to relativize these data by the mean, PC-ORD found the mean value for each column, and subtracted that value from each individual value. The above matrix is the resulting matrix after relativization.

ii) Column Summary:

```
Summary of relativization by mean (COLUMNS)

             Summary of:     5 Species      N =      5 Stands
-----------------------------------------------------------------------------------------------
 Num.  Name         Mean      Stand.Dev.        Sum     Minimum     Maximum          S     E      H      D`
-----------------------------------------------------------------------------------------------
    1 A          0.000      1.581          0.000     -2.000      2.000          4  0.923  1.280 0.7000
    2 B          0.000      15.81          0.000     -20.00      20.00          4  0.923  1.280 0.7000
    3 C         0.9537E-07  3.647         0.4768E-06 -4.600      4.400          5  0.791  1.273 0.6994
    4 D          0.000      1.581          0.000     -2.000      2.000          4  0.923  1.280 0.7000
    5 E          0.000      15.81          0.000     -20.00      20.00          4  0.923  1.280 0.7000
-----------------------------------------------------------------------------------------------
   AVERAGES:   0.1907E-07  7.686         0.9537E-07 -9.720      9.680        4.2  0.897  1.279 0.6999
```

Sum of Column A = 0.000
Sum of Column B = 0.000
Sum of Column C = 0.0000004768 (practically = 0)
Sum of Column D = 0.000
Sum of Column E = 0.000

- Relativize by adjusting by rank: What did PC-ORD do?

i) Resulting Matrix:

5	Stands				
5	Species				
	Q	Q	Q	Q	Q
	A	B	C	D	E
s1	1	1	3	5	5
s2	2	2	2	4	4
s3	3	3	4	3	3
s4	4	4	1	2	2
s5	5	5	5	1	1

In order to relativize these data by their rank, PC-ORD changed the values within each cell to the cell's rank from lowest value (1) to highest value (5). The above matrix is the resulting matrix after relativization.

ii) Column Summary:

```
Summary of relativization by rank (Columns)

             Summary of:     5 Species      N =     5 Stands
----------------------------------------------------------------------------------------------------
 Num.  Name       Mean     Stand.Dev.       Sum    Minimum    Maximum        S     E     H     D`
----------------------------------------------------------------------------------------------------
    1 A             3.000      1.581       15.0000      1.000      5.000     5  0.926  1.490 0.7556
    2 B             3.000      1.581       15.0000      1.000      5.000     5  0.926  1.490 0.7556
    3 C             3.000      1.581       15.0000      1.000      5.000     5  0.926  1.490 0.7556
    4 D             3.000      1.581       15.0000      1.000      5.000     5  0.926  1.490 0.7556
    5 E             3.000      1.581       15.0000      1.000      5.000     5  0.926  1.490 0.7556
----------------------------------------------------------------------------------------------------
   AVERAGES:    3.000      1.581         15.00      1.000      5.000        5.0 0.926  1.490 0.7556
----------------------------------------------------------------------------------------------------
```

Sum of Column A = 15.0000
Sum of Column B = 15.0000
Sum of Column C = 15.0000
Sum of Column D = 15.0000
Sum of Column E = 15.0000

- Relativize by adjusting by standard deviate: What did PC-ORD do?

i) Resulting Matrix:

5	Stands				
5	Species				
	Q	Q	Q	Q	Q
	A	B	C	D	E
s1	-1.264911	-1.264911	0.1096817	1.264911	1.264911
s2	-0.6324555	-0.6324555	-0.712931	0.6324555	0.6324555
s3	0	0	0.6580902	0	0
s4	0.6324555	0.6324555	-1.261339	-0.6324555	-0.6324555

s5	1.264911	1.264911	1.206499	-1.264911	-1.264911

In order to relativize these data by their standard deviates, PC-ORD changed the values by subtracting the mean of the column from each value, then dividing the value by the standard deviation of the value away from that same mean. The above matrix is the resulting matrix after relativization.

ii) Column Summary:

```
Summary of relativization by standard deviate (Columns)

             Summary of:     5 Species      N =      5 Stands
----------------------------------------------------------------------------------------------------
 Num.  Name         Mean      Stand.Dev.        Sum     Minimum     Maximum         S     E      H     D`
----------------------------------------------------------------------------------------------------
    1 A         0.2384E-07  1.000        0.1192E-06 -1.265       1.265          4  0.923  1.280 0.7000
    2 B         0.2384E-07  1.000        0.1192E-06 -1.265       1.265          4  0.923  1.280 0.7000
    3 C          0.000      1.000         0.000     -1.261       1.206          5  0.791  1.273 0.6994
    4 D        -0.2384E-07  1.000       -0.1192E-06 -1.265       1.265          4  0.923  1.280 0.7000
    5 E        -0.2384E-07  1.000       -0.1192E-06 -1.265       1.265          4  0.923  1.280 0.7000
----------------------------------------------------------------------------------------------------
   AVERAGES:     0.000      1.000         0.000     -1.264       1.253        4.2 0.897  1.279 0.6999
----------------------------------------------------------------------------------------------------
```

Sum of Column A = 0.0000001192 (practically 0)
Sum of Column B = 0.0000001192 (practically 0)
Sum of Column C = 0.000
Sum of Column D = -0.0000001192 (practically 0)
Sum of Column E = -0.0000001192 (practically 0)

2) PCA Analysis – Interpretation:

Load the upwell_PCA.xls file and check the "summary of the column data (the variables)".

Report the skewness for the five variables:

```
                                Skewness
-----------------------------------
    1 time            _0.000___
    2 MEI             _-0.460__
    3 PDO             _-0.140__
    4 upwell36        _0.846___
    5 upwell39        _-0.021__
```

Report the percent of "empty" cells: __0.333_____

Based on these results, would you proceed with the PCA analysis? Why / Why not?

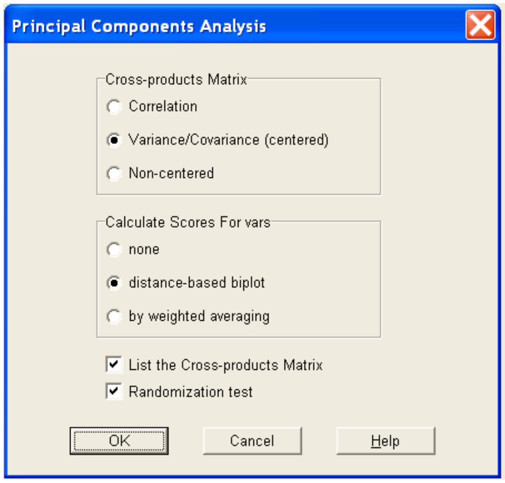
One of the assumptions of PCA is that there are few zeroes within the dataset. As only 0.333% of cells are "empty" or contain zeroes (which is less than 20%, which is the rule), this initial condition is met for proceeding with PCA analysis. It is difficult to use PCA with zero-rich datasets as they highly violate normality conditions and linearity assumptions (like they have high skewness). As the skewness values are all between -1 and 1, it means that the data is highly NOT skewed, and therefore this second condition is met for proceeding with PCA analysis. Therefore, **I WOULD proceed with PCA analysis.**

Also report the “sum” of each of these five variables:

```
                                Sums
----------------------------------
    1  time           _479040.0000________
    2  MEI            _-115.6________
    3  PDO            _-76.01________
    4  upwell36       _1289.0000________
    5  upwell39       _5076.0000________
```

On the basis of these totals, which variable do you think will have the largest effect determining the Euclidean distances between samples? Explain why:

It seems that the variable of “time” will have the largest effect on determining Euclidian distances between samples. This is because the values in this column (and the column total) are so much larger than any of the other values that this column will weigh heavily on determining any distance measures between variables. It seems likely that the data must be relativized before any type of PCA can be started.



Principal Components Analysis

Cross-products Matrix
○ Correlation
● Variance/Covariance (centered)
○ Non-centered

Calculate Scores For vars
○ none
● distance-based biplot
○ by weighted averaging

☑ List the Cross-products Matrix
☑ Randomization test

Perform a PCA using the recommended settings presented in lecture:

Note:
Set up the randomization tests for 999 runs
Save the “results.txt” file and look for the following results:

```
     -  Find the amount of variance explained
        by the first 5 axes:
```

```
VARIANCE EXTRACTED, FIRST  5 AXES
---------------------------------------------------------------
                                                Broken-stick
AXIS    Eigenvalue   % of Variance  Cum.% of Var.  Eigenvalue
---------------------------------------------------------------
  1      5458.660       81.808         81.808      3047.112
  2      1171.457       17.556         99.365      1712.610
  3        31.730        0.476         99.840      1045.359
  4         8.790        0.132         99.972       600.526
  5         1.872        0.028        100.000       266.900
---------------------------------------------------------------
```

Based on this table, which PCA axis explains more variability than would be expected by chance? Explain Why?

The PCA axis which explains more variability than would be expected by chance is Axis 1 only. This is because the Eigenvalue for Axis 1 is larger than the Broken-stick Eigenvalue (which is the eigenvalue produced by chance) for only this one column. Eigenvalues are what explain the variance, which is why we can look at these values for each axis to determine the amount of variability.

How come we explain 100% of the variance with 5 axes?

It is actually very rare that 100% of the variance is explained. For 100% of the variance to be explained, there would need to be as many axes as there are variables that are trying to explain the patterns. In this case, the Upwell_PCA document has 5 variables, and our results show us 5 axes, so therefore 100% of the variance is explained.

These are the "loadings of the variables" in the axes:

```
   -  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
vars               1           2           3           4           5
time          0.0171     -0.0134     -0.9987      0.0435      0.0121
MEI           0.0040      0.0100     -0.0353     -0.9175      0.3960
PDO           0.0007      0.0038     -0.0284     -0.3951     -0.9182
upwell36      0.5460     -0.8375      0.0203     -0.0077     -0.0004
upwell39      0.8376      0.5462      0.0074      0.0088     -0.0011
-----------------------------------------------------------------------------
```

Which are the two variables with the strongest loadings (coefficients) loading in axis 1. Interpret what this means (how are these variables influencing axis1?):
The 2 variables with the strongest loadings in axis 1 are upwell36 and upwell39. Each variable has a coefficient within each axis which contributes to the overall value of the axis. As all coefficients within axis 1 would be multiplied by each other to determine the value of the axis, the 2 variables with the largest coefficients would therefore have the largest influence on the axis. Therefore, since Upwell36 and Upwell39 have the largest values, any point that lies on this axis will be most influenced by these 2 variables.

Which are the two variables with the strongest loadings (coefficients) loading in axis 2. Interpret what this means (how are these variables influencing axis2?):
The 2 variables with the strongest loadings in axis 2 are also upwell36 and upwell39. Each variable has a coefficient within each axis which contributes to the overall value of the axis. As all coefficients within axis 2 would be multiplied by each other to determine the value of the axis, the 2 variables with the largest coefficients would therefore have the largest influence on the axis. In this case, even though upwell36 has a negative coefficient in axis 2, it still has a strong influence on the overall value of axis 2, just in the opposite direction as it did in Axis 1.

Based on this interpretation, where would you expect the sample from June 1999 (s99.458) - the maximum upwelling recorded during this time series - to fall on the plot of axis 1 and axis 2 (Hint: Sample lies on positive / negative end of axis 1? On positive / negative end of axis 2?) Explain:
It is likely that the sample from June 1999 (the maximum upwelling record) would fall all the way on the right of axis 1 (as both upwelling coefficient values are positive), and on axis 2, the value should fall on the negative end of axis 2, but not all the way at the negative side. This is because the 2 upwelling coefficients in the 2nd axis are -0.8375 and 0.5462. Therefore, the negative value has a slightly stronger loading than the positive value, and therefore the location of the point should be slightly negative for axis 2.

You can check the scores of this sample (#162 in the list of COORDINATES OF samples) in axis 1 and axis 2. The scores are: for axis 1: _289.6390_ and for axis 2: _-43.5085__

Check the randomization results below, and calculate the p values, for the two axes, using the equation: 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.

```
RANDOMIZATION RESULTS
            999 = number of randomizations
--------------------------------------------------------------------------
      Eigenvalue          Eigenvalues from randomizations
        from        ----------------------------------------
Axis  real data       Minimum        Average        Maximum         p *

   1    5458.7         4179.0         4201.9         4401.5      ____0.001____

   2    1171.5         2227.8         2426.6         2450.2      ____1.000____
--------------------------------------------------------------------------
```

For Axis 1, n = 0 because the eigenvalue from the real data (1304600) is larger than the maximum eignevalue from all randomizations (1073800). Therefore, there are no randomizations with an eigenvalue equal to or larger than that of the observed eigenvalue. N = 999 because there were 999 randomizations run. So, p = (0+1)/(999+1) = **0.001 = p for axis 1**
For Axis 2, n = 999 because every random eigenvalue (minimum value of 510350) is either larger than or equal to the observed eigenvalue (279980). N remains 999. So, p = (999+1)/(999+1) = **1.000 = p for axis 2**

Finally, use three "stopping rules", to determine how many PCA axes to use. For each method, listed below, enter in the spaces provided the number of axes (starting with 1) that meet the rule. For instance, if axis 1and axis 2 meet the criterion Rnd-Lambda, write down "2".

```
APPLICATION OF STOPPING RULES
--------------------------------------------------------------------------
Last useful     Rule
   axis        acronym     Explanation (see Peres-Neto, Jackson & Somers 2005)
____________________________________________________________________________

__1__       Rnd-Lambda  Observed eigenvalue as compared to randomization p values

__1__         Avg-Rnd    Observed eigenvalue compared to average eigenvalue
                          from randomizations

__1__           BS      Observed eigenvalue compared to broken-stick eigenvalue
```

To check for the outliers, you can use "summary > outlier analysis". Make sure you select Euclidean Distance as you distance measure. Ask for those outlier samples that are over 2 SD units away and request for all outputs (ranks / graph). Report the following information for the top two outliers:

```
-------------------------------------------
         ENTITY        AVERAGE     STANDARD
 RANK     NAME         DISTANCE   DEVIATIONS
-------------------------------------------
    1    s95.125      336.49670     6.28844
    2    s99.458      296.31989     5.24428
```

Finally, look at the graph using the “graph > graph ordination > 2d” menu. View “simple scatterplot” and look for these outliers? Label these two samples to identify them. Copy and paste the image here:

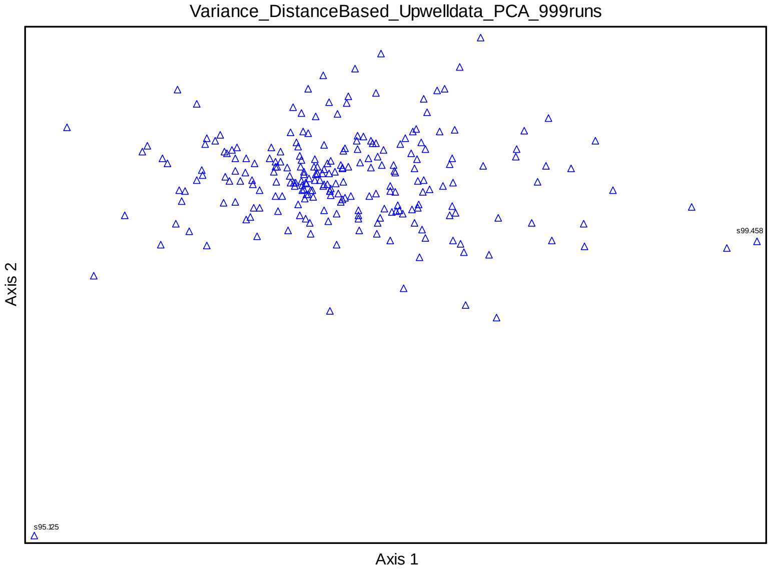


Figure 1. Simple scatterplot of PCA ordination graph of data from Upwell_PCA document. The 2 points which are labeled, s95.125 and s99.458, have the 2 largest outlier values.

View “main plot” and look for the correlations of “time” with both axes:
r (with axis 1): r = +0.218___
r (with axis 2): _r = -0.079___

Using “Statistics > Correlations with Main Matrix”, calculate the correlations of each variable with each axis. What is the correlation of “time” with axis 3? Report the r here: _-0.972__

Note that the third axis is not significant (based on the Eigenvector), but “time” is very strongly correlated with it. Can you explain this result? (Hint: is “time” correlated with any other four oceanographic variables – check correlation matrix scatterplots).

Time is not strongly correlated with either axis 1 or axis 2. In fact, time is not strongly correlated with any of the four oceanographic variables (see Figure 2 below). Therefore, “time” is its own factor, and is in fact has the strongest loading by far on axis 3 (value of -0.9987), and none of the other oceanographic variables have a strong loading with axis 3.

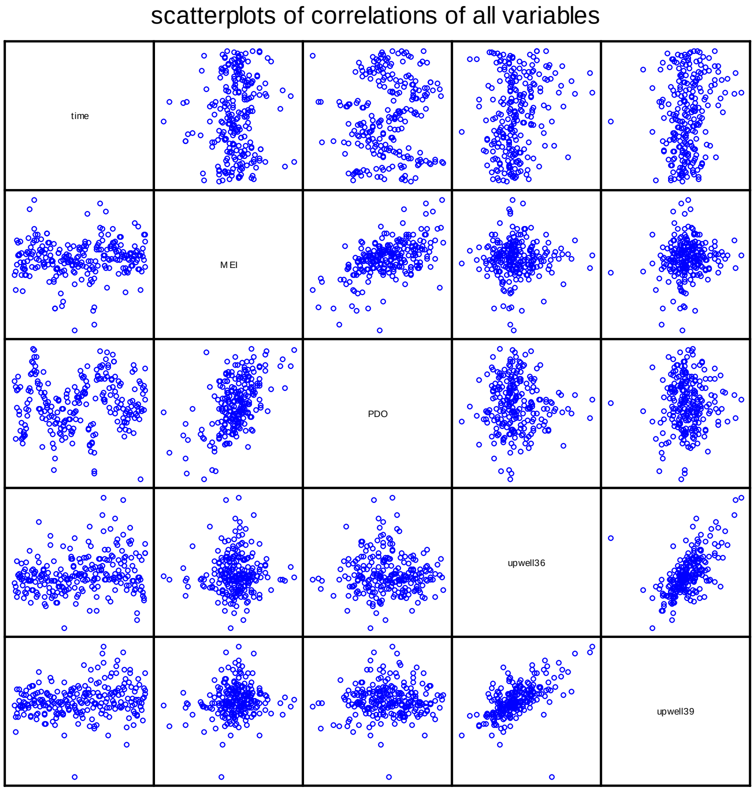


Figure 2. Correlations of all 5 oceanographic variables being considered in the PCA analysis. It appears as though the time is not correlated with any of the other oceanographic variables.

3) PCA Analysis – the influence of data relativization:

Load the “upwell_PCA.xls” file and perform a relativization by maximum.

Individual constants added to all columns to remove negative numbers prior to performing relativization by maximum.   Column 1 = no constant added

Column 2 = +12 (min value was -11.75)   Column 3 = +6 (min value was -5.429)

Column 4 = +135 (min value was -131)   Column 5 = +300 (min value was -296)

Report the missing values in the eigenvector table below:

```
      VARIANCE EXTRACTED, FIRST  5 AXES
---------------------------------------------------------------
                                                   Broken-stick
AXIS    Eigenvalue   % of Variance  Cum.% of Var.  Eigenvalue
---------------------------------------------------------------
  1         0.041          46.856         46.856         0.040
  2         0.030          33.583         80.439         0.023
  3         0.011          12.317         92.755         0.014
  4         0.006           7.236         99.991         0.008
  5         0.000           0.009        100.000         0.004
---------------------------------------------------------------
```

Why are the eigenvalues from this table much lower than the eigenvalues from the previous example? Explain.

The data have been transformed by adding constants to make each column all nonnegative values and then relativized to the maximum. This process limited the maximum values of the columns to one. This reduced the eigenvalues considerably, compared to the original data set.

These are the “loadings of the variables” in the axes:

```
   -  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
vars                     1           2           3           4           5
time                0.0013      0.0037      0.0007      0.0006      1.0000
MEI                 0.5087      0.0160     -0.8415      0.1810     -0.0003
PDO                 0.8557     -0.1003      0.5056     -0.0455     -0.0010
upwell36            0.0329      0.8126      0.1560      0.5606     -0.0035
```

```
upwell39      0.0887      0.5739     -0.1090     -0.8068     -0.0017
----------------------------------------------------------------------------
```

Which are the two variables with the strongest loadings (coefficients) loading in axis 1.

The two variables with the strongest loadings in axis 1 are MEI and PDO.

Interpret what this means (how are these variables influencing axis1?):

Relativizing the data set by the maximum value allowed for a more even influence of all of the variables, the magnitude of the changes is now more consistent. With all variables being close to equal (removing the impact of the magnitude of the changes) indicates that MEI and PDO have the greatest influence on the pattern in the data set.

Which are the two variables with the strongest loadings (coefficients) loading in axis 2.

The two variables with the strongest loadings in axis 2 are upwell36 and upwell39.

Interpret what this means (how are these variables influencing axis2?):

The two upwelling variables influence the variability explained by axis 2 the most and they are correlated meaning they impact axis in the same direction.

Check the randomization results below, and calculate the p values, for the two axes, using the equation: `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.
```

```
RANDOMIZATION RESULTS
            999 = number of randomizations
-------------------------------------------------------------------------
      Eigenvalue         Eigenvalues from randomizations
        from      ----------------------------------------
Axis  real data      Minimum       Average       Maximum       p *
   1   0.41233E-01   0.33275E-01   0.33797E-01   0.35615E-01  0.001000
   2   0.29552E-01   0.20368E-01   0.22139E-01   0.24849E-01  0.001000
   3   0.10839E-01   0.15781E-01   0.18226E-01   0.20148E-01  1.000000
```

Finally, use three "stopping rules", to determine how many PCA axes to use. For each method, listed below, enter in the spaces provided the number of axes (starting with 1) that meet the rule. For instance, if axis 1and axis 2 meet the criterion Rnd-Lambda, write down "2".

```
APPLICATION OF STOPPING RULES
-----------------------------------------------------------------------------
Last useful     Rule
   axis        acronym     Explanation (see Peres-Neto, Jackson & Somers 2005)
```

2	Rnd-Lambda	Observed eigenvalue as compared to randomization p values
2	Avg-Rnd	Observed eigenvalue compared to average eigenvalue from randomizations
2	BS	Observed eigenvalue compared to broken-stick eigenvalue

Finally, briefly describe how axis 1 and 2 have changed, as you performed the data relativization (Compare results from questions 2 and 3).

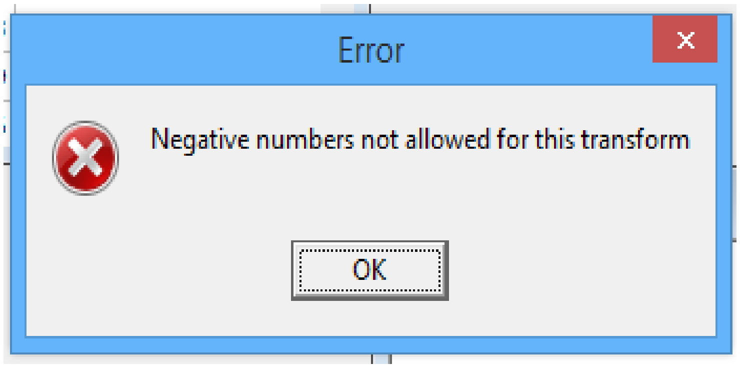
Prior to relativizing the data set Time appeared to have the greatest influence on axis 1 and pattern observed in the data set. When the data were unrelativized the variables with greatest magnitudes had the greatest influence on the variability observed in the data as demonstrated by the highest eigenvalues. Once relativized all of the variables are the same order of magnitude and the variables impact is based on influence not influence and the order of magnitude of the variable parameter. The variables with the smallest range (MEI and PDO) are now highlighted as having the greatest influence on the pattern in the data set.

Compared to the PCA with the raw data, the drivers for axis 1 and 2 have switched between MEI – PDO and the two upwelling indices. However, TIME remains the main variable driving axis 3. There was also a slight change in the p values for axis 2, and for the Arc-Rnd result for axis 2. The relativization of the "raw" data did not manage to de-emphasize the importance of TIME, but it did alter the relative say of the upwelling indices and the oceanographic indices.

4) PCA Analysis – the influence of data relativization: general relativization

Load the upwell_PCA.xls file and perform a general relativization by columns using p = 1.

Note:
This cannot be done because there are negative values in the data. You can go two routes: relativize with p = 2 (ideal for Euclidean, but does not ensure variables sums are equal to 1) or transform the data and relative with p=1 (ideal for making variable sums = 1). I will illustrate the result involving the data transformations



Negative numbers not allowed for this transform

Because a general relativization cannot be performed if any value is negative, you can add the absolute value of the minimum value of the matrix (-296) to EVERY cell, so that all values either greater than or equal to 0. Then, the general relativization can be completed.
Thus, the general relativization completed for this problem using two methods:

p=1 with constants added to eliminate negative numbers and p=2.

The change of p value changes the method to compute the distance measure: the general relativization using p=1 uses Sorensen method whereas using p=2 uses Euclidean method.

The reason why these eigenvalues are so much lower than the eigenvalues from the previous example is because by doing a general relativization decreases the magnitude of the variables, thereby giving them equal weight within the analysis. The major difference is that the factor of "time" is no longer as influential as it now has equal weight to all other variables. Therefore, the eigenvalues all become much smaller after a general relativization.

NOTE: you can add different values to each column, to make all values positive.

Column 1 = no constant added

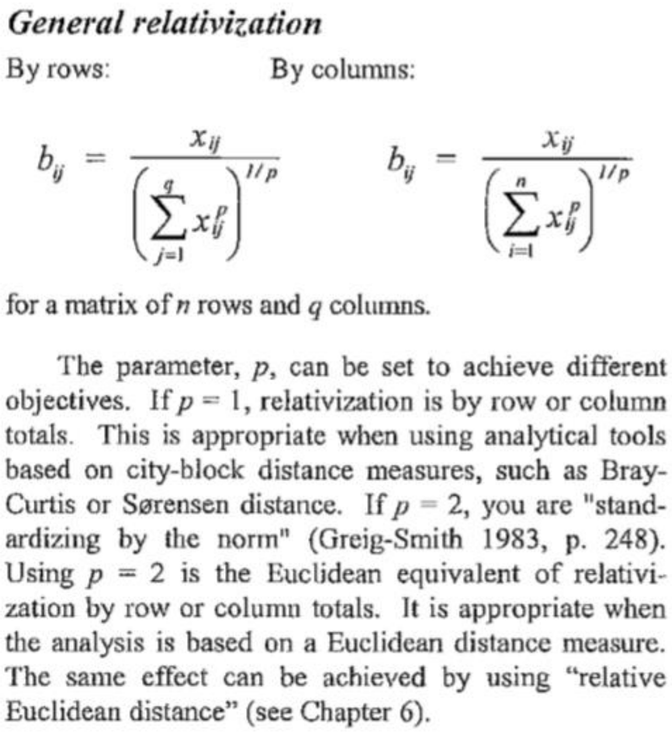
Column 2 = +12 (min value was -11.75)
Column 3 = +6 (min value was -5.429)

Column 4 = +135 (min value was -131)
Column 5 = +300 (min value was -296)

REMEMBER:

THIS IS WHAT GENERAL RELATIVIZATION DOES

(P = 1 OR P = 2)



General relativization

By rows:

$$b_{ij} = \frac{x_{ij}}{\left(\sum_{j=1}^{q} x_{ij}^{p}\right)^{1/p}}$$

By columns:

$$b_{ij} = \frac{x_{ij}}{\left(\sum_{i=1}^{n} x_{ij}^{p}\right)^{1/p}}$$

for a matrix of n rows and q columns.

The parameter, p, can be set to achieve different objectives. If $p = 1$, relativization is by row or column totals. This is appropriate when using analytical tools based on city-block distance measures, such as Bray-Curtis or Sørensen distance. If $p = 2$, you are "standardizing by the norm" (Greig-Smith 1983, p. 248). Using $p = 2$ is the Euclidean equivalent of relativization by row or column totals. It is appropriate when the analysis is based on a Euclidean distance measure. The same effect can be achieved by using "relative Euclidean distance" (see Chapter 6).

REMEMBER:

THIS IS WHAT RELATIVIZATIONBY MAXIMUM DOES: $b_{ij} = x_{ij} /\mathrm{xmax}_j$

Report the missing values in the eigenvector table below:

P=2

```
      VARIANCE EXTRACTED, FIRST   5 AXES
---------------------------------------------------------------
                                                   Broken-stick
AXIS    Eigenvalue   % of Variance  Cum.% of Var.  Eigenvalue
---------------------------------------------------------------
  1         0.007        41.586        41.586         0.007
  2         0.006        37.325        78.911         0.004
  3         0.002        11.917        90.829         0.003
  4         0.001         9.171       100.000         0.001
  5         0.000         0.000       100.000         0.001
```

P=1 with individual constants added to columns 2 through 5 to make positive

```
VARIANCE EXTRACTED, FIRST   5 AXES
---------------------------------------------------------------
                                                   Broken-stick
AXIS    Eigenvalue   % of Variance  Cum.% of Var.  Eigenvalue
---------------------------------------------------------------
  1         0.000        45.050        45.050         0.000
  2         0.000        37.666        82.717         0.000
  3         0.000        10.645        93.361         0.000
  4         0.000         6.636        99.998         0.000
  5         0.000         0.002       100.000         0.000
---------------------------------------------------------------
```

Why are the eigenvalues from this table much lower than the eigenvalues from the previous example? Explain.

The values resulting from a general relativization (especially using p=1) are much smaller than those than resulting from a relativization by maximum when you have a large data set. The formulas used to calculate these three relativizations listed above show that with a large enough data set b_{ij} will be much smaller using the general relativization due to the summation term in the denominator. The smaller the actual values prior to running the PCA the smaller than eigenvalues that will be calculated. Eigenvalues indicate the amount of variability, if you remove some of the variability than the eigenvalues go down. Relativizing the data by the maximum value will yield values between 0 and 1 for a data set with all nonnegative values.

Using a general relativization with p=2 will result in values that can be much smaller when dealing with a large data set. Using a general relativization with p=1 the PCA returns eigenvalues equal to zero indicating the amount of variability in the data set is so low that the PCA calculates the variables have no influence on any of the axis generated.

These are the “loadings of the variables” in the axes:

P=2:

```
   -   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
vars               1           2           3           4           5
time          0.0005      0.0002      0.0002      0.0002      1.0000
MEI           0.4694     -0.5269     -0.6704      0.2293      0.0000
PDO           0.4115     -0.5734      0.6985     -0.1178     -0.0002
upwell36      0.5390      0.5022      0.2034      0.6449     -0.0005
upwell39      0.5654      0.3761     -0.1457     -0.7195     -0.0002
-----------------------------------------------------------------------------
```

P=1 with constant added:

```
---------------------------------------------------------------------------

                              Eigenvector
vars               1           2           3           4           5
time         -0.0018      0.0005     -0.0002      0.0004      1.0000
MEI          -0.0601      0.5528      0.8040     -0.2107     -0.0002
PDO          -0.0306      0.8289     -0.5545      0.0674     -0.0006
upwell36     -0.9167     -0.0787     -0.1127     -0.3752     -0.0015
upwell39     -0.3939      0.0345      0.1827      0.9002     -0.0010
---------------------------------------------------------------------------
```

Which are the two variables with the strongest loadings (coefficients) loading in axis 1.

Upwell39 and upwell36 have the strongest loadings in axis 1 using either PCA method.

Interpret what this means (how are these variables influencing axis1?):

Upwell39 and upwell36 have the greatest influence on the variability explained by axis 1. The changes in upwell39 and upwell36 are most closely aligned with variations of axis 1. The upwell36 and upwell39 axes are closest in ‘space’ to axis 1.

Which are the two variables with the strongest loadings (coefficients) loading in axis 2.

PDO and MEI have the strongest loadings in axis 2 using either PCA method.

Interpret what this means (how are these variables influencing axis2?):

MEI and PDO have the greatest influence on the variability explained by axis 2. The changes in MEI and PDO are most closely aligned with variations of axis 2. The MEI and PDO axes are closest in ‘space’ to axis 2.

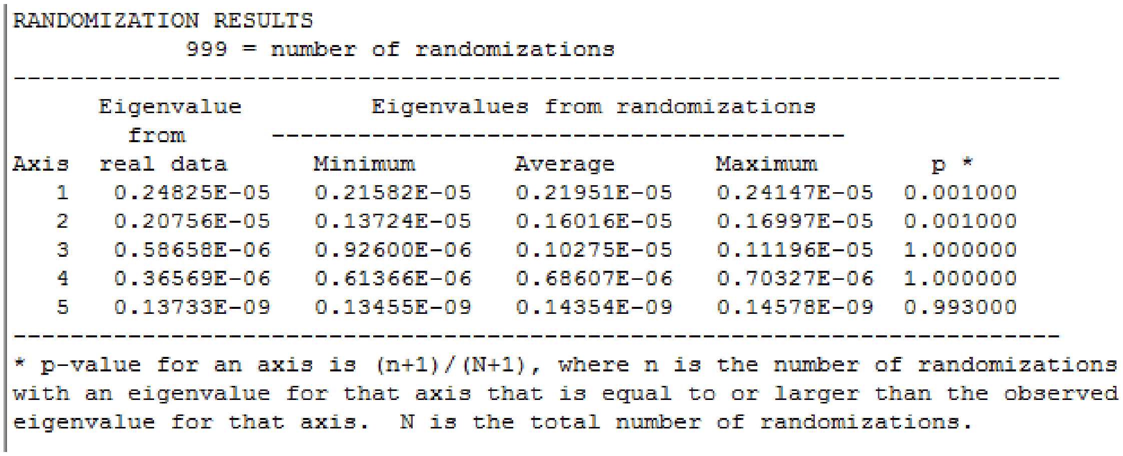
Check the randomization results below, and calculate the p values, for the two axes, using the equation: `p-value for an axis is = (n+1)/(N+1),`

```
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.
```

P=2:

```
RANDOMIZATION RESULTS
            999 = number of randomizations
-------------------------------------------------------------------------
      Eigenvalue          Eigenvalues from randomizations
        from       ----------------------------------------
Axis  real data       Minimum        Average        Maximum        p *
   1   0.66646E-02   0.41451E-02   0.46069E-02   0.54006E-02  0.001000
   2   0.59818E-02   0.37646E-02   0.41808E-02   0.46619E-02  0.001000
   3   0.19099E-02   0.34465E-02   0.38292E-02   0.41661E-02  1.000000
```

P=1 with constant added:



```
RANDOMIZATION RESULTS
            999 = number of randomizations
-------------------------------------------------------------------------
      Eigenvalue          Eigenvalues from randomizations
        from       ----------------------------------------
Axis  real data       Minimum        Average        Maximum        p *
   1   0.24825E-05   0.21582E-05   0.21951E-05   0.24147E-05  0.001000
   2   0.20756E-05   0.13724E-05   0.16016E-05   0.16997E-05  0.001000
   3   0.58658E-06   0.92600E-06   0.10275E-05   0.11196E-05  1.000000
   4   0.36569E-06   0.61366E-06   0.68607E-06   0.70327E-06  1.000000
   5   0.13733E-09   0.13455E-09   0.14354E-09   0.14578E-09  0.993000
-------------------------------------------------------------------------
* 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.
```

Finally, use three “stopping rules”, to determine how many PCA axes to use. For each method, listed below, enter in the spaces provided the number of axes (starting with 1) that meet the rule. For instance, if axis 1and axis 2 meet the criterion Rnd-Lambda, write down “2”.

```
APPLICATION OF STOPPING RULES
----------------------------------------------------------------------------
Last useful    Rule
   axis       acronym     Explanation (see Peres-Neto, Jackson & Somers 2005)
____________________________________________________________________________

2    Rnd-Lambda  Observed eigenvalue as compared to randomization p values

2    Avg-Rnd    Observed eigenvalue compared to average eigenvalue
                      from randomizations

0      BS      Observed eigenvalue compared to broken-stick eigenvalue
```

P=1: Broken-stick values are all zero, not counting any axes as significant.

P=2: For axis 1 the broken-stick value was the same as the random eigenvalue.

Finally, briefly describe how axis 1 and 2 have changed, as you performed the data relativization (Compare results from questions 2, 3 and 4).

With no relativization, the magnitude of change across the variables measured influences the pattern observed and impacts the influence of each variable. Time has the greatest magnitude, followed by upwell39 and upwell36, so these variables may show a falsely significant pattern. The PCA using the raw data showed a single significant axis (supported by all the stopping criteria) most strongly correlated with upwell36 and upwell39. MEI and PDO showed no correlation with axis 1, suggesting that the influence of MEI and PDO was overshadowed by the other variables, which varied by at least one order of magnitude greater than MEI and PDO.

When the data were first transformed by adding constants and then relativized to the maximum the PCA results were different. The broken-stick eigenvalue, p-value criteria, and randomization mean eigenvalue criteria still indicated two axes were significant, but the eigenvalues were reduced. The two variables most correlated with axis 1 were MEI and PDO. The two variables most correlated with axis 2 were upwell36 and upwell39. By using the maximum relativization, the magnitude of the variable measurements was scaled, thus allowing MEI and PDO to have the greatest influence on the PCA pattern, followed by upwell36 and upwell39.

When the data were first transformed by adding constants and then relativized using the general relativization the results were slightly different. The broken-stick eigenvalue analysis did not indicate any axes were significant (with either p=1 or p=2) BUT the p-value criteria and the randomization mean eigenvalue test indicated two significant axes. All of the eigenvalues calculated were much smaller than both of the previous PCA. The two variables most correlated with axis 1 were upwell36 and upwell39. The two variables most correlated with axis 2 were MEI and PDO.

Axis 1: In Q2, upwell36 and upwell39 have the greatest influence on axis 1 which is the same result from Q4. In Q3, MEI and PDO have the greatest influence on axis 1.

Axis 2: In Q2, upwell36 and upwell39 had the greatest influence on axis 2 which was the same results as in Q3. In Q4, MEI and PDO have the greatest influence on axis 2.

Based on the method of relativization, the order of influence on axis 1 and 2 switched (between Q3 and Q4).

The main take home lesson of this exercise is that the results PCA provides are influenced by how you relativize the data (unrelativized vs general relativization vs relativization by maximum), and by what constant you use during the data transformation to ensure all data meet the relativization criteria (> or = 0). When a single constant was applied to all columns (+300) only one axis of significance was observed using PCA. When individual constants were added to each column that were just slightly greater than the lowest negative number, two axes of significance was observed using PCA.

4) Critical reading of the literature:

Read the three posted PCA papers (Mantua et al. 1997, Nelson et al. 2004, Bonaiuto et al. 2003) and report the following:

*** Mantua et al. 1997:**

- Did paper report how many samples / variables were measured? How many?

The paper does report 6 different variables used in the analysis, but seems to break them up into further variables, possibly up to a total of 16. It doesn't report the number of samples.

- Did the paper report the number of "empty" cells? How many?

No, but it was reported that missing SOI values were estimated from a linear regression.

- Did the paper quantify normality of the variables? How?

The authors report that each SST, stream flow, and air temperature time series were normalized with respect to the 1947-95 time period, all prior to compositing the data together for PCA. Then, the mean for this time period was removed. Also, histograms show normalized amplitudes of selected regional climate time series with PDO signatures.

- Did the paper quantify cross-correlations between variables?

Yes, as shown in Table 1 of the paper.

- Did the paper use any data relativizations? If yes, briefly explain: No.

*** Nelson et al. 2004:**

- Did paper report how many samples / variables were measured? How many?

Yes, number of samples = 1912 otolith-aged northern anchovies (but 1836 were used in further analyses). Number of variables = 11 morphometric measurements, plus 8 other independent variables totaling 19 variables. However, the PCA used only the 11 morphometric variables.

- Did the paper report the number of "empty" cells? How many?

Not reported. But, the authors said they only used the sets of samples that had complete measurements, so the number of "empty cells" becomes irrelevant.

- Did the paper quantify normality of the variables? How?

Morphometric traits were log-transformed for size and sex to reach normality. The residuals from the regression of female gonad wet weight vs. dry weight had a bimodal distribution. Branching protocol was used to treat the outliers.

- Did the paper quantify cross-correlations between variables?

Yes, there were 2 general patterns reported presented in Table 4 – 1) condition was positively correlated with the anchovies' body lengths and depths, and there was a negative correlation with jaw length and orbit/preorbit length. 2) GSI slope has a negative correlation with size and GSI (size and GSI are positively correlated). Positively correlated with year (and negatively with temperature) are body length, body depth, GSI slope, and condition. Jaw length is significantly negatively correlated with body depth and condition. Anal-fin-base is negatively correlated with CalCOFI line, depth, and temperature. Size has a positive correlation with offshore distance, depth, and age.

- Did the paper use any data relativizations? If yes, briefly explain:
The authors used a regression to remove the variance due to size before the PCA was run.

*** Bonaiuto et al. 2003 (focus on scale 1 only):**
- Did paper report how many samples / variables were measured? How many?
Yes, number of samples = 312 residents in 7 different Roman neighborhoods. There are a total of 22 variables used in scale 1 – described in Table 1.

- Did the paper report the number of "empty" cells? How many? No.

- Did the paper quantify normality of the variables? How? No.

- Did the paper quantify cross-correlations between variables?
Yes, a preliminary Oblimin rotation was run to test for correlations. When the value of the correlation was <0.20, a Varimax roation was run to obtain the simple structure. After new PCAs were run, Cronbach's Alpha was calculated, and items showing a low inter-correlation with the factor or which were consistently lowering the alpha were eliminated.

- Did the paper use any data relativizations? If yes, briefly explain:
No, not discussed.

5) **Reporting of results in the literature:** Read the three posted PCA papers (Mantua et al. 1997, Nelson et al. 2004, Bonaiuto et al. 2003) and report the following:

*** Mantua et al. 1997:**
- Did paper report PC axes loadings of variables? Were "important" variables identified? How?
The axes loadings were not reported, but some "important" variables were identified as mentioned briefly throughout the paper. At the end of the "Data and methodology" section, the paper reports that the leading PC for sea level pressure in the North Pacific was NPPI, while the leading PC for SST is the PDO.

- Did the paper report the % of explained variance by each axis?
No.

- How many axes were used to explain the pattern?
Not reported.

- Did the paper discuss how the number of axes was selected? If yes, explain:
Not reported.

- Did the paper report p values? If yes, report how many permutations were used.
P-values were reported, but for the intervention model which was run, not for the PCA. The number of runs was not reported.

*** Nelson et al. 2004:**
- Did paper report PC axes loadings of variables? Were "important" variables identified? How?
Yes, the five most important factors were body length, jaw length, anal-fin-base length, body depth, and orbit/preorbit length. These were identified as they accounted for 72% of the total variance alone. The loadings which were important were identified with asterisks next to the values in Tables 2 and 4.

- Did the paper report the % of explained variance by each axis?
The 5 components with eigenvalues greater than 1.0 accounted for 72% of the variance. The % explained by each axis alone was reported in Table 2, in the last row.

- How many axes were used to explain the pattern?
5 components with eigenvalues greater than 1.0.

- Did the paper discuss how the number of axes was selected? If yes, explain:
Yes, if the eigenvalues were greater than 1.0.

- Did the paper report p values? If yes, report how many permutations were used.
P-values for ANOVAs and for analysis of covariance are reported, but not for PCA.

*** Bonaiuto et al. 2003 (focus on scale 1 only):**
- Did paper report PC axes loadings of variables? Were "important" variables identified? How?
Yes, PC axes loadings were reported in Table 1. The authors eliminated those items which had loadings smaller than 0.45, which help to identify the "important variables."

- Did the paper report the % of explained variance by each axis?
Yes, "building aesthetics" = 50.8%, "building density" = 8.7%, and "building volume" = 7.3%.

- How many axes were used to explain the pattern?
3 – "Building aesthetics," "building density," and "building volume"

- Did the paper discuss how the number of axes was selected? If yes, explain:
No. items with loadings of <0.45 or presenting high loadings in more than one factor were eliminated. Otherwise there is no discussion of how the number of axes were selected.

- Did the paper report p values? If yes, report how many permutations were used.

No, p-values not reported. They used Cronbach Alpha values instead to look at the consistency but did not report significance