Advanced Biometry (**MARS6300**) - Homework 4 **Name:** _____KEY_____

Distributed: Wednesday, March 2, 2016 **Due:** Wednesday, March 16, 2016

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#4" and label all files with you're a suffix including your name (e.g., MARS6300_hw4_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 NMDS.
B) To investigate the effects of species on NMDS results.
C) To perform and interpret NMDS analyses.
D) To critically evaluate the NMDS literature.
E) To perform and interpret Polar Ordination analyses.

To complete this homework, you will need:
- Instruction file: "BIOL6090_hw4.doc" (open with word file) – turn in
- "IndianOcean_seabirds.xls" data file: (open with excel) – do not turn in
  (Note: matrix 1: "bird data" sheet, and matrix 2: "env data" sheet)

1) **Assessing Species / Sample Dominance (1 point):**
   Import matrix 1 and matrix 2 from the excel file.

   Use the column / row summary tool in PC_ORD and calculate column / row totals for the data in matrix 1. Paste the results below:

Column_Row_Summary_Bird_data

Summary of: 16 plots N = 42 species

Num.	Name	Mean	Stand.Dev.	Sum	Minimum	Maximum	S	E	H	D`
1	plot1	2.381	8.005	100.0100	0.000	44.310	7	0.748	1.456	0.7135
2	plot2	2.381	7.803	100.0000	0.000	37.500	5	0.885	1.424	0.7266
3	plot3	2.381	10.009	99.9900	0.000	58.970	5	0.654	1.053	0.5654
4	plot4	2.332	6.142	97.9600	0.000	23.980	14	0.733	1.935	0.8150
5	plot5	2.381	6.096	100.0100	0.000	34.230	18	0.755	2.181	0.8239
6	plot6	2.382	11.388	100.0300	0.000	73.670	21	0.382	1.163	0.4448
7	plot7	2.381	7.143	100.0100	0.000	43.280	19	0.684	2.015	0.7671
8	plot8	2.380	11.505	99.9700	0.000	74.610	22	0.366	1.132	0.4332
9	plot9	2.381	6.108	100.0000	0.000	29.350	20	0.673	2.017	0.8232
10	plot10	2.380	5.450	99.9800	0.000	21.790	15	0.802	2.171	0.8543
11	plot11	2.309	10.735	96.9700	0.000	69.020	10	0.473	1.089	0.4737
12	plot12	2.381	13.738	100.0000	0.000	89.020	4	0.324	0.449	0.2023
13	plot13	2.381	13.659	100.0000	0.000	88.370	3	0.392	0.431	0.2112
14	plot14	2.381	13.870	100.0000	0.000	90.000	6	0.265	0.475	0.1875
15	plot15	2.381	12.335	100.0000	0.000	79.780	7	0.417	0.811	0.3524
16	plot16	2.381	12.545	100.0100	0.000	80.750	6	0.375	0.672	0.3310
	AVERAGES:	2.373	9.783	99.68	0.000	58.66	11.4	0.558	1.280	0.5453

		Skewness	Kurtosis
1	plot1	4.199	19.511
2	plot2	3.720	14.048
3	plot3	5.060	27.109
4	plot4	2.939	7.672
5	plot5	4.115	19.227
6	plot6	6.284	40.370
7	plot7	4.999	27.759
8	plot8	6.332	40.875
9	plot9	3.189	10.668
10	plot10	2.726	6.923
11	plot11	6.151	39.123
12	plot12	6.418	41.660
13	plot13	6.386	41.350
14	plot14	6.452	41.975
15	plot15	6.325	40.781
16	plot16	6.247	40.007
	Averages:	5.096	28.691

672 cells in main matrix
Percent of cells empty = 72.917
Matrix total = 0.15949E+04
Matrix mean = 0.23734E+01
Variance of totals of plots = 0.78300E+00
CV of totals of plots = 0.89%

S = Richness = number of non-zero elements in row
E = Evenness = H / ln (Richness)
H = Diversity = - sum (Pi*ln(Pi)) = Shannon`s diversity index
D = Simpson`s diversity index for infinite population = 1 - sum (Pi*Pi)
where Pi = importance probability in element i (element i relativized by row total)

Column_Row_Summary_Bird_data

Summary of: 42 species N = 16 plots

Num.	Name	Mean	Stand.Dev.	Sum	Minimum	Maximum	S	E	H	D`
1	WCPT	7.169	8.963	114.7100	0.000	29.350	10	0.874	2.012	0.8459
2	WISP	2.974	5.588	47.5900	0.000	20.530	9	0.736	1.618	0.7307
3	SPPT	6.027	10.071	96.4300	0.000	28.210	9	0.739	1.623	0.7739
4	YNAL	5.595	17.118	89.5200	0.000	69.020	7	0.428	0.833	0.3891
5	LISH	2.017	3.755	32.2800	0.000	12.500	6	0.846	1.515	0.7345
6	LTJA	0.375	0.688	6.0000	0.000	2.250	5	0.894	1.439	0.7405
7	SOSH	0.112	0.384	1.8000	0.000	1.530	2	0.610	0.423	0.2550
8	GRSH	0.032	0.127	0.5100	0.000	0.510	1	0.000	0.000	0.0000
9	BBAL	0.941	2.643	15.0600	0.000	10.600	5	0.599	0.965	0.4754
10	SOAL	0.841	2.598	13.4600	0.000	10.440	4	0.546	0.758	0.3789
11	MAPN	0.632	1.508	10.1100	0.000	4.930	3	0.904	0.993	0.6037
12	BLPT	1.336	2.873	21.3700	0.000	10.560	5	0.780	1.256	0.6664
13	BBSP	2.921	5.866	46.7400	0.000	22.450	8	0.716	1.489	0.7013
14	CAPT	0.406	0.810	6.5000	0.000	2.990	5	0.870	1.399	0.7047
15	WAAL	0.940	1.145	15.0400	0.000	2.700	9	0.909	1.997	0.8506
16	NGPT	0.289	0.481	4.6200	0.000	1.500	5	0.963	1.550	0.7751
17	KESH	0.074	0.295	1.1800	0.000	1.180	1	0.000	0.000	0.0000
18	GHAL	0.300	0.594	4.8000	0.000	1.970	5	0.836	1.345	0.7077
19	KEPT	0.545	1.356	8.7200	0.000	4.920	5	0.634	1.020	0.5746
20	WHPT	0.283	0.665	4.5300	0.000	2.430	5	0.709	1.142	0.6145
21	BRSK	0.162	0.411	2.5900	0.000	1.620	4	0.769	1.066	0.5605
22	KIPN	0.156	0.364	2.4900	0.000	1.350	4	0.816	1.131	0.6164
23	ROPN	0.222	0.840	3.5600	0.000	3.370	3	0.211	0.232	0.1018
24	GBSP	0.161	0.407	2.5800	0.000	1.310	5	0.600	0.965	0.5644
25	WBSP	0.169	0.486	2.7000	0.000	1.920	3	0.727	0.799	0.4519
26	WFSP	0.477	1.389	7.6300	0.000	5.130	2	0.913	0.633	0.4406
27	LMSA	0.649	1.185	10.3900	0.000	3.610	6	0.850	1.524	0.7425
28	SGPT	0.360	0.717	5.7600	0.000	2.040	5	0.819	1.318	0.7048
29	DPSP	0.509	1.400	8.1400	0.000	5.570	5	0.607	0.977	0.4940
30	PRSP	16.298	26.346	260.7700	0.000	74.610	6	0.916	1.641	0.7844
31	GWPT	25.874	39.160	413.9800	0.000	90.000	7	0.860	1.673	0.8033
32	BRPT	2.531	5.916	40.4900	0.000	20.300	3	0.933	1.024	0.6173
33	AUSH	1.087	3.101	17.3900	0.000	11.140	2	0.942	0.653	0.4605

34	COSH	1.339	5.358	21.4300	0.000	21.430	1	0.000	0.000	0.0000
35	FFSH	1.258	2.657	20.1200	0.000	8.140	4	0.877	1.216	0.6759
36	MSPT	2.299	7.778	36.7800	0.000	31.250	3	0.480	0.527	0.2667
37	BUPT	2.437	9.358	38.9900	0.000	37.500	2	0.234	0.162	0.0735
38	WTTR	0.062	0.248	0.9900	0.000	0.990	1	0.000	0.000	0.0000
39	JFPT	0.032	0.127	0.5100	0.000	0.510	1	0.000	0.000	0.0000
40	WNPT	0.6250E-02	0.2500E-01	0.1000	0.000	0.1000	1	0.000	0.000	0.0000
41	DKTE	3.494	11.262	55.9000	0.000	44.310	2	0.736	0.510	0.3287
42	WTSH	6.293	20.400	100.6800	0.000	80.750	3	0.492	0.540	0.3217
AVERAGES:		2.373	4.916	37.97	0.000	16.37	4.3	0.628	0.952	0.4888

		Skewness	Kurtosis
1	WCPT	1.183	1.456
2	WISP	2.568	7.390
3	SPPT	1.449	1.149
4	YNAL	3.846	15.792
5	LISH	2.176	4.809
6	LTJA	1.801	3.240
7	SOSH	3.812	15.522
8	GRSH	4.000	16.709
9	BBAL	3.683	14.785
10	SOAL	3.814	15.592
11	MAPN	2.388	5.487
12	BLPT	2.600	7.888
13	BBSP	2.816	9.420
14	CAPT	2.531	7.634
15	WAAL	0.734	-0.724
16	NGPT	1.479	1.910
17	KESH	4.000	16.709
18	GHAL	2.067	4.341
19	KEPT	2.805	8.456
20	WHPT	2.779	8.443
21	BRSK	3.374	12.882
22	KIPN	2.820	8.918
23	ROPN	3.985	16.620
24	GBSP	2.534	5.946
25	WBSP	3.544	13.856
26	WFSP	3.058	10.026
27	LMSA	2.003	3.762
28	SGPT	1.853	2.656
29	DPSP	3.576	14.064
30	PRSP	1.539	1.908
31	GWPT	0.983	-0.364
32	BRPT	2.402	6.013
33	AUSH	2.901	8.789
34	COSH	4.000	16.709
35	FFSH	2.081	3.893
36	MSPT	3.904	16.138
37	BUPT	3.989	16.647
38	WTTR	4.000	16.709
39	JFPT	4.000	16.709
40	WNPT	4.000	16.709
41	DKTE	3.624	14.231
42	WTSH	3.692	14.709
Averages:		2.867	9.608

672 cells in main matrix
Percent of cells empty = 72.917
Matrix total = 0.15949E+04
Matrix mean = 0.23734E+01
Variance of totals of species = 0.57424E+04
CV of totals of species = 199.55%

S = Richness = number of non-zero elements in row
E = Evenness = H / ln (Richness)
H = Diversity = - sum (Pi*ln(Pi)) = Shannon`s diversity index
D = Simpson`s diversity index for infinite population = 1 - sum (Pi*Pi)
where Pi = importance probability in element i (element i relativized by row total)

********** Analysis completed **********

First, lets look at the row summaries. (This is the summary of plots)

- Are there any empty plots (without any species recorded)? Which one? _There were no empty plots__

All 16 plots have at least 1 "zero" value, and all plots have a maximum value within the row that is larger than zero. Therefore there is no "empty" plot – each has at least one of the 42 species recorded.

- What metric (hint: letter variable) tells you how many species are found per plot? S, as this stands for richness, which is the number of non-zero elements in the entire row___ What is the range of species found per plot? _3 species (found in plot 13) ranging to 22 species (found in plot 8), therefore the range is 19 species_

- How are the four metrics of species "diversity" related to each other? Calculate pair-wise cross correlations (hint, use CORREL in Excel) of the four variables: `S, E, H, D`

Table 1. Correlation coefficient values for two-way comparisons of Species Richness (S), Evenness (E), Shannon's Diversity Index (H), and Simpson's Diversity Index for Infinite Population (D').

	S	E	H	D'
S	1			
E	0.181097	1		
H	0.665352	0.821193	1	
D	0.512414	0.922854	0.968139	1

Is Shannon species diversity more closely driven by species richness or evenness? Explain: Shannon's Diversity Index (H) is more closely driven by Evenness (E) as the correlation coefficient for these two parameters is higher than when compared with species richness (S; 0.821 vs. 0.665). Therefore, E explains a higher proportion of the variance in H than S does.

Is Simpson species diversity more closely driven by species richness or evenness? Explain: Simpson's Diversity Index for Infinite Population (D') is more closely driven by Evenness (E) as the correlation coefficient for these two parameters is higher than when D' is compared with species richness (S; 0.923 vs. 0.512). Therefore, E explains a higher proportion of the variance in D' than S does.

Next, lets look at the column summaries. (This is the summary of species data)

- Are there any empty columns (any "absent" species)?

There are no columns (species) that are "empty" or "absent" – every column has at least one cell with a value other than zero.

- What species (indicate four letter acronym) is the most numerous: _GWPT (even though it only has 7 cells with data in that column and WCPT has 10 cells with values in that column, GWPT has a much larger sum – 413.98, compared to 114.71 of WCPT)____

How large is its sum of abundance (birds / km 2) across all 16 samples: _413.98 birds/km^2 _____

- What species (indicate four letter acronym) are only sighted in one sample: _
- GRSH, KESH, COSH, WTTR, JFPT, WNPT____

- How many (%) cells are empty? 72.917%____

- What are the mean skewness for species: 2.867_ and for plots: _5.096_

- On the basis of the inspection of the data, would you use PCA or NMDS for its analysis? Could a transformation yield normal distributions, given the high proportion of "zeros"? Explain:

As one of the rules for using PCA is that the percent of cells left "empty" or with zeroes must be at 20% or less, and with the skewness for both species and plots not between -1 and 1, it is likely that the dataset is not normal, and if there are many outliers, all of these factors will greatly affect the outcome of a PCA analysis in a negative way. Therefore, I would recommend using NMDS.

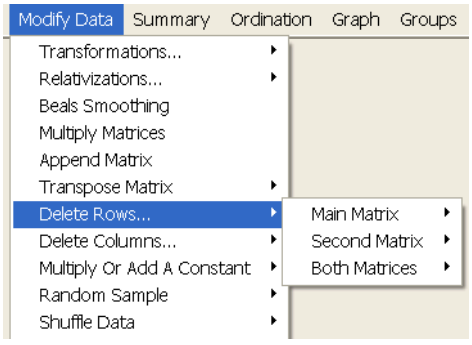
Performing a transformation would not necessarily yield normal distributions due to the problem of there being so many zeros. If one was to do a log transformation, a constant would need to be added first as the log of zero cannot be done. If the value of the constant is larger than 1, the skewness values of both plots and species would remain quite large, still indicating a PCA should not be run, especially as the data remains non-normal. If a smaller value than 1 is added to each cell, taking the log of these cells will result in a negative number, and it does not make sense for species abundance data to be negative. Doing a power transformation wouldn't work as the values that are zero would remain zero. Doing a presence/absence transformation also would not work as not only would the zero values remain zero, but the data would no longer represent abundance as it currently does. And, an arcsine transformation wouldn't make sense as it is used for normalizing proportional data, which this species abundance data is not.

- Why would you remove the plots / species from the dataset you identified above?

These would be removed as they are rare species (no plots would be removed), and as outliers they could greatly affect the distances or variances within species calculated by the statistical analysis, and therefore present false significant patterns within the data due to these outliers. Leaving in these rare species greatly increases the skewness.

Moreover, if the species are found in only one sample, their presence does not help to connect samples and develop an intertwined community.

- To remove the plots / species, do the following:



- Remove the species first, using the "delete columns" command. Request that species with only one presence to be removed. How many species were discarded: __6__

- Copy and paste the "results.txt" output here:

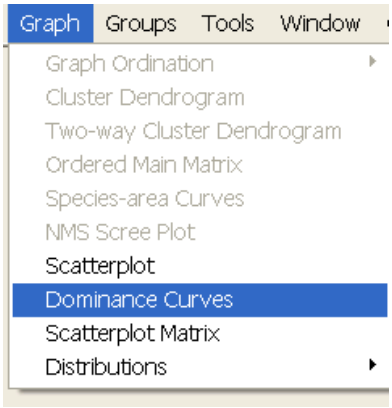
```
****************************** Data Modification ******************************
PC-ORD, 5.10
15 Mar 2011, 14:54

Deletion of     6 columns:
GRSH     KESH     COSH     WTTR     JFPT     WNPT
***************************** Operation completed *****************************
```

- Next, check again if there are any "empty" samples after removing these species. Are there any empty plots?

After doing another row/column summary after removing those species, there are still no empty plots (as there weren't any before). The minimum number of samples per plot is now 3 (like it was before), and the minimum number of samples per species is now 2. The number of empty cells overall has dropped to 69.444% (from 72.917%).

2) Selecting the species to be analyzed (1 point):
The species in the dataset vary greatly, on basis of their density (birds / km 2). Use the "dominance curves" command to assess the species relative abundance.



Inspect the "results.txt" file. You can save the "results.txt" file and open with Excel. Use these data to test whether the ranks of species abundance and the ranks of species frequency are correlated. Report r: _0.379151___

Paste the curve of "sum" versus "rank abundance" here:

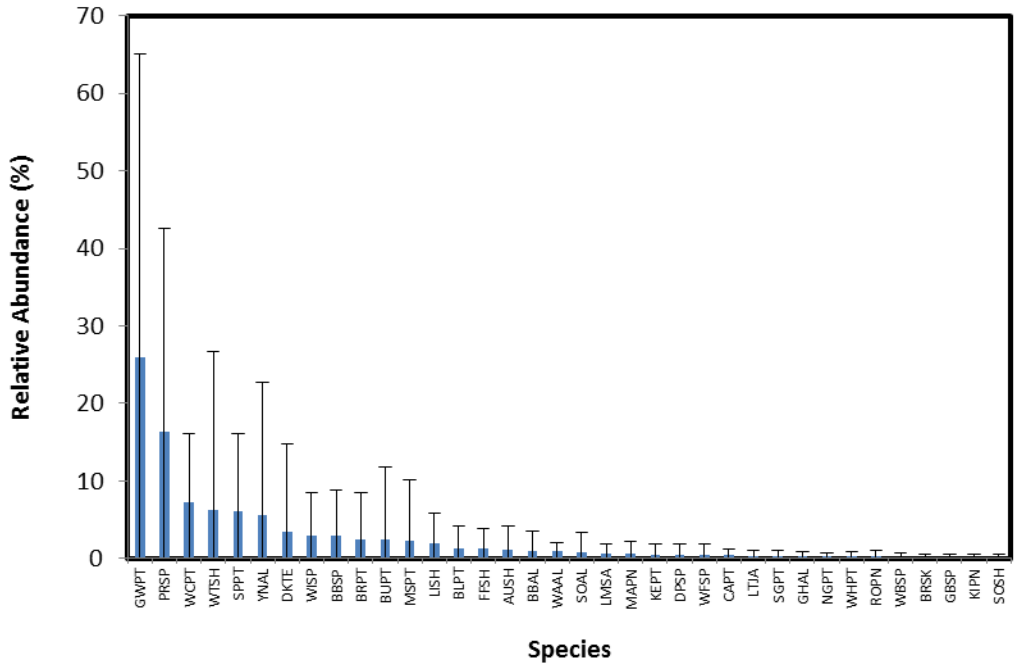


Figure 1. Graph of the total sum of abundance of each bird species vs. its rank in abundance. The Error bars depict 1 S.D.

Using the data in the "results.txt" file to calculate proportions of abundance for each species. Paste an excel table, showing the ranked (from most abundant to least abundant) species (column 1) and their relative abundance (column 2) and the cumulative relative abundance (column 3).
Table 2. List of 36 bird species ranked in decreasing order of abundance (Sum) and showing their relative abundance (%), and cumulative relative abundance (Cumm).

Species	Sum	%	Cumm
GWPT	413.98	26.36445848	26.36445848
PRSP	260.77	16.60722701	42.9716855
WCPT	114.71	7.305345748	50.27703124
WTSH	100.68	6.411840379	56.68887162
SPPT	96.43	6.141177669	62.83004929
YNAL	89.52	5.701111946	68.53116124
DKTE	55.9	3.560010699	72.09117194
WISP	47.59	3.030785495	75.12195743
BBSP	46.74	2.976652953	78.09861039
BRPT	40.49	2.578619556	80.67722994
BUPT	38.99	2.483091541	83.16032148
MSPT	36.78	2.342346932	85.50266842
LISH	32.28	2.055762887	87.5584313
BLPT	21.37	1.36095579	88.91938709
FFSH	20.12	1.28134911	90.2007362
AUSH	17.39	1.107488123	91.30822433

BBAL	15.06	0.959101272	92.2673256
WAAL	15.04	0.957827566	93.22515316
SOAL	13.46	0.857204723	94.08235789
LMSA	10.39	0.661690718	94.7440486
MAPN	10.11	0.643858822	95.38790743
KEPT	8.72	0.555336195	95.94324362
DPSP	8.14	0.518398696	96.46164232
WFSP	7.63	0.485919171	96.94756149
CAPT	6.5	0.413954732	97.36151622
LTJA	6	0.382112061	97.74362828
SGPT	5.76	0.366827578	98.11045586
GHAL	4.8	0.305689649	98.41614551
NGPT	4.62	0.294226287	98.7103718
WHPT	4.53	0.288494606	98.9988664
ROPN	3.56	0.226719823	99.22558622
WBSP	2.7	0.171950427	99.39753665
BRSK	2.59	0.16494504	99.56248169
GBSP	2.58	0.164308186	99.72678988
KIPN	2.49	0.158576505	99.88536638
SOSH	1.8	0.114633618	100

How many species – together – make up 95 % of all the bird abundance? _It takes 21 species until the 95th percentile of all bird abundance is reached._

How many species – together – make up 100 % of all the bird abundance? _It takes all 36 species to make up all 100% of the bird abundance._

How many species make up, at least 1% of the total bird abundance? _16_

How many species make up, at least 5% of the total bird abundance? _6_

Make a plot, showing the relative abundance of the different species and report the descriptive statistics for those species-specific proportions of abundance. Based on these graph and the descriptive statistics, what thresholds would you use to identify "numerous", "common" and "rare", species? Explain your rationale:

There are no clear guidelines – as evidenced in the literature. What is obvious is that not all species have the same abundance, and the community is dominated by a few numerous species.

You could use the median of their % abundances to define common species (relative abundance above median for all species) and uncommon species (relative abundance below median for all species).

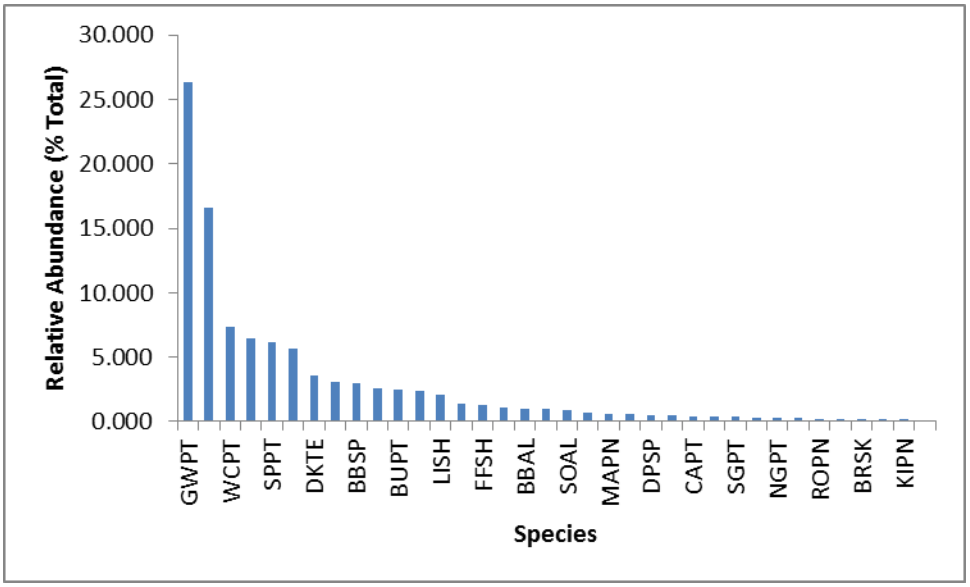
You could use the mean +/- 2 SDs of their % abundances to define three groups of species: numerous (relative abundances above mean + 2 SDs), rare species (relative abundances below mean - 2 SDs), and common species (relative abundances between the mean + / - 2 SDs).

Or you could define species to be numerous if they have % abundances greater than 5% of the community, to be common if they have % abundances between 5% and 1%, and to be rare if they have % abundances below 1%.

For these community data, these are the descriptive statistics:

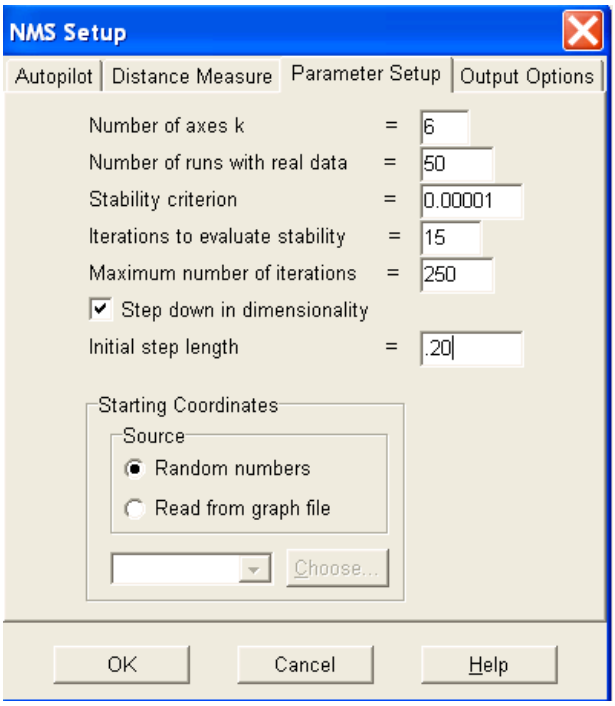
Mean	2.78
Median	0.91
Standard Deviation	5.13
Kurtosis	13.75
Skewness	3.52
Range	26.25
Minimum	0.11
Maximum	26.36
Sum	100
Count	36

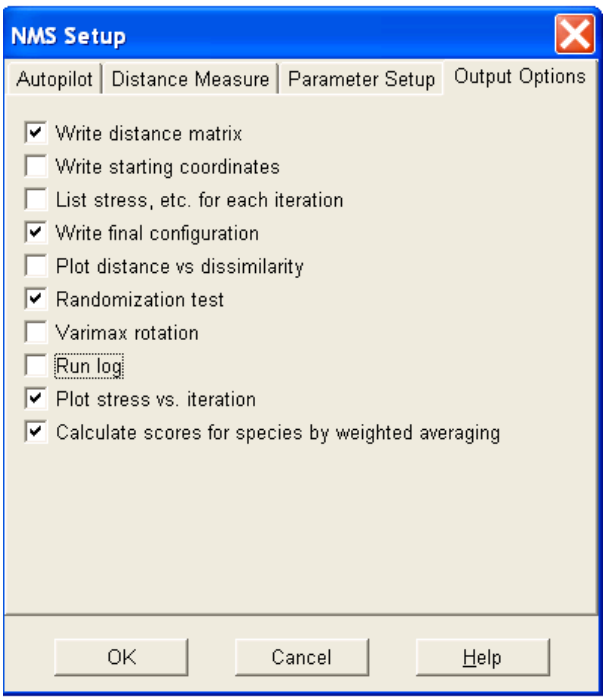
For these community data, this is a plot of the relative abundance of the species:



3a) NMDS analysis with entire dataset - 16 plots and 36 species:

Use the following parameter set-up and output:





What distance measure would you use? Why?
I would choose the "Relative Sorensen" distance measure as it already has relativization built within it, does not care if the data is already normalized or not, and doesn't care if there are a lot of zeros or joint absences in the data. Relative Sorenson is based on relativization by maximum, which ensures that each sample will be equally weighted within the analysis, setting each species as a proportion of the potential maximum.

How many runs will you select to get a p value as small as 0.001? Explain?
Because the equation to find the p-value is p = (n+1)/(N+1), where n = number of randomized runs with a final stress that is equal to or less than the observed minimum stress, and N = number of randomized runs, it would make sense to do **999 runs** (so that if n = 1 and N = 999, p can equal 0.001).

At the end of the "results.txt" file you will find the recommended solution. How many dimensions does PC-ORD recommend? _ 2 Dimensions _

Copy and paste the NMS scree plot here:

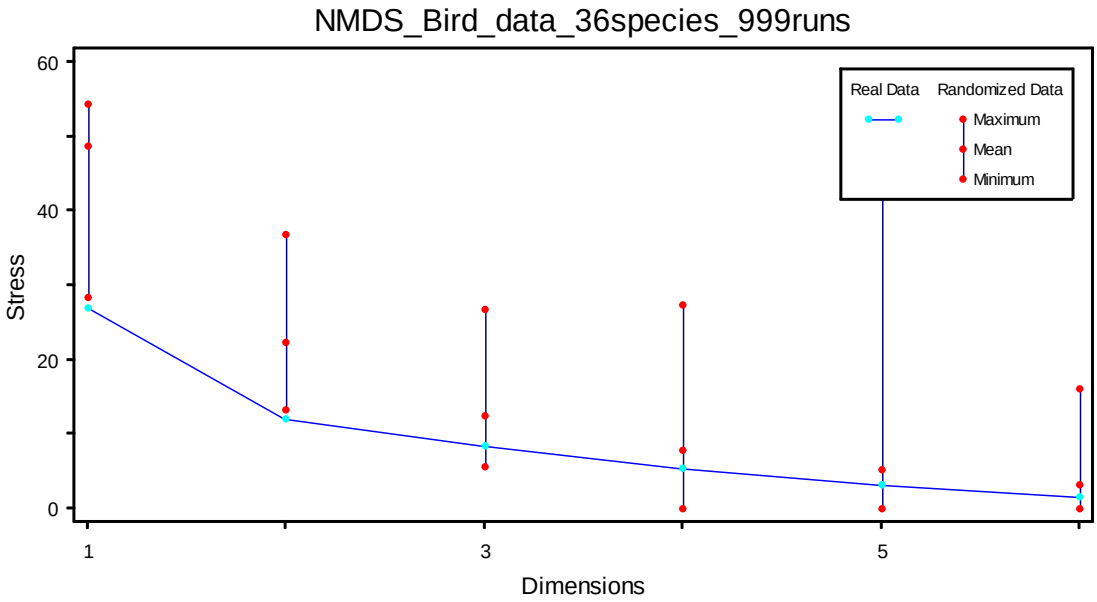


Figure 2. NMS Scree plot of 999 randomizations run for the "Bird data" using the 36 species which were observed in more than 1 plot.

Does this result make sense with the two "criteria" we discussed in class? Paste the `STRESS IN RELATION TO DIMENSIONALITY (Number of Axes)` table and explain your answer.
This result makes sense for both of these 2 "criteria." For the p-value criteria, the p-value must be < 0.05 to accept the axis. In the case of 2 axes, or 2 dimensions, the p-value is 0.001, which is < 0.05. However, this is the case for 3 of the 6 axes (dimensions), so we must look at another criterion as well. The 2nd criteria is that additional dimensions or axes will be included until the stress is reduced by less than or equal to a value of 5, aka the "stress reduction rule." As the difference in stress level between Axis 1 and 2 is greater than 5 in the "minimum" stress column, we must look at the stress level between Axis 2 and 3, which is not greater than 5. Therefore, this "criteria" would tell us to choose 2 dimensions or 2 axes, which is what PC ORD has recommended we use.

STRESS IN RELATION TO DIMENSIONALITY (Number of Axes)

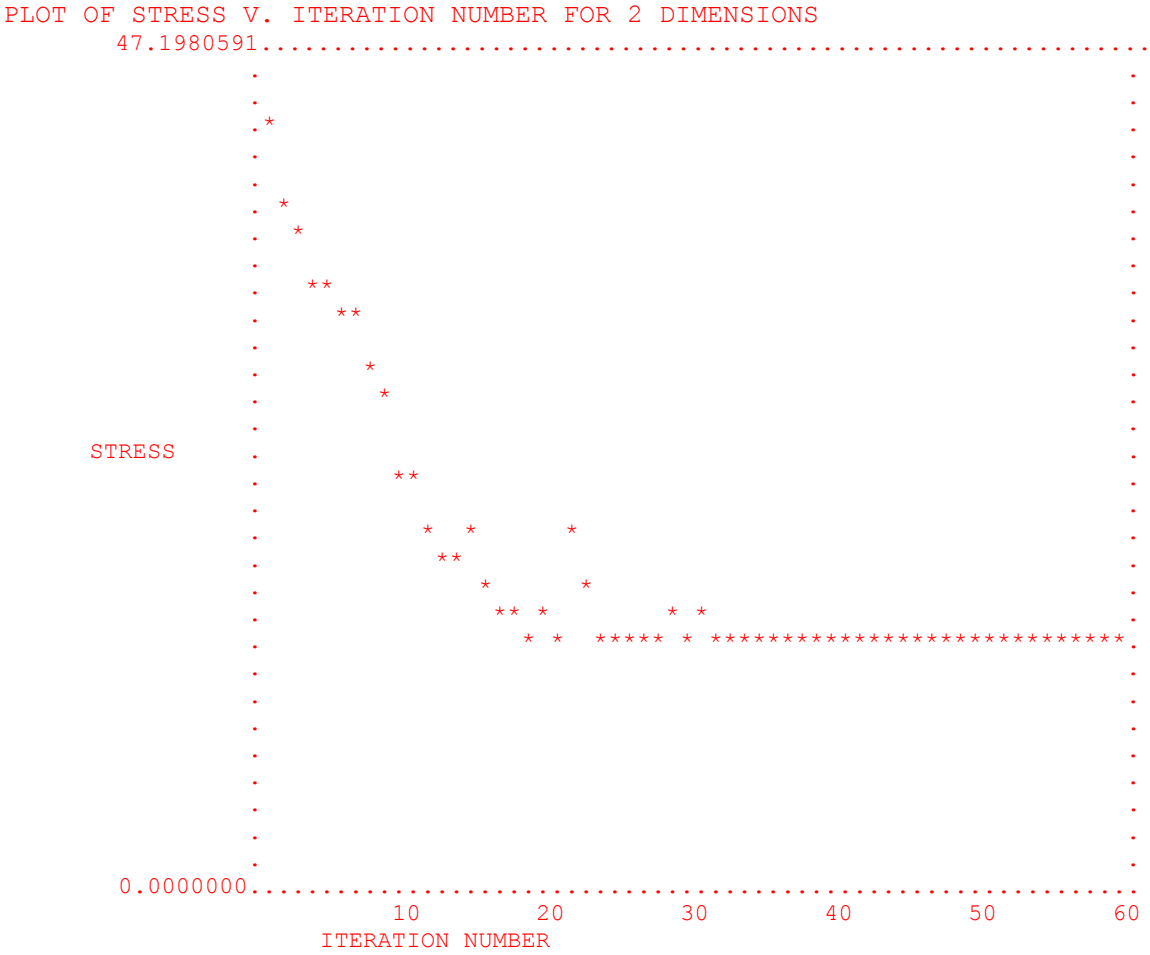
	Stress in real data 50 run(s)			Stress in randomized data Monte Carlo test, 999 runs			
Axes	Minimum	Mean	Maximum	Minimum	Mean	Maximum	p
1	26.989	43.722	54.506	28.337	48.782	54.372	0.0010
2	12.007	15.018	34.127	13.349	22.271	36.937	0.0010
3	8.391	9.466	14.153	5.724	12.554	26.790	0.0210
4	5.360	6.442	18.241	0.001	7.904	27.338	0.0550
5	3.146	4.968	10.921	0.003	5.140	45.301	0.0630
6	1.558	4.544	14.416	0.004	3.207	16.046	0.0700

p = proportion of randomized runs with stress < or = observed stress

i.e., p = (1 + no. permutations <= observed)/(1 + no. permutations)

If we did not use the stress reduction rule, how many axes would the best solution have on the basis of the significance level ($p < 0.05$) of the randomizations ?
If ignoring the stress reduction rule, the best answer would have been 3 axes, since we want the minimum number of significant axes.

Paste the PLOT OF STRESS V.ITERATION NUMBER for recommended dimension answer (Remember, your answers may vary slightly from those of your colleagues):



Did the iterations stop before 250? Why / why not?
Yes, the iterations stopped before 250. This is because the stabilization criterion had been met at 60 iterations. We told PC ORD to stop the iterations when either a) that stabilization criterion had been met, or b) if it is not met, just stop at 250 iterations, the maximum limit we set.

Create a plot, using the "Graph \ Graph Ordination" command. Add environmental vectors – scaled to 100%.

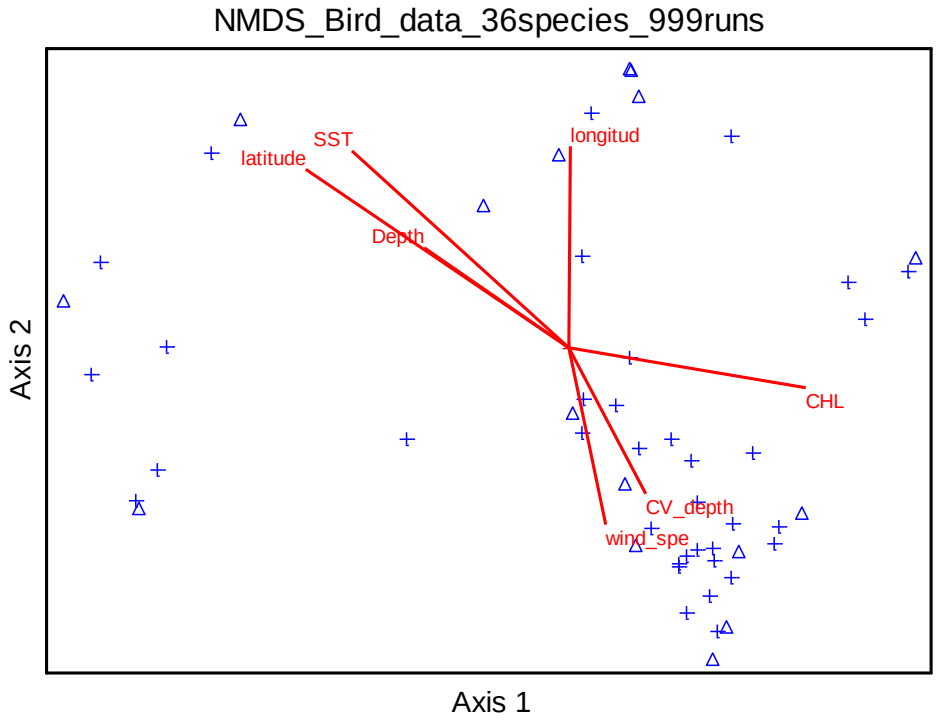


Figure 3. An ordination plot for the NMDS of the Bird Data (999 runs) with environmental vectors scaled to 100%.

Request amount of variance in distance matrix explained. Copy and paste the results below:

```
Coefficients of determination for the correlations between ordination
distances and distances in the original n-dimensional space:

            R Squared
Axis    Increment    Cumulative
 1        .266          .266
 2        .427          .694

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    -0.194      96.2

Number of entities = 16
Number of entity pairs used in correlation = 120
Distance measure for ORIGINAL distance: Relative Sorensen
```

Report the orthogonality of the two axes:

The orthogonality is 96.2%. (If this was 100%, it would be a 90 degree angle between the 2 axes, showing they are completely independent of each other).

```
Axis pair      r      Orthogonality,% = 100(1-r^2)
  1 vs 2    -0.194      96.2
```

Request the correlations with the environmental variables. Copy and paste the results below:

```
Pearson and Kendall Correlations with Ordination Axes   N= 16
```

Axis:	1			2		
	r	r-sq	tau	r	r-sq	tau
longitude	.053	.003	-.033	.687	.472	.450
latitude	-.785	.616	-.633	.648	.419	.517
wind_speed	.297	.088	.127	-.644	.414	-.312
SST	-.714	.510	-.661	.679	.461	.460
CHL	.745	.554	.559	-.307	.094	-.220
Depth	-.582	.339	-.533	.483	.233	.383
CV_depth	.426	.181	.333	-.586	.343	-.617

Interpret these results:

- Which environmental variables best explain the two main axis of variation?

Look at the tau values as these are non-parametric, rank correlations. The tau values with the highest absolute values are the ones which contribute the most to the values of the axes. Therefore, axis 1 is best described by the environmental variables of latitude (with an absolute tau value of 0.633), and SST (with an absolute tau value of 0.661). CHL and Depth have absolute tau values (0.559 and 0.533, respectively) that are just slightly smaller, so also have a large influence on this axis. Axis 2 is best described by CV_Depth (absolute tau value of 0.617) and latitude (absolute tau value of 0.517). Close behind these 2 variables are SST (absolute tau value of 0.460) and longitude (absolute tau value of 0.450).

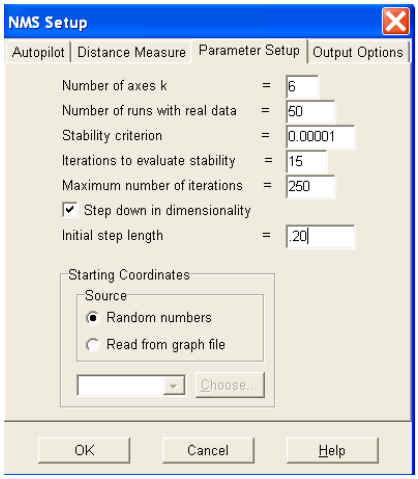
3b) MNDS analysis of reduced dataset – without "rare" species: 16 plots and 21 species (so total proportion of all birds sighted is over 95%)

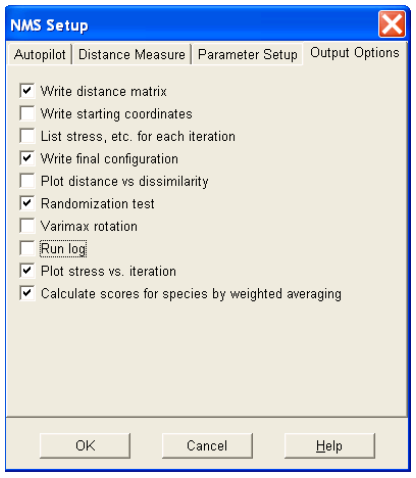
Note: Remove the following species:

BRSK	CAPT	DPSP	GBSP	GHAL
KEPT	KIPN	LTJA	NGPT	ROPN
SGPT	SOSH	WBSP	WFSP	WHPT

All samples still have bird counts - there are no empty plots.

Use the same parameter set-up and output as in question 3a





What distance measure would you use? Why?

I would use Relative Sorensen because the average skewness and kurtosis for these data are still very high, indicating that they are non-parametric. Also, I would want to use the same distance measure I used previously, to ensure the results were comparable with those from analysis 3a.

How many runs will you select to get a p value as small as 0.001? Explain?

Because the equation to find the p-value is p = (n+1)/(N+1), where n = number of randomized runs with a final stress that is equal to or less than the observed minimum stress, and N = number of randomized runs, it would make sense to do **999 runs** (so that if n = 1 and N = 999, p can equal 0.001). Also, I would want to use the same number of runs I used previously, to ensure the results were comparable with those from analysis 3a.

At the end of the "results.txt" file you will find the recommended solution. How many dimensions does PC-ORD recommend? __2 Dimensions__

Copy and paste the NMS scree plot here:

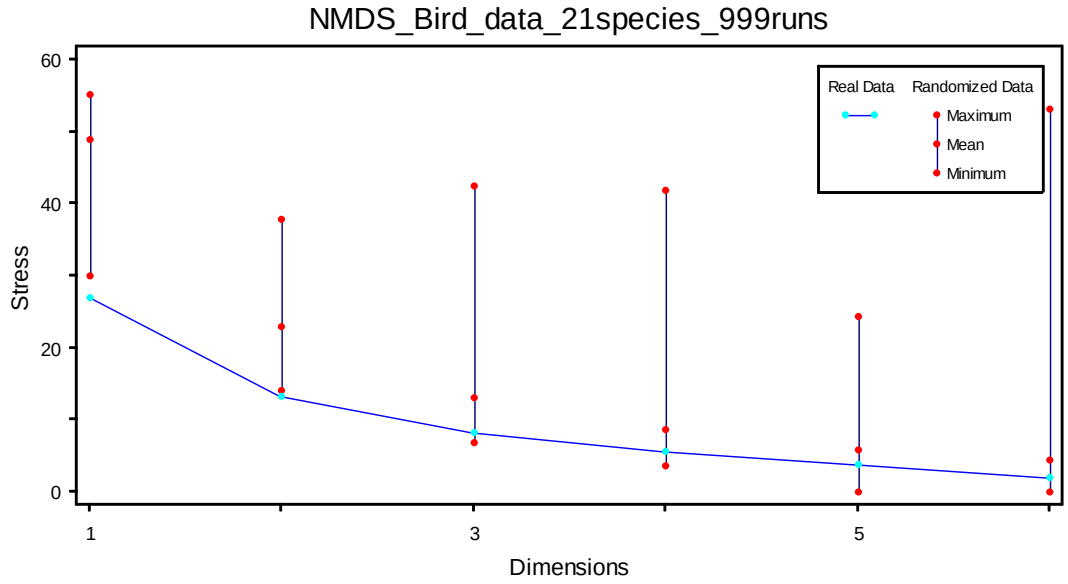


Figure 4. NMS Scree plot of 999 randomizations run for the "Bird data" using only the 21 most prevalent species.

Does this result make sense with the two "criteria" we discussed in class? Paste the `STRESS IN RELATION TO DIMENSIONALITY (Number of Axes)` table and explain your answer.
This result does make sense for both "criteria." For the p-value criterion, the p-value must be < 0.05. In the case of 2 axes, or 2 dimensions, the p-value is 0.001, which is < 0.05. However, this is the case for 3 of the 6 axes (dimensions), so we must look at another criterion as well. The 2nd criterion is that additional dimensions or axes will be included until the minimum stress is reduced by less than or equal to a value of 5, aka the "stress reduction rule." As the difference in minimum stress level between Axis 1 and 2 is greater than 5, we must look at the minimum stress level between Axis 2 and 3, which is slightly less than 5. Therefore, this "criterion," along with the p-value criterion, tell us to select 2 dimensions, which is what PC ORD recommends!

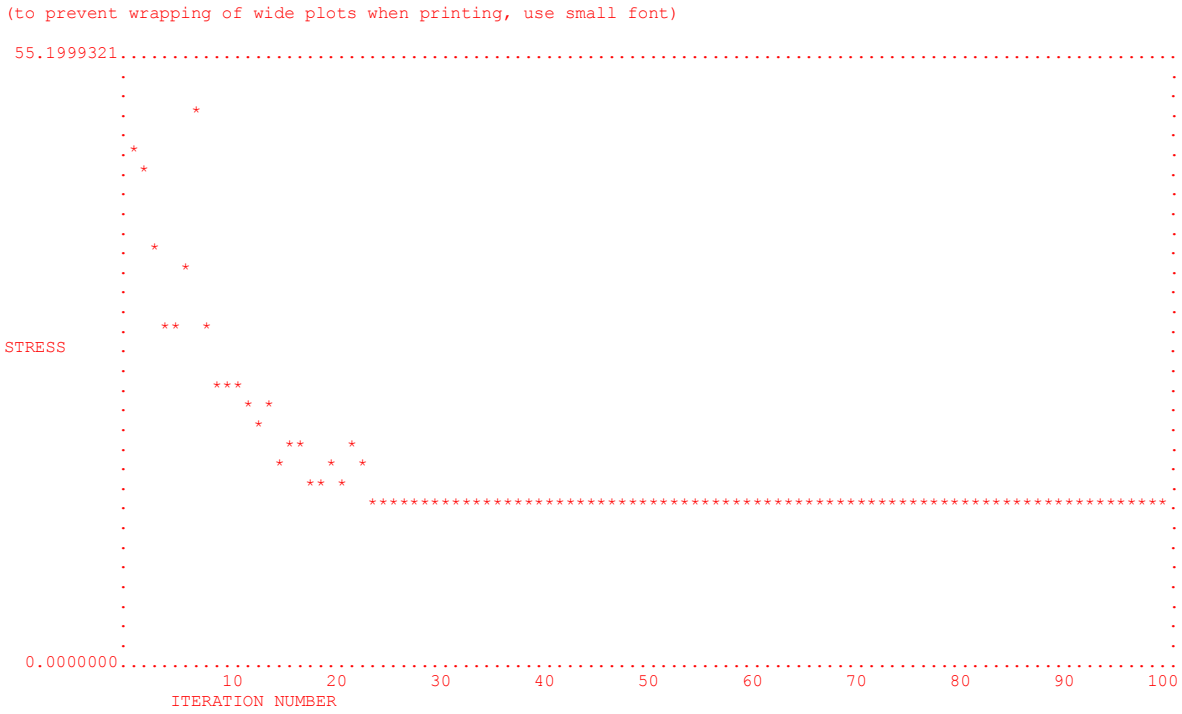
```
STRESS IN RELATION TO DIMENSIONALITY (Number of Axes)
-----------------------------------------------------------------------
          Stress in real data          Stress in randomized data
               50 run(s)               Monte Carlo test,  999 runs
```

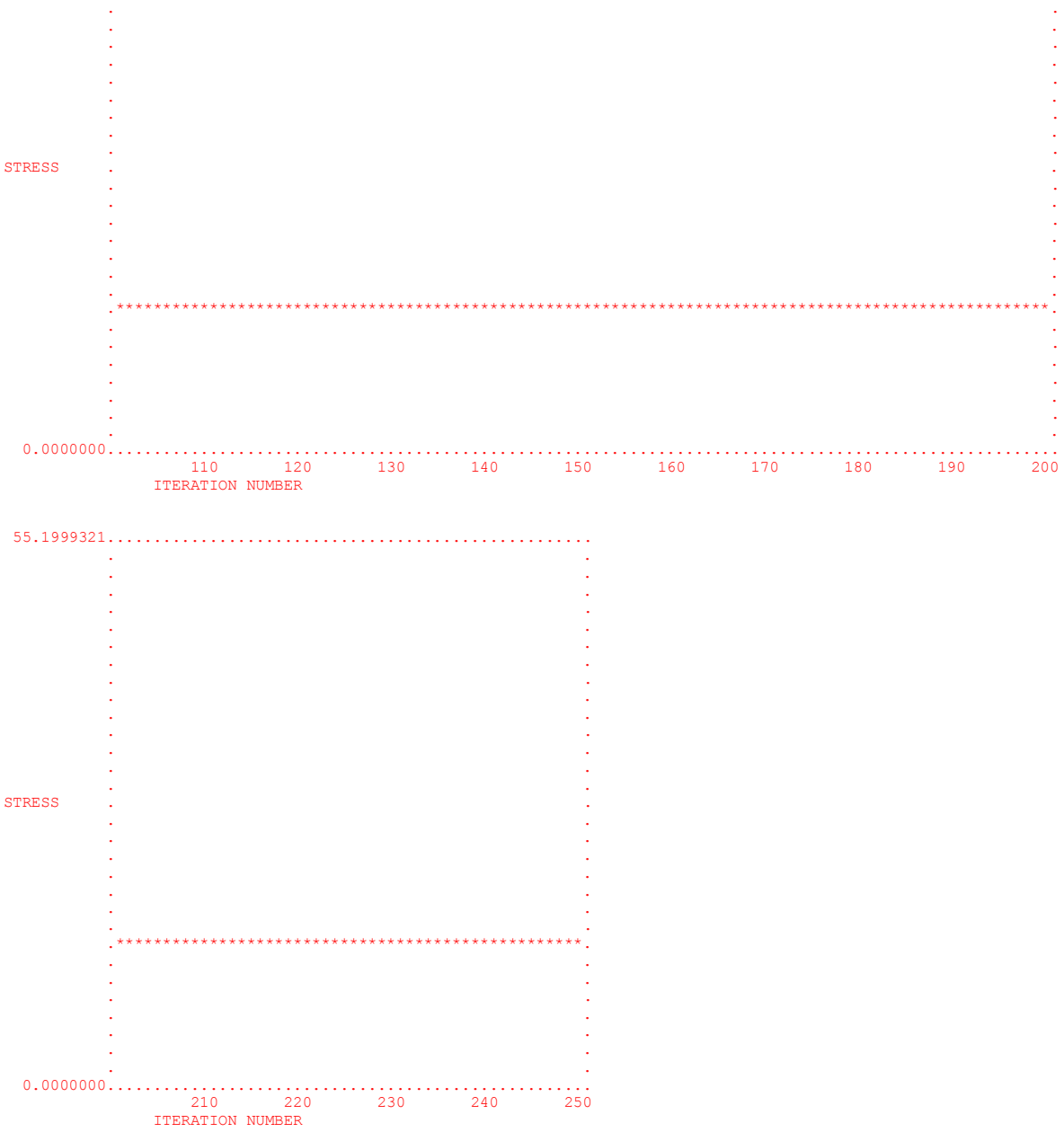
	Minimum	Mean	Maximum	Minimum	Mean	Maximum	p
Axes							
1	27.048	43.528	54.736	29.954	48.921	55.154	0.0010
2	13.217	16.205	22.751	14.104	22.952	37.843	0.0010
3	8.329	11.528	14.081	6.880	13.167	42.402	0.0150
4	5.684	9.196	13.832	3.607	8.706	41.793	0.0720
5	3.870	7.901	12.971	0.001	5.858	24.461	0.1560
6	2.109	6.393	9.437	0.000	4.411	53.154	0.1170

p = proportion of randomized runs with stress < or = observed stress
i.e., p = (1 + no. permutations <= observed)/(1 + no. permutations)

If we did not use the stress reduction rule, how many axes would the best solution have on the basis of the significance level ($p < 0.05$) of the randomizations ?
If ignoring the stress reduction rule, the best answer would have been 3 axes, as you want the minimum number of axes while not going over a level of $p = 0.05$.

Paste the PLOT OF STRESS V.ITERATION NUMBER for recommended dimension answer (Remember, your answers may vary slightly from those of your colleagues):
PLOT OF STRESS V. ITERATION NUMBER FOR 2 DIMENSIONS





Did the iterations stop before 250? Why / why not?

No, the iterations did not stop before 250. This is because the stabilization criterion had not yet been met by this time. However, we told PC ORD to stop the runs when either a) that stabilization criterion had been met, or b) if it is not met, just stop at 250 iterations, our maximum limit that we set.

Create a plot, using the "Graph \ Graph Ordination" command.
Add environmental vectors – scaled to 100%.

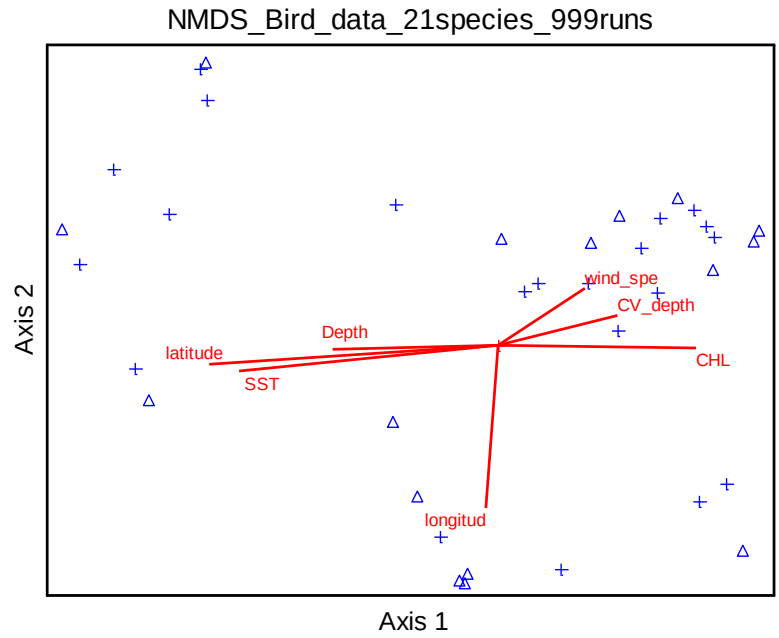


Figure 5. An ordination plot for the NMDS of the Bird Data (999 runs) with environmental vectors scaled to 100%.

Request amount of variance in distance matrix explained. Copy and paste the results below:

```
Coefficients of determination for the correlations between ordination
distances and distances in the original n-dimensional space:

            R Squared
Axis    Increment    Cumulative
 1        .389          .389
 2        .306          .695

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     0.009      100.0

Number of entities = 16
Number of entity pairs used in correlation = 120
Distance measure for ORIGINAL distance: Relative Sorensen
```

Request the correlations with the environmental variables. Copy and paste the results below:

```
Pearson and Kendall Correlations with Ordination Axes   N= 16

Axis:                1                      2
               r     r-sq     tau      r     r-sq    tau
```

longitude	-.192	.037	-.233	-.684	.468	-.433
latitude	-.911	.830	-.833	-.229	.053	-.100
wind_speed	.499	.249	.194	.406	.165	.194
SST	-.863	.744	-.845	-.275	.076	-.075
CHL	.752	.566	.610	-.076	.006	-.034
Depth	-.689	.475	-.667	-.116	.013	-.167
CV_depth	.585	.342	.533	.292	.085	.300

- Did the results change from what you found in question 2? In particular, how did the amount of variance explained change? Does this make sense? Hint: As you remove "rare" species, how does this affect the dissimilarities between samples and the total variance in the community?

The amount of variance increased in question 3b when the rare species (outliers) were removed from the analysis, but only just barely – the variance increased from a value of 0.694 to 0.695 (increase of 0.001). It makes sense that the variance SHOULD increase as rare species are removed as the dissimilarities between samples would be greatly decreased, meaning that the total explained variance amongst all species would go up.

Latitude and SST are still the drivers of Axis 1 (largest absolute tau values) with CHL and Depth close behind. In this case, these absolute values are larger, showing they have more of an influence on the value of the axis, likely because the rare species, or outliers, have been removed and no longer are affecting the axes as much. Axis 2 is still driven greatly by CVDepth, the same as in question 3a, but now longitude has the largest absolute tau value of any environmental parameter in question 3b once the outliers have been removed. Additionally, the orthogonality has increased in question 3b (from 96.2% to 100%), showing that once the outliers have been removed, the 2 axes are completely independent of each other.

3c) MNDS analysis of reduced dataset – without "rare" species. Using the criteria you selected before to identify and remove the rare species, identify and remove additional species from your list.

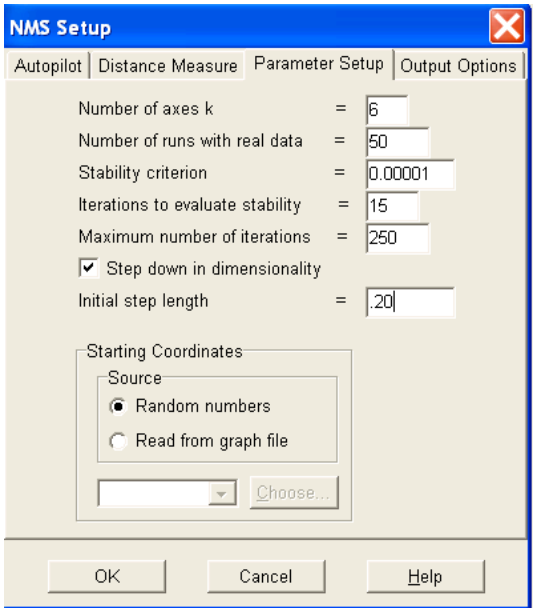
Note: I defined those species that accounted for less than 1% of the community as rare. Thus, I had to remove five more species from the dataset and was left with a sample of 16 species that, together accounted for 91.3% of all of the birds that were sighted.

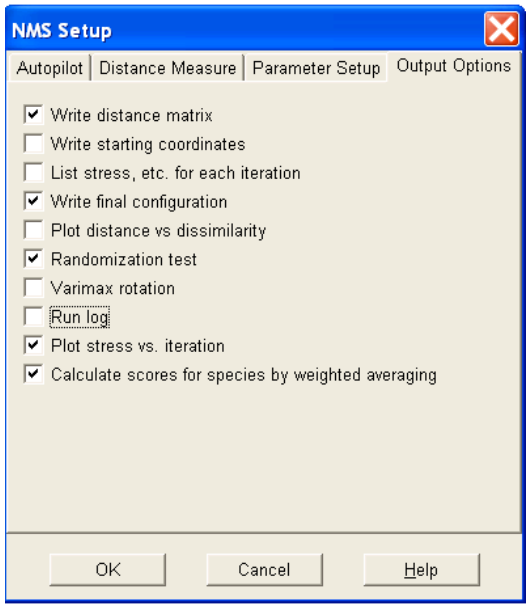
List the species you will remove from the dataset: BBAL, WAAL, SOAL, LMSA, MAPN

Report how many species and plots are included in the analysis (Hint: Did you lose any plots that were left empty when you removed those "rare" species ? ____16 species remain__

All samples still have bird counts - there are no empty plots.

Use the same parameter set-up and output as in question 3a and 3b





What distance measure would you use? Why?

I would use Relative Sorensen because the average skewness and kurtosis for these data are still very high, indicating that they are non-parametric. Also, I would want to use the same distance measure I used previously, to ensure the results were comparable with those from analysis 3a.

How many runs will you select to get a p value as small as 0.001? Explain?

Because the equation to find the p-value is p = (n+1)/(N+1), where n = number of randomized runs with a final stress that is equal to or less than the observed minimum stress, and N = number of randomized runs, it would make sense to do **999 runs** (so that if n = 1 and N = 999, p can equal 0.001). Also, I would want to use the same number of runs I used previously, to ensure the results were comparable with those from analysis 3a.

At the end of the "results.txt" file you will find the recommended solution. How many dimensions does PC-ORD recommend? _____2 Dimensions_____

Copy and paste the NMS scree plot here:

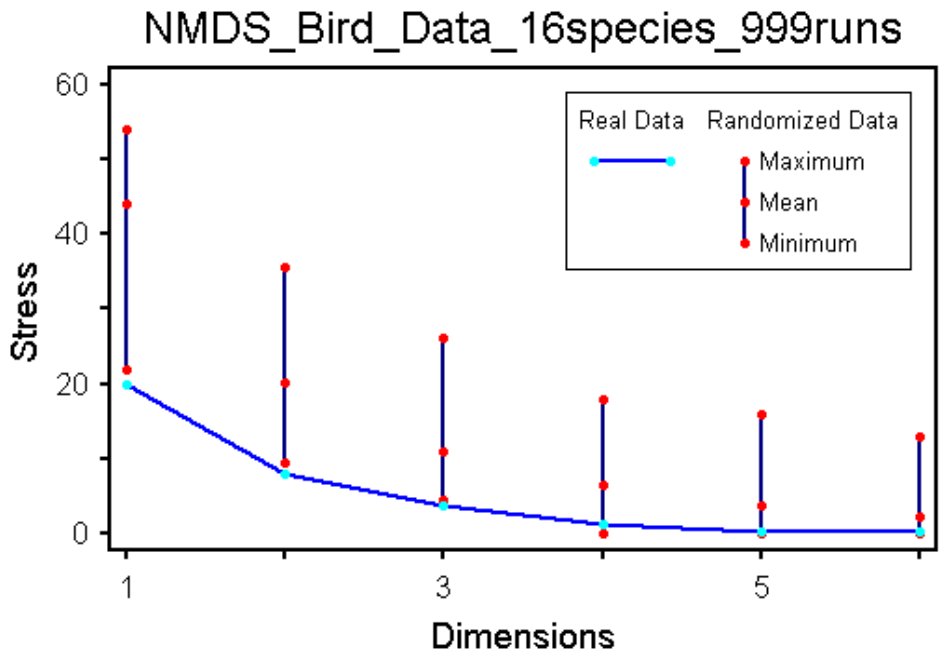


Figure 6. NMS Scree plot of 999 randomizations run for the "Bird data" using only the 16 most prevalent species.

Does this result make sense with the two "criteria" we discussed in class? Paste the `STRESS IN RELATION TO DIMENSIONALITY (Number of Axes)` table and explain your answer.

Yes, this makes sense because the minimum stress is not reduced by more than 5 by adding another axis, and the p-value for axis 2 is less than 0.05.

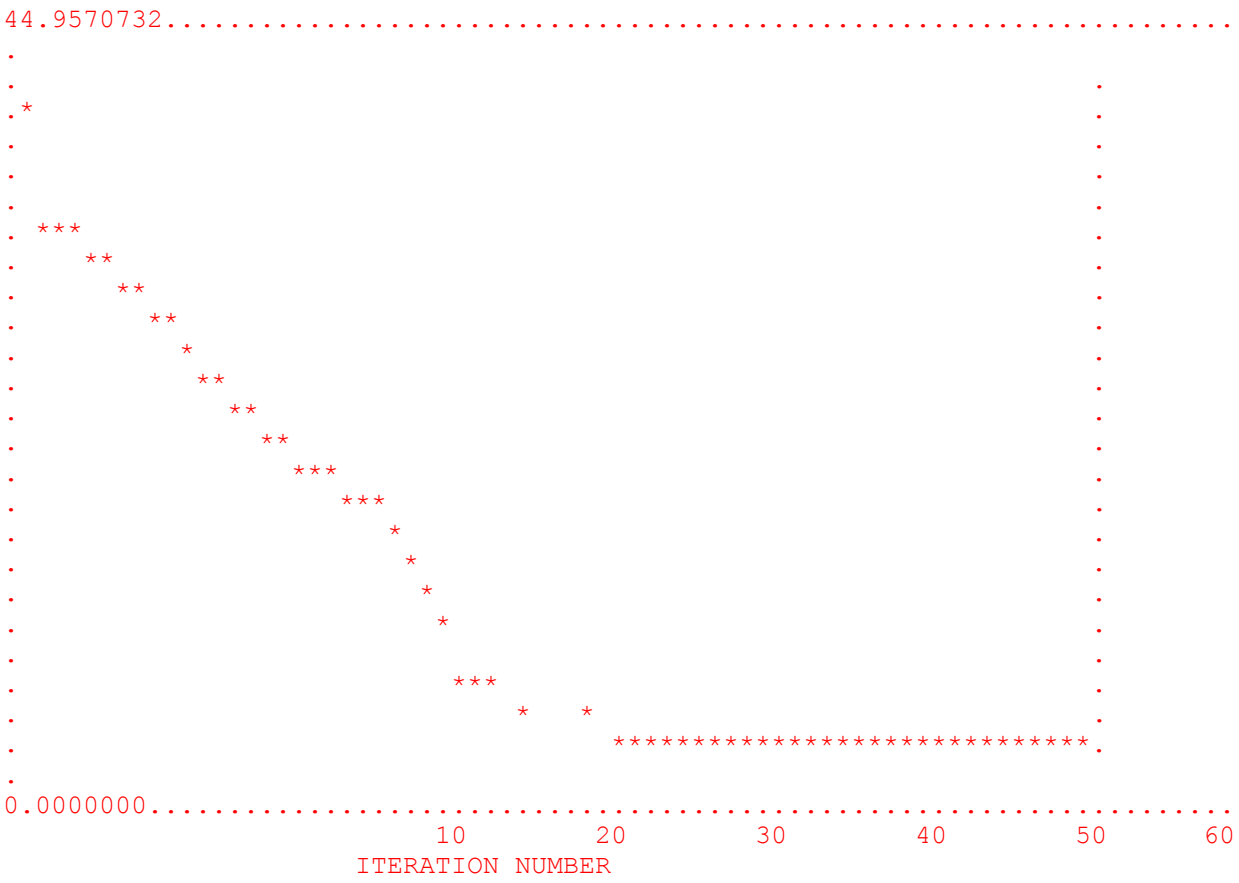
```
STRESS IN RELATION TO DIMENSIONALITY (Number of Axes)
-----------------------------------------------------------------------
          Stress in real data          Stress in randomized data
               50 run(s)                Monte Carlo test,  999 runs
      ------------------------  ---------------------------------
Axes  Minimum     Mean  Maximum  Minimum     Mean  Maximum      p
-----------------------------------------------------------------------
   1   19.796   26.829   46.547   21.842   44.175   53.980    0.0010
   2    7.935    9.412   12.609    9.579   20.175   35.667    0.0010
   3    3.790    4.198    6.953    4.535   11.071   26.106    0.0010
   4    1.178    1.511    2.775    0.003    6.409   17.863    0.0070
   5    0.369    0.842    1.596    0.011    3.827   15.874    0.0310
   6    0.209    0.493    1.065    0.012    2.166   12.914    0.0340
-----------------------------------------------------------------------
p = proportion of randomized runs with stress < or = observed stress
i.e., p  = (1 + no. permutations <= observed)/(1 + no. permutations)

Randomizing data resulted in undefined distances.
Data were reshuffled for this reason a total of    15 times.


Conclusion:  a 2-dimensional solution is recommended.
```

If we did not use the stress reduction rule, how many axes would the best solution have on the basis of the significance level ($p < 0.05$) of the randomizations ?
If ignoring the stress reduction rule, the best answer would have been 6 axes, since we want the minimum number of significant axes.

Paste the PLOT OF STRESS V.ITERATION NUMBER for recommended dimension answer (Remember, your answers may vary slightly from those of your colleagues):



Did the iterations stop before 250? Why / why not?

Yes, it stopped at 50 iterations because instability reached 0

Create a plot, using the "Graph \ Graph Ordination" command. Add environmental vectors – scaled to 100%.

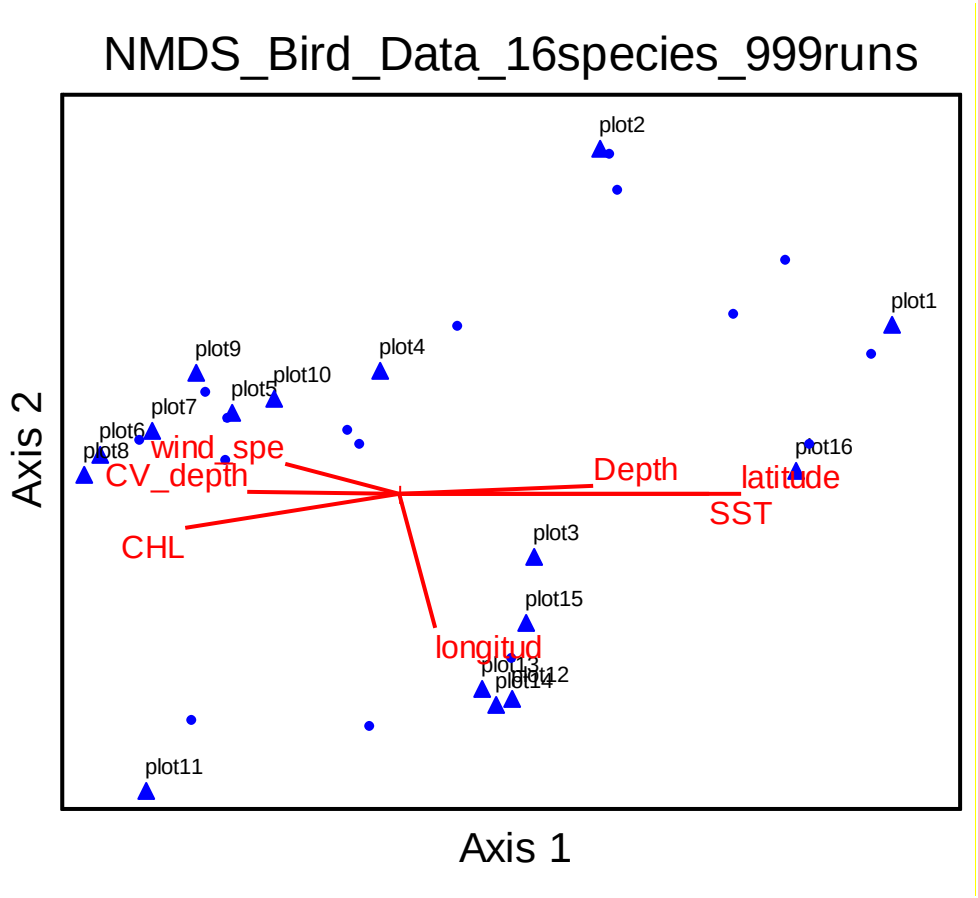


Figure 7. An ordination plot for the NMDS of the Bird Data (999 runs) with environmental vectors scaled to 100%.

Request amount of variance in distance matrix explained. Copy and paste the results below:

```
Coefficients of determination for the correlations between ordination
distances and distances in the original n-dimensional space:

            R Squared
Axis    Increment    Cumulative
 1        .413          .413
 2        .275          .687

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     0.105      98.9

Number of entities = 16
Number of entity pairs used in correlation = 120
Distance measure for ORIGINAL distance: Relative Sorense
```

Request the correlations with the environmental variables. Copy and paste the results below:

Axis:	1			2		
	r	r-sq	tau	r	r-sq	tau
longitud	.298	.089	.233	-.577	.333	-.383
latitude	.920	.847	.833	.008	.000	.083
wind_spe	-.530	.281	-.228	.273	.075	.279
SST	.875	.766	.845	-.040	.002	.109
CHL	-.728	.529	-.576	-.294	.086	-.186
Depth	.694	.482	.667	.145	.021	-.017
CV_depth	-.613	.376	-.500	.066	.004	.150

Did the results change from what you found in question 3a? In particular, how did the amount of variance explained change? Does this make sense? Hint: As you remove "rare" species, how does this affect the dissimilarities between samples and the total variance in the community?

The amount of variance increased in question 3b when the rare species (outliers) were removed from the analysis, but only just barely – the variance decreased from a value of 0.694 to 0.687 (decrease by 0.007). It does not makes sense that the variance SHOULD decrease as rare species are removed as the dissimilarities between samples would be greatly decreased, meaning that the total explained variance amongst all species would go up.

Latitude and SST are still the drivers of Axis 1 (largest absolute tau values) with CHL and Depth close behind. In this case, these absolute values are larger, showing they have more of an influence on the value of the axis, likely because the rare species, or outliers, have been removed and no longer are affecting the axes as much. Axis 2 is still driven greatly by CVDepth, the same as in question 3a, but now longitude has the largest absolute tau value of any environmental parameter in question 3c once the outliers have been removed. Additionally, the orthogonality has decreased in question 3c (from 100% to 98.9%), showing that the 2 axes are not completely independent of each other.

4) Document your data modifications:

Create a log, where you describe the manipulations you did to accomplish questions 2 and 3 above. A diagram is not needed, but make sure you provide a detailed sequential description of the steps that you took and how each step changed the dataset you were working with.

Steps	Steps Taken	How the Dataset Changed
1	Bird data was uploaded as the main matrix into PC ORD	Main matrix (16 plots, 42 species) appeared
2	Environmental data was uploaded as the secondary matrix into PC ORD	Secondary matrix (16 plots, 7 variables) appeared
3	Used PC ORD to do a row/column summary for main matrix	Dataset did not change
4	Correlation done of the 4 metrics of "diversity"	Dataset did not change
5	Number of empty rows and columns in Matrix 1 found to be 0 - all rows/columns have at least 1 non-zero value in them	Dataset did not change
6	Based on the large skewness, number of zeros, and non-normality, an NMDS was chosen to be run	Dataset did not YET change
7	The 6 species with only 1 value (in Matrix 1) were removed from the analysis to remove "outliers" which could disrupt NMDS analysis	Main matrix now has 16 plots, 36 species
8	A dominance curve of the main matrix was completed to assess the correlation of ranks of species abundance and ranks of species frequency	Dataset did not change
9	A table was created ranking each of the remaining 36 species in order from largest overall sum (most abundant) to smallest overall sum (least abundant)	Dataset did not change
10	It was determined that 21 of the 36 species make up 95% of all bird abundance	Dataset did not change
11	An NMDS was run for the entire dataset (16 plots, 36 species) with 6 axes, 50 runs with real data, stability criterion of 0.00001, 15 iterations to evaluate stability, max number of 250 iterations, initial step length of 0.2, and 999 randomization runs, using Relative Sorensen distance measure	PC ORD recommends a 2-D solution
12	Using the 2 "criteria", it was found that only 1 of the 2 criteria (p-value, and not stress reduction rule) meets the 2-D solution	Dataset did not change
13	Plot of stress vs. iteration number showed that the iterations stopped around 60 as the stability criterion was met	Dataset did not change
14	The amount of variance in distance matrix, orthogonality, and correlations with environmental variables were requested	Dataset did not change
15	15 more species removed from the main matrix so that 16 plots and 21 species remained (the 21 species making up 95% of the total abundance)	Main matrix now has 16 plots, 21 species
16	An NMDS was run for this new dataset (16 plots, 21 species) with 6 axes, 50 runs with real data, stability criterion of 0.00001, 15 iterations to evaluate stability, max number of 250 iterations, initial step length of 0.2, and 999 randomization runs, using Relative Sorensen distance measure	PC ORD recommends a 2-D solution
17	Using the 2 "criteria", it was found that both of the criteria (p-value and stress reduction rule) meet the 2-D solution	Dataset did not change
18	Plot of stress vs. iteration number showed that the iterations stopped at 250, as the stability criterion was not met and needed to run to the max	Dataset did not change
19	The amount of variance in distance matrix, orthogonality, and correlations with environmental variables were requested	Dataset did not change

5) Critical reading of the literature (1 point): Read these papers (Lopez et al. (2014), Brodeur et al. (2011), Yildirim et al. (2010), Aswani (2005)) and report the following:

* Lopez et al. 2014:

- Describe what type of data were analyzed? Where they quantitative / categorical / both?

This study used NMS to analyze the geographic variation in seal persistent organic pollutant concentrations and associations with environ- mental characteristics. Both the POP data and the environmental data analyzed were quantitative, except for the "island with most adjacent monk seal home range or core area", which is categorical.

- How many axes were involved in the "best answer"? How were they selected?

2 axes based on the stress reduction criterion of McCune & Grace (2002)

- Did the paper report p values for the axes? If they did, how many randomizations were used?

Yes, 1,000 randomizations resulted in a p value of 0.004.

- Did the paper report the variance explained by the axes? How large was the total % variance?

The proportion of the observed variance explained by the first axis was 1.9% and the second was 60.4%, with a total of 62.3% of variance explained.

* Brodeur et al. 2011:

- Describe what type of data were analyzed? Where they quantitative / categorical / both?

This study analyzed species-level differences in taxonomic composition among samples from the different gear types and from coho and Chinook salmon prey taxa. The prey taxa are categorical data, as were the tows/stomachs used as samples, but biomass was quantitative.

- How many axes were involved in the "best answer"? How were they selected?

The paper did not specify, but the figure from NMDS appeared to have 2 axes.

- Did the paper report p values for the axes? If they did, how many randomizations were used?

No they reported p value for the analysis of similarity

- Did the paper report the variance explained by the axes? How large was the total % variance?

The variance explained was not reported.

* Yildirim et al. 2010:
- Describe what type of data were analyzed? Where they quantitative / categorical / both?
3 samples of fecal matter subjected to small subunit rRNA tag sequencing from each of three different old-world monkeys – black-and-white colobus, red colobus, and red-tailed guenon, for a total of 9 points (Q). A total of 136,750 sequences were analyzed. After applying removal criteria, 41% of total sequences in 27F data set removed, 37% of the 534R and 22% of the 27F-534R datasets removed as well. 4 different NMDS tests were run – one based on the 534R dataset, one on the 27F dataset, one on the 27F-534R dataset, and one on an enriched 27F-534R dataset. All data was quantitative (Q).

- How many axes were involved in the "best answer"? How were they selected?
Not directly stated in the article, but each figure (3-6) shows in the upper corner that it was a 2-D NMDS, or 2 axes. It does not state how these axes were selected, however.

- Did the paper report p values for the axes? If they did, how many randomizations were used?
Not stated.

- Did the paper report the variance explained by the axes? How large was the total % variance?
Not reported.

* Aswani 2005:
- Describe what type of data were analyzed? Where they quantitative / categorical / both?
The proportion of members within each village with overlapping rights to estates next door (Q), cultural knowledge into agree/disagree format (C).

- How many axes were involved in the "best answer"? How were they selected?
2 axes. Axis 1 shows a progression from the "intimate" to the "alien." The 2nd axis shows a ranking from what is benevolent to that which is malevolent. The authors base this answer on the final stress level, or the stress reduction rule.

- Did the paper report p values for the axes? If they did, how many randomizations were used?
Not reported.

- Did the paper report the variance explained by the axes? How large was the total % variance?
Not reported.