COMPAH DOCUMENTATION

Eugene D. Gallagher
Associate Professor
Environmental, Coastal & Ocean Sciences Department
University of Massachusetts at Boston
100 Morrissey Blvd.
Boston MA 02125-3393

E-mail: Eugene.Gallagher@umb.edu
Web: http://www.es.umb.edu/edgwebp.htm
Voice: (617) 287-7453; Fax: (617) 287-7474
Updated: 10/4/99

TABLE OF CONTENTS

Page

List of Tables 2

List of Figures 3

Fast Introduction 3

Introduction 4
- ftp availability 5
- Register for updates 5
- Test data files 5
- Transformations 6
- Standardizations 6
  - Transformations vs. standardizations 6
  - Hypergeometric standardization 7
- Dropping species (Ndrop) & samples 9
- Comments on similarity & dissimilarity measures 9
  - Semimetrics & metrics 11
  - Synonymous indices: % Similarities 11
  - Families of indices 13
  - Canberra-Metric 15
  - Morisita & Morisita-Horn 16
  - NESS, NNESS & CNESS 17
  - Clustering species using CNESS 22
  - What NESSm sample size should you use? 22
  - Fractional data & hypergeometric probabilities 23
- Combinatorial strategies 24
  - Lance & Williams' Combinatorial equation 24
  - Compatible indices and strategies 25
  - Combinatorial strategies and indices 27

Running COMPAH96 28
- Input options 28
  - COMPAH input file format 28
    - Format statements (29)
    - Condensed and Presence/Absence forms (29)
    - Station & species labels (31)
    - List-directed input (31)
    - Full form list-directed input (32)
    - Condensed & Presence/Absence list-directed input (32)

Converting Matlab™ -mat files to COMPAH input . . . 34
Reading and clustering similarity matrices generated by Matlab™ . . . 34
From Quattro Pro to Matlab to COMPAH . . . 35
Reading an ASCII lower triangular similarity matrix . . . 35
Converting Matlab™ sparse matrices to ASCII files for COMPAH input . . . 36
Starting COMPAH96 . . . 37
Changing parameters on the COMM.DAT file or from the console . . . 37
Output options . . . 38
Printing standardized data and Matlab™ -mat files . . . 38
Matlab™ -mat files (39)
Printing a lower triangular similarity matrix . . . 39
Printing a similarity matrix as a Matlab™ .mat file . . . 39
Gower's Principal Coordinates Analysis (PCoA) (40)
Printing an AMOVA lower triangular similarity matrix . . . 41
Printing trees . . . 41
Tree format (41)
Tree examples (43)
Making graphics files from trees (44)
Reducing the output from large data sets . . . 45
Stopping the program . . . 46

Other Programs . . . 46

COMPAH, Matlab™ , and PCA-H . . . 47

References . . . 48

Technical Appendix . . . 49
Printing this document . . . 49
Comments on computation time and precision . . . 49
CNESS & NNESS use double precision math . . . 50
The 1996 sorting bug and processing time . . . 50
VAX FORTRAN modifications . . . 50
Upgrades (& Bug Fixes) . . . 51
Differences between COMPAH95 & COMPAH96 . . . 51
List of modifications . . . 52
Run-time and compilation errors . . . 53
Compiling the source code with WATCOM F77/32 . . . 54
Virtual Memory Management using DOS4GVM . . . 56
COMPAH96 and Windows . . . 57

Index . . . 58

LIST OF TABLES

Table 1 Pielou's (1984) data on testdata.exe and options needed to reproduce the text figures. . . . 5
Table 2 COMPAH's transformation options, set by editing the **COMM.DAT** file or interactively. . . . 6
Table 3 COMPAH's Standardization options, set by editing the **COMM.DAT** file or interactively at the start of a COMPAH analysis. The standardized data can be printed to a separate file. . . . 6
Table 4 COMPAH's (dis)similarity indices. Only 1 through 16 are shown on the **COMM.DAT** file, but the other similarity indices can still be used. . . . 9
Table 5 COMPAH's families of similarity and dissimilarity indices . . . 13
Table 6 Comparison of COMPAH similarities with Magurran's Example 15. . . . 19
Table 7 COMPAH's combinatorial sorting strategies. . . . 24
Table 8 Values of the parameters for Lance & Williams' (1967) combinatorial equation (Equation 17) . . . 25
Table 9 Different versions of COMPAH and their system requirements . . . 55

LIST OF FIGURES

Page

Figure 1. The geometry of CNESS. The positions of samples are plotted in species space according to their hypergeometric probabilities. The position of sample points at random sample sizes from 1 to 16 are shown. The chord distances are shown for m=10. These distances can be viewed by connecting a vector from the origin to the $H_{ij|m}$ coordinate and then marking the point 1 unit from the origin. The Euclidean distances among these points, the projections of the samples onto the unit hypersphere, are the CNESS distances. These sample coordinates on the hypersphere are the row-normalized elements of the **H** matrix. 21

Figure 2. A single-spaced tree produced by **PTREE=1**. With HP pcl drivers, print the tree with a Courier New or Courier T1 7.5 pitch (non-proportionally spaced font). Linespacing was set to 1.5 43

Figure 3 A single-spaced tree, composed of graphic characters, produced by **PTREE=2**. This tree can be printed with the Courier New 7.5 pitch font for non-postscript printers and with Courier T1 or Lotus Linedraw (available with Adobe Type Manager) for encapsulated postscript printers. 43

Figure 4 A double-spaced tree produced by **Ptree=22**. This tree can be printed using either single-spacing or ½- line spacing (this tree produced with 0.9 line spacing, Courier New 9.0 font) for non-postscipt printers and with Lotus Linedraw (available with Adobe Type Manager) for encapsulated postscript printers. 44

Figure 5. The same tree from the previous display, converted to a vector image file using WordPerfect 8.0 and Corel Presentations 8. One way to get a vector based-tree tree is to scan the tree as a bitmapped image (bmp or tif) and convert it to a vector-based graphic with Corel's "trace bitmap" tool . The lettering must be retyped in Presentations. 45

Figure 6. The same tree from the previous display, with the colors changed to blue, the size reduced and with WordPerfect borders and fills.. 45

FAST INTRODUCTION

The following instructions are for those who want to run their data quickly:

1. Place the COMPAH96.EXE, COMM.DAT, TEST.DAT and DOS4GW.EXE files on your hard disk in the same folder (subdirectory). If you obtained the program as self-executing zip files, then copy COMPAH.EXE and TESTDATA.EXE (self-extracting zip files) to a folder on your hard disk. In Windows, double-click on these files to extract them. In DOS, just type COMPAH or TESTDATA at the command line prompt to start the self-extraction.
2. In DOS, type COMPAH96 at the DOS prompt to run COMPAH96. In Windows, double-click on COMPAH96 using Explorer or program manager. The program will prompt you for the name of input files. Enter TEST.DAT at the prompt as the name of the input file. Hit return at most prompts to use the default parameters. Type TEST.OUT as the name of the output file. Type yes to overwrite the existing TEST.OUT program or type a new name, *e.g.*, TEST2.out
3. Examine the output and modify your data to fit COMPAH input format. You must save your data files as an ASCII text file for input to COMPAH. Notepad uses this as a default, and all other word processors have ASCII text as an output option. Change the COMM.DAT file to fit the type of analysis that you wish to run and save as an ASCII text file.

INTRODUCTION

COMPAH stands for COMbinatorial Polythetic Agglomerative Hierarchical clustering. COMPAH96 is an interactive clustering package for personal computers using DOS or Windows and can be easily modified to run with VAX FORTRAN. COMPAH clusters using Lance & Williams' (1967) combinatorial sorting equation. I adapted the PC version of this program from an earlier version called COMPAH written by Don Boesch and described in Boesch (1977). Several of the major algorithms are based on Anderberg (1973), but his algorithms have been greatly modified to enhance computation speed and memory allocation. The COMPAH input format and space allocation strategy is that used in Cornell University's ORDIFLEX program.

Barbara Hecker (Falmouth MA) and Fred Grassle (Rutgers University) made some minor changes to Boesch's program. I updated the original FORTRAN code to conform to standard FORTRAN77. I compiled COMPAH96 using WATCOM FORTRAN F77/32 and placed comments in the source code showing the changes needed to compile the program with VAX FORTRAN (Version 5.00) and Microsoft FORTRAN. I distribute this program and source code free of charge. Contact me at the above address if you experience any difficulties. I continually upgrade the program with new features, so you may want to check my web page for updates:
http://www.es.umb.edu/edgwebp.htm

The program includes options for six data transformations, fifteen standardizations, clustering of species (=variables) or samples, 30 similarity/dissimilarity indices, and Lance & Williams' eight Combinatorial clustering methods. COMPAH can calculate NNESS, a modification of Grassle & Smith's (1976) NESS (a similarity index for clustering samples). The program also includes CNESS, an acronym for Chord-distance Normalized Expected Species Shared. CNESS is a metric distance version of NNESS that I developed in May 1992. The first description of CNESS and NNESS appears in Trueblood *et al.* (1994). I also describe the differences here.

COMPAH prints a cluster diagram that is adequate for viewing the cluster patterns (PTree=1). A higher quality tree can be printed using PTree=2 or 22, but this tree uses ASCII graphics characters that cannot be printed on computer mainframe line printers. The program provides the exact clustering levels for producing publication-quality trees. One COMPAH option (PrSmM=4 or 5) prints the similarity or dissimilarity matrix in lower triangular form (with no main diagonal elements) in a separate file for export to other programs. There are also options to export COMPAH data to MATLAB™ and other programs.

I cannot guarantee the results of COMPAH, but to my knowledge all results are correct. The last bug that I've found in COMPAH was in 1996. This bug was very unlikely to have affected any results. This logical error, described in the technical appendix (p. 49) had been in all previous versions of COMPAH but is corrected in COMPAH96.

I welcome any comments from users of the program and will answer any questions about problems analyzing data. Please send me part of your data file so I can identify the problem.

ftp availability

This document, the FORTRAN source code, test data sets and a variety of compiled and linked executable COMPAH programs (described in the Table 55 on p 55) are available by anonymous ftp via Gallagher's home page: http://www.es.umb/edgwebp.htm

Register for updates

Please send me an E-mail message if you use COMPAH. I do not charge for COMPAH, and I distribute the source code with the program. In return, I will send you a notification when I update COMPAH.

Test data files

I have enclosed copies of Pielou's (1984) example data on the distribution files. Table 1 shows the combination of COMPAH options needed to reproduce her clusters.

Table 1 Pielou's (1984) data on testdata.exe and options needed to reproduce the text figures.

Pielou's Figure	Data File	Stand-Simop-ClusM	Comment
Fig 2.3	ECPMAT1.DAT	0-7-3	Single-linkage clustering with Euclidean distance.
Fig 2.4	ECPMAT2.DAT	0-7-3	"
Fig 2.5	ECPMAT1.DAT	0-7-4	Complete linkage=farthest-neighbor with Euclidean distance
Fig 2.6	ECPMAT1.DAT	0-13-8	Centroid clustering with Euclidean distance squared. Fig. uses √(clustering level), available by console prompt.
Fig 2.12a	ECPMAT4.DAT	0-7-8	Centroid clustering of Euclidean distances. Pielou switched 5 & 6, an error in her Fig. COMPAH cannot display the non-monotonic fusion of sample 11.
Fig 2.12b	ECPMAT4.DAT	0-12-8	Centroid clustering of Geodesic metric.
Fig 2.15a	ECPMAT5.DAT	0-7-3	Nearest-neighbor clustering of Euclidean distance.
Fig 2.15b	ECPMAT5.DAT	0-14-3	Single-linkage clustering with Marczewski-Steinhaus distance (1-Jaccard's)
Fig 2.17A	ECPMAT6.DAT	0-13-8	TREE printed with √(CLUSTER LEVEL)
Fig. 2.17B		0-13-5	TREE printed with √(CLUSTER LEVEL)
Fig. 2.17C		0-7-1	
Fig 2.17D		0-7-2	

Transformations

COMPAH has six transformation options, shown in Table 2. Many papers have used 4th-root transformed data with Bray-Curtis similarity. To specify such an analysis, use TRANS=4, SIMOPT=6 (see Table 4, 7, 7, 8, p. 9, 24, 38).

Table 2 COMPAH's transformation options, set by editing the **COMM.DAT** file or interactively.

Trans	Action	Trans	Action
0	No transformation	4	4th root (x)
1	log10 (x+1) transform	5	Boolean, converts quantitative data to 0 or 1
2	square root (x)	6	ROUND, Rounds fractional data to nearest integers
3	cube root (x)		

COMPAH can print the transformed data in COMPAH format, which is identical to the Cornell University ORDIFLEX data input. To print the transformed data, set PrTrD=1 in the **COMM.DAT** file or change PrTrD at the console prompt.

Standardizations

Transformations vs. standardizations

Transformations can be performed on each element of a data matrix independent of the other elements. Standardizations modify the elements to meet boundary conditions based on all of the elements in a row, column or both the row and column.

COMPAH has fifteen standardization options, shown in Table 3.

Table 3 COMPAH's Standardization options, set by editing the **COMM.DAT** file or interactively at the start of a COMPAH analysis. The standardized data can be printed to a separate file.

STAND	Action	STAND	Action
0	None	8	TOTAL ON SAMPLES (divide by each sample's total);
1	MAX ON SPECIES (Divide by largest species abundance);	9	SIMULTANEOUS (divide by SQUARE ROOT (sample total* species. total);
2	MEAN ON SPECIES (Divide by mean species abundance);	10	HYPERGEOMETRIC (RND) (=Probability of sampling species j in sample i with a random draw of m individuals without replacement; m is set by NESSm). Species abundance data are rounded to the nearest integers)
3	MEAN ON SAMPLES (Divide by mean sample abundance);	11	RANGE (Divide each abundance by RANGE for each species)

Table 3 COMPAH's Standardization options, set by editing the **COMM.DAT** file or interactively at the start of a COMPAH analysis. The standardized data can be printed to a separate file.

STAND	Action	STAND	Action
4	SPECIES NORMALIZATION (The sum of squared abundances of each species across samples ill equal 1.0);	12	HYP (RND) & NrmSt. Round species data to nearest integers. Convert to hypergeometric probabilities (STAND=10), then normalize by station (STAND=5).
5	NRMST; STATION NORMALIZATION After, station or sample normalization, the sum of squared species abundances in each sample will equal 1.0. This is equivalent to projecting the station vectors onto the unit hypersphere in species space.	13	Z-standardization. Each species abundance is converted to a Z-normal deviate (subtract mean species abundance and divide by sample standard deviation).
6	STANDARD DEVIATION (divide by each species' standard deviation);	14	HYPERGEOMETRIC (CONT). As STAND=10, but species data not rounded to nearest integers. The continuous ln(gamma) distribution used to calculate factorials for fractional data.
7	TOTAL ON SPECIES (divide by each species' total);	15	HYPERGEOMETRIC (CONT) & NormSt. As STAND=12, but species data not rounded to nearest integers. The continuous ln(gamma) distribution used to calculate factorials for fractional data.

Clifford & Stephenson (1975) and Boesch (1977) describe the first nine standardizations. I discuss the hypergeometric standardization in the next section.

The standardization choices are set in the comm.dat file, and the standardization options can be set interactively at the start of a COMPAH run via the console. After standardizing the data, COMPAH will prompt the user asking whether it should use an additional standardization. If the answer is **y** or **Y**, COMPAH displays the command option file for selection of a new standardization option. This option allows the user to use several standardizations on the same data. For example, the user might wish to do a sample normalization (STAND=5) followed by centering by species (STAND=2). COMPAH prints the name of only the last standardization on the output.

Hypergeometric standardization

For cluster analysis with NNESS or CNESS, STAND=0 should be used. However, there are occasions when the user may wish to calculate the hypergeometric statndardizations independently of their use in calculating NNESS and CNESS. STAND=10, 12, 14 or 15 perform hypergeometric standardizations. The hypergeometric standardization converts the abundance of species j in sample i ($DATA_{ij}$) to the probability of sampling species j in sample i with a random draw of 'NESSm' individuals without replacement. Hypergeometric probabilities are calculated using the ratios of binomial coefficients, $\binom{N}{k}$. The binomial coefficient in the numerator provides the number of ways of drawing NESSm individuals from sample i without sampling species j. The denominator provides the total number of ways of drawing NESSm individuals from sample i. One minus this ratio is the probability of sampling species j in sample i with a random draw of NESSm individuals. COMPAH calculates hypergeometric

probabilities in double precision using the natural log of the gamma distribution. The formula for calculating these hypergeometric probabilities H_{ij} is shown in equation 1:

$$H_{ik\,/m} = 1 - \frac{\binom{Total_{i.} - x_{ik}}{m}}{\binom{Total_{i.}}{m}}$$

$$= 1 - \left[\frac{\frac{(Total_{i.} - x_{ik})!}{m!\ (TOTAL_{i.} - x_{ik} - m)!}}{\frac{Total_{i.}!}{m!\ (TOTAL_{i.} - m)!}} \right]. \qquad \textbf{(1)}$$

where, $Total_{i.}$ = *the sample total.*
x_{ik} = *the abund. of species k in sample i.*
m = *NESSm=No. of ind. to be drawn at random.*
! = *factorial.*

The random sample size m, called NESSm in COMPAH, can be set in advance in the **COMM.DAT** file (the ASCII file containing default parameters) or interactively at the console prompt. NESSm must be an integer. COMPAH rounds all species abundance data to the nearest integers before calculating hypergeometric probabilities.

The hypergeometric probabilities in equation (1) can be combined in a matrix to produce what I call the **H** matrix:

$$\mathbf{H} = \begin{bmatrix} H_{11|m} & H_{12|m} & \cdots & H_{1K|m} \\ H_{21|m} & H_{22|m} & \cdots & H_{2K|m} \\ \vdots & \vdots & \ddots & \vdots \\ H_{I1|m} & H_{I2|m} & \cdots & H_{IK|m} \end{bmatrix}. \qquad \textbf{(2)}$$

where, H_{ij} = *Hypergeometric probabilites.*
= *Prob. (species k in sample i | m).*
m = *random sample size = NESSm.*
K = *number of species.*
I = *number of samples.*

The **H** matrix can be used to calculate diversity, similarity and distance matrices. Hurlbert (1971) and Smith & Grassle (1977) describe the use of hypergeometric probabilities in calculating Sanders' (1968) rarefied species diversity, $E(S_n)$. Rarefied species diversity is simply the row sum of **H**. Grassle & Smith (1976) describe how hypergeometric probabilities s can be used to calculate the NESS faunal similarity index based on the expected species shared between random draws of individuals from samples. These expected species shared estimates, in a matrix form called **ESS**, are calculated by **ESS = H*H'**, where **H** is the station x sample matrix shown in Equation (2).

Dropping species (Ndrop) & samples

COMPAH has an option to drop rare species. By setting Ndrop in the **COMM.DAT** file to 1, COMPAH will drop all species that occur in only one sample. COMPAH automatically drops all species that do not occur in any samples and drops all stations with 0 individuals. The only exception to this is if Euclidean distance, Euclidean distance squared, or 1/(E.D.^2) are specified. I sometimes use COMPAH to create lower-triangular matrices of Euclidean distances among samples for input to other programs, and I wanted to retain samples that have 0 distance between them.

I never drop species when using NNESS or CNESS. 'Singleton species', or species that occur in only a single sample can play a significant role in the clustering patterns using NNESS and CNESS.

Comments on similarity & dissimilarity measures

COMPAH can calculate 30 similarities or dissimilarities, listed in Table 4. The beginning user should consult a statistical ecology text for recommendations on the choice of similarity or dissimilarity indices. Clifford & Stephenson (1975) and Boesch (1977) describe the first nine indices. The user could consult these references for the equations used in COMPAH (or look at the FORTRAN code in SUBROUTINES SIM1-SIMG). Legendre & Legendre (1983 Chapter 6, 1998 Chapter 7) provide lists of similarity and dissimilarity indices. Pielou (1984) describes the chord distance, geodesic metric, Euclidean distance squared metric, percentage dissimilarity (=1-Bray-Curtis) and percentage remoteness. Magurran (1988) provides worked examples for many similarity and dissimilarity indices. I discuss NESS and CNESS below.

Table 4 COMPAH's (dis)similarity indices. Only 1 through 16 are shown on the **COMM.DAT** file, but the other similarity indices can still be used. I have included a few references for each similarity index. Column 3, **L&L**, shows the designation from Legendre and Legendre (1983, 1998).

SIMOPT	Name	L&L	Description
			Binary
1	J	S_7	Jaccard's similarity (complement is metric)
2	S	S_8	Sørensen's presence/absence [=DICE=Czekanowski's binary]
3	Oc	S_{14}	Ochiai (Boesch 1977)
			Quantitative
4	r		Pearson's r
5	CM	D_{10}	Canberra Metric
6	B		Bray-Curtis similarity [=Pielou's (1984) percentage similarity] S_{17} in Legendre & Legendre (1983) (attributed to Steinhaus) (complement is not a metric)
7	E	D_1	Euclidean Distance (a metric)
8	M		Morisita (1959) similarity; Boesch (1977), Grassle & Smith (1976) (complement is not a metric)
9	%		Sanders' (1960) %-age similarity (dominance affinity) (Boesch 1977) (complement not metric)

Table 4 COMPAH's (dis)similarity indices. Only 1 through 16 are shown on the **COMM.DAT** file, but the other similarity indices can still be used. I have included a few references for each similarity index. Column 3, **L&L**, shows the designation from Legendre and Legendre (1983, 1998).

SIMOPT	Name	L&L	Description
10	N		NewNESS=NNESS; New version of Grassle & Smith's (1976) NESS described in Trueblood et al. (1994) (complement is not a metric)
11	C	D_3	Orloci's chord distance; Orloci (1978), Pielou (1984) = Euclidean distance standardized by stand vector (Orloci 1967, Greig-Smith 1983) (a metric).
12	G	D_4	Geodesic Metric
13	E2		Euclidean distance squared. (not a metric)
14	MS		Marczewski-Steinhaus (=1-Jaccard's Index) (Pielou 1984) (a metric)
15	CN		CNESS, Chord-distance Normalized Expected Species shared (Trueblood *et al.* 1994, defined below) (a metric)
16	MH		Morisita-Horn Magurran (1988, p. 95) (complement not a metric)
17	PD	D_{14}	Percentage dissimilarity (Pielou 1984, p. 44; =1-Percentage similarity = Bray-Curtis dissimilarity
18	PR		Percentage remoteness (Pielou 1984, p. 44. =1-Ruzicka similarity) (a metric)
19	RU		Ruzicka similarity (Pielou 1984, p. 44; Boesch 1977, p. 26) (complement [SIMOPT=18] is a metric)
20	E^{-2}		(Euclidean distance)$^{-2}$ for physical structure matrices, not species counts. I use this option to convert x,y Cartesian coordinates to inverse Euclidean distance squared matrices needed for some spatial statistics analysis. A semimetric (Legendre & Legendre 1998, p. 277)
21	CC		COMPLEMENT OF CNESS, √(2)-CNESS.
22	ED^{-1}		1/Euclidean distance for physical structure matrices
23	CN^2		CNESS^2, the square of CNESS distances among samples
24	CB	D_7	City-block or Manhattan metric
25	CT		Cos (theta). Cosine of theta, where theta is the angle between species or sample vectors (Pielou 1984, p. 48-49)
26	CN		CNESS calculated with the continuous ln(gamma) distribution. Fractional data not rounded prior to calculation of CNESS. Identical to SIMOPT=15 for integer data. (metric)
27	N		NNESS calculated with the continuous ln(gamma) distribution. Fractional species abundance data not rounded prior to calculation of CNESS. Identical to SIMOPT=10 for integer data.(Semimetric)
28	R2		$\cos^2\theta$ = cosine squared theta dissimilarity
29	K	S_{18}	Kulczynkski's similarity coefficient
30	KC		1-Kulczynkski's similarity coefficient, Kulczynski Complement

Semimetrics & metrics

Most of the similarity and dissimilarity indices in COMPAH are only semimetrics. Metrics must always meet the triangular inequality axiom. The triangular inequality states that for any three samples, i, j, and k:

$$dissimilarity_{ik} \leq (dissimilarity_{ij} + dissimilarity_{jk}). \quad \textbf{(3)}$$

For similarity indices with a range of 0 to 1.0:

$$1 - similarity_{ik} \leq ([1 - similarity_{ij}] + [1 - similarity_{jk}]). \quad \textbf{(4)}$$

A simple data set that shows violations of the triangular inequality in many indices is:

```
    5    3Legendre's semimetric test data*10, p. 200                  SV
(3(1X,F3.0))
 10.  0. 10.
 10.  0. 10.
 10.  0. 10.
  0. 10. 10.
  0. 10. 10.
Species Y1
Species Y2
Species Y3
Species Y4
Species Y5
Sample X1
Sample X2
Sample X3
```

Pielou (1984, p. 43) discusses the advantages of using metrics instead of semimetrics. For example, Torgerson's metric scaling (=Gower's 1966 Principal coordinates analysis=PCoA) works better with metric distances. A PCoA of semimetric distances often produces negative eigenvalues that complicate interpretation (Gower 1967). Even non-metric multidimensional scaling (NMDS) can produce a better low-dimension fit to data (lower STRESS) if the similarity or dissimilarity index meets the triangular inequality axiom.

Often, a metric equivalent exists for a semimetric index. For example, Ruzicka's similarity is a metric equivalent of Bray-Curtis similarity (Pielou 1984). CNESS is a metric equivalent of Grassle & Smith's (1976) NNESS.

Synonymous indices: % Similarities

Ecologists often use different names to describe the same index. For example, the following names are all synonyms for the same index: Czekanowski's binary, Sorensen's index, and the Dice index (Clifford & Stephenson 1975, p. 55). COMPAH calls it Sorenson's index.

Sometimes, ecologists call different similarity indices by the same name. For example, COMPAH's % similarity (SIMOPT=9) is different from Pielou's (1984) percentage similarity (SIMOPT=6). Pielou's percentage similarity is known by several synonyms: Czekanowski's quantitative index, Sorensen's quantitative index (Magurran 1988, p. 95), Legendre & Legendre's (1983, 1998) Steinhaus index (S_{17}) and COMPAH's Bray-Curtis similarity index (SIMOPT=6). Legendre & Legendre (1998, p. 265) describe the history of this index, which they call S_{17}:

> *"...the best known is a coefficient attributed to the Polish mathematician Steinhaus by Motyka (1947). This measure has been rediscovered a number of times; its one-complement is known as the Odum or Bray-Curtis coefficient. It is sometimes incorrectly attributed to Czekanowski …"*

This index, usually called the Bray-Curtis similarity, is shown in Equation 5:

$$\begin{aligned} \textit{Bray–Curtis similarity} &= \textit{Steinhaus similarity} = \textit{Pielou's \% Similarity.} \\ &= \textit{Sorensen's quantitative} = \textit{Czekanowski's similarity.} \end{aligned}$$

$$SIM_{ij} = \frac{2\sum_{k=1}^{S}\min(x_{ik}, x_{jk})}{\sum_{k=1}^{S}(x_{ik} + x_{jk})}. \qquad \textbf{(5)}$$

where, S = *Number of species.* x_{ik} = *Abundance of species k in sample i.*

Legendre & Legendre's (1983) Czekanowski dissimilarity (their D_8) is not programmed in COMPAH96. Czekanowski's quantitative dissimilarity is related to both % similarity and the Manhattan or city-block metric (shown below). It can be described as:

$$D_{ij} = \frac{1}{S}\sum_{k=1}^{S}|x_{ik} - x_{jk}| \; . \qquad \textbf{(6)}$$

where, D_{ij} = *Czekanowski's* (1909) *mean character difference for samples i and j.*
S = *Number of species.*
x_{ik} = *Abundance of species k in sample i.*

Kulczynski's similarity (SIMOPT=29) resembles Pielou's % similarity and COMPAH's % similarity (shown below) and can be regarded as intermediate between the two:

$$\textit{Kulczynski's Similarity} = SIM_{ij} = \frac{1}{2}\left(\frac{\sum_{k=1}^{S}\min(x_{ik}\ ,\ x_{jk})}{\sum_{k=1}^{S}x_{ik}} + \frac{\sum_{k=1}^{S}\min(x_{ik}\ ,\ x_{jk})}{\sum_{k=1}^{S}x_{jk}} \right). \qquad \textbf{(7)}$$

where, S = *Number of species.*
x_{ik} = *Abundance of species k in sample i.*

The complement of Kulczynski's similarity ($1\text{-}SIM_{ij}$) is available as SIMOPT 30.

COMPAH's % Similarity, which is the one used routinely by benthic ecologists, is different from Pielou's (1984) % similarity. Sanders (1960) introduced COMPAH's % similarity to benthic ecologists under the name "dominance affinity." COMPAH's % similarity (STAND=0 and SIMOPT=9) equals Pielou's percentage similarity only if the species abundances are first standardized by the sample totals (STAND=8 and SIMOPT=6) (see Boesch 1977, p. 26):

$$\%\ \mathrm{SIM}_{ij} = \sum_{k=1}^{S} \min \left(\frac{x_{ik}}{\sum_{k=1}^{S} x_{ik}} , \frac{x_{jk}}{\sum_{k=1}^{S} x_{jk}} \right) . \qquad \textbf{(8)}$$

where, x_{ik} = *Abund. of species k in sample i;* S = *Number of species.*

Families of indices

The indices in COMPAH can be grouped into families of indices. Within a family, all indices can be calculated from the others through simple transformations or standardizations of the data. The families are described in Table 5. The Minkowski metric family includes Euclidean distance and the city-block metric. Minkowski metrics are solved with the same equation:

$$\begin{gathered} \textit{Minkowski Metrics} \\ D_{ij} = \left(\sum_{k=1}^{S} |(x_{ij} - x_{jk})|^{R} \right)^{R^{-1}} . \\ R = 1 \\ \textit{Manhattan metric} = \textit{City-block metric.} \\ R = 2 \\ \textit{Euclidean distance.} \end{gathered} \qquad \textbf{(9)}$$

Chord distance and CNESS are Minkowski metrics, being the Euclidean distances among rows of the data or H matrix after row normalization (STAND=5). One of these families is a class of geometric indices. All similarities and dissimilarities can probably be interpreted geometrically, but the listed set of indices can be defined explicitly as distances or angles in geometric displays.

Table 5 COMPAH's families of similarity and dissimilarity indices (SIMOP value in parentheses)

THE JACCARD FAMILY (METRIC)		
	Presence/absence	**Quantitative**
Similarity	Jaccard (1)	Ruzicka (19)
Dissimilarity	Marczewski-Steinhaus (14)	Percentage remoteness (18)

Table 5 COMPAH's families of similarity and dissimilarity indices (SIMOP value in parentheses)

THE SORENSEN FAMILY (SEMIMETRIC)		
UNTRANSFORMED	**Presence/absence**	**Quantitative**
Similarity	Sorensen=Dice (2)	Bray-Curtis (6), Kulczynski (29)
Dissimilarity	None	% dissimilarity (17), 1-Kulczynski (30)
AFTER NORMALIZATION BY SAMPLE TOTAL		Percent similarity (9)
AFTER HYPERGEOMETRIC STAND. WITH ROUNDING		NNESS (10), NESS
AFTER HYPERGEOMETRIC STAND., NO ROUNDING		Morisita-Horn (16), NNESS (27)
THE MINKOWSKI METRIC FAMILY		
		Manhattan metric (24), Euclidean distance (7)
AFTER SAMPLE NORMALIZATION (STAND=5)		Chord distance (SIMOPT=11)
THE PROBABILISTIC MEASURES		
METRIC	**Presence/absence**	**Quantitative**
Similarity	√2-CNESS (21) after Boolean Transform (TR=5)	√2-CNESS (21)
Dissimilarity	CNESS after Boolean (TR=5,SIM=15) Ochiai (1)	CNESS (15 or 26)
Semimetric	**Presence/absence**	**Quantitative**
Similarity	NNESS after Boolean (TR=5, SIM=10), Morisita-Horn (16), Morisita (8)	NNESS (10 or 27)
THE GEOMETRIC FAMILY		
Dissimilarity	Chord distance (11), CNESS (15), Euclidean distance (7), Geodesic metric (12)	
Similarity	NNESS (10) Cos(theta), Pearson's r (4)	

You can calculate some similarity measures using more than one method. COMPAH calculates Sorensen's index (=DICE, SIMOPT=2) by combining the Boolean transform (TRANS=5) and Bray-Curtis similarity (SIMOPT=6). COMPAH calculates Jaccard's similarity by combining the Boolean transform with Ruzicka's quantitative similarity.

Orloci's chord distance (SIMOPT=11) is the same as normalizing the data by station (STAND=5) and then calculating the Euclidean distances among stations (SIMOPT=7). You can calculate CNESS by first using a Hypergeometric standardization (STAND=10 or 14), then calculate chord distances (SIMOPT=11). An alternate method is to specify STAND=12 or 15, then calculate Euclidean distances (SIMOPT=7). Pearson's r among variables can be calculated by standardizing the data to z-deviates (STAND=13), followed by calculating cos(theta) among variables (BYOPT=1, SIMOPT=25).

Canberra-Metric

COMPAH contains two options for calculating the Canberra-metric dissimilarity. This index, developed by Lance & Williams (1966) and discussed by Clifford & Stephenson (1975, pp. 58-60) is calculated as:

$$CM_{ij} = \frac{1}{S} \sum_{k=1}^{S} \frac{|x_{ik} - x_{jk}|}{x_{ik} + x_{jk}} \quad . \qquad \textbf{(10)}$$

where, CM_{ij} = *Canberra metric for samples i and j.*
S = *Number of species.*
x_{ik} = *Abundance of species k in sample i.*

Option 1: Double-zero comparisons. If both samples being compared have zero abundances for a species, then **the default is not to count such double-zero comparisons in 1/S**. Clifford & Stephenson **do** count such double-zeros in **1/S**, reducing the dissimilarity between pairs of such samples. The default can be changed at a keyboard prompt.

Option 2: When only 1 of a pair of samples contains zero abundance, then the fraction $\{|x_{ij}-x_{ik}|/(x_{ij}+x_{ik})\}$ is unity. Clifford & Stephenson (1975, p. 59-60) state:

> *"...a pair of elements with values of 1000 and 0, respectively, will appear equally dissimilar to a pair with values of 0.1 and 0.0. This would appear to be unreasonable and the problem may be circumvented by replacing those zero values, which enter into one of the above terms by a small positive number. Choice of the number is arbitrary but it should be smaller than any of the recorded values in the population under study and its efficiency may be judged by trial and error. Stephenson et al. (1972) used a value one-fifth of the lowest entry in their data matrices"*

The default is to keep zeros in the calculations. If Canberra-metric (SIMOPT=5) is called, then COMPAH will prompt whether single zeros in the Canberra metric calculation should be replaced with a value 1/5th of the minimum absolute value of non-zero elements in the full data set. This value, called CMZERO, will be printed in the COMPAH output.

The following data file (in COMPAH **S**tandard-order **C**ondensed form) shows some differences obtained using the 4 possible methods of calculating the Canberra metric:

```
   6    3Cliff. & Stephenson p. 59, data to test Canberra metric   XXSC
(I2,1X,3(I3,F5.0))                                                 3
1   1  10.  2   5.
2   1   5.  2   1.
3   1   0.  2   1.
4   1   1.  2   0.
5   1   0.  2   0.  3 1.0
6   1 1.E4  2  10.
```

Shown below are four lower triangular distance matrices for the three sites described by the above data. The first lower triangular matrix is calculated using the default Canberra metric approach, the second matches the example shown in Clifford & Stephenson (1975, p. 60).

Canberra metric (CMZERO= .0000,DOUBLE 0's NOT COUNTED)

```
          1       2
   2    .800
   3   1.000   1.000
```

Canberra metric (CMZERO= .0000,DOUBLE 0's COUNTED)

```
          1       2
   2    .666
   3    .833    .833
```

Canberra metric (CMZERO= .2000,DOUBLE 0's NOT COUNTED)

```
          1       2
   2    .666
   3    .843    .777
```

Canberra metric (CMZERO= .2000,DOUBLE 0's COUNTED)

```
          1       2
   2    .555
   3    .703    .647
```

Morisita & Morisita-Horn

COMPAH's Morisita similarity (SIMOPT=8), based on Morisita (1959, cited in Boesch 1977), is different from Magurran's (1988) Morisita-Horn index. Boesch (1977, p. 29) defines the Morisita index used in COMPAH:

$$MS_{ij} = 2 \sum_{k=1}^{S} \frac{x_{ik}\, x_{jk}}{(\lambda_i + \lambda_j)\, N_{i.}\, N_{j.}}$$

where, MS_{ij} = *Morisita similarity for samples i and j.*
S = *Number of species.*
x_{ik} = *Abundance of species k in sample i.*
$N_{i.}$ = *Total individuals in sample i.*

$$\lambda_i = \frac{\sum_{k=1}^{S} x_{ik}\,(x_{ik} - 1)}{N_{i.}\,(N_{i.} - 1)} .$$

λ_i=*Simpson's unbiased diversity estimator.*

(11)

The Morisita-Horn index is defined by Magurran (1988, p. 95) and is COMPAH's SIMOPT=16:

$$MHS_{ij} = 2 \sum_{k=1}^{S} \frac{x_{ik}\, x_{jk}}{(d_i + d_j)\, N_{i.}\, N_{j.}} .$$

$$\begin{aligned} \text{where, } S &= \text{Number of species.} \\ x_{ik} &= \text{Abundance of species } k \text{ in sample } i. \\ N_{i.} &= \text{Total individuals in sample } i. \end{aligned} \qquad \textbf{(12)}$$

$$d_i = \frac{\sum_{k=1}^{S} x_{ik}^2}{N_{i.}^2} .$$

d_i=Simpson's (biased) diversity estimator.

Magurran (1988, p. 95) describes some properties of the Morisita-Horn index:

> *"Wolda (1981) investigated a range of quantitative similarity indices and found that all but one, the Morisita-Horn index (Worked Example 15) were strongly influenced by species richness and sample size. A disadvantage of the Morisita-Horn index however is that it is highly sensitive to the abundance of the most abundant species."*

The following sections describe NESS, NNESS, and CNESS, which can be regarded as more generalized forms of the Morisita and Morisita-Horn indices. However, these newer indices can be adjusted, by altering NESSm, to emphasize the importance of rarer species in the data.

NESS, NNESS & CNESS

Grassle & Smith (1976) introduced the **NESS** similarity index. **NESS** stands for normalized expected species shared. I have corrected **NESS** so it can properly analyze samples containing singleton species. I call the new version **NNESS** (SIMOPT=10). SIMOPT=27 calculates NNESS without rounding fractional species counts to integers. I also developed a metric version of NESS called CNESS (chord-normalized expected species shared) and included it in COMPAH (SIMOPT=15 or 26).

Grassle & Smith (1976, p. 16) state that NESS at m=1 is the same as Morisita's similarity. The original NESS (at NESSm=1) is Morisita's (1959) similarity, but NNESS (at NESSm=1) is not. Like Grassle & Smith's (1976) NESS, Morisita's (1959) equation cannot properly calculate the similarity between samples containing species represented by only single individuals (singletons). Pairs of samples composed only of singleton species would require a division by zero in both the original NESS and Morisita's index. Morisita's original index and NESS both could produce very high similarities (often exceeding 1.0) if pairs of samples contain the same singleton species. These shared singleton species contribute to the numerator, increasing the similarity, but not the denominator used to normalize the index. For example, a hypothetical data set can be constructed in which two samples contain the following species abundances:

	Sp. 1	Spp. 2->37	Sp. 38	Totals:
Sample i	1	1	1	39
Sample J	1	1	0	38

This data set is called **ROULETTE.DAT** on the distribution diskette, since the problem is analogous to calculating the similarity or dissimilarity between American and European roulette wheels. The European roulette wheel lacks species **00**. COMPAH produces a run time error (divide by 0) if Morisita (SIMOPT=8) [or the original NESS] similarity is used.

The Morisita-Horn (SIMOPT=16) and NNESS with NESSm=1 to 37 (SIMOPT=10, NESSm=1 to 37) similarity between i and j are both 0.987. Sorensen's (SIMOPT=2), Ochiai's (SIMOPT=3) and Bray-Curtis (SIMOPT=6) and cos(theta) similarity are also 0.987. CNESS (NESSm=1 to 37) and chord-distance between the two samples are both 0.163.

A severe problem with Morisita's original index can be appreciated if the abundance of Sp. 1 in both samples is increased from 1 to 2. The Morisita similarity between the samples becomes **19.5**. The NESS similarity, calculated using Grassle & Smith's (1976) original normalizing equation is 19.4737. The reason for these high values is that the original Morisita and Grassle-Smith indices normalized their indices by the number of species expected in disjunct random draws of individuals from the same sample. Singleton species can never be present in 2 disjunct random draws from the same sample. However, these singleton species can be represented in the numerator of the Morisita and Grassle-Smith equations. The numerator represents the expected species shared between 2 random draws from different samples. The Morisita-Horn and NNESS similarity between the 2 samples is 0.988.

Faunal similarities among samples are always lower with NNESS than the original NESS. As shown in Trueblood *et al.* (1994), the differences are minor for temperate benthic data. However, if samples contain many singleton species, the differences between NESS and NNESS can be very large, and NESS can be larger than 1.0, its theoretical maximum.

Magurran (1988) provides examples against which I checked the COMPAH's accuracy. Using data from Magurran's Example 15, COMPAH calculated the similarities shown in Table 6.

The Morisita-Horn similarity between the two samples in Magurran's Example 15 is different from the NNESS, m=1 similarity, calculated with SIMOPT=10. The difference is due to COMPAH rounding fractional data. Magurran's example 15 contains many fractional species abundances (*e.g.*, 1.4 mallards). The standard NNESS and CNESS algorithms (SIMOPT=10 and SIMOPT=15) round these fractions to the nearest integers. COMPAH now contains NNESS and CNESS algorithms that calculate factorials using the continuous ln(gamma) distribution. With SIMOPT=27, NNESS (at m=1) always equals the Morisita-Horn index (SIMOPT=16), and with SIMOPT=26, CNESS at m=1 always equals Orloci's chord distance (SIMOPT=16).

Table 6 Comparison of COMPAH similarities with Magurran's Example 15. The first four similarity values, calculated by COMPAH, match those in Magurran (1988, p. 166-167). Magurran does not provide estimates of NNESS or the old Morisita index. The original Morisita index (SIMOPT=8) produces a similarity equal to the original Grassle & Smith (1976) NESS.

Index	COMPAH SIMOPT	COMPAH Similarity	Magurran Similarity
Jaccard	1	0.462	0.462
Sorensen (=Dice)	2	0.632	0.632
Bray-Curtis	6	0.444	0.444
Morisita-Horn	16	**0.813**	**0.813**
Morisita-Horn Rounded using Trans=6	16	**0.802**	-
NNESS, NESSm=1	10	**0.802**	-
NNESS, NESSm=1 Abundances not rounded	27	**0.813**	-
Morisita (1959)	8	0.905	-

In May 1992, I developed the CNESS distance measure and added it to COMPAH. CNESS, a metric, is short for chord-distance normalized expected species shared. NNESS, Morisita's, and Sorensen's indices are semimetrics. Like Orloci's chord distance, CNESS has a range between 0 and $\sqrt{2}$. CNESS has a similar geometric interpretation as Orloci's chord distance metric. The equations for NESS and CNESS are both based on hypergeometric probabilities, the probability of sampling species k in sample i or j with a random sample size of NESSm. The formula for calculating these probabilities was shown earlier. These probabilities, calculated with Equ. 1 can be arranged into a sample-by-species matrix, shown in Equation 2. The Expected Species Shared matrix, **ESS**, is the sum of squares and cross products matrix of the hypergeometric probability matrix **H**:

$$\mathbf{ESS} = \mathbf{H} * \mathbf{H}'.$$

$$\text{where, } H_{ik|m} = 1 - \frac{\binom{Total_{i.} - x_{ik}}{m}}{\binom{Total_{i.}}{m}} .$$

$$= 1 - \left[\frac{\frac{(Total_{i.} - x_{ik})!}{m! * (TOTAL_{i.} - x_{ij} - m)!}}{\frac{Total_{i.}!}{m! * (TOTAL_{i.} - m)!}} \right] . \qquad \textbf{(13)}$$

$Total_{i.}$ = *the sample total.*

x_{ik} = *the abundance of species k in sample i.*

m = *NESSm* = *Number of individuals to be drawn at random.*

! = *a factorial.*

The elements of **ESS** represent the Expected Species Shared in random draws of NESSm individuals from two samples. These elements can be combined to form both the NNESS and CNESS indices. The main diagonal elements of the matrix, the ESS_{ii} and ESS_{jj} elements, can be regarded as diversity estimates for samples i and j. Both NESS and CNESS normalize the faunal similarity among samples by the average species diversity. NNESS, a modified form of Grassle & Smith's (1976) NESS is:

$$NNESS_{ij|m} = \frac{ESS_{ij|m}}{\frac{1}{2}(ESS_{ii|m} + ESS_{jj|m})} . \qquad \textbf{(14)}$$

CNESS uses the same ESS values, but normalizes the expected species shared by the geometric mean rather than the arithmetic mean of sample diversities:

$$CNESS_{ij|m} = \sqrt{2 - \left(\frac{2\ ESS_{ij|m}}{\sqrt{ESS_{ii|m}\ ESS_{jj|m}}} \right)} . \qquad \textbf{(15)}$$

NNESS and CNESS are families of similarity measures, based on the expected species (ESS) shared between random draws of individuals from samples i and j without replacement. Increasing the random sample size increases the sensitivity of NNESS and CNESS to rare species.

Figure 1 shows the geometry of CNESS. At a random sample size of 1, the hypegeometric probabilities are simply the frequencies of each species in a sample and CNESS is just the chord distance among samples. With increasing random sample sizes, the hypergeometric probabilities increase towards 1.0 as shown in Fig. 1.

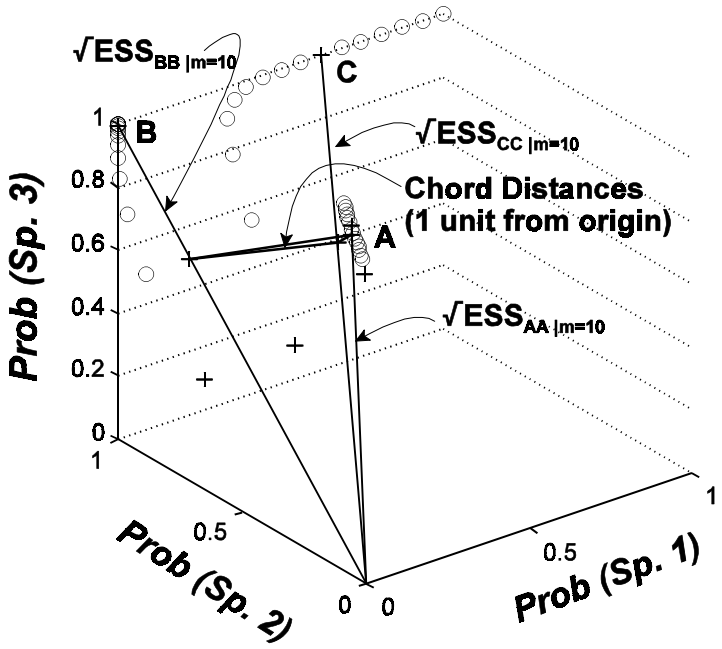

Figure 1. The geometry of CNESS. The positions of samples are plotted in species space according to their hypergeometric probabilities. The position of sample points at random sample sizes from 1 to 16 are shown. The chord distances are shown for m=10. These distances can be viewed by connecting a vector from the origin to the $H_{ij\,|\,m}$ coordinate and then marking the point 1 unit from the origin. The Euclidean distances among these points, the projections of the samples onto the unit hypersphere, are the CNESS distances. These sample coordinates on the hypersphere are the row-normalized elements of the **H** matrix.

The clustering patterns of NESS and CNESS are very similar as are the results obtained using non-metric multidimensional scaling using these indices (Trueblood *et al.* 1994).

There are three different ways of generating CNESS distances. The simplest is to select CNESS (SIMOPT=15 or 26) and set the NESSm random sample size. If NESSm=1 and the data are integers, CNESS is Orloci's chord distance (=Cavalli-Sforza & Edwards' chord distance). For species abundance data that are not integers, CNESS should be calculated with SIMOPT=26 (see fractional data section below). A second way is to convert species abundance data to hypergeometric probabilities (STAND=10 or 14) and calculate the distance between samples using Orloci's chord distance (SIMOPT=11). Using this method, the user can have COMPAH save the hypergeometric probabilities as a COMPAH data file and a MATLAB™ -mat file for further analysis (PrStD=2 or 3, see below). The final method is to convert

the data to hypergeometric probabilities, followed by a station normalization (STAND=12). Then the CNESS distance between stations is the Euclidean distance between the standardized data.

There are differences obtained using the three different methods to calculate NESS and CNESS. Historically, NESS was calculated by rounding fractional data to the nearest integers. SIMOPT=10 [NESS] and SIMOPT=15 [CNESS] and STAND=10 and STAND=12 retain this historical feature. This feature usually has the effect of reducing the importance of rare species (especially if the data analyzed are the means of a set of replicate samples).

Clustering species using CNESS

To cluster species using hypergeometric probabilities, first standardize the data to hypergeometric probabilities, followed by station normalization (STAND=12). Then, calculate the similarity among species using Pearson's product-moment correlation (SIMOPT=4). As discussed in Trueblood *et al.* (1994), this clustering uses the same species affinities displayed in Gabriel's (1971) covariance biplot of CNESS distances among stations. Pearson's r is merely the cos(θ) between species in the Gabriel (1971) covariance biplot, where θ is the angle between species vectors after the data have been transformed to Z-deviates (STAND=13). Lance & Williams (1967) noted that the average clustering strategies are probably not appropriate for Pearson's r. I recommend using single-linkage clustering, but complete-linkage clustering would also be appropriate for Pearson's r. The following **COMM.DAT** file performs the R-mode clustering of species corresponding to the Q-mode clustering using CNESS, NESSM=10. Trueblood *et al.* (1994) used NESSm=15 for their data.

```
 1  NDROP   nn:  Drop species that occur in no more than NDROP samples     0.000
 2  Trans   0-6:0=none;1=log10;2=sqrt;3=cube root;4=4th rt;5=Boolean       0.000
 3  Stand   0;1MxSp;2MnSp;3MnSt;4NrSp;5NrSt;6SD;7TSp;8TSt;9Si;10Hy;12H&N   12.00
 4  R1,Q2   1=>Cluster species (R mode); 2 =>Cluster samples (Q mode)      1.000
 5  SIMOP   1J;2S;3Oc;4r;5CM;6B;7E;8M;9%;10N;11C;12G;13E2;14MS;15CN;16MH   4.000
 6  ClusM   1-8;1(UPGMA)2(WPGMA)3(SLnk)4(CLnk)5(Md)6(Flx)7(ISSQ)8(Cntrd)   3.000
 7  Beta    Flexible average sorting BETA parameter                       -.250
 8  NESSM   "m" for NNESS,CNESS,and Hy; Samples w.< "m" ind are dropped    10.00
 9  PrEdD   1 => Print edited data;0=> Don't print;2 => Print & Save       1.000
10  PrTrD   1 => Print transformed data matrix;  0 => Don't print          1.000
11  PrStD   0 => Don't print;1=>Print; 2 =>Print & Save 3=>MATLAB.mat      0.000
12  PrSmM   1=Print SIM;2=Dont print;3=Dont calc.;4=print&save;5=Save      1.000
13  PrnCl   0=Don t do or print results of clustering; 1 => Cluster        1.000
14  PTREE   0 No tree;1 I- tree, ss;2 ├ tree, ss;21 I-, ds;22 ├ , ds       22.00
15  #Runs   1=> 1 run; 2=> A second analysis requested; 0=> Stop now       1.000
```

What NESSm sample size should you use?

There is no right answer to that question for all data. The choice of sample size is important. NNESS at m=1 is the Morisita-Horn index. With increasing NESSm sample sizes, NNESS reduces the influence of the abundant species in the data set, making the index more sensitive to the rarer species in the community. However, at large NESSm, NNESS and NESS converge to Sorensen's (=DICE=Czekanowski binary) presence/absence similarity index (SIMOPT=2). At large NESSm, CNESS converges to an index that is related to the 1-complement of the Ochiai presence/absence similarity index (SIMOPT=3):

$$\mathrm{CNESS}_{NESSM=\infty} \approx \sqrt{2\ (1\ -\ \mathit{OCHIAI\ SIMILARITY})}. \qquad \textbf{(16)}$$

To analyze the effects of very large NESSm on CNESS, request a Boolean transform of the data (TRANS=5) and then calculate Orloci's chord distance (SIMOPT=11). This is equivalent to CNESS at

an infinite NESSm. This converts species abundance data to (0,1) variables. CNESS at infinite sample size is similar but not identical to the Ochiai presence/absence distance measure.

Kenkel & Orloci (1986) show that chord distance combined with non-metric multidimensional scaling does a good job recovering the patterns in complex simulated ecological data.

I have done many analyses showing that often only the most abundant species contribute to the variance in CNESS distances among samples with NESSm=1. Trueblood *et al.* (1994) found that with NESSm=15, the CNESS distance matrix was highly correlated (using Kendall's tau) with both CNESS with NESSm=1 (=Orloci's chord distance) and CNESS with NESSm=150. The Kendall's tau correlation between CNESS with NESSm=1 and CNESS with NESSm=150 was close to zero. Trueblood's temporal data set included both rare species and species that on some dates constituted over 90% of the individuals in a sample. At a NESSm of 150, the probability of sampling at least 1 of these abundant species in every sample was always close to 1.0. Thus, at large NESSm these abundant species contributed very little to the variance in CNESS distances among stations. Neither NESSm=1 nor NESSm="minimum sample total" is appropriate if the goal of the analysis is to produce an index sensitive to both the rare and abundant species.

Trueblood *et al.* (1994) describe a method to find an appropriate NESS and CNESS sample size. By comparing the Kendall's *tau* rank order correlation of distances calculated with NNESS and CNESS, a NESSm size can be found which has a correlation that is roughly the same between CNESS (NESSm=1) and CNESS (max NESSm). A MATLAB™ m.file, called findcnm.m, will find this consensus NESSm sample size. This m.file is distributed at both the Rutgers ftp site and the UMASS/Boston web sites (see p. 5) in both MATLAB™ 3.5 and 4.2 versions (DDTMAT35.EXE and DDTMAT42.EXE). With temperate, shallow-water benthic data, CNESS with NESSm=10 to 25 produces a metric that is *sensitive to both the rare and abundant species in a sample*. This was Grassle & Smith's (1976) goal in developing NESS.

Grassle & Smith's (1976) formula for calculating the original NESS index set the upper limit for NESSm at half the minimum sample total. Their formula calculated the expected species shared between two disjunct random draws of size NESSm from the same sample **without replacement**. The NNESS calculation replaces the NESSm individuals before the 2nd draw, so that the effective upper limit for NNESS is the minimum sample total.

Fractional data & hypergeometric probabilities

Using standard equations for calculating factorials, species abundance data should be integers. However, COMPAH is not restricted to using integers for calculating hypergeometric probabilities because it calculates factorials using the continuous gamma (Γ) distribution: (Γ (n+1) = n!). To make COMPAH consistent with earlier algorithms, COMPAH can round fractional data to the nearest integer before calculating sample sums and factorials (using SIMOPT=10, 15 or STAND=10 or 12). Standardization options 14 and 15 calculate hypergeometric probabilities for fractional data, and SIMOPT 26 and 27 calculate CNESS and NNESS with fractional data. While factorials are defined only for integers, COMPAH uses the continuous natural log of the gamma distribution to calculate these factorials.

For temperate benthic data, the effects on NNESS or CNESS of using the continuous gamma distribution instead of rounding species counts is usually small. With the fractional data in Magurran's example 15, Orloci's chord distance between stations is 0.610, but CNESS (with NESSm=1) after rounding is 0.629.

If CNESS is calculated with the continuous ln(gamma) distribution (SIMOPT=26 or STAND=14 and SIMOPT=11), then CNESS at NESSm=1 is also 0.610.

Never convert data based on small samples (*e.g.*, numbers/0.05 m^2) to a larger sample size (numbers/m^2). This conversion will give undue weight to species represented by single individuals, especially at large NESSm. These singleton species, instead of being represented as 0 and 1 individuals per sample, will now be recorded as 0 and 20 individuals per sample. The contribution of these species to the overall variance in CNESS distances will be exaggerated.

Combinatorial strategies

There are eight Combinatorial sorting strategies available in COMPAH, called by the parameter ClusM in as shown in Table 7. The strategy can also be set interactively at the start of each COMPAH analysis.

Table 7 COMPAH's combinatorial sorting strategies. The acronyms UPGMA, WPGMA, etc. were introduced by Sneath & Sokal (1973).

ClusM	Sorting strategy	ClusM	Sorting strategy
1	Unweighted pair-group mean average (UPGMA)	5	Median;
2	Simple average (WPGMA)	6	Flexible average sorting;
3	Single linkage (Nearest-neighbor);	7	Incremental sums of squares;
4	Complete linkage (Farthest-neighbor);	8	Centroid

Lance & Williams' Combinatorial equation

COMPAH clusters samples using Lance & Williams' (1967) Combinatorial equation, which allows the similarity of clustered groups to be calculated without requiring recalculation of the original similarities or distances:

$$d_{hk} = \alpha_i\, d_{hi} + \alpha_j\, d_{hj} + \beta\, d_{ij} + \gamma\, |d_{hi} - d_{hj}|. \quad \textbf{(17)}$$

where, d_{hk} = *distance between group h and fused group k, composed of i and j.*
d_{hi} = *distance between groups h and i.*
d_{hj} = *distance between groups h and j.*
α, β, γ = *parameters set for specific strategies.*

Boesch (1977) describes this equation, which also can be found in most books on classification or multivariate analysis in ecology (*e.g.*, Sneath & Sokal 1973, Clifford & Stephenson 1975, Orloci 1978, Pielou 1984). Table 8, adapted from Boesch's (1977) Table 2, provides the parameters used in the Combinatorial equation to perform the various clustering strategies.

Table 8 Values of the parameters for Lance & Williams' (1967) combinatorial equation (Equation 17), where n_h, n_i and n_j are the numbers of entities in groups h, i, and j and n_k is the number of entities in group k resulting from the fusion of i and j (*i.e.*, $n_k = n_i + n_j$).

Method	α_i	α_j	β	γ
Single linkage (Nearest neighbor)	$\frac{1}{2}$	$\frac{1}{2}$	0	$-\frac{1}{2}$
Complete linkage (Furthest neighbor)	$\frac{1}{2}$	$\frac{1}{2}$	0	$\frac{1}{2}$
Group Average UPGMA	$\frac{n_i}{n_k}$	$\frac{n_j}{n_k}$	0	0
Simple Average WPGMA	$\frac{1}{2}$	$\frac{1}{2}$	0	0
Centroid (Unweighted centroid) UPGMC	$\frac{n_i}{n_k}$	$\frac{n_j}{n_k}$	$-\alpha_i \alpha_j$	0
Median (Weighted centroid) WPGMC	$\frac{1}{2}$	$\frac{1}{2}$	$-\frac{1}{4}$	0
Flexible (β is the flexible average sorting parameter)	$\frac{1-\beta}{2}$	$\frac{1-\beta}{2}$	1	0
Incremental Sums of Squares Ward's	$\frac{n_h + n_i}{n_h + n_k}$	$\frac{n_h + n_j}{n_h + n_k}$	$\frac{-n_h}{n_h + n_k}$	0

Compatible indices and strategies

Lance & Williams (1967) described the properties that a compatible sorting strategy should have:

> *"Compatible or incompatible. A compatible strategy is one in which measures calculated later in the analysis are of exactly the same kind as the original inter-element measures; they have the same dimensions (if any),* ***are subject to the same constraints [emphasis added]****, and can be illustrated by an exactly comparable model. An incompatible strategy is one in which some at least of these properties are lost; the ensuing difficulties in interpretation render incompatible strategies undesirable." (Lance & Williams 1967, p. 374).*

The clustering levels produced by the clustering program should not exceed the expected range for the similarity or dissimilarity index. For example, if a similarity index has a range between 0 and 1, the clustering level should not be negative. Similarly, the final clustering level calculated for unbounded dissimilarity measures like Euclidean distance should not exceed the greatest dissimilarity observed among the objects being clustered.

The following COMPAH sorting methods are incompatible with many of COMPAH's similarity or dissimilarity indices: **median, flexible average, incremental sums of squares, and centroid sorting**. Flexible average sorting with Lance & Williams' (1967) cluster-intensity coefficient (Beta) set at -0.25 is a common clustering strategy in ecology, but is incompatible with virtually all similarity or dissimilarity indices. All four of these strategies can produce negative clustering levels using similarity indices that have 0 as a theoretical minimum. They can also can produce clustering levels exceeding the maximum distance possible for the dissimilarity measure. These methods are often used because they produce coherent clusters (*i.e.*, space-dilating clusters [Lance & Williams 1967]), but these clusters are produced at tremendous cost in interpretation of the clustering levels.

The following simple data set reveals this pathological flaw in these four Combinatorial methods.

```
    3    3Test data to show flaws in some sorting strategies          TV
(3(F1.0,1X))
1 1 0
1 1 0
0 0 1
Species A
Species B
Species C
Sample. 1
Sample. 2
Sample. 3
```

Using Jaccard's similarity (SIMOPT=1), COMPAH produces the following lower-triangular similarity matrix:

```
          1       2
   2   1.000
   3    .000    .000
```

Using all clustering methods, groups 1 and 2 fuse at a clustering level of 1.0 as expected. UPGMA, WPGMA, Single linkage, and Complete-linkage sorting fuse group 3 with the combined 1 & 2 group at a similarity of 0.0 as expected. However, centroid, median and flexible average (Beta=-0.25) sorting fuse group 3 with the fused 1 & 2 group at a clustering level of -0.25. Incremental sums of squares fuses group 3 and 1 & 2 at a level of -0.333. These fusion levels are lower than the theoretical minimum for the similarity index.

Obviously, getting a negative clustering level for a (1->0) similarity measure or a clustering level > 1 for a (0->1) dissimilarity measure violates the second of Lance and Williams' three conditions for a compatible strategy.

Using Euclidean distance squared as the dissimilarity measure, the following lower-triangular matrix is produced:

```
          1       2
   2    .000
   3   3.000   3.000
```

All sorting methods except flexible average and incremental sums of squares fuse group 3 with 1 & 2 at the expected distance of 3.00. Median and Centroid sorting appear compatible with Euclidean distance squared dissimilarity (Grieg-Smith 1983 p. 204-205), Euclidean distance and the Geodesic metric.

Flexible average sorting (with β=-0.25) fuses group 3 with 1 & 2 at a distance of 3.75 and incremental sums of squares fuses them at a value of 4.00. Grieg-Smith (1983, p. 205) lists the combination of Euclidean distance squared and flexible sorting as a compatible combination, but clearly it is not since 3.75 exceeds the maximum dissimilarity observed. Flexible average sorting is incompatible with every COMPAH similarity or dissimilarity index, in that the final clustering level can fall outside the range of dissimilarities or similarities observed.

The problem of incompatibility is not restricted to samples that share no species. In one large data set (293 samples), the lowest similarity among 293 samples was 0.5 using Jaccard's index. However, all four of these combinatorial sorting strategies produced negative clustering levels for the largest clusters. With Euclidean distance squared dissimilarity, flexible average sorting produced a final fusion level of 47.8 even though the maximum Euclidean distance squared between the two most dissimilar samples was 26.0.

Flexible average sorting can produce discrete clusters when other methods produce badly chained clusters (Clifford & Stephenson 1975, Boesch 1977, Legendre & Legendre 1998, p. 337). However, this virtue is more than offset by difficulties of interpretation. In one recent analysis using percentage dissimilarity (=1-Bray-Curtis similarity, SIMOPT=17), I observed the final fusion among clusters at a level of 3.3. Percentage dissimilarity has a maximum value of 1.0. How can one interpret how different two clusters are when they fuse at a level of 3.3, when the original dissimilarity index has a maximum value of 1.0? Do the samples in the two fusing clusters share any species at all? It is impossible to say, given the incompatiblilty of the index and sorting method.

Combinatorial strategies and indices

In addition to problems of compatibility, some combinations of similarity and sorting strategies are not combinatorial. Lance & Williams (1967, p. 374) define what they mean by "combinatorial." When two groups (i) and (j) fuse in a cluster analysis, they behave as a new group (k) with n_k ($n_k=n_i+n_j$) elements. With a combinatorial strategy, all future clustering levels can be derived from the original (i) and (j) similarities or dissimilarities. The original data do not need to be stored after the original similarity or dissimilarity matrix is calculated.

There is some confusion in the literature about which combinations of similarity or dissimilarity measures and clustering methods are combinatorial. Furthest-neighbor and nearest-neighbor clustering methods are combinatorial with all similarity and dissimilarity indices (Lance & Williams 1967). Lance and Williams (1967) showed that centroid clustering was Combinatorial for squared Euclidean distance. Boesch (1977), Clifford & Stephenson (1975, p. 116) and Legendre & Legendre (1983, p. 238), recommend centroid clustering only with squared Euclidean distance. Pielou (1984) uses the geodesic metric with centroid clustering. Orloci (1978) argues that practically any similarity measure is admissible with average linking clustering, his term for centroid clustering.

Boesch (1977) states that Euclidean distances or standardized Euclidean distance can be used with incremental sum of squares clustering. However as shown above, incremental sums of squares clustering is incompatible with Euclidean distance and other metrics, producing clustering levels greatly exceeding the maximum distance among samples.

Jardine & Sibson (1968) reviewed the combinatorial methods proposed by Lance & Williams (1967) and concluded that only single-linkage clustering met their criteria for a successful clustering strategy.

Williams (1971) while commending Jardine & Sibson's rigor, concluded that the rigorous mathematical criteria should be relaxed allowing the use of the "average" clustering strategies.

RUNNING COMPAH96

Input options

COMPAH input file format

Your must format your data an a text file using COMPAH format. COMPAH format is nearly identical to the data input format used by the Cornell ORDIFLEX clustering program. The distribution diskette(s) should contain files labeled TEST.DAT, TESTC.DAT and TESTP.DAT and files labeled ECPMAT*.DAT (*=1:6) taken from Pielou (1984). These files show the major forms of data input used by COMPAH. The user can call the data file using DOS paths (*e.g.*, c:\data\TEST.DAT).

The first line of the DATA file is a header file containing the number of variables (SPECIES) in the 1st 5 columns, the number of samples (STATIONS) in columns 6-10, and a descriptive title (optional but length < 56 characters) in columns 11-66. If species labels are **not** provided, type an **X** in **column 69**. If sample labels are not provided, type an **X** in **column 70**. Columns 71-72 of the first line of the data file tell COMPAH the form of the following data.

Column 71 of the 1st line of the data file indicates whether the data are in **S**tandard or **T**ransposed order. An **S** or **T** must be found in column 71. Type an **S** in column 71 of the first line of the DATA file for **STANDARD ORDER** (species=rows, stations=columns). COMPAH refers to a data set with samples as rows and species as columns as **TRANSPOSED ORDER**. Type a **T** in column 71 for transposed order.

Column 72 indicates the form of the data. COMPAH can read files in FULL FORM (**S**tandard), **C**ONDENSED FORM or **P**RESENCE/ABSENCE FORM. Indicate the form of data with an **S**, **V**, **C** or **P** in column **72** on the first line of the data file. A blank, **V** or ***** in **column 72** implies Full FORM. A blank in column 72 indicates that the default (8F9.2) format should be used. A **V** indicates that a format statement describing the full form data follows on the second line of the data file. A ***** indicates that list-directed input/output will be used (Note list-directed input added to COMPAH in March '95; in earlier versions * was treated as V). List-directed input is described below. An **L** in column 72 indicates that a lower triangular similarity or dissimilarity matrix will be entered (see below).

Most users use **FULL FORM** input with data in the form of a species-by-station matrix (**S**tandard order) or a sample-by-species matrix (**T**ransformed order). All Condensed form and Presence/Absence data files must have a format statement for the second line of the data file.

COMPAH reads all abundance data as REAL variables (coded for using Floating point, or F format) and sample and species numbers for **C**ondensed & **P**resence/Absence form as integer variables (see TESTP.DAT and TESTC.DAT on the distribution diskette). The program includes a default **(8F9.2)** data format for FULL form input if Column 72 on the first record is blank. This default is only used if **SS**, **'s** ', **TS**, or '**T** ' are typed in columns 71-72. I usually change this default format by specifying a variable format with a **V** in column 72 of the first line of the data file. The new variable format should be the second line on the data file. If **C**ondensed or **P**resence/absence is indicated by typing a **C** or **P** in column

72, then a variable format is assumed and the 2nd line of the data file must contain the format statement which is always enclosed in parentheses.

Format statements

Users unfamiliar with FORMAT statements should consult any FORTRAN users guide or text. In brief, the format statement, enclosed in parentheses, must be in **columns 1-60** of the second line of the data input file if there is a **V**, **C**, or **P** in column 72 of the first line. This format statement indicates the number of columns assigned to each variable, the number of digits to the right of the decimal point (a typed decimal point overrides this format), and the number of blank spaces between variables (*e.g.*, 4X for 4 blank spaces). The following 3 Format statements describe the same set of data:

```
(F3.1,1x,F3.1,1x,F3.1,1x,F3.1)
(3(F3.1,1x),F3.1)
(F3.1,3(1x,F3.1))
```

The data could be typed as:

```
1.0 2.0 1.1 3.2
```

which is safer and easier to interpret than typing these same data as:

```
 10  20  11  32
```

Explicit typing of the decimal point overrides the format statement's specification of where the decimal point should be. Also, large numbers can be typed in scientific notation (1.E6 instead of 1000000, 1.E-3 instead of 0.001), overriding the format statement's F format. The scientific notation must fall within the columns specified by the FORMAT statement.

You can code the same data on two lines using the slash editing character (/). The slash means "move to the next line." For example, many data sets require more than one line of data per sample or species. The above data could be coded on 2 lines using (F3.1,1X,F3.1,/,F3.1,1X,F3.1):

```
1.0 2.0
1.1 3.2
```

Don't use slash (/) formatting if you can possibly avoid it. Since each line has the same format, the slash (/) isn't necessary, nor is it good programming practice to use it. It is better to program multiple lines of data using the simpler format (F3.1,1X,F3.1). If COMPAH expects 4 species in a sample (columns 1-5 of the first data line) and the format describes fields for only 2 species per line, COMPAH will automatically move to the next line or lines until all the data have been read. The simplest method for entering multiple lines of data for a station or species is to format each row of data using the same format and enter the format statement for the **longest** row of data without using the slash (/).

Condensed and Presence/Absence forms

Condensed and **P**resence/absence form input are set by typing a **C** or a **P**, respectively, in **column 72** of the first line of the data input file. Both forms require a format statement in columns 1-60 of the second line (record) of the data file (or a * for list-directed input). Station numbers and Species numbers are coded as integer variables (*e.g.*, **13** not **13.**). The data can be either **S**tandard order (read data for multiple samples for a single species on each row) or **T**ransposed order (read data for multiple species for a single

sample on each row). Indicate the order using an **S** or **T** in **column 71** of the first line of the data file. For both forms, the first variable read on a line is the either the integer species number (**S**tandard order) or Sample number (**T**ransposed order).

In Standard order-Condensed Form (**SC** in columns 71-72 of first data line), the data are in the form of couplets. The first variable read on each row is the species number (in **I**nteger format) followed by couplets of INTEGER sample numbers and REAL abundances. Format the REAL abundance variables using the FORTRAN E, F, or G format.

The maximum number of couplets on a row is typed as an integer (*e.g.*, 6 not 6.0) in columns 61-70 on the 2nd line of the data file, after the format statement. The integer in columns 61-70 does not need to be right justified, but other integers should be right justified in their format fields unless the BN edit descriptor is included in the FORMAT statement (see a FORTRAN manual for a description of BN). Note that if the integer in columns 61-70 is **J**, then the format statement should be formatted for 2***J**+1 variables. If the format statement describes fields for fewer than **J** variables, COMPAH will move from one line of a data file to the next to find the number of couplets listed in columns 61-70. This usually produces and input error. If the format statement provides fields for at least **J** variables, then a carriage return after a couplet terminates the read statement for that row of data even if fewer than **J** couplets have been read.

In **T**ransposed order-**C**ondensed form (**TC** in columns 71-72 of the first line of the data file), the station number is the first integer variable read on each line, followed by couplets of INTEGER species number followed by the REAL abundance of that species in that sample.

In **S**tandard order-**P**resence/absence form (**SP** in columns 71-72), the integer species number is the first variable read on each row of the data file starting on the 3rd line (after the header and format lines). Following this species number is a list of sample numbers in which that species is found. As in Condensed form, the maximum number of samples (**S**tandard order) or species (**T**ransposed order) on a single line of data should be typed as an integer (*e.g.*, 6 not 6.0) in columns 61-70 on the 2nd line of the data file, after the format statement. This integer does not need to be right justified.

In **T**ransposed order-**P**resence/absence form, the INTEGER sample number is read first followed by the integer species numbers for species found in that sample. Versions of COMPAH compiled after March 1, 1995 allow up to 100 couplets or species numbers per row of data. There is an upper limit on the number of characters that can be included in a single row of data. Many FORTRAN compilers have a default of 256 characters. I have increased this to 1000 characters on versions of COMPAH compiled after 1 March 1997 (see p. 90 in WATCOM F77/32 User's Guide for instructions on changing the recl= specifier in OPEN. COMPAH uses 'sequential' data access, and Watcom allows recl=xxxx to be set for sequential access. The strict FORTRAN77 standard allows recl to be changed only for direct access. This would only be an issue if you wish to compile COMPAH on a compiler adhering to the strict standard).

In both **C**ondensed and **P**resence/Absence form, the data are read until a blank data line is encountered. This blank line should precede the Species or Station Labels (if present). If there are no labels, you must still include a blank line after the last line of data. In Standard or Full form, there should not be a blank line before the species or sample labels. Species labels always precede station labels in both **S**tandard and **T**ransposed order.

For an example of **C**ondensed form input see the Canberra-Metric section on page 9 (or TESTC.DAT on the distribution diskette). The example on page 9 doesn't include species or station labels, but does include a blank line after the last row of data. Here is an example of presence/absence form, note the blank line of spaces (not simply a Hard Return) which terminates the data read statement:

```
   10    4Present/Absence data (Clifford & Stephenson, Table 6.1)     SP
(I2,1X,4I2)                                                          4
 1  1 2 3
 2  1
 3  2 4
 4  1 2 3 4
 5  4
 6  1
 7  2
 8  4
 9  1 2 3
10  1

Attribute 1
Attribute 2
Attribute 3
Attribute 4
Attribute 5
Attribute 6
Attribute 7
Attribute 8
Attribute 9
Attribute 10
Entity A
Entity B
Entity C
Entity D
```

Station & species labels

The COMPAH documentation below describes the format of the data file. Type the species labels (when present) before the sample labels for both transposed and standard order data files. The first label must be on the first line after the sample and species data. You don't need to add quotation marks.

If you don't have station or species labels you must type XX in columns 69 and 70 of the first record of the data file. If columns 69 and 70 of the first line of the data file are blank, then COMPAH expects both species and station labels. COMPAH uses only the first 20 characters of the labels in the output.

About half a dozen FORTRAN statements need to be changed in the source code to increase the length of variable names allowed. For every 3 characters added to the Label length, a clustering level is lost in the final dendrogram. The limit is set by the 132 character widths of mainframe computers (but this isn't a necessary limit for PC's). At present, COMPAH can produce a cluster diagram with 35 clustering levels.

List-directed input

Most word processors and text editors indicate horizontal position in inches, not columns. This makes counting characters to create accurate FORMAT statements difficult. I added list-directed input (sometimes called free format) to COMPAH in 1995 to eliminate the need for FORMAT statements in DATA input files.

Full form list-directed input

In Full form, there are 2 ways to request list-directed input. Type a ***** in column 72 of the first line of the data file (note, versions of COMPAH prior to 3/95 interpreted a * in column 72 as indicating **V**ariable format). The second way to request list-directed input for Full form data is to type **V** in column 72 of the first line of the data file and place an ***** anywhere in columns 1-60 of the 2nd line of the data file.

In **Full form** list-directed input, COMPAH reads data elements until it has read I*J variables, where I is the number of species (column 1-5 of the first line) and J is the number of stations. Variables should be separated by at least one space, a comma, or a carriage return. Multiple spaces between variables are interpreted as a single space. You can enter data using the *r***c* convention, where r is an unsigned non-zero integer and c is a numeric constant. Carriage returns are optional, but a single line of data can't exceed 400 characters. The following three data files produce identical input to COMPAH:

```
   10   12Test Data(Full Form, Transposed Order,Variable format)    XXTV
(F1.0,9(1X,F1.0))
9 9 0 0 0 9 0 0 0 0
9 9 9 0 0 9 0 0 0 0
9 9 9 9 0 9 0 0 0 0
9 9 9 9 9 9 0 0 0 0
0 9 0 0 0 0 0 0 0 0
0 9 0 9 0 9 0 0 0 0
0 9 9 9 0 9 9 0 0 0
0 0 0 9 9 9 0 0 0 0
0 9 0 0 0 9 9 0 0 0
0 9 0 0 0 9 9 9 0 0
0 9 0 0 0 9 0 9 9 0
0 0 0 0 0 0 0 0 0 0

   10   12Test Data(Full Form, Transposed Order, List-Directed)     XXTV
  *
9 9 0 0 0 9 0 0 0 0
9 9 9 0 0 9 0 0 0 0
9 9 9 9 0 9 0 0 0 0
9 9 9 9 9 9 0 0 0 0
0 9 0 0 0 0 0 0 0 0
0 9 0 9 0 9 0 0 0 0
0 9 9 9 0 9 9 0 0 0
0 0 0 9 9 9 0 0 0 0
0 9 0 0 0 9 9 0 0 0
0 9 0 0 0 9 9 9 0 0
0 9 0 0 0 9 0 9 9 0
0 0 0 0 0 0 0 0 0 0

   10   12Test Data(Full Form, Transposed Order, List-Directed)     XXT*
9,9,3*0,9,4*0,3*9,0,0,9,4*0,4*9,0,9,4*0,6*9,5*0,9,9*0,9,0,9,0,9,5*0,3*9
0,9,9,6*0,3*9,5*0,9,3*0,9,9,4*0,9,3*0,3*9,3*0,9,3*0,9,0,9,9,11*0
```

Condensed & Presence/Absence list-directed input

Condensed form and **P**resence/absence form, list-directed input is tricky since there aren't fixed numbers of variables to be entered. For both forms, list-directed input is requested by typing ***** anywhere in columns 1-60 of the 2nd line of the data file. Column 61-70 of the 2nd line of the data file must contain an integer describing the maximum number of items (samples or species numbers in **P**resence/Absence) or couplets (**C**ondensed Form) per row.

If the integer J is typed in column 61-70 and there are fewer than J items on a row, the row must be terminated with a slash (/) preceded by a space. The slash is optional if the row contains J items or couplets.

The blank line normally used to terminate the reading of data in **P** and **C** form input is passed over in List-directed input. To terminate the list-directed data read statement, replace the blank line with a line containing 0 as the first element. The following pair of condensed-form data files read the same data into COMPAH96:

```
   15   11Condensed Mode Test DATA SET                                  XXTC
 (I2,1X,6(I3,F9.2))                                                      6
 1   1     9.00  2     8.00  3     6.00  4     3.00  5     5.00  6     2.00
 1   7     3.00  9     2.00 11     4.00
 2   1     8.00  2     9.00  3     6.00  4     5.00  5     4.00  7     4.00
 2   9     2.00
 3   1     3.00  2     8.00  3     2.00  4     6.00  5     9.00  8     5.00
 3   9     4.00 11     2.00
 4   1     5.00  2     7.00  3     7.00  4     6.00  5     9.00  7     6.00
 4   9     5.00 11     2.00
 5   1     6.00  4     6.00  5     7.00  6     3.00  7     9.00  8     2.00
 5   9     6.00 10     2.00 11     5.00
 6   3     2.00  4     4.00  5     7.00  6     5.00  7     8.00 10     7.00
 6  11     7.00 12     5.00 14     5.00
 7   1     5.00  4     5.00  5     4.00  6     6.00  7     7.00  9     5.00
 7  10     6.00 11     8.00 12     6.00 13     7.00 14     4.00
 8   5     6.00  6     4.00  7     6.00  8     2.00 10     6.00 11     8.00
 8  12     4.00 13     4.00 14     8.00
 9   4     4.00  6     3.00  7     4.00  9     2.00 10     7.00 11     8.00
 9  13     6.00 14     8.00
10   4     1.00  5     2.00  7     3.00  8     2.00  9     5.00 10     6.00
10  11     7.00 12     3.00 13     5.00 14     9.00
```

(Don't Forget to fill a line with spaces to terminate the reading sequence- a carriage return won't do!)

In the following condensed mode list-directed input file, the COMPAH READ procedure is terminated when a row containing a 0 as the first element is read.

```
   15   11Condensed Mode Test DATA (List-directed input)              XXTC
*                                                                     6
1 1 9. 2 8. 3 6. 4 3. 5 5. 6 2.
1 7 3. 9 2. 11 4. /
2 1 8. 2 9. 3 6. 4 5. 5 4. 7 4.
2 9 2. /
3 1 3. 2 8. 3 2. 4 6. 5 9. 8 5.
3 9 4. 11 2. /
4 1 5. 2 7. 3 7. 4 6. 5 9. 7 6.
4 9 5. 11 2. /
5 1 6. 4 6. 5 7. 6 3. 7 9. 8 2.
5 9 6. 10 2. 11 5. /
6 3 2. 4 4. 5 7. 6 5. 7 8. 10 7.
6 11 7. 12 5. 14 5. /
7 1 5. 4 5. 5 4. 6 6. 7 7. 9 5.
7 10 6. 11 8. 12 6. 13 7. 14 4. /
8 5 6. 6 4. 7 6. 8 2. 10 6. 11 8.
8 12 4. 13 4. 14 8. /
9 4 4. 6 3. 7 4. 9 2. 10 7. 11 8.
9 13 6. 14 8. /
10 4 1. 5 2. 7 3. 8 2. 9 5. 10 6.
10 11 7. 12 3. 13 5. 14 9. /
0 0 0. /
```

The following two Presence/Absence form data files read the same data into COMPAH. Note that the READ sequence is terminated in the second file by the row containing a 0 as the first element.:

```
   10    4Present/Absence data (Clifford & Stephenson, Table 6.1)   XXSP
(I2,1X,4I2)                                                        4
 1  1 2 3
 2  1
 3  2 4
 4  1 2 3 4
 5  4
 6  1
 7  2
 8  4
 9  1 2 3
10  1
```

```
   10    4Presence/Absence data (List-directed input)               XXSP
*                                                          4
1 1 2 3 /
2 1 /
3 2 4 /
4 1 2 3 4 /
5 4 /
6 1 /
7 2 /
8 4 /
9 1 2 3 /
10 1 /
0 0 /
```

Converting Matlab™ -mat files to COMPAH input

COMPAH can read binary mat.files containing data directly. COMPAH looks for the *.mat extension created with a MATLAB™ save statement and assumes the DATA are in **STANDARD ORDER** (variables or species as rows, samples as columns). MATLAB™ usually stores these matrices as double-precision variables (MATLAB SAVE type 0000). If the MATLAB™ matrices are large and contain only integers, MATLAB™ saves these as 16-bit variables (Integer*2) variables using SAVE type 0040. [These formats are discussed in MATLAB's External Interface Guide]. If your MATLAB™ program saves data in forms other than 0000 or 0040, COMPAH won't read the files. The title of the COMPAH job will be assigned the matrix name. Only 1 matrix at a time should be stored in each MATLAB™ .mat file for clustering. Labels are not provided, but the data file can be printed out by COMPAH and labels added to the ASCII data file produced.

Reading and clustering similarity matrices generated by Matlab™

MATLAB™ can readily calculate similarity matrices, but I haven't yet written a MATLAB program to cluster samples. MATLAB™ 's limited ability to print script characters makes it a poor choice for producing cluster diagrams. COMPAH96 is set up to read the upper triangular elements of a station-by-station similarity matrix and cluster them using the cluster options set in the **COMM.DAT** file. This input file should have the *.mat extension. If this *.mat file is a vector, COMPAH assumes it is the upper triangular portion of a distance or similarity matrix and immediately moves to the clustering section. If this file is a matrix with minimum dimension greater than 1, COMPAH assumes that this MATLAB .mat file is a data matrix. It will perform all of the data analyses specified in the **COMM.DAT** file. The following statements will create a vector of upper triangular elements in MATLAB™ that can be read and clustered by COMPAH:

```
% if DATA is a data matrix and
% CNESS is an m.file that calculates CNESS
C=cness(DATA,1);
I=find(~tril(ones(size(C))));
V=C(I);
save compain V
% This statement will create a binary *.mat file
% called COMPAIN.MAT that can
% be read and clustered using COMPAH.
```

From Quattro Pro to Matlab to COMPAH

Here are some tricks that I find useful in loading very large data into COMPAH from Corel's Quattro Pro 8 via MATLAB™. The largest matrix possible in a single sheet of Quattro Pro is 8192 x 256. This entire matrix can be set up for clustering in a few minutes using the following commands:

1. Highlight the cells in Quattro Pro 8 for input to MATLAB™
2. Click on Tools, Data tools, extract to file (Short-cut keystrokes: alt-d, d, x)
3. Click on File type, change to ASCII TAB delimited file. Type in the file name. Quattro will append a *.txt extension.
4. If the file created in the previous step is output.txt, go into MATLAB™ and type load **output.txt**
5. The entire data set will be loaded into MATLAB™ and placed in a matrix called output.
6. To cluster this binary matrix:
   a. You will need to save the matrix in a double precision MATLAB™ binary file. MATLAB™ will use an integer format if all of the data are integers, as they often are with species abundance matrices. Find a non-zero entry in the matrix, say output(1,3), and add the smallest double precision number to it: **output(1,3)=output(1,3)+eps;** This will ensure that the matrix gets stored double precision, the only *.mat format that COMPAH reads.
   b. In MATLAB™, type save anyname.mat output. The data should be in standard order, species as rows. The output matrix will be saved in a binary mat file.
   c. In COMPAH, use anyname.mat as the data input. If you want to add labels to the clusters, then use PrEdD=2 (see p. 38) or PrStD=2 (see p. 46) to print out the binary *.mat file out as an ASCII COMPAH data file, then add the labels with a Word Processor.

By the way, a 8192 x 256 matrix takes over 4 hours to cluster using CNESS (using COMP96WB.exe, 64MB RAM, 200 MHZ Pentium). The output is 498 pages long, including the 182 p cluster diagram.

If one is careful counting columns, it is possible to go directly from Quattro Pro or other spreadsheets directly into COMPAH. I'd recommend using list-directed input (see p. 31).

Reading an ASCII lower triangular similarity matrix

COMPAH96 can read a lower triangular similarity or dissimilarity matrix using list-directed input. The following, called LT.DAT, on the data distribution diskette, is an example of a lower triangular distance matrix. Previous versions of COMPAH were able to process lower triangular similarity or dissimilarity matrices only if they were converted to binary MATLAB™ *.mat files (see previous section).

```
8    1   Stocks inverted MWRA clustering data                    X SL
0.657
1.080  1.070
1.030  0.964 1.390
1.130  0.880 1.020 1.380
0.961  0.868 0.768 1.370 0.815
0.990  0.760 0.925 1.170 0.883 0.579
1.310  1.360 1.190 1.400 1.290 1.160 1.220
T1
T2
T3
T4
T5
T6
T7
T8
```

The number of items to be clustered should be provided in columns 1-5. The number of stations must be 1. The data will be read in a row-wise fashion. You could have the entire lower triangular matrix in a long vector, with one distance per row. Don't use a format statement. Just separate the lower triangular elements with spaces or hard returns. If labels are available for the items to be clustered, column 70 should be left blank. Column 71 is not used for lower triangular input. Column 72 of the 1st line of the data file should contain an **L**, which indicates a lower-triangular matrix. COMPAH will check the similarity or dissimilarity options in the **COMM.DAT** file to determine if the matrix is a similarity or dissimilarity matrix.

Converting Matlab™ sparse matrices to ASCII files for COMPAH input

For data sets stored as MATLAB™ sparse matrices, you can use MATLAB's save -ASCII convention. Here is an example of how to create a MATLAB™ sparse matrix from a full matrix and how to read this matrix into MATLAB™ :

```
DATA % a species by sample matrix in Standard order (rows=species)
DATAsp=sparse(DATA);
save datfile DATAsp -ascii
```

The resulting text file will look something like this:

```
   11       1  2.0000000e+000
   14       1  1.0000000e+000
   16       1  1.0000000e+000
```

11 is the column (=species) number, **1** is the station number and **2.0000000e+000** is the abundance of species 11 in sample 1. This can be quickly converted to COMPAH condensed standard data input. Don't forget to put a blank line between the last data record and the first species label. The 1 on the right on the second line indicates that there is only 1 couplet written per record. This 1 can appear anywhere in columns 61-70

```
  361  750George's Bank ssp4sp data set                              SC
(2I8,F16.0)                                                    1
      11       1  2.0000000e+000
      14       1  1.0000000e+000
      16       1  1.0000000e+000
      26       1  2.2000000e+001
      43       1  5.0000000e+000
      48       1  8.4000000e+001
      67       1  1.0000000e+000
     106       1  1.0000000e+000
                   ...
     355     750  1.0000000e+000
     356     750  3.0000000e+000
     361     750  0.0000000e+000

A026
A040
A056
A080
```

(followed by the other species and station labels)

Starting COMPAH96

Call the program from DOS by typing COMPAH96 (or COMP961G, or COMP963G, or COMP966G) at the system prompt. When prompted for the DATA file, type in your DOS or ASCII text file's name. When prompted for the command file, hit return, and COMPAH will load the options from the **COMM.DAT** text file from the default diskette or subdirectory. If you wish, you may type the name of another text file containing a different list of command options having the same format as the **COMM.DAT** file. COMPAH's prompts were written so that a Hard Return calls the most frequently used defaults.

The **COMM.DAT** file should usually be on the same subdirectory from which the user calls COMPAH. The program will check the default drive or subdirectory for the **COMM.DAT** file. The user can type a full path name for the **COMM.DAT** file when prompted by COMPAH. For example, the following file names would be valid within COMPAH: C:\COMPAH**COMM.DAT** or C:\DATA\EPCMAT1.DAT.

Changing parameters on the COMM.DAT file or from the console

The default parameters for a cluster analysis are stored on a DOS text file called **COMM.DAT**. **COMM.DAT** is usually on the calling drive or subdirectory. An alternate file and set of parameters also can be specified from the console when prompted (*e.g.*, B:**COMM.DAT**). The user can create separate **COMM.DAT** files, formatted as DOS or ASCII text files. For example, separate **COMM.DAT** files could correspond to standard R-mode and Q-mode analyses.

The **COMM.DAT** FILE is in (I2,2x,A5,3X,A60,2x,F6.3) format. The following is a listing of the **COMM.DAT** file distributed with the program:

```
 1  NDrop   nn:  Drop species that occur in no more than NDROP samples     0.000
 2  Trans   0-4:0=none;1=log10;2=sqrt;3=cube root;4=4th rt;5=Boolean       0.000
 3  STAND   0None;1MxSp;2MnSp;3MnSt;4NrmSp;5NrmSt;6SD;7TSp;8TSt;9Si;10Hy   0.000
 4  R1,Q2   1=>Cluster species (R mode); 2 =>Cluster samples (Q mode)      2.000
 5  SIMOP   1J;2D;3Oc;4r;5CM;6B;7E;8M;9%;10N;11C;12G;13E2;14MS;15CN;16MH   15.00
 6  ClusM   1-8;1(UPGMA)2(WPGMA)3(SLnk)4(CLnk)5(Md)6(Flx)7(ISSQ)8(Cntrd)   1.000
 7  Beta    Flexible average sorting BETA parameter                       -.250
 8  NESSM   "m" for NNESS,CNESS,and Hy; Samples w.< "m" ind are dropped    10.00
 9  PrEdD   1=> Print edited data;0=> Don't print;2 => Print & Save       1.000
10  PrTrD   1 => Print transformed data matrix;  0 => Don't print         0.000
11  PrStD   0 => Don't print;1=>Print; 2 =>Print & Save 3=>MATLAB.mat     0.000
12  PrSmM   1=Print SIM;2=Dont print;3=Dont calc.;4=print&save;5=Save     1.000
13  PrnCl   0=Don t do or print results of clustering; 1 => Cluster       1.000
14  PTree   0 No tree;1 I- tree, ss;2 ├ tree, ss;21 I-, ds;22 ├ , ds      1.000
15  #Runs   1=> 1 run; 2=> A second analysis on these data requested      1.000
```

COMPAH prompts the user at the beginning of each run asking whether the user wishes to change the defaults. The user can change any or all options interactively. At the end of each run an additional run will begin if #Runs=2. COMPAH will prompt the user asking whether it should read a new **COMM.DAT** file. If the answer is no, COMPAH prints the options from the previous run for upgrading.

The transformations, standardizations, similarity measures, and clustering methods specified in the **COMM.DAT** file are described fully in Boesch (1977). These options are described in previous sections of this document. The data are first transformed and then standardized.

Output options

The output file is a DOS or ASCII text file. You can edit this file using a word processor. COMPAH formats the output for a 132-character-wide printer, the standard on mainframe computer systems. The output contains the line-printer commands for mainframe computers in column 1 of the output. A `0' in carriage position 1 indicates a double space. A `1' in carriage position 1 indicates a page break. I usually load the entire 132-column output into my word processor, reformat the output using a small or fine font, and replace all occurrences of '**<Hard Return>1**' with page breaks and '**<Hard Return>0**' with **<Hard Return><Hard Return>**.

This program is compatible with mainframe computer compilers and printers. I have described the changes required to compile COMPAH with VAX and WATCOM FORTRAN compilers.

If PrEdD=2, COMPAH prints the edited data file to disk in COMPAH format. This can be very useful if you want to print out the data read from a binary MATLAB™ *.mat file (Read the *.mat file & immediately print it out as an ASCII file in COMPAH format).

Printing standardized data and Matlab™ -mat files

COMPAH has several options for printing standardized data. These can be set in advance by editing the PrStD row in the **COMM.DAT** file or they can be changed interactively from the computer console during a COMPAH analysis. PrStD=1 will print the standardized data with species and station labels in the COMPAH output file.

If PrStD = 2, the standardized data file will be printed to a disk file in COMPAH input format. This option is useful for generating the matrix of hypergeometric probabilities or log-transformed data. COMPAH prints an ellipsis (...) at the end of a row of data if the data record continues on the next line.

The ellipsis is ignored by the FORTRAN FORMAT statement provided for reading the data. The ellipsis is required for reading an ASCII data file into MATLAB™ . MATLAB™ has difficulty reading long rows of data, but no difficulty reading many columns of data of length <60 characters. If your data contain many species, print the standardized data using the Standard format. COMPAH asks whether it should print this data file in Standard or Transposed order [S/T]. A species x station matrix is standard order.

Matlab™ -mat files

I find it difficult to get MATLAB™ to read very large ASCII files. For that reason, I added **PrStD=3 & 4** which print the single-precision standardized data as a binary MATLAB™ -mat file. The standardized matrix will be in STANDARD order, that is, species=rows & stations=columns. If PrStD=3 is requested, the data will also be written to the COMPAH output file. With PrStD=4, the standardized data are not written to the text output file.

If the MATLAB™ option is requested, COMPAH asks for a filename to store the matrices. You should add a .MAT extension to this DOS file name (*e.g.*, TESTOUT.MAT) to indicate that this is a binary mat file. COMPAH asks for a 4-character name for each matrix stored (*e.g.*, DATA, H010, H001). In the VAX version, only 3-character matrix names are allowed (*e.g.*, DAT, H10, H01). Spaces count. After loading a DATA.MAT file created by COMPAH into MATLAB™ (*e.g.*, load TESTOUT), MATLAB™ will not recognize a variable called 'H1'. The user would have to type in 'H1 space space' to fill in the trailing 2 spaces. Note also that MATLAB™ is case sensitive. To MATLAB, H001 is not the same as h001.

Request TRANS=0 and STAND=0 and PrStD=4 if you merely wish to translate a DOS data file in COMPAH format to a binary MATLAB™ -mat file.

COMPAH stores the matrices from multiple runs on a single *.mat file, so you must provide different names for the matrices. Each matrix must be given a different 4-character name (*e.g.*, H001, H010). After exiting COMPAH, all of the matrices in this .mat file can be loaded into MATLAB™ using the MATLAB™ LOAD command (*e.g.*, load testout).

After storing the standardized data matrix in a binary mat file, the program will prompt you asking whether you also want the data printed as an ASCII file (*i.e.*, a DOS text file). It is good to check whether the data have been transferred correctly to MATLAB™ .

Printing a lower triangular similarity matrix

If PrSmM is set to 4 or 5, COMPAH will print the similarity matrix in lower triangular form as a separate file. Many other computer programs, like the Bell System KYST2A program, can read this lower triangular matrix. COMPAH will prompt you for a filename for this similarity matrix. If #RUNS=2, COMPAH appends additional similarity matrices to the same output file. This file contains a FORTRAN FORMAT statement in G format which most FORTRAN compilers recognize. This lower triangular matrix is very large for large data sets.

Printing a similarity matrix as a Matlab™ .mat file

COMPAH96 can write any similarity or distance matrix to a binary .mat file for processing by MATLAB™ . I use this option to perform principal coordinates analysis on distance matrices (see next section). To request this option use PrntSM=6. The output .mat file, when loaded into MATLAB™ , will

be a single long vector. This vector can be converted to a full distance matrix using the following MATLAB™ statements (these statements assume that the distance matrix was stored by COMPAH96 in a *.mat file called DISSIM.MAT using the variable name DIST).

```
load dissim  % This loads the matrices stored on DISSIM.MAT
               % DISSIM.MAT was created by COMPAH using PrntSM=6.
               % COMPAH prompts for the name of the mat-file and
               % the name of the matrix to be stored (e.g., DIST)
lv=floor(sqrt(2*length(DIST)))+1;
D=zeros(lv,lv);I=find(~tril(ones(lv,lv)));
D(I)=DIST;D=D+D';
```

Please note that the main diagonal of this D matrix will be all zeros. This is appropriate for dissimilarity indices. If your matrix is a similarity matrix, like NNESS, you must add one more statement to change the main diagonal to all ones: D=D+eye(size(D));

Gower's Principal Coordinates Analysis (PCoA)

Here is how easy it is to perform a principal coordinates analysis, also known as Torgerson's metric scaling, on the D matrix generated above. Using MATLAB™, Just type **Z=pcoa(D)**. This calls the following MATLAB™ program pcoa.m. The sample coordinates are found in the sample-by-dimension Z matrix.

```
function [Z,lambda,VAR,CVAR,Q]=pcoa(Dist);
% Gower's Principal Coordinate Analysis: NO GRAPHS
% Adapted from Marcus's PRCRND2.m program in Reyment &
% Joreskog's Applied Factor Analysis.
% format:  [Z,lambda,VAR,CVAR,Q]=PCOA(Dist);
% where, Dist is any dissimilarity matrix.
%        Z=principal coordinates
% Similarity matrices should be converted to
% dissimilarity matrices (e.g., Dist=ones(N,N)-NNESS) prior
% to calling PCOA(Dist);
% output:
%     lambda=eigenvalues
%     Z=Station coordinates for positive eigenvalues.
%       (n.b., for non-metric indices only the first few columns
%              of Z may be useful - check the lambda vector for negative
%              eigenvalues!)
%     VAR, CVAR=Variance and cumulative variance
%     Q=standardized association matrix.
% Refs: Gower, J. C. 1966.  Biometrika 53:325.
% L Marcus's appendix in Reyment & Joreskog.  1993. Applied
% Factor Analysis, Cambridge U. Press., Gower (1987, p. 28)
% Adapted by E. Gallagher Environmental Sciences Program
% Internet: Gallagher@umbsky.cc.umb.edu, revised: 6/2/95
[N,N]=size(Dist);
Dist=Dist.^2;  % Must convert Euclidean distances to Dist.^2
               % prior to transformation in order to have PCO distances
               % match Euclidean distances, see discussion of equ. (4)
               % in Gower (1966)
cmean=mean(Dist); % column means of squared distance matrix
cm=cmean(ones(N,1),:); % creates order N matrix, each row=cmean;
% create matrix of grand means: gm
gm=mean(cmean);gm=gm(ones(N,1),:);gm=gm';gm=gm(ones(N,1),:);
Q=-.5*(Dist-cm-cm'+gm);
[V,S]=eig(Q);  % eigenanalysis of Q.
% these three statements sort eigenvalues in ascending order and
% sort eigenvectors accordingly:
```

```
[lambda,k]=sort(diag(S));
lambda=flipud(lambda); % rearrange to descending order
V=V(:,k); % sort eigenvectors in descending order
V=fliplr(V); % eigenvectors corresponding to sorted lambda
VAR=abs(lambda)/sum(abs(lambda))*100; % modulus
CVAR=cumsum(VAR);
posl=find(lambda>0);
Z=V(:,posl)*diag(sqrt(lambda(posl)));
```

Printing an AMOVA lower triangular similarity matrix

If PrSmM is set to 7, COMPAH will print as a separate file a distance or similarity matrix in lower triangular form **with main diagonal**. The first record of this similarity matrix contains a list of station numbers from 1:N. This is the input form required by Excoffier *et al.*'s (1992) AMOVA program for Windows. AMOVA, short for Analysis of Molecular Variation, performs a Type II ANOVA showing the % of variation attributable to a nested hierarchy of samples (replicate samples within stations, stations within regions, region *vs.* region). AMOVA performs significance tests based on random permutations of the similarity matrix. COMPAH will prompt you for a filename for this similarity matrix.

The AMOVA model requires the square of metric distances. There are only two squared metric distances available in COMPAH96 (SIMOPT=13, Euclidean distance squared; SIMOPT=23, CNESS^2). If either of these indices are requested, PrntSM=7 prints the distances without further transformation. If a similarity index is used, PrntSM=7 will calculate the square of the 1-complement of the similarity index (1-Similarity index). If a dissimilarity index is used (other than SIMOPT=13 or 23), then PrntSM=7 will square this dissimilarity before printing the AMOVA distance-squared matrix.

The main diagonal of the lower triangular AMOVA matrix will be filled with 0's.

If #RUNS=2, COMPAH will request a new DOS output file name for each new AMOVA distance matrix. The present upper limit of AMOVA is 256 samples (earlier versions of AMOVA could only handle about 200 samples).

Printing trees

Tree format

COMPAH prints the exact clustering levels at which groups fuse. COMPAH then segments the range of clustering levels into classes. COMPAH will prompt the user for the number of classes. The default is 25 classes and the maximum is 35 classes. Since each class takes 3 output columns to print and labels take 27 columns, 35 classes are the maximum that can be printed on mainframe 132-column-width line printers. COMPAH's dendrograms or trees may have several items clustering at the same class. COMPAH writes a table showing the distance or similarity level corresponding to each of the classes. COMPAH can also space the classes using log10 based classes. Table 10 lists the tree printing options.

Table 10 COMPAH's Tree printing options. Only a handful of the hundreds of Windows True type fonts can print the graphics characters used in COMPAH (see text).

PTree	Description	Requirements
0	No tree printed.	
1	Line-printer tree, composed of '---I' symbols. Single-spaced.	Any screen or printer
2	Graphics characters, single-spaced output	Screen, LaserJet printer (& some dot matrix printers)
11	Line-printer tree, composed of '---I' symbols. Single-spaced output. Log10 spaced classes.	Any screen or printer
12	Graphics characters, Single-spaced output. Log10 spaced classes.	Screen, LaserJet (& some dot matrix printers)
13	Line printer tree, Double-spaced output. Log10 spaced classes.	Any printer
14	Graphics characters, Double-spaced output. Log10 spaced classes.	Screen, LaserJet (& some dot matrix printers)
21	Line-printer tree, composed of '---I' symbols. Double-spaced output.	Any screen or printer
22	Graphics characters, Double-spaced output.	Screen, LaserJet (& some dot matrix printers)

The standard tree (PTree=1), suitable for any output device, is composed of ASCII dashes (--), and I's (I). COMPAH can print a high resolution tree using ASCII graphics characters 179 through 218. For viewing with most computer screens and output on many printers (but not EPSON dot matrix printers or mainframe line printers), you can request PTree=2, 12, or 22. Then, the tree will be constructed with these ASCII graphics characters. The VAX VMS FORTRAN compiler was incapable of writing these ASCII codes (this may have changed since 1994).

To aid interpretation, COMPAH can print the TREE diagrams using log10-scaled clustering levels (PTREE options 11-14). COMPAH still prints the exact clustering levels on the output. If Euclidean distance squared is the similarity (SIMOP=13), COMPAH prompts the user from the console asking whether it should print the TREE using the SQRT(clustering level). Pielou (1984) used this convention. COMPAH can cluster but not accurately display non-monotonic clustering levels (*e.g.*, Pielou's (1984) Fig. 2.18). These non-monotonic clustering levels are also called cluster reversals and inverted cluster levels.

Double-spaced trees can be edited in most Word Processors to produce publication-quality figures. The trick is to find a font that contains the graphic characters and uses identical widths for each character. Lately, I've been using the Adobe Type 1 Courier font to print trees, but the True type Courier New font also usually works. Lotus Linedraw and IBMPCDos fonts also print the trees.

Different fonts may be needed on different printers. I have used four HP printers to print trees: an HP Laserjet 4M & Copyjet (with PCL & eps drivers), and three HPDeskjets (550c, 722c & 820). With non-postscript print drivers, the WordPerfect Courier New font prints the trees well. This True Type font doesn't print properly with the Adobe postscript drivers. For postscript, I switch to the Adobe Type 1 Courier font and set the kerning to normal. This font is available with Adobe Type Manager, part of the Adobe Acrobat 3.0, which sells for $49 (Academic Price).

Tree examples

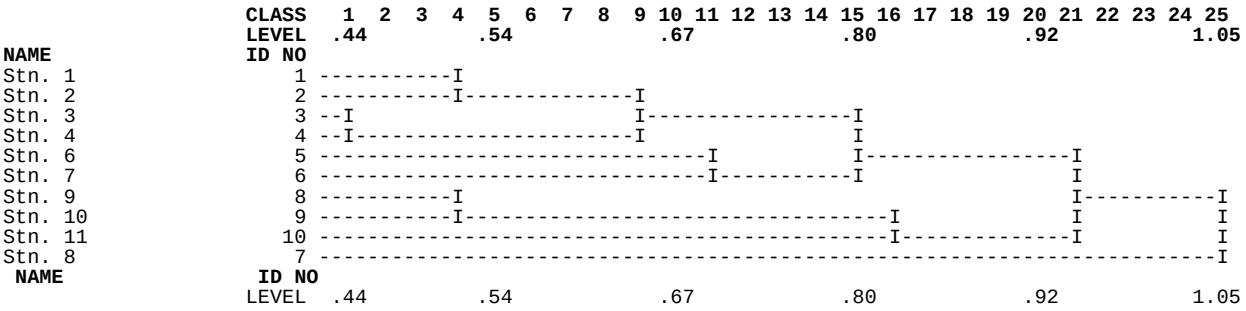

Figure 2. A single-spaced tree produced by **PTREE=1**. With HP pcl drivers, print the tree with a Courier New or Courier T1 7.5 pitch (non-proportionally spaced font). Linespacing was set to 1.5

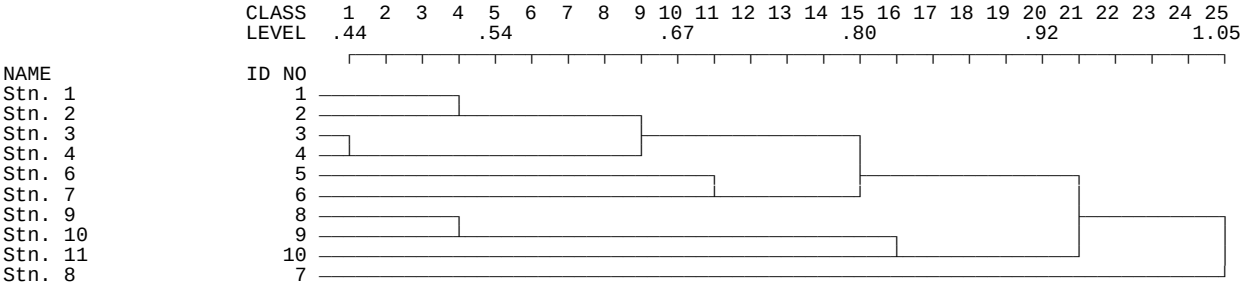

Figure 3 A single-spaced tree, composed of graphic characters, produced by **PTREE=2**. This tree can be printed with the Courier New 7.5 pitch font for non-postscript printers and with Courier T1 or Lotus Linedraw (available with Adobe Type Manager) for encapsulated postscript printers.

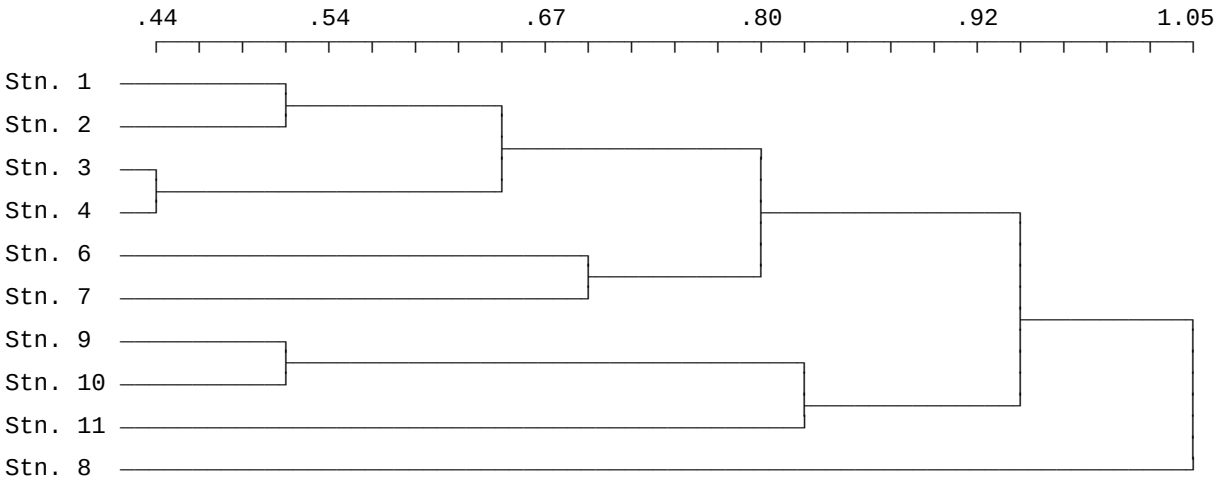

Figure 4 A double-spaced tree produced by **Ptree=22**. This tree can be printed using either single-spacing or ½- line spacing (this tree produced with 0.9 line spacing, Courier New 9.0 font) for non-postscipt printers and with Lotus Linedraw (available with Adobe Type Manager) for encapsulated postscript printers.

It is always a challenge to find a combination of font and printer that will print these trees properly. The font and printer must print with no kerning (each character must take up exactly the same amount of horizontal space). Only a small number of True type and Adobe Type 1 fonts will both print with no kerning and have the ASCII graphics characters (e.g., ┤, ├, ┐) necessary to print the trees. Back in the good old DOS days, it was easy. WordPerfect 4x or 5.x with the Courier New font printed the graphic characters in the trees perfectly. Now with Windows' WYSIWYG, the combination of fonts and printers must be chosen carefully. I've been using the Adobe Type 1 Courier or Lotus LineDraw fonts, available as part of the Adobe Acrobat package (which sells for $50 academic) to print the trees using any of the HP printers named above or printing to pdf's using Adobe Acrobat. Of the Windows TrueType fonts, Courier New and MSLineDraw work well. I've had great difficulty getting MSWord to print the trees.

Making graphics files from trees

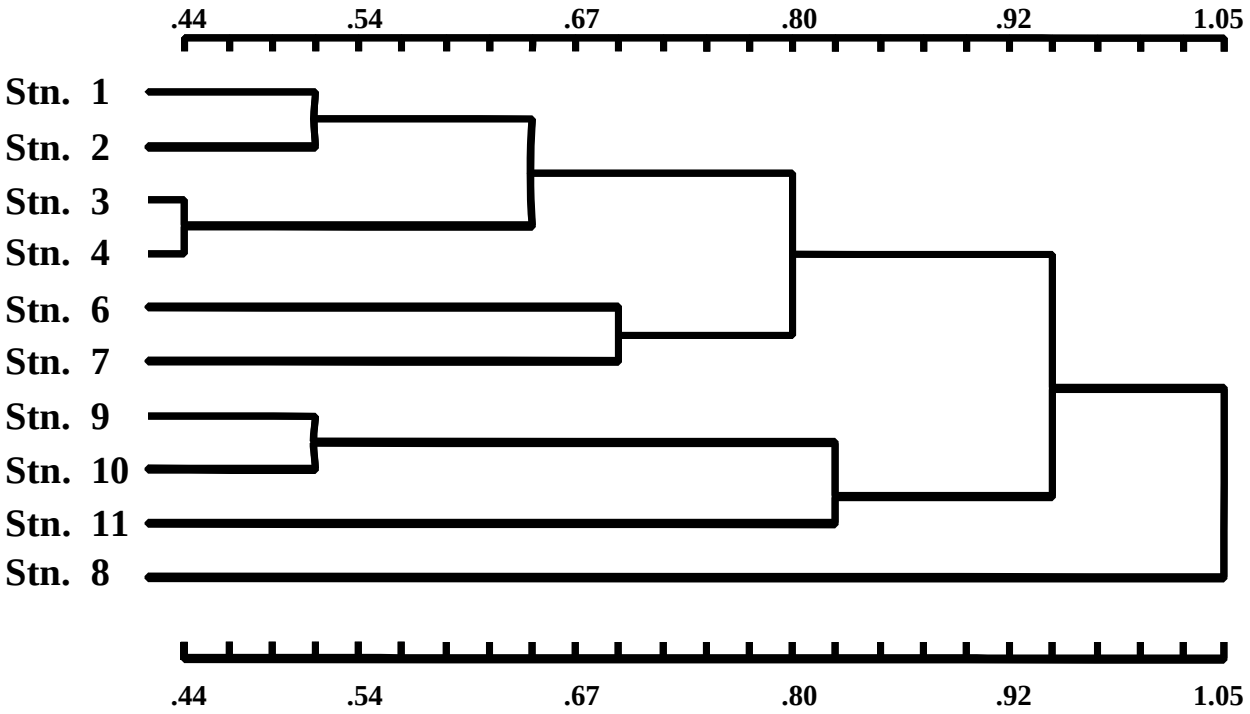

Figure 5. The same tree from the previous display, converted to a vector image file using WordPerfect 8.0 and Corel Presentations 8. One way to get a vector based-tree tree is to scan the tree as a bitmapped image (bmp or tif) and convert it to a vector-based graphic with Corel's "trace bitmap" tool . The lettering must be retyped in Presentations.

As shown in Figure 5
, trees of graphics characters, can be made into a vector-based graphic.

Figure 6 shows the same vector-based graphic used in Fig. 6. It only took two minutes to change the size and color of the tree since the graphics are now vector-based.

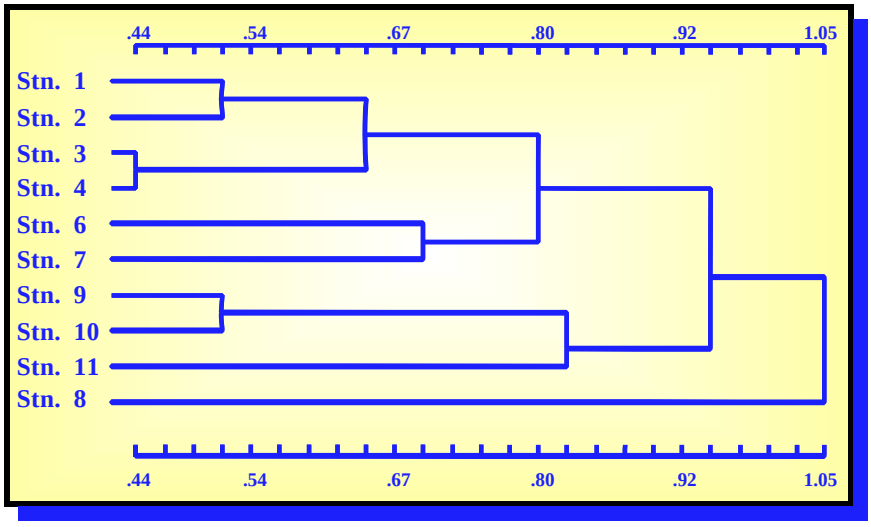

Figure 6. The same tree from the previous display, with the colors changed to blue, the size reduced and with WordPerfect borders and fills..

Reducing the output from large data sets

The output from large data sets can be very large (several MB of disk space). You can set the **COMM.DAT** file to suppress much of the output. Be especially wary of printing out huge lower triangular similarity matrices. The lower triangular matrix for a 250 sample Q-mode analysis takes 32 pages to print. Here are some calls in the **COMM.DAT** file (described below) that produce a short output:

```
 9  PrEdD   1=> Print edited data;0=> Don't print;2 => Print & Save        0.000
10  PrTrD   1 => Print transformed data matrix;  0 => Don't print          0.000
11  PrStD   0 => Don't print;1=>Print; 2 =>Print & Save 3=>MATLAB.mat      0.000
12  PrSmM   1=Print SIM;2=Dont print;3=Dont calc.;4=print&save;5=Save      2.000
```

Stopping the program

COMPAH will continue to analyze the data as long as #RUNS in the **COMM.DAT** file is set to 2. Change #RUNS at the console prompt to 1 to end the program after the next analysis (or 0 to end immediately).

If you must stop the program in progress, type the Ctrl and Break keys simultaneously. Once calculations are being done within a subroutine, Ctrl & Break won't work. If all else fails, turn off the computer or use Alt-Ctrl-Delete to stop the program.

OTHER PROGRAMS

I distribute FORTRAN programs to calculate the NNESS similarity index, called NEWNESS, and Hurlbert's (1971) rarefied species number, $E(S_n)$, called ESNNEW. These programs also use the COMPAH data format. I also wrote a FORTRAN program called MATCOMP.FOR that converts MATLAB™ mat-files to COMPAH data input format. As described above, this feature is now built into COMPAH (data files ending in *.mat will be analyzed as MATLAB™ binary mat files). It is possible to calculate Hurlbert's diversity for a sample using COMPAH by standardizing the data to Hypergeometric probabilities (STAND=10), storing the standardized data as a COMPAH data input file (PrStD=2), and reading the data back into COMPAH. The sample total calculated using Hypergeometric probabilities, which COMPAH prints for every analysis, is Hurlbert's $E(S_n)$. I have a separate PC-program called ESNNEW.FOR that calculates the $E(S_n)$ values for any n.

For users owning MATLAB™ , I distribute m.files for calculating the original NESS (Grassle & Smith 1976), NNESS (the modified version of NESS, see Trueblood *et al.* 1994), & CNESS. I wrote a MATLAB™ program called findcnm.m to calculate the Kendall's tau correlation among CNESS and NESS matrices (see Trueblood *et al.* 1994). I also distribute programs to calculate the Sanders-Hurlbert rarefaction curves for samples (Sanders 1968, Hurlbert 1971), a jackknifed estimator for Hurlbert's $E(S_n)$ for replicated samples, Gower's (1966) principal coordinates and Procrustes analysis, and factor analysis. I wrote an m.file that performs a complete ordination of samples based on CNESS, a technique called PCA-H [Principal component analysis of Hypergeometric probabilities] (see Trueblood *et al.* 1994). Many of these files are included in the DDTMAT35.EXE and DDTMAT42.EXE self-executing zip files, which are available by anonymous ftp (see p. 5).

For correspondence analysis, I use Greenacre's (1984) correspondence analysis program SIMCA, which Greenacre sells for about $70.00. For non-metric multidimensional scaling, I use the Bell System labs KYST2A program, developed by Kruskal, Young & Seery and described by Carroll (1987). The lower triangular similarity or dissimilarity matrices generated by COMPAH (PrSmM=4 or 5) can be edited in a matter of minutes for KYST2A input. The KYST2A and other Bell System programs are available by anonymous ftp (use a web search program to find the current location).

COMPAH, MATLAB™ , AND PCA-H

I maintain COMPAH primarily for my own research. I'm pleased if others find it useful. I use COMPAH mainly to complement my own work on PCA-H, an ordination technique. Trueblood *et al.* (1994) describes **PCA-H**, which stands for **P**rincipal **C**omponents **A**nalysis of Hypergeometric **P**robabilities. **PCA-H** performs a metric scaling of the sample distances calculated using the CNESS metric. PCA-H is an ordination technique that produces a metric scaling of samples, in which the distances among samples in the 1-, 2-, 3- or higher dimensional displays is a best least-squares fit to the underlying CNESS distances among samples. The results of the metric scaling portion of PCA-H are mathematically identical to a Gower's Principal Coordinates Analysis of CNESS distances (see p. 40). There is a tremendous advantage of PCA-H analysis in that the exact contribution of each species to the CNESS distances can be calculated and displayed graphically using the Gabriel (1971) Euclidean biplot display (see Trueblood *et al.* 1994). Here is an abbreviated listing of how COMPAH and MATLAB™ can be used to complete a full PCA-H analysis of a standard sample x species data matrix. The MATLAB™ m.files necessary to perform these analyses are available from both the Rutgers anonymous ftp site and Gallagher's web page (see p. 5):

- Format data for COMPAH input using any of COMPAH's input formats. I usually use list-directed input. The major hassle in setting up a data input file is counting the columns in record 1. All standard Windows-based wordprocessors display the cursor position in inches (or mm's), not in columns. I still edit data files and programs in WordPerfect 5.1 for DOS, which has the option for displaying column numbers (WP4.2 display type) instead of inches. Alternatively, there are shareware programming editors like E! available on the web which display units in column numbers rather than inches.
- On the first pass through COMPAH, I only want COMPAH to rewrite the data as a MATLAB™ binary *.mat file. I set PrStD=4 to write the DATA as a MATLAB™ mat file (see p. 39). I usually set NESSm=1 so that no samples are dropped (samples containing < NESSm individuals are dropped before printing the standardized data). To ensure that no samples are dropped in the first pass through, you can set NESSm=0 and SIMOPT=7 (Euclidean distance). COMPAH stops sample deletion with this pair of commands. Note that COMPAH writes the DATA matrix in Standard order (Species as rows, Samples as columns).
- Enter MATLAB™ . Load the MATLAB™ .mat file and transpose it to sample x species order (I wrote all of my MATLAB™ m.files to process sample x species matrices)
- Run findcnm.m. This MATLAB™ m.file, written by Gallagher, will find a consensus NESSm sample size that produces a CNESS index that is equally correlated (with Kendall's τ) with NESSm=1 and NESSm=(minimum sample total). This will produce an index that is sensitive to both the rare and abundant species. Trueblood *et al.* (1994) shows a display generated by findcnm.m.
- Run COMPAH with CNESS and the NESSm sample size found by the findcnm.m program I almost always use UPGMA clustering (SIMOPT=1, see p. 9, 24, 38) for Q-mode analysis.
- Perform an R-mode cluster analysis of the data, using **COMM.DAT** parameters shown above in the section **"Clustering species with CNESS"** (p. 22).

- Run a PCA-H analysis. In addition to the standard binary mat.file containing the data, my MATLAB™ PCA-H programs require files to label stations and species. I usually quickly edit the end of the COMPAH data input file to create two MATLAB™ text matrices containing 3- or 4-character species labels and station labels, called SPLAB and STLAB:

```
SPLAB=['sp1';'sp2';'sp3'];STLAB=['st 1';'st 2';'st 3'];
```

- After a MATLAB™ PCA-H analysis, the data file can be edited so that only the important species are included in an R-mode cluster analysis. For example, the R-mode cluster diagram in Trueblood *et al*. (1994) shows only the eleven species that contribute at least 2% of the variation to CNESS distances among samples. To produce this cluster diagram, I saved the lower triangular portions of the MATLAB™ **PCA-H XR** matrix (the matrix of standardized Hypergeometric probabilities) and clustered this matrix with Pearson's r and single-linkage clustering as described in the section **"Clustering species with CNESS"** (p. 22) The MATLAB™ *.mat file containing the XR matrix will not have the corresponding species labels. These can be printed out as an ASCII file from MATLAB™ and the species labels added with a Word Processor. The resulting R-mode cluster analysis will have a direct one-to-one correspondence with the angles among vectors in the covariance biplot produced by the PCA-H analysis. The Pearson's r values can be converted to angles, given the relationship that Pearson's r=cos(θ) among species vectors, with the angle being θ.

REFERENCES

Anderberg, M. R. 1973. Cluster analysis for applications. Academic Press, New York and London. *[4]*

Boesch, D.F. 1977. Application of numerical classification in ecological investigations of water pollution. Report prepared for US EPA Ecological Research Series (EPA-600/3-77-033). Available as PB269 604 from: National Technical Information Service. U.S. Dept. of Commerce. Springfield VA 22161. *[4, 7, 9, 13, 16, 24, 27, 38, 48, 52]*

Carroll, J. D. 1987. Some multidimensional scaling and related procedures devised at Bell laboratories with ecological applications. Pp. 65-138 *in* P. Legendre and L. Legendre, eds., Developments in Numerical Ecology. Springer Verlag, Berlin. *[46]*

Clifford, H. T. and W. Stephenson. 1975. An introduction to numerical classification. Academic Press, New York. *[7, 9, 11, 15, 16, 24, 27, 31, 34, 48]*

Excoffier, L, P. E. Smouse, and J. M. Quattro. 1992. Analysis of molecular variance inferred from metric distances among DNA haplotypes: application to human mitochondrial DNA restriction data. Genetics *131*: 479-491.*[41, 48, 52]*

Gabriel, K. R. 1971. The biplot graphic display of matrices with application to principal component analysis. Biometrika *58*: 453-467. *[22, 47]*

Gower, J. C. 1966. Some distance properties of latent root vector methods used in multivariate analysis. Biometrika *53*: 325-338. *[11, 40, 46]*

Gower, J. C. 1967. Multivariate analysis and multidimensional geometry. The Statistician *17*: 13-28. *[11]*

Grassle, J. F. and W. Smith. 1976. A similarity measure sensitive to the contribution of rare species and its use in investigation of variation in marine benthic communities. Oecologia *25*: 13-22. *[4, 8, 9, 10, 11, 17, 18, 23, 46, 52]*

Greenacre, M. J. 1984. Theory and application of correspondence analysis. Academic Press, Orlando Fl. *[46]*

Grieg-Smith, P. 1983. Quantitative plant ecology. U. California Press, Berkeley. *[10, 26, 27]*

Hurlbert, S. M. 1971. The non-concept of species diversity: a critique and alternative parameters. Ecology *52*: 577-586. *[8, 46, 48, 52]*

Jardine, N. and R. Sibson. 1968. The construction of hierarchic and nonhierarchic classifications. Computer J. *11*: 177-184. *[27, 28]*

Kenkel, N. C. and L. Orloci. 1986. Applying metric and nonmetric multidimensional scaling to ecological studies: some new results. Ecology *67*: 919-928. *[23]*

Lance, G. N. and W. T. Williams. 1966. Computer programs for hierarchical polythetic classification ("similarity analyses"). Comput. J. *9*: 60-64. *[15]*

Lance, G. N. and W. T. Williams. 1967. A general theory of classificatory sorting strategies. I. Hierarchical systems. Computer journal. *9*: 373-380. *[4, 22, 24, 25, 26, 27]*

Legendre, L. & P. Legendre 1983. Numerical Ecology. Elsevier, Amsterdam. *[9, 11, 12, 27]*

Legendre, P. and L. Legendre. 1998. Numerical Ecology Second English edition. Elsevier, Amsterdam. [9, 10, 11, 27]

Magurran, A. E. 1988. Ecological diversity and its measurement. Princeton University Press, Princeton. *[9, 10, 11, 16, 17, 18, 19, 23]*

Morisita, M. 1959. Measuring of interspecific association and similarity between communities. Mem. Fac. Sci. Kyushu Univ. Ser. E. (Biol.) *3*: 65-80. *[9, 16, 17, 19]*

Orloci, L. 1967. An agglomerative method for classification of plant communities. J. Ecol. *55*: 193-205. *[10]*

Orloci, L. 1978. Multivariate analysis in vegetation research, 2nd ed. Junk Publishers, The Hague, Netherlands. *[10, 24, 27, 52]*

Pielou, E. C. 1984. The interpretation of ecological data. A primer on classification and ordination. John Wiley, New York. *[5, 9, 10, 11, 12, 13, 24, 27, 28, 42, 52]*

Sanders, H. L. 1960. Benthic studies in Buzzards Bay. III. The structure of the soft-bottom community. Limnol. Oceanogr. *5*: 138-153. *[9, 13]*

Sanders, H. L. 1968. Marine benthic diversity: A comparative study. Amer. Natur. *102*: 243-282. *[8, 46]*

Smith, W. and J. F. Grassle. 1977. Sampling properties of a family of diversity measures. Biometrics *33*: 283-292. *[8]*

Sneath, P. H. A. and R. R. Sokal. 1973. Numerical taxonomy - the principles and practice of numerical classification. W. H. Freeman, San Francisco, 573 pp. *[24]*

Trueblood, D. D., E. D. Gallagher, and D. M. Gould. 1994. The three stages of seasonal succession on the Savin Hill Cove mudflat, Boston Harbor. Limnol. Oceanogr. 39: 1440-1454. *[4, 10, 18, 21, 22, 23, 46, 47, 48]*

WATCOM FORTRAN 77^{32} Optimizing Compiler User's Guide 4th Ed. WATCOM International Corporation. Waterloo Ontario, Canada

Williams, W. T. 1971. Principles of clustering. Ann. Rev. Ecol. Syst. *2*: 303-326. *[28]*

TECHNICAL APPENDIX

Printing this document

This document, called COMPAD96.wpd, was written with Wordperfect 8.0 with many formatting features set as styles, saved as sty8-compapdf.wpd. You only need to use the styles file if you overwrite my styles with your own. The index is generated with a concordance file called indx-compa.wpd. I also distribute a Microsoft Word 6.0a version (COMPAD96.DOC), but this documentation does not have the full graphics and equations found in the WordPerfect document.

Comments on computation time and precision

After reading the DATA and **COMM.DAT** file, COMPAH works only with RAM and accesses disk files only to write results. Most cluster analyses are performed in seconds on 386, 486 & Pentium computers. A 1300-sample x 20 species data set can be analyzed with COMPAH dimensioned with WORK=1e6. This analysis takes 14 minutes on a VAX 11/780 mainframe computer, slightly less time on a 486 PC, and is completed in less than a minute with a Pentium 90.

COMPAH uses double precision math for calculations within subroutines but passes results between most subroutines with the single-precision WORK array. This reduces the size of the executable file, but with reduced accuracy. Redimensioning this WORK array to double precision would reduce by half the size of data set that COMPAH could analyze.

CNESS & NNESS use double precision math

STAND=14 and STAND=15 and SIMOPT=26 (CNESS) and SIMOPT=27 (NNESS) calculate hypergeometric probabilities without rounding data to the nearest integers. Most calculations are performed and variables saved in COMPAH with single-precision. The calculation of hypergeometric probabilities is performed with double precision (approximately 13 significant figures). If SIMOPT = [10, 15, 26 or 27] the NESS and CNESS similarities and distances are calculated in full double precision from DATA to final similarity or distance matrix. CNESS can be calculated by requesting STAND=12 (HYPERGEOMETRIC probabilities (rounded data) & Norm. standardization)or STAND=15 (HYP (cont) & Norm. standardization) and then calculating Euclidean distance. Due to a programming quirk, this calculation is 4 to 100 times faster (increasing with the square of matrix size) than calculating CNESS directly (using SIMOPT=15 or SIMOPT=26), however precision is lost as the double precision hypergeometric probabilities are passed as single-precision variables to the Euclidean distance routine. This loss of precision can result in slight differences in clustering patterns.

The 1996 sorting bug and processing time

Centroid clustering will take much longer than other clustering options. All versions of COMPAH were written so that the whole lower triangular matrix is searched completely only once during sorting. This led to a rare bug in versions of COMPAH prior to COMPAH96 (see differences between COMPAH95 and COMPAH96 on p. 51). With centroid clustering, the entire lower triangular matrix must be searched at each cycle to find the next pair of items to cluster. With the other cluster methods, only a few rows need to be searched.

I don't recommend using the WATCOM Virtual Memory Manager (VMM) unless absolutely necessary. Request a version of COMPAH96 that can be run with your RAM. Using a swap file increases the computation time about ten-fold.

VAX FORTRAN modifications

I used to maintain a VAX VMS version of COMPAH, but I no longer even try to get a version running with VAX or Digital FORTRAN. I still include statements in the source code that are needed to compile using VAX VMS FORTRAN. However, there are several features of COMPAH which were not available on VAX FORTRAN, especially the option of writing the graphics characters, the intrinsic gammaln function (for calculating the natural log of the gamma distribution) necessary to print publication-quality trees and the ability to read and write MATLAB™ *.mat files.

Upgrades (& Bug Fixes)

Differences between COMPAH95 & COMPAH96

The major difference between COMPAH95 and COMPAH96 is in the main clustering algorithm: SUBROUTINE CLUSTR. Previous versions of COMPAH, including the original COMPAH, had a logical error which could produce inverted clustering levels with UPGMA clustering (SIMOPT=1), a mathematical impossibility. Karen Stocks at Rutgers found this bug during summer 1996. The CNESS distance matrix that produced the anomalous result is shown on p. 36. COMPAH96 now allows such ASCII lower-triangular matrices to be read using list-directed input (see p. 31) and clustered. They could be read into previous versions of COMPAH only as binary MATLAB™ mat files.

The logical error in previous versions of COMPAH was due to a modification of Anderberg's (1973) original clustering algorithm (Subroutine CLUSTER in Anderberg's Appendix F). After calculating a similarity or dissimilarity matrix, COMPAH's SUBROUTINE CLUSTR found the extreme pair of similarities (maximum similarity) or dissimilarities (minimum dissimilarity) in each row of the lower triangular matrix. The extreme of these row extremes was found to identify the first two items for clustering.. After these two items (samples or species) were fused, the new similarities or dissimilarities between this fused group and the remaining items were calculated using Lance & Williams' (1967) combinatorial formula. COMPAH retained in memory the row extremes for each row of the lower triangular similarity or dissimilarity matrix and the column number for this extreme row element. On subsequent clustering cycles, COMPAH searched a row to find a new extreme only if the row was one of the two groups fused on the previous pass, or the row's previous extreme was from a column corresponding to one of the items fused on the previous pass. These two criteria speed up the program tremendously, because instead of searching the entire matrix, only a handful of rows are reexamined on each cycle. These row extremes were rechecked to find the new global extreme for the remaining items in the cluster analysis. Unfortunately, the logic behind this algorithm is flawed. COMPAH did not check all of the elements of a full similarity or dissimilarity matrix. If a row's extreme value was found in a column exceeding the row number (*i.e.*, not an element in the lower triangular matrix), then COMPAH could, under very rare circumstances, miss the true global extreme. COMPAH could cluster items with similarities or dissimilarities that were not the extreme for the entire matrix. The lower triangular matrix, shown on p. 36, produces an erroneous clustering pattern. The following listing shows the flawed clustering pattern from previous versions of COMPAH and the correct clustering pattern from COMPAH96 using the Stocks data on p. 36.

All previous COMPAH's- Note the inverted clustering pattern (.847->.955->.946)

NO. GROUPS	LEVEL	GROUPS			ITEMS INCLUDED							
7	0.579	6 AND	7	:	6	7						
6	0.657	1 AND	2	:	1	2						
5	0.847	3 AND	6	:	3	6	7					
4	0.955	1 AND	3	:	1	2	3	6	7			
3	0.946	1 AND	5	:	1	2	3	5	6	7		
2	1.22	1 AND	4	:	1	2	3	4	5	6	7	
1	1.28	1 AND	8	:	1	2	3	4	5	6	7	8

COMPAH96 - The correct solution to the Stocks lower triangular matrix.

NO.	GROUPS	LEVEL	GROUPS			ITEMS INCLUDED							
	7	0.5790	6 AND	7	:	6	7						
	6	0.6570	1 AND	2	:	1	2						
	5	0.8465	3 AND	6	:	3	6	7					
	4	0.9060	3 AND	5	:	3	5	6	7				
	3	0.9674	1 AND	3	:	1	2	3	5	6	7		
	2	1.217	1 AND	4	:	1	2	3	4	5	6	7	
	1	1.276	1 AND	8	:	1	2	3	4	5	6	7	8

List of modifications

I began working on COMPAH in 1984. The first version was compiled to run on PC's with the original Microsoft FORTRAN compiler. It took nearly an hour flipping diskettes in and out of my original IBM PC to recompile the program. The following list of modifications have been made on the program since 1990.

7/90 I corrected my programming error in the SUBROUTINE STAND. Earlier versions of this PC versions were incapable of performing data standardization properly.

5/92
1) I added Orloci's chord-distance metric (see Orloci 1978), the geodesic metric (see Pielou 1984), squared Euclidean distance, the Marczewski-Steinhaus distance (=1-Jaccard, see Pielou 1984), and a metric distance version of NESS called CNESS (Gallagher in preparation).
2) I changed the NESS sample size cutoff so the user can use m values up to the sample total. Grassle & Smith (1976) recommend only m values less than ½ of the sample totals.
3) To better view cluster diagrams using distance measures, I added a log-scale option for the TREE diagrams.
4) I added an option for printing the similarity matrix in a separate file (PrSmM=4,5). This saves some text editing time when using the similarity matrix in other programs (*e.g.*, NMDS)
5) I added STAND=10, which converts raw species counts to their hypergeometric probabilities. This standardized hypergeometric probability matrix can be printed out in COMPAH format (PrStD=2). Orloci's chord distance (SIMOPT=11) calculated with data standardized to hypergeometric probabilities is CNESS. This data file can be read by COMPAH as a separate data file. The sample sum of the hypergeometric probabilities equals Hurlbert's $E(S_m)$ diversity for the sample. I distribute a separate program called ESNNEW which will calculate Sanders-Hurlbert's $E(S_n)$ diversity for a variety of sample sizes in one run.
6) I changed the rounding of sample totals in NNESS (Subroutine SIMA). Instead of rounding the sample sum, I now sum the rounded species abundances. This makes a slight difference (tenths of a NESS unit) if the original data were non-integer.

10/92 I added Pielou's (1984) Percentage dissimilarity, percentage remoteness, and Ruzicka similarity (also described by Boesch 1977). I added the Boolean transform and the option to perform multiple standardizations on the same data.

11/92 I added PrStD=3 to print binary MATLAB™ -mat files.
Added SAVEMAT.FOR for reading the standardized data matrix directly into a Matlab™ -mat file. One *.mat file will be opened per run. All 4-character named variables will be stored in this binary file. The raw data can be stored by setting standardization and normalization options to 0 and saving with PrStD=3

1/93 Added 1/(Euclidean distance^2) distance measures. These distance measures are useful to generate spatial distance matrices for calculating Mantel statistics for spatial pattern. Since (0,0) spatial coordinates are possible, I cut off the sample delete call for this similarity index. Otherwise, samples with (0,0) coordinates would be dropped. $1/d^2$ needed for preparing spatial and temporal data for Mantel-type spatial analysis.
-I added complement of CNESS as a similarity measure

4/93 I changed the DDLOG look-up table in function fact. The old table wasn't in full double precision accuracy. I also replaced Stirling's approximation to the ln(factorial) with Press et al.'s gammln subroutine.

12/93 Changed the way memory is allocated to NSI in Subroutines SIMA and SIMF. MATLAB™ mat-routine working for VAX-compiled versions of COMPAH. Changed calculations within subroutines to double precision.

1/94 Added PTree>1, STAND=11-13, SIMOPT=25

2/94 Changed calculation of CNESS and NNESS and HYP Standardization if (Totali-Abundij)=NESSm. Added SIMOPT=26 & 27, added MENUS

9/94 Compiled COMPAH96 with WATCOM FORTRAN 77/32. Added PrntSM=7 which writes a distance^2 matrix in a form readable by Excoffier *et al.'s (1992) AMOVA program*

3/95 Added list-directed I/O, corrected error in SUBROUTINE READCM that could produce an infinite loop if an error was found in input.
6/95 Fixed minor bug that prevented NESSM values > 2 digits on console input.
7/95 Completely revamped calculations of NNESS & CNESS, greatly speeding up the calculations
10/96 Added new subroutine CLUSTR to COMPAH, correcting a bug identified by Karen Stocks at Rutgers. Added the option to enter and cluster lower triangular similarity or dissimilarity matrices.
11/96 Added Kulczynski similarity (SIMOPT 29 & 30)
3/97 Increased recl=1000 for data input
7/98 Changed output format for printing standardized data to 6G14.7 to allow more significant digits in PrEdD=2 & PrStD outputs.
7/98 Compiled compwb.exe, which will cluster an 8192 x 256 matrix.
5/99 Changed recl=1000 to recl=2000 in open statements in response to a user's request. This will permit records with up to 2000 characters to be read by compah.

Run-time and compilation errors

1. **WORK arrays too small**. If the data set is too large to fit in the memory allocated to the program, the size of the WORK vectors and the value of LIMIT can be increased in the COMPAH96.FOR source code.
2. **R6002 floating point not loaded** If your PC does not have a numerical co-processor and if the source code was compiled without the /FPi option, then this error will appear when COMPAH is called. The solution is to recompile with the Microsoft /FPi and /G0 options and the Microsoft LLIBFORE.LIB FORTRAN library or use the existing COMPAH96.EXE file on a computer with a numerical co-processor. Note, this is an old Microsoft FORTRAN compiler error.
3. **M6201: MATH -sqrt: DOMAIN error**. COMPAH attempted to take the square root of a negative number.
4. **File does not exist.** If the **COMM.DAT** file is not found on the default drive, the program will print an error message and stop. Copy the **COMM.DAT** file to the drive or subdirectory from which COMPAH96 is called.
5. **INAPPROPRIATE CLUSTERING METHOD** If an incompatible cluster method and similarity or distance measure are specified, the program will print an error message and prompt the user to change the parameter. For example, if the species are to be clustered (R-mode clustering), NESS is not an appropriate index.
6. **"End-of-file encountered"** This is a common error if the data format statement is wrong (record 2 of data file) or if the program is expecting to read species or station labels and none are there. If there are no labels, an X must be typed in columns 69 and 70 of the data file.
7. ****** WARNING**** CENTROID CLUSTERING MAY NOT BE COMBINATORIAL WITH CHOSEN SIMILARITY.** This warning will be printed with all but the geodesic metric, chord distance metric and Euclidean distance squared metric. The user can continue with centroid clustering or reset the similarity or cluster method options by typing a Y or y at the console prompt.
8. **DISK FULL** The COMPAH96 output file can be very large. If you receive this error, create more space on your calling drive, or cut some of the output options. The equivalent VAX error is: **disk quota exceeded**
9. **M6103: MATH - floating point error: divide by 0**. The only time I've seen this error is when trying to calculate Morisita similarity for samples composed only of singleton species. Send me a copy of the data that produced the error if you ever receive it for another type of problem.
10. **Negative factorial**. Produced sometimes when trying to calculate Hypergeometric probabilities from fractional data. Check your data file to ensure that species counts can be truncated to integers.
11. **DOS/4GW Fatal error: Loader failed -- LINEXE_LOADER** This WATCOM error is given if the WATCOM-compiled COMPAH program is too large for the RAM on your 386 or 486 computer. Recompile with a smaller version or find a computer with more RAM.
12. **%LINK-E-EXPAGQUO. exceeded page file quota** This VAX error is produced on my VAX account if I try to link COMPAH with WORK dimensioned >1000000. This system- and account-specific limit could probably be overridden.
13. **FORT-W-INCHASOU Invalid character in source treated as blank.** This VAX FORTRAN compilation error is produced because the graphics characters used to print high resolution trees in COMPAH's SUBROUTINE TREE are not recognized by VAX FORTRAN. This is a non-fatal error, and the program compiles and runs fine. PTREE=2, 12, 14 and 22 are unavailable in the VAX-compiled COMPAH.
14. **UNEXPECTED INTERRUPT.** This error occurred just after calling COMPAH96.EXE. COMPAH96 is compiled using Rational System's DOS4GW.EXE DOS extender program. The current version of COMPAH96 is compiled with WATCOM's F77/32 compiler and the programs will not work with earlier versions of DOS4GW.EXE. MATLAB, for example is bundled with an earlier DOS4GW.EXE extender and earlier versions of COMPAH96 were sent with an earlier version of DOS4GW.EXE. Make sure that the latest DOS4GW.EXE program (Compiled 1/11/94, size=254,566)

is on the calling directory, or the path statement leads to this version before others, or call compah96 with the correct subdirectory listed (if DOS4GW.EXE was on a directory called c:\WATCOM\bin), type C:\WATCOM\bin\dos4gw compah94

15. .***ERR*IO-6**. This error results from an improperly formatted data file, especially if the format statement doesn't match the format of the data. Correct the FORMAT statement or use list-directed input (type a * in line 2 of the data file).
16. ***ERR* IO-11 input item does not match the data type of list variable** This error results if the data are of a different type from that expected from the format statement. For example COMPAH expected an integer variable, but read a variable containing a decimal point. The equivalent Microsoft error is **F6101 - invalid integer**. COMPAH94 was expecting an integer when reading a data file and either a character or decimal point was read.
17. ***ERR* IO-21 invalid STATUS specifier for given file**. The equivalent Microsoft FORTRAN error is **F6415 file already exists**. You get this error if you try to open a file that doesn't exist (*e.g.*, you mistyped the name) or save to a file that should exist, but doesn't. Make sure that the **COMM.DAT** and data file are on the calling directory used for COMPAH or that these files are on a DOS PATH in the autoexec.bat file. You can type in the full path name for files if you don't want to change the default directory or DOS Path (*e.g.*, type: c:\mydata**COMM.DAT**).
18. ***ERR* IO-27 formatted record or format edit descriptor is too large for record size** The maximum length of a record in WATCOM FORTRAN is 256 characters. The FORMAT statement on the second line of the data file described a data input line exceeding this limit. I have increased this recl to 400 characters in versions of COMPAH after 3/1/95 (This limit increased to 1000 in March 1997).
19. ***ERR* PC-02 missing or misplaced opening parenthesis**. A closing parenthesis was found in a statement where parenthesis are not expected. This error results from an incorrect format statement on the 2nd line of the data file.
20. **DOS/4GW fatal error (1307) not enough memory.** I receive this error if the COMPAH version requires more memory than is available. Solution: Use a COMPAH version compiled with a smaller work array, or use the DOS4GW Virtual Memory Manager, see p. 56.
21. **NO MORE WORKSPACE. INCREASE WORKSPACE BY XXXXX AND RECOMPILE**. COMPAH allocates all data and intermediate data matrices to one large single-precision vector (called WORK- it is accessed as IWORK & RWORK with an equivalence statement - allowed in WATCOM but not in stricter FORTRAN77 compilers) This error message is produced if there isn't enough space to complete the calculations. There are now 6 versions of COMPAH, one Windows and 5 DOS based. The five DOS based versions have workspaces ranging from 100,000 to 10,000,000. Use one of these DOS versions for huge data. Email me and I can send you a Windows version with a larger work space. The present Windows version was compiled to run with only 8 MB RAM (but can take advantage of more RAM).
22. **DOS/4GW fatal error (1307): not enough memory**. Error received in DOS mode running COMP96TG.exe with 64 MB RAM. Solution: use the windows versions which use Windows memory management (create swap files).

Compiling the source code with WATCOM F77/32

The earliest PC versions of COMPAH were written and compiled using Microsoft FORTRAN. In August 1994, I recompiled COMPAH with the WATCOM™ F77/32 FORTRAN compiler (Version 9.5). This compiler creates an executable file that uses the high memory found in 386 and 486-based DOS computers.

The different versions of COMPAH are shown in Table 9. There really is no limit to the size of data matrix that can be analyzed with COMPAH. COMPAH stores all intermediate and final results on one long vector, called WORK. To increase the size of problem that can be analyzed, the WORK array has to be dimensioned to a larger size. With the WORK dimensioned at 6 million, my Pentium 90 computer with 32 MB RAM can cluster a 2000-sample by 230-species matrix with CNESS and UPGMA clustering in less than 10 minutes. My home 486DX with 16 MB RAM can run either COMP963G.EXE and COMP96WN.EXE to cluster a 2000-sample by 10-species data set in about 5 minutes.

Table 9 Different versions of COMPAH and their system requirements. All analyses were run using a 20-species data set. Labels take up little memory compared to the large lower-triangular similarity or dissimilarity matrix. WATCOM FORTRAN's VMM (see p. 56) allows a hard-disk swap file to be used with the DOS COMPAH versions to supplement a 386 or 486 computer's RAM. The data maxima were estimated with a 20 spp matrix. The first 4 versions run in DOS (which can be called from Windows) using DOS4GW.EXE. The last 2 versions require Windows.

Program Name	WORK Vector	System Requirements	Data Maxima
COMPAH96.EXE	100000	Any DOS computer. 640 KB RAM, Hard drive	≈400 samples
COMP961G.EXE	1000000	Any DOS or Windows computer (8-12 MB RAM)	>500 samples
COMP963G.EXE	3000000	About 12 MB RAM (or use VMM, see p. 56)	>1000 samples (2000 samples if NSpecies=5)
COMP966G.EXE	6000000	About 24 MB RAM (or use VMM, see p. 56)	>2000 samples
COMP96WN.EXE	2200000	Windows, 12 MB RAM	≈1000 samples
COMP96WB.EXE	44000000	Windows, 64MB RAM	≈10,000 samples

The DOS versions of COMPAH96 require the Rational Systems royalty-free program DOS4GW.EXE, which manages high-level memory. I copied DOS4GW.EXE onto the same zipped file containing COMPAH96.EXE. Note that you should use the DOS4GW.EXE program compiled on 1/11/94 (Size: 254556). COMPAH96 does not run with earlier versions of the DOS4GW.EXE program (note MATLAB™ and other programs distribute DOS4GW.EXE, so check your DOS Path so that the proper version is called).

To run these programs type: COMPAH96, COMP961G, or whatever at the DOS prompt or double click on this the COMP96XX.exe file in the Windows Program manager.. The program DOS4GW.EXE must be on the calling subdirectory or on the DOS PATH (set in the DOS Autoexec.bat file) for the DOS versions.

If you try to run a COMPAH96 executable file that has been compiled with a WORK array too large for your RAM, you will receive a DOS4GW error:

> **DOS/4GW fatal error: Loader failed -- LINEXE_LOADER**, or
> **DOS/4GW fatal error (1307) not enough memory**

These errors are given if the WATCOM-compiled COMPAH program is too large for the RAM on your computer. If you are running COMPAH out of Windows (as a DOS executable file), you will immediately be bounced back into Windows. You have four options if you get this error:

- Use a smaller version, *e.g.*, use COMP961G.EXE instead of COMP963G.EXE
- Find a computer with more RAM
- Use your hard disk to supplement your RAM using the WATCOM/Rational Systems VMM program (instructions below),
- Use the Windows version of COMPAH COMP96WN.EXE which uses Windows memory management to create a hard disk swap file.

Virtual Memory Management using DOS4GVM

DOS4GW can be set to use a Virtual Memory Manager (**VMM**) program to create a SWAP file on your hard disk to augment your RAM. This very large SWAP file, called DOS46VM.SWP, will remain on the root directory of your hard disk unless the option deleteswap is used. I don't recommend using VMM unless necessary. Using time-intensive disk read and write statements slows down COMPAH incredibly (at least 10-fold on my PC).

Here is the statement needed to set the WATCOM VMM program. Invoke this statement before calling COMPAH96: **set DOS4GVM=1** This sets the DOS4GW Virtual Memory Manager to its default settings. These defaults require 4MB of RAM and a hard disk with 16 MB of memory for storing the DOS4GVM.swp file. The default also leaves this 16 MB swap file on the root directory. If you want this space for other programs, delete this swap file after running COMPAH.

Here are some of the options that you can use to create a virtual memory configuration that matches your RAM and hard-disk space. This information is taken from the WATCOM F77/32 Users guide (pp. 325-326). The format of the call to VMM is: **set DOS4GVM=option[#value]] [option[#value]]**

A # is used since the DOS command shell will not accept '='. Here are the VMM options:

MINMEM	The minimum amount of RAM managed by VMM. *The default is 512K bytes*
MAXMEM	The maximum amount of RAM managed by VMM. *The default is 4MB*
SWAPMIN	The minimum or initial size of the swap file. If this option is not used, the size of the swap file is based on VIRTUALSIZE (see below)
SWAPINC	The size by which the swap file grows.
SWAPNAME	The swap file name. *The default name is "DOS4GVM.SWP". Be default the file is in the root directory of the current drive.* Specify the complete path name if you want to keep the swap file somewhere else.
DELETESWAP	Whether the swap file is deleted when your program exits. *By default the swap file is not deleted.* Program startup is quicker if the file is not deleted.
VIRTUALSIZE	The size of the virtual memory space (Swap file plus allocated memory). *The default is 16MB*

You can change the default virtual memory options in two ways:

1. Specify different parameter values as arguments to the DOS4GVM environment variable: **set DOS4GVM=deleteswap maxmem#8192** A # is used since the DOS command shell will not accept '='.
2. Create a configuration file with the filetype extension ".VMC", and call that as an argument to the DOS4GVM environment variable: **set DOS4GVM=@NEW4G.VMC**

The ".VMC" file contains VMM parameters and settings. Comments are permitted. Comments on lines by themselves are preceded by an exclamation point(!). Comments that follow option settings are

preceded by white space. Do not insert blank lines or spaces between symbols; processing starts at the first blank line. A blank space in a statement number is interpreted as the start of a comment. The following file will allow a 16 MB RAM PC to run COMP966G.EXE. If the file is called NEW4G.VMC, type **set dos4gvm=@new4g.vmc** before calling COMP966G

```
!Sample .VMC file
!This file shows the parameter values that I use
!for my home computer with 16 MB RAM to run COMP966G
!This file will run comp966g.exe.  Note that there can
!be no spaces between symbols in statement numbers.
minmem=512                the minimum memory
maxmem=12000              maximum RAM to be used
virtualsize=24592         Total memory (RAM plus Hard disk)
!To store the swap file in a directory called SWAPFILE, add
!swapname=c:\swapfile\dos4gvm.swp
!ensure that you have sufficient space on your hard drive
swapname=c:\temp\dos4gvm.swp
!To delete the swap file automatically when the program
!exits, add deleteswap.  Note that there will be a few
!second pause at the end of a program run as the swapfile is
!deleted
deleteswap
```

After a COMPAH analysis with virtual memory management with the deleteswap option, there will be a few second pause while the swap file is deleted. If deleteswap is not specified, then you may wish to delete the very large swap file from your root directory or the directory set by swapname.

I don't know how to shut off VMM after using "SET DOSVGM=1" Additional cluster analyses that don't require VMM will be slowed down incredibly if VMM is in place (see next section). I reboot my computer to get rid of VMM.

I can invoke VMM from within Windows, by writing a small DOS batch file to set the DOS4GVM parameters. However, after a successful COMPAH run, I often get a general protection fault that exits me from Windows 3.1. If you must use VMM to run very large datasets (2000 samples) on a computer with less than 12-16 MB RAM, then exit to DOS first.

COMPAH96 and Windows

The DOS-based COMPAH96 uses the Rational System DOS extender DOS4GW.EXE. COMPAH96 can be called as a DOS application from within Windows 3.1 or Windows 95 with no problems. I don't own Windows98, so I haven't tried it on that OS. There shouldn't be a problem running any of the versions on Windows 98. I cluster using the Windows version of COMPAH96 (comp96wn.exe). If you use the DOS versions of COMPAH96 (COMP961G...COMP966G.exe), you might want to ensure that the DOS4GW.EXE and **COMM.DAT** files are on your DOS PATH (in your autoexec.bat file) statement if Windows calls the program from a directory not containing them.

WATCOM FORTRAN allows the compilation of a FORTRAN programs to run under the Windows memory supervisor. I have compiled COMP96WN.EXE and COMP96WB.EXE, which run under a windowed environment using the Windows high-memory management system. Windows 3.1 must be running in enhanced mode. It will run in Windows 95 with no changes. A WATCOM I/O Window is created for COMPAH's console input and output. Due to the proportional spaced font used for most Windows text and because not as many characters can be displayed in the Input window, the console prompts aren't as clear using COMP96WN. I've included a modified **COMM.DAT** file called

COMMW.DAT which makes the console input prompts a little clearer. The Windows version of COMPAH will expect the COMMW.DAT file as the default data parameter input file. COMP96WN also runs more slowly than the DOS version (about ½ the speed on large data sets). There are two advantages of COMP96WN (and COMP96WB.EXE) over the DOS version (COMPAH96). COMP96WN.EXE does not use the Rational Systems DOS extender DOS4GW.EXE. Also, larger WORK arrays can be used with the Windows version of COMPAH. For example, if I try to run COMPAH96 with WORK dimensioned to 2.2 million on a 12 MB PC, I get an error. COMP96WN will run this program using a Windows hard disk swap file. A similar option is available with the DOS version by using **set DOS4GVM=1** (see above).

INDEX

Adobe
Acrobat 43, 44
AMOVA 41, 52
Biplot 22, 47, 48
ClusM
Sorting strategies 5, 22, 24, 38
Cluster reversals 42
Combinatorial equation 24, 25
COMM.DAT
command file . . . 3, 6-9, 22, 34, 36-38, 45-47, 49, 53, 54, 57
COMMW.DAT 58
COMP961G.EXE 55, 56
COMP963G.EXE 54-56
COMP966G.EXE 55, 57
COMP96WB.EXE 35, 55, 57, 58
COMP96WN.EXE 54-58
COMPAH96.EXE 3, 53, 55
Condensed form 15, 28, 30-32
Couplets 30, 32
Covariance biplot 22, 48
Dendrogram 31
Diversity
E(Sn) 8, 46, 52
DOS path command 54, 55, 57
DOS/4GW fatal error 53-55
DOS4GVM
Virtual Memory Manager 56-58
DOS4GW 3, 53-58
ESS
Expected Species Shared 8, 19, 20
Euclidean distance . . . 5, 9, 10, 13, 14, 22, 25-27, 41, 42, 47, 50, 52, 53
Factor analysis 40, 46
Format statement 28-30, 36, 39, 53, 54
fractional data 6, 7, 10, 18, 21-23, 53
Full form 28, 30, 32
H matrix 8, 13, 21
Indices
Bray-Curtis 6, 9-12, 14, 18, 19, 27
Canberra metric 9, 15, 16
Chord distance . . . 9, 10, 13-15, 18, 19, 21-23, 52, 53
city-block 10, 12, 13
CNESS . . 4, 7, 9-11, 13-15, 17-24, 35, 38, 41, 46-48, 50-54
cos(θ) 22, 48
Czekanowski's 9, 11, 12
geodesic metric 5, 9, 10, 14, 26, 27, 52, 53
Kulczynski's 12
Manhattan metric 10, 14
Minkowski 13, 14
Morisita 9, 10, 14, 16-19, 22, 49, 53
NESS 4, 8-10, 14, 17-23, 46, 50, 52, 53
NNESS . . 4, 7, 9-11, 14, 17-20, 22, 23, 38, 40, 46, 50, 52, 53
Percentage dissimilarity 9, 10, 27, 52
Percentage similarity 10, 11, 13
Ruzicka 10, 13, 52
Sorensen 14, 19
Steinhaus 5, 9-13, 52
kerning 43, 44
Labels
species 28, 30, 31, 48
station 30, 31, 37, 38, 48, 53
List-directed input 28, 29, 31-35, 47, 51, 54
MATLAB
*.mat file 34, 35, 38-40, 47, 48, 52
Matlab m.files
findcnm.m 23, 46, 47
pcoa.m 40
NMDS 11, 52
non-metric multidimensional scaling 11, 21, 23, 46
Ordiflex 4, 6, 28
ordination 46, 47, 49
Partial PCA 43
PCoA 11, 40
Postscript printers 43, 44
Presence/absence form 28-32, 34
Presentations 8.0 45
Principal coordinates analysis 11, 39, 40, 47
Printer drivers
HP Deskjet 550c 43
HP Laserjet 4M 43
Procrustes analysis 46
Quattro Pro 8 . . . . . 35, 48
Q-mode 22, 37, 45, 47
Rotations

oblique 7, 13, 14, 22, 52
Roulette 18
R-mode 22, 37, 47, 48, 53
Semimetric 10, 11, 14
Singleton species 17, 18, 24, 53
Sorting strategies 24, 26, 27, 49
centroid 5, 24-27, 50, 53
complete linkage 5, 24, 25
farthest-neighbor 5, 24
flexible average 22, 24-27, 38
incompatible 25-27, 53
Incremental Sum Squares 22, 27, 38
median 24-26
nearest-neighbor 5, 24, 27
single linkage 24-26
UPGMA 22, 24-26, 38, 47, 51, 54
WPGMA 22, 24-26, 38
Standard order 28-31, 34-36, 39, 47
Standardization 6, 7, 15, 23, 50, 52
hypergeometric 7, 15
Transformation 6, 40, 41
Transposed order 28-30, 32, 39
Triangular inequality 11
Typesetting
Adobe Type Manager 43, 44
Courier New font 42-44
Courier T1 font 43
IBMPCDOS font 42
Lotus LineDraw font 42-44
True type font 43
VMM
Virtual Memory Manager 50, 54-57
WATCOM F77/32 FORTRAN 4, 30, 38, 49, 50, 52-57
Windows 3.1 57
Windows 95 57
WordPerfect 43-45, 47, 49
5.1 47
8.0 . . . . . 45, 49
styles 49
WORK vector 55
WYSIWYG 44