1973

Data Processing Methodolgies

U.S. International Biological Program

IBP

DESERT BIOME

US/IBP ANALYSIS OF ECOSYSTEMS

MODELS

RM 73-51

Data Processing

Data Processing Methodologies

US/IBP Analysis of Ecosystems

Desert Biome

Data Processing Methodologies

September 1973

TABLE OF CONTENTS

FIGURES

TABLES

Preface

The intention of this report is to meaningfully relate computer data processing to the research of the ecologist working within the Desert Biome. To the "computer shy" ecologist it points out the advantages of automated data processing in his research. To the computer-using ecologist it suggests alternative modes of analysis and increased efficiency.

The computer analysis programs used by the Desert Biome Data Processing Group (DBDPG) represent a substantial resource both in program logic and personnel skilled in interpretation. As a package, they provide maximum information by means of analysis from many directions. These programs were acquired or written by the DBDPG as the need or anticipated need arose.

Programmers Alma Olsen and Kim Marshall deserve credit for programs written by the DBDPG. Brien Norton, Assistant to the Directorate, and programmer Verne King guided the formulation of the Data Set Abstract Search System. James MacMahon suggested that the species diversity program (SPECDIV) be written and provided the formulation of diversity indices.

Requests for analyses within the Desert Biome should be directed to:

Charles Romesburg,
Coordinator Data Processing
Desert Biome
Utah State University
Logan, Utah 84322
(752-4100, Ext.: 7619)

For the interested reader outside the Desert Biome, requests will be honored for additional information pertaining to these programs as well as for acquisition of particular programs.

Charles Romesburg,
Coordinator Data Processing

September, 1973

1.0 Introduction

This report maintains that better ecological research results when the investigator records data in computer readable form (punched cards or magnetic tape). Research projects usually generate large amounts of data and the only reasonable mode of analyses is via the computer. This process leads to greater efficiency and increases the possibility of insight since data sets once established can be subject to alternative forms of analysis using existing programs.

For example, a project measuring plant cover generates a "raw" data set containing the cover dimensions by species, plot, and date. A computer program can then calculate percent cover from the individual cover dimensions to create a "reduced" data set. Both the raw and reduced data sets, stored on magnetic tape, can be analyzed from a number of different directions using computer analysis programs.

Section 2.0 describes procedures in establishing computer readable data sets. Kinds of data analyses (programs) a researcher may request are covered in Section 3.0. The identification of data sets stored in the Desert Biome data bank relevant to an investigator's needs may be made through the Data Set Abstract Search System, described in Section 4.0.

The DBDPG is to be regarded as a resource useful for consulting, execution of analyses, interpretation of computer printouts, and locating specific data stored in the data bank. Since DBDPG personnel are trained primarily in computer programming and analytical methods, the process of defining data sets, deciding on appropriate analyses, validating hypotheses, and, in general, finding insights, are the responsibilities of the the ecologist.

2.0 Establishment of Computer Readable Data Sets

Getting raw data into computer readable form usually seems to the uninitiated more effort than the expected rewards are worth and, indeed, the process is more structured and hence more time consuming than entering data into field notebooks. After going through the process one or more times, however, it becomes much less of a chore. The steps needed to establish computer readable data sets are described in detail in the Desert Biome Report: Data Bank and Data Processing. The main steps, necessary to understand the relation to data set analysis, follow.

Data is entered into the data bank by recording data on a Data Set Coding Form, an example of which appears in Figure 1. Each row represents a punched card while the column headings identify where different information is to be placed within the punched card. These forms are completed by the DBDPG from information supplied by the investigator on a Data Set Description Form (not shown).

From the Data Set Coding Form with data transcribed, cards are punched. These cards are then entered as a data set into the data bank with the card images stored on magnetic tape. The data set created is referenced by a unique seven character Data Set Code (DSCode), and all requests for data set analyses necessarily include references to the DSCode.

A computer printout of the stored magnetic tape card images processed for readability is called a Data Set Listing. The Data Set Coding Form appearing in Figure 1, with data entered, generates the example Data Set Listing shown in Figure 2 (only the first page of the listing is shown.

Additional information describing experimental procedures used to generate a DSCode is provided by a Data Set Abstract. In the DBDPG all abstracts are stored on magnetic tape apart from the DSCodes they describe. Figure 3 shows the Data Set Abstract corresponding to the data set listing appearing in Figure 2.

Figure 1

Example of a Data Set Coding Form (Reduced)

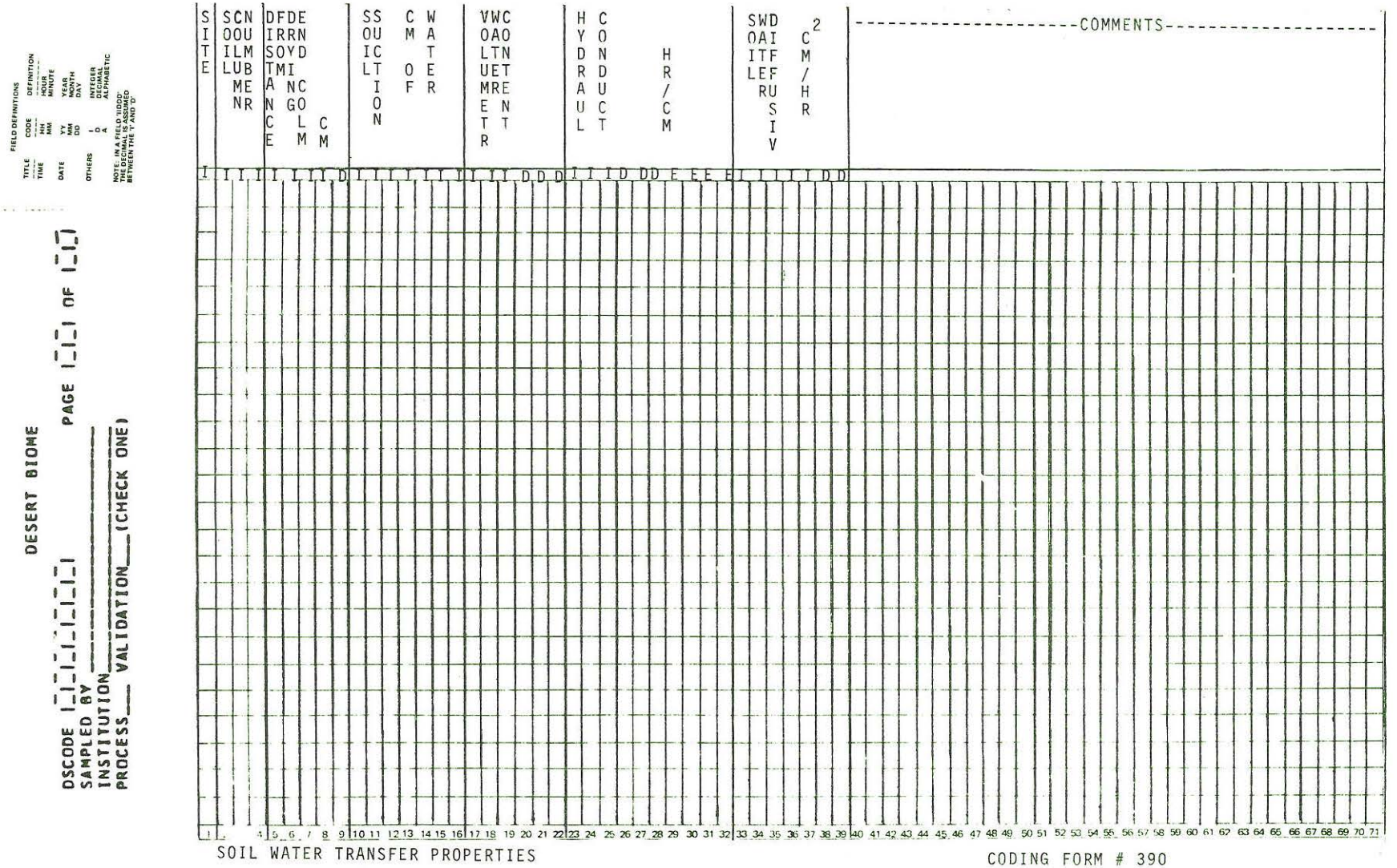


Figure 2

Example of a Data Set Listing

```
A30ST01                                        -                 73-08-30
S SCN DFDE   SS    C   VWC      HC          SWD     <-----------COMMENTS----------->
I OUO IRRN   OU    M   DAO      YD          OAI
T ILM SOYD   IC        LTN      DN          ITF   2
F LOR IMI    LT    OW  UET      RD        C LEF   C
  ME  A NG    I    FA  MRE      AU        M  RU   M
  NR  + GO    O     I  E N      OC        /   S   /
      C  LC   N     E  T T      LT        H   I   H
      F  MM         R  R                  R   V   R
------------------------------------------------------------------------------------     SEQUENCE
I III IIII.) IIIIIII IIIDDD IIIDDDEEEE IIIIIDD
------------------------------------------------------------------------------------
1  2   2.5      79.  0.205  1.499E-02  25.21  SONOITA SANDY LOAM LESS THAN  *        ST01    5
1  2   2.5      95.  0.198  8.021E-03  18.70  2MM FRACTION                           ST01   10
1  2   2.5     107.  0.194  4.227E-03  12.77                                         ST01   15
1  2   2.5     119.  0.191  2.387E-03   9.15                                         ST01   20
1  2   2.5     130.  0.188  1.611E-03   7.18                                         ST01   25
1  2   2.5     138.  0.186  1.228E-03   6.08                                         ST01   30
1  2   2.5     148.  0.184  9.446E-04   5.17                                         ST01   35
1  2   2.5     157.  0.183  5.844E-04   3.82                                         ST01   40
1  2   2.5     271.  0.167  1.930E-04   1.78                                         ST01   45
1  2   2.5     376.  0.159  8.137E-05   1.11                                         ST01   50
1  2   2.5     431.  0.156  4.878E-05   0.77                                         ST01   55
1  2   2.5     515.  0.151  3.002E-05   0.65                                         ST01   60
1  2   2.5     750.  0.142  1.756E-05   0.51                                         ST01   65
1  2   2.5    1150.  0.133  9.910E-06   0.56                                         ST01   70
1  2   2.5    1300.  0.131  8.212E-06   0.61                                         ST01   75
1  2   2.5    1400.  0.129  6.836E-06   0.59                                         ST01   80
1  2   2.5    1500.  0.128  4.972E-06   0.61                                         ST01   85
1  2   2.5    2000.  0.122  2.697E-06   0.46                                         ST01   90
1  2   2.5    3800.  0.110  6.492E-07   0.17                                         ST01   95
1  2   2.5   10500.  0.094  2.035E-07   0.09                                         ST01  100
1  2   2.5   18560.  0.086  1.337E-07   0.09                                         ST01  105
1  2   2.5   33600.  0.078  8.497E-08   0.21                                         ST01  110
1  2   2.5   50000.  0.073  6.464E-08   0.08                                         ST01  115
1  2   5.0      58.  0.218  3.090E-02  38.08                                         ST01  120
1  2   5.0      68.  0.211  1.756E-02  27.06                                         ST01  125
1  2   5.0      75.  0.207  9.954E-03  17.96                                         ST01  130
1  2   5.0      82.  0.204  5.994E-03  12.50                                         ST01  135
1  2   5.0      88.  0.201  4.142E-03   9.90                                         ST01  140
1  2   5.0      93.  0.199  3.127E-03   8.64                                         ST01  145
1  2   5.0      99.  0.196  2.424E-03   7.66                                         ST01  150
1  2   5.0     104.  0.195  1.516E-03   6.41                                         ST01  155
1  2   5.0     182.  0.178  4.971E-04   2.87                                         ST01  160
1  2   5.0     276.  0.167  1.943E-04   1.73                                         ST01  165
1  2   5.0     333.  0.162  1.135E-04   1.27                                         ST01  170
1  2   5.0     396.  0.158  7.280E-05   1.10                                         ST01  175
1  2   5.0     503.  0.152  5.010E-05   1.01                                         ST01  180
1  2   5.0     555.  0.149  3.986E-05   0.91                                         ST01  185
1  2   5.0     636.  0.146  3.212E-05   0.88                                         ST01  190
1  2   5.0     676.  0.145  2.627E-05   0.82                                         ST01  195
1  2   5.0     738.  0.143  1.778E-05   0.62                                         ST01  200
------------------------------------------------------------------------------------
0 000 00000 1111111 111222 2222222333 3333333 444444444455555555556666666666677
1 234 56789 0123456 789012 3456789012 3456789 01234567890123456789012345678901
------------------------------------------------------------------------------------
SOIL WATER TRANSFER PROPERTIES                          CODING FORM #390
```

Figure 3

Example of a Data Set Abstract

```
DISCIPLINE AND CODING FORM NUMBERS
     A3 (STU)                A390

PROJECT TITLE
     EVALUATION OF CRITICAL SOIL PROPERTIES NEEDED TO PREDICT SOIL-WATER
     FLOW UNDER DESERT CONDITIONS

ABSTRACT TITLE
     SOIL-WATER TRANSFER PROPERTIES

INVESTIGATORS
     DR. L. F. STOLZY (PRINCIPAL INVESTIGATOR, DEPT. OF SOIL SCIENCE,
     UNIVERSITY OF CALIFORNIA, RIVERSIDE, CALIFORNIA 92502 TELEPHONE
     714-787-5112)
     DR. J. LETEY (PRINCIPAL INVESTIGATOR - 714-787-5116)
     G. F. MEHUYS (RESEARCH ASSISTANT - 714-787-5113)

GEOGRAPHIC INFORMATION
     ROCK VALLEY
     SANTA RITA

PARAMETERS
     SOIL SUCTION (CM OF WATER), VOLUMETRIC WATER CONTENT (CM**3/CM**3),
     HYDRAULIC CONDUCTIVITY (HR/CM), SOIL-WATER DIFFUSIVITY (HR/CM**2)

TIME OF SAMPLING
     01 MAR 72                    (SOIL SAMPLES TAKEN IN MARCH, MAY, AND NOVEMBER 1972)

EXPERIMENTAL
     METHODS
     TRANSIENT OUTFLOW METHOD FOR EVALUATION OF HYDRAULIC CONDUCTIVITY AND
     SOIL-WATER DIFFUSIVITY (WEEKS, ET AL., 1967). SOIL SUCTION MEASURED
     WITH TENSIOMETERS AND PSYCHROMETERS.
     EXPERIMENTAL DESIGN
     BOTH UNDISTURBED AND LOOSE SAMPLES WERE TAKEN AT THE SANTA RITA,
     TUCSON SITE. UNDISTURBED CORES WERE 10 CM IN DIAMETER BY 30 CM LONG.
     LOOSE SAMPLES WERE TAKEN TO THE 1.20 M DEPTH AT EVERY 30 CM INTERVAL.
     BECAUSE OF THE EXTREME STONINESS OF THE ROCK VALLEY, NEVADA SOIL,
     ONLY LOOSE SAMPLES WERE COLLECTED, TO A DEPTH OF 50 CM. LOCATIONS FOR
     SAMPLING WERE CHOSEN AT RANDOM WITHIN EACH VALIDATION SITE.
     SOIL COLUMNS 10 CM I.D. BY 30 CM LONG, EITHER RECONSTITUTED IN THE LAB
     FROM LOOSE SAMPLES OR TAKEN IN SITU, WERE USED THROUGHOUT.
     CITATIONS
     WEEKS, L. V., AND S. J. RICHARDS. 1967. SOIL WATER PROPERTIES
     COMPUTED FROM TRANSIENT FLOW DATA. SOIL SCI. SOC. AMER. PROC.
     31-721-725.

SUPPORTING DOCUMENTS
     PROPOSAL 1971: 5.6.0.-1
     RESEARCH MEMORANDUM 73-43
```

3.0 Data Set Processing

In general, data set processing requires that either a specific program be written or the use of a pre-written "canned" program. This report emphasizes the latter for the reason that there is much less to be said about specific programs. In fact, specific programs constitute a substantial portion of the programming within the DBDPG.

The questions of what data to collect and how the data should be analyzed are interrelated, a choice for one constraining the possible choices for the other. The intention in presenting computer analysis techniques is not to suggest that data should be collected in such a way that these techniques can be used but to suggest new directions of approach of which the investigator might be unaware.

Each technique is minimally described by what it will accomplish. Where it is felt understanding will benefit, an example is given accompanied by an illustration of the computer output. Each program is given a name followed by a reference to its origin.

Section 3.1 covers the process of submitting requests for analysis. Section 3.2 presents data display techniques where statistical inference to a larger parent population is infeasible. Programs for commonly occurring problems in making statistical inferences are given in Section 3.3. Other programs available for standard statistical analysis, numerical analysis, and optimization problems are listed in Section 3.4.

3.1 Data Set Processing Request

Because geographic location often prohibits direct contact between investigator and the DBDPG it is important that written instructions for analysis be complete. There are several general rules which lead to good communication. The request should include text explaining what should be calculated and the DSCodes from which the raw data are obtained. Within each DSCode the data fields applying to the calculation should be specified by column numbers. The request should procede in clear, step-wise fashion and if possible, the calculations should include a numerical example. Figure 4 is regarded as a clearly presented request for a specific program analysis.

Figure 4

Example of Written Request for Program Analysis

TO: Charles Romesburg
Desert Biome
Utah State University
Logan, Utah 84322

FROM: Russell P. Balda
Department of Biology
Northern Arizona University
Flagstaff, Arizona

Please find enclosed a total of 14 coding forms, #17. The first 9 are for trapping period 9, the last 5 for period 10. Data will always be analyzed by trapping period so it is important to keep the periods separate.

Before we spend any more time preparing our raw data and transposing it to the coding forms, we would like to analyze this data to make certain we are doing everything correctly. Please note that we used two spaces from the TAXON CODE fields (24 & 25, 36 & 37, 48 & 49, 60 & 61) for a new entry, i.e. frequency. Also note that the plant data for some mammal capture-sites will not fit on one card but will need to be placed on two cards. Where a second card is necessary, the first 18 fields are left blank, meaning the information in the remaining fields goes with the capture-site immediately above. This is necessary because the number of plant species per capture-site is highly variable.

I hope your staff will be able to put these data on cards, write the appropriate program, and then run the analysis for us. The biggest problem will be my written directions, but I will try to be as clear as possible. We would like the data to be analyzed in a number of different ways.

Part A: For each capture-site, we would like a print-out showing relative density, relative frequency, relative dominance, the sum of the above three, percent dominance for the herbaceous plants, and percent dominance for the overstory. This analysis is to be done as follows for the capture-site of each mammal. On all plant species with a taxon code number from 0001 to 0299 do the following:

1. Relative density = $\frac{\text{Total density of each species}}{\text{Total density of all species}} \times 100$

For example, on sheet one the first capture-site has five plant species with densities of 7,1,1,3,1. The total density is 13. Thus, the relative density of each plant species is as follows:

Plant Species 0046 - 7/13 x 100 = 53.84
" " 0021 - 1/13 x 100 = 7.69
0013 - 1/13 x 100 = 7.69
0047 - 3/13 x 100 = 23.07
0041 - 1/13 x 100 = 7.69

2. Relative dominance = $\frac{\text{Total dominance of a species}}{\text{Total dominance of all species}} \times 100$

All measurements in the dominance categories for plants with taxon codes between 0001 and 0299 are in hundredths of feet so the decimal is always 0.00. For example, again using the first capture-site on sheet one, the five plant species had dominances of 3.09 ft., 0.53 ft., 0.22 ft., 0.98 ft., 0.12 ft. This totals to 4.94 ft.

Plant Species 0046 - $\frac{3.09}{4.94} \times 100 = 62.55$

" " 0021 - $\frac{0.53}{4.94} \times 100 = 10.72$

" " 0013 - $\frac{0.22}{4.94} \times 100 = 4.45$

" " 0047 - $\frac{0.98}{4.94} \times 100 = 19.83$

" " 0041 - $\frac{0.12}{4.94} \times 100 = 2.42$

3. Relative Frequency = $\frac{\text{Frequency of a species}}{\text{Frequency of all species}}$ X 100

Using the data from the first capture-site on sheet one, the relative frequency is calculated as follows. The frequency values for the five plant species total to 12, therefore:

Plant Species
0046 - 7/12 x 100 = 58.33
0021 - 1/12 x 100 = 8.33
0013 - 1/12 x 100 = 8.33
0047 - 2/12 x 100 = 16.66
0041 - 1/12 x 100 = 8.33

4. Importance Value is the sum of 1., 2., and 3., above; for example: For the five plant species at the capture-site used above.

0046 - 53.84 + 62.55 + 58.33 = 174.72
0021 - 7.69 + 10.72 + 8.33 = 26.74
0013 - 7.69 + 4.45 + 8.33 = 20.47
0047 - 23.07 + 19.83 + 16.66 = 59.56
0041 - 7.69 + 2.42 + 8.33 = 18.44

5. Percent herbaceous dominance is calculated as follows:

% Herbaceous Dominance = $\frac{\text{Total dominance for all species}}{11.00}$ X 100

The total dominance figure is the same one used in A2. above. In the example it is 4.94 ft. The 11 is a constant; it is the maximum number of feet that could be covered with vegetation. Thus the percent herbaceous dominance for our example is:

$\frac{4.94}{11.00}$ x 100 = 44.90%

6. Percent overstory dominance is to be calculated only for plants with taxon codes from 0300 to 0599. This is the only calculation in which these plants are used. Also, the dominance field for these plants is the only one with entries. For an example, check the second capture-site as the first one has no overstory plants present.

Plant 0301 has 05.6 ft.
Plant 0302 has 00.4 ft.
Plant 0304 has 01.4 ft.

Note the placement of the decimal is 00.0

The total is: 7.4 ft.

% overstory dominance = $\frac{7.4}{21}$ x 100 = 35.23%

The 21 is the maximum number of feet that could be covered by these plants.

The printout for Part A., as calculated above, should read something like the example given below for the first capture-site on Page 1. Each capture-site should be listed in succession as on the coding forms.

09 - PERBAI - CF10 - 1

Plant Sp.	R. Den.	R. Dom.	R.F.	I.V.
0046	53.84	62.55	58.33	172.72
0021	7.69	10.72	8.33	26.74
0013	7.69	4.45	8.33	20.47
0047	23.07	19.83	16.66	59.56
0041	7.69	2.42	8.33	18.44

percent herbaceous dominance = 44.90%
percent overstory dominance = 00.00%

Part B. For each species of mammal we need the above information totaled by habitat (1 or 2; not 3 in this data set) and sampling period. For example, we would like these data for all PERBAI during sampling period 9 in habitat 1. The calculations will be similar to those done above.

1. Relative Density = $\frac{\text{Total density of each plant species at all PERBAI}}{\text{Total density of all plant species at all PERBAI}} \times 100$

2. Relative Dominance = $\frac{\text{Total dominance of each plant species at all PERBAI}}{\text{Total dominance of all plant species at all PERBAI}} \times 100$

3. Relative Frequency = $\frac{\text{Total frequency of each species at all PERBAI}}{\text{Total frequency of all species at all PERBAI}} \times 100$

4. Importance Value is the same as done in Part A.; the sum of 1., 2., and 3., above.

5. Percent herbaceous dominance is calculated differently since, rather than use 11, we must multiply 11 by the number of capture-sites before dividing by the total dominance figure as calculated in B2. For example, if there are 10 PERBAI, then percent herbaceous dominance is:

$$\frac{\text{Total dominance of all plant species at all PERBAI}}{11 \times 10} \times 100$$

6. We also need to know additional information about the absolute dominance provided by each species per species of mammal. This is an additional calculation not performed in Part A. Simply use the total density of each plant species at all PERBAI as was done for B1., and total dominance figures for each plant species as was done for B2., and simply print it out (see below).

7. Percent oversotry dominance is calculated as explained in A6, with the modifications discussed in B6. It is done only on plants with taxon codes from 0300 to 0599. However, rather than divide by 21, we must multiply 21 by the number of capture-sites, then divide by the total diminance figure of all these plants.

8. We need to know how much overstory was contributed by each different species of plant with a taxon code between 0300 and 0599. This is done by adding up the dominance figures for each species of plant separately and then dividing by 21 times the number of capture-sites. The total of these percentages for each species will sum to the figure obtained in B7.

The print out for Part B should read as follows for each species of mammal:

09 - PERBAI - 1

Plant Sp.,	R. Den. (B1)	R. Dom. (B2)	R.F. (B3)	I.V. (B4)

Percent herbaceous dominance = ____% (B5)

Plant Sp.,	Total density of each plant species (As described in B6)	Total Dominance

Percent overstory dominance = ____% (B7)

Plant Sp.,	% Overstory Dominance

The above analysis is to be done for each species of mammal by habitat type. For example, DIPMER occurs in both habitat 1., and 2. The analysis needs to be done separately for each habitat. Please note that habitat 3 was not represented in any of the data submitted this time. In the future, however, we will be submitting data with capture-sites in 3. When this occurs we will want analysis of all three habitats separately and then habitats 2 and 3 combined as if they were the same.

Part C. For each habitat studied, (in this case 1 and 2), we need exactly the same data as was calculated in Part B. Everything is to be done exactly the same as for Part B, but this time the species of mammal can be ignored. In B5, B7, and B8, where number of capture-sites by species was multiplied by either 11 or 21, this time simply substitute number of capture-sites of all mammal species per habitat. Thus, the print out will be in two sections, one for habitat 1, and one for habitat 2. If the animals had been captured in habitat 3, then the print out for Part C would need to be in four sections: One section for habitat 1, one for habitat 2, one for habitat 3, and one for habitat 2 and 3 combined. The print out for each of these sections should follow the format shown for Part B except for the headings which should show sampling period and habitat. The four headings for the data included here would be:

09 - ALL species - 1
09 - ALL species - 2
10 - ALL species - 1
10 - ALL species - 2

This is all the analysis we can envision for these data and all others to be sent at a later date on coding form #17. However, it would be wise to put these data on magnetic tape because we will most likely want other information as we proceed with the project.

I realize the problem of trying to write a compact program for these data based on my written instructions. When problems arise please call and we can work them out verbally, as I realize my shortcomings in trying to convey my instructions by mail.

If possible, after the cards are punched we would appreciate having the coding forms back for our reference. Thank you for your help and cooperation in this matter.

3.2 Non-inferential Data Analyses

3.2.1 Data Sorts and Summaries

Often it is desirable to rearrange and/or summarize raw data. By example, data are usually recorded on the Data Set Coding Form by sequential date of collection, site of observation, taxon, replicate, observer, etc. It may be more meaningful, however, to list the data by a different hierarchy, e.g., perhaps by taxon first followed by replicate, date, site, and observer. In addition, it may be useful to count the number of occurrences of each taxon, that is, summarize the raw data. Rearrangements and summaries of large data sets are all but impossible without using the computer.

The DBDPG has created a series of programs to accomplish this under the name SORT. The procedure is illustrated for a data set which is in part shown in Figure 5. These data were summarized by counting the occurrences of each taxon by size class, with the output appearing in Figure 6.

Figure 5

Portion of Data Set Used to Illustrate SORT Analysis

A3UMLD1 73-05-24

D S C O D E	S I T E	DATE	A O R L	TAXON CODE	D U P	S IG ZR EO U P	N U MI BN R	SG IR ZO EU P	DW RE YI G H T	SEQUENCE	
AAAAAAA	A	YYMMDD	A	IAAAAAAI	II	IIII			DDDDD		
A3UMLD1	A	700607	L	6BAETRI	02	0014				MLD1	5
A3UMLD1	A	700607	L	6BAETRI	03	0005				MLD1	10
A3UMLD1	A	700607	L	6BAETRI	04	0008				MLD1	15
A3UMLD1	A	700607	L	6BAETRI	05	0006				MLD1	20
A3UMLD1	A	700607	L	5CHI	03	0005				MLD1	25
A3UMLD1	A	700607	L	5CHI	04	0003				MLD1	30
A3UMLD1	A	700607	L	6HYDOCC1	05	0002				MLD1	35
A3UMLD1	A	700607	L	6HYDOCC1	09	0004				MLD1	40
A3UMLD1	A	700607	L	6HYDOCC1	10	0013				MLD1	45
A3UMLD1	A	700607	L	6PIS	03	0001				MLD1	50
A3UMLD1	A	700607	L	6SIM	05	0001				MLD1	55
A3UMLD1	B	700607	L	6ARGVIV	04	0001				MLD1	60
A3UMLD1	B	700607	L	6ARGVIV	12	0001				MLD1	65
A3UMLD1	B	700607	L	6ARGVIV	15	0002				MLD1	70
A3UMLD1	B	700607	L	6BAETRI	02	0016				MLD1	75
A3UMLD1	B	700607	L	6BAETRI	03	0002				MLD1	80
A3UMLD1	B	700607	L	6BAETRI	04	0006				MLD1	85
A3UMLD1	B	700607	L	6BAETRI	05	0005				MLD1	90
A3UMLD1	B	700607	L	6BAETRI	06	0004				MLD1	95
A3UMLD1	B	700607	L	5CHI	02	0004				MLD1	100
A3UMLD1	B	700607	L	5CHI	03	0022				MLD1	105
A3UMLD1	B	700607	L	5CHI	04	0010				MLD1	110
A3UMLD1	B	700607	L	5CHI	05	0004				MLD1	115
A3UMLD1	B	700607	L	5CHI	06	0001				MLD1	120
A3UMLD1	B	700607	L	5EMP	03	0001				MLD1	125
A3UMLD1	B	700607	L	6HYAAZT	04	0002				MLD1	130
A3UMLD1	B	700607	L	6HYAAZT	06	0002				MLD1	135
A3UMLD1	B	700607	L	6HYD 2	03	0005				MLD1	140
A3UMLD1	B	700607	L	6HYDOCC1	05	0001				MLD1	145
A3UMLD1	B	700607	L	6HYDOCC1	06	0003				MLD1	150
A3UMLD1	B	700607	L	6HYDOCC1	08	0001				MLD1	155
A3UMLD1	B	700607	L	6HYDOCC1	09	0001				MLD1	160
A3UMLD1	B	700607	L	6HYDOCC1	10	0012				MLD1	165
A3UMLD1	B	700607	L	6PIS	03	0001				MLD1	170
A3UMLD1	B	700607	L	6SIM	04	0001				MLD1	175
A3UMLD1	B	700607	L	6SIM	05	0002				MLD1	180
A3UMLD1	C	700607	L	5CHI	03	0015				MLD1	185
A3UMLD1	C	700607	L	5CHI	04	0030				MLD1	190
A3UMLD1	C	700607	L	5CHI	05	0002				MLD1	195
A3UMLD1	C	700607	L	5CHI	06	0001				MLD1	200
0000000	0	011111	1	11112222	22	2222			33333		
1234567	8	901234	5	67890123	45	6789			01234		

BENTHOS NMBRS-WGTS CODING FORM #33

Figure 6

Example Output from SORT Analysis

TAXON	SIZE	NUMBER
ARGVIV	04	1
ARGVIV	12	1
ARGVIV	15	2
BAETRI	02	41
BAETRI	03	66
BAETRI	04	52
BAETRI	05	28
BAETRI	06	4
CER	07	7
CER	09	1
CHI	02	13
CHI	03	72
CHI	04	70
CHI	05	25
CHI	06	3
DUBGTU	04	1
DUBGTU	05	1
DUBGTU	06	1
DYT	03	1
EMP	03	1
HELELO	08	1
HYAAZT	01	3
HYAAZT	02	7
HYAAZT	03	3
HYAAZT	04	4
HYAAZT	05	2
HYAAZT	06	2
HYD	203	5
HYDOCC	105	3
HYDOCC	106	3
HYDOCC	108	1
HYDOCC	109	5
HYDOCC	110	25
LIMFRI	12	3
OPTDIV	02	2
PIS	01	3
PIS	02	20
PIS	03	5
PIS	04	1
SIM	04	1
SIM	05	3

3.2.2 Visual Data Display Analyses

Two of the most used modes of visual data display in science are histograms and "curve drawing". Two programs have been written by the DBDPG to automate this display process. Histograms are presented by program HIST which computes the frequency of occurrence for attributes of interest by class interval (Figure 7). Program GRAPH plots points for x-y variables and although automated point connection is not provided this is easily accomplished by hand (Figure 8). For both programs, scaling of axes is optionally input by the user or computed automatically.

Figure 7

Example Histogram Drawn by HIST Program

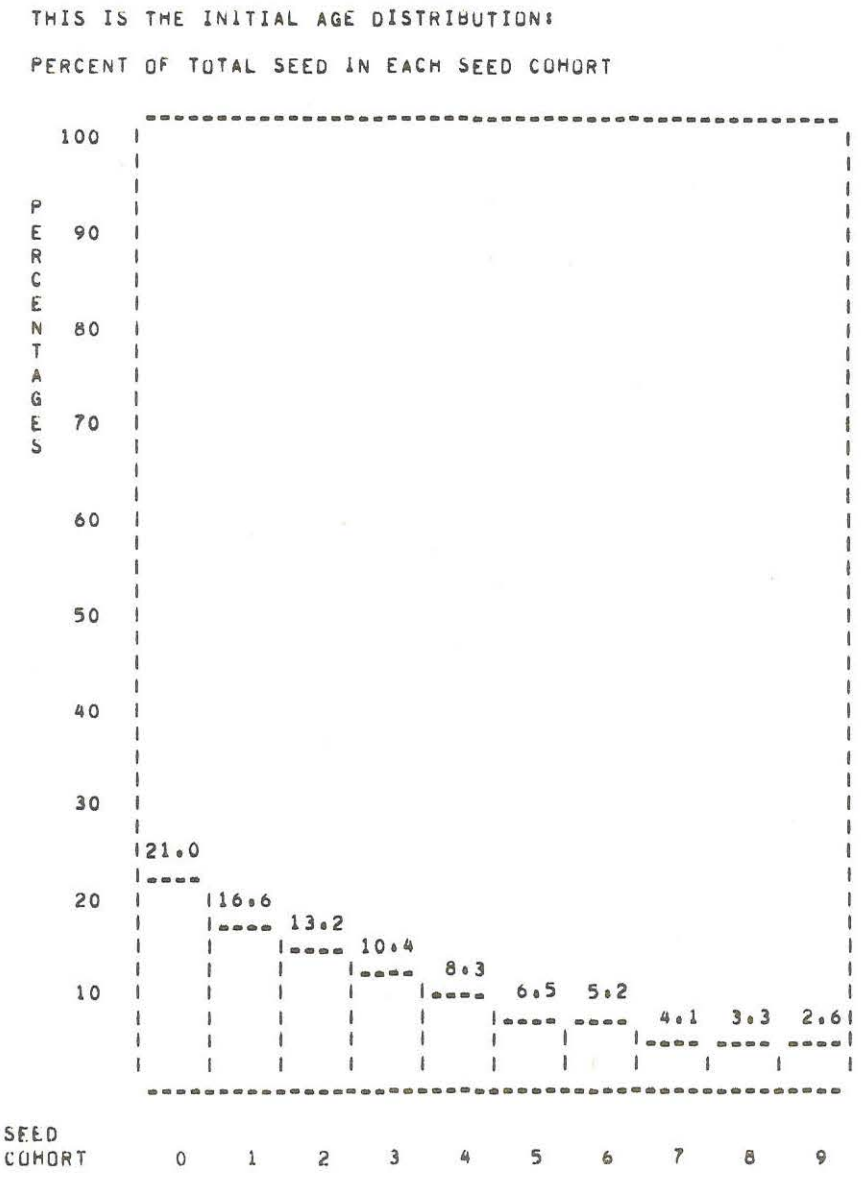
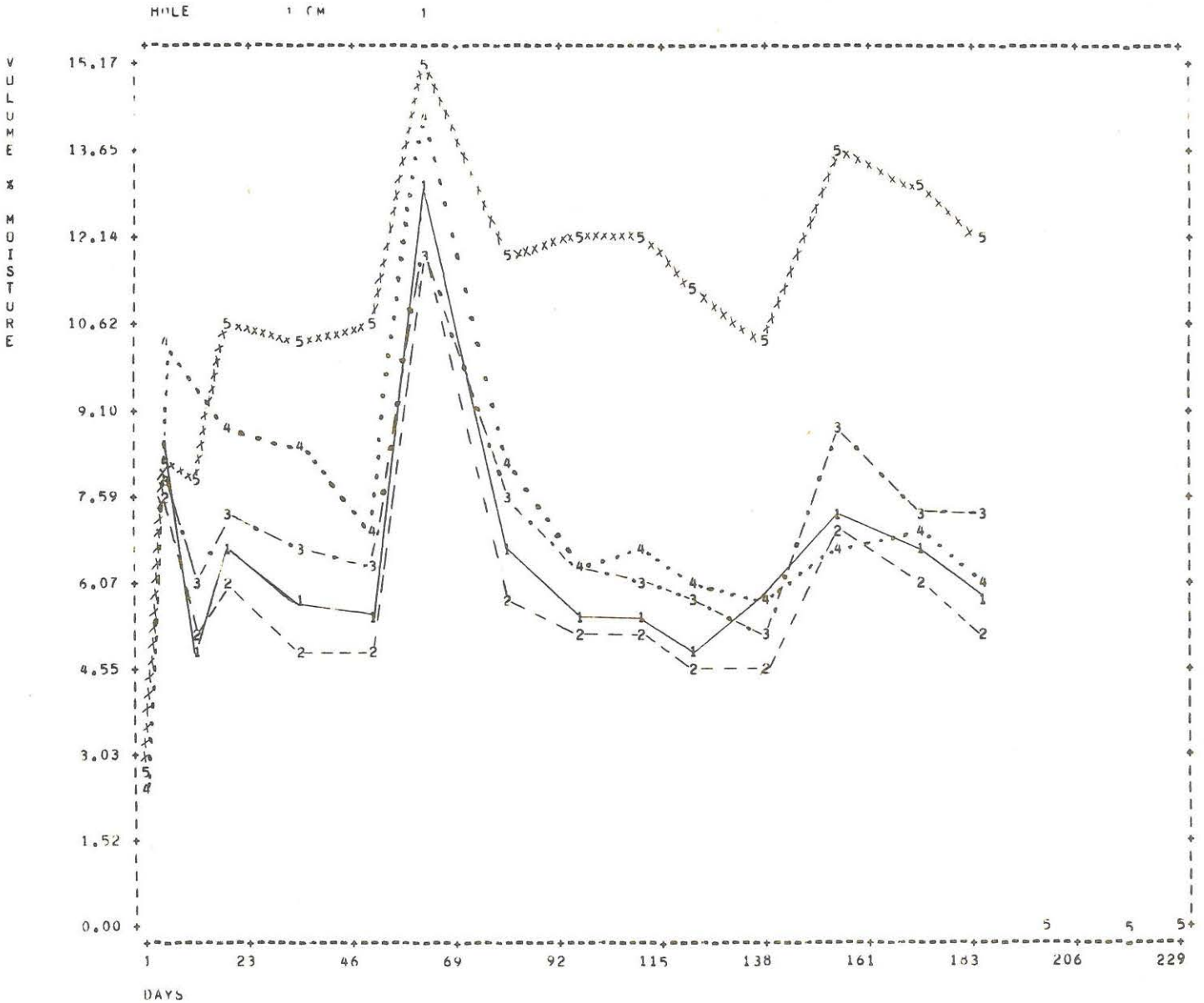


Figure 8

Example Plot of Points From GRAPH Program (Reduced)



3.2.3 Classification and Ordination Analyses

Classification and ordination techniques are useful for revealing fundamental underlying structure of multivariate data. The data basic to each consists of measurements made for a set of attributes across a set of "individuals". An individual is the entity the measurement is made on while an attribute is what is measured. The problem is to understand the relation among individuals given the attribute measurements for each. Most frequently attention is directed to finding those individuals which are either alike or dissimilar.

The data are usually arranged in matrix form with individuals being the columns and attributes the rows. A given column-row entry is a datum giving information on the attribute for the individual. For example, individuals might be taxa and attributes measurements on taxa such as biomass, size, etc., or individuals could be sites with attributes specified by measurements on taxa found there. In the first case interest might be focused on distinguishing among taxon and in the second case among sites.

Classification is concerned with the ordering of individuals into groups on the basis of their attribute relationship. Ordination techniques are used to reduce the dimensionality of the attribute space by replacing the attribute set with a new and smaller set; often the reduced space allows for relations among individuals to be identified. In general, the basic data matrix can be subject to both classification and ordination treatment, and this is the usual procedure. The results of multivariate data analysis are seldom black and white but subtle shades of gray. It is usually beneficial to treat the problem from several directions and integrate the results to arrive at a conclusion.

The multivariate programs used within the DBDPG are:

Classification	Program Name	Reference
Generalized Cluster Analysis	MINT	(Rohlf, 1971)
Minimum Dispersion Cluster Analysis	MDISP	(Goldstein and Grigal, 1972)
Minimum Information Cluster Analysis	MINFO	(Goldstein and Grigal, 1972)
Ordination		
Principal Coordinate Analysis	MINT	(Rohlf, 1971)
Nonmetric Multidimensional Scaling	MDSCAL	(Kruskal, 1964)

Only the classification programs are discussed here and illustrated by example although a complete analysis would include running the ordination programs.

Brief descriptions of the classification programs follow:

MINFØ (Goldstein and Grigal, 1972)

This method considers each individual as a separate group at the start of the algorithm. During each clustering cycle, a pair of groups is joined which results in the minimum increase in mutual information, where information is in accordance with Shannon's (1949) definition. The clustered pair then becomes a new group (cluster) and the process continues until all individuals are contained within a single group.

MDISP (Godstein and Grigal, 1972)

This method is similar to MINFO except that clustering occurs for that pair of groups which provide the minimum increase in within-group dispersion.

MINT (Rohlf, 1971)

This program has a number of user options for "similarity coefficients" and "clustering methods". Similarity coefficients define alikeness among individuals or groups while the various clutering methods represent choices for forming clusters, i.e., discriminating for alikeness. The options follow Sokal and Sneath (1963, 1973):

Similarity Coefficient (interval data)

1. correlation coefficient
2. average Euclidean distance
3. average Manhattan distance
4. variance-convariance matrix

Similarity Coefficient (qualitative data)

1. simple matching coefficient
2. Jaccard coefficient
3. Dice coefficient
4. Yule coefficient

Clustering Methods

1. weighted pair-group method using arithmetic averages
2. unweighted pair-group method using arithmetic averages
3. single linkage
4. complete linkage
5. centroid linkage
6. flexible linkage

The provision for qualitative data is useful when it is desirable or of necessity to measure attributes on a present-absence basis.

An example problem using the MINT program with 10 individuals and 47 attributes illustrates cluster analysis. The individuals in this case are streams while the attributes are the presence-absence (coded as 1 or 0, respectively) of aquatic invertebrate taxa. The data matrix is shown in Table 1. The simple matching coefficient of similarity and the single linkage method of clustering is employed. The output is shown in the form of a dendrogram, or tree diagram, in Figure 9. Horizontal lines in the tree represent nodes where diverse subgroups are joined. Heterogeneity is greatest

at the bottom of the tree where each group is separate, consisting of one individual. At the top of the tree there is a single group containing all individuals. As one procedes from the bottom to the top of the tree individuals will group to form clusters. A measure of similarity of clustering groups is the distance from the bottom of the tree to the node of the clustering group. The smaller this distance the more alike are the members clustering. Thus, in Figure 9 individuals 1 and 3 cluster first and are considered more alike than, say, the cluster of group 1,3,9 with individual 7.

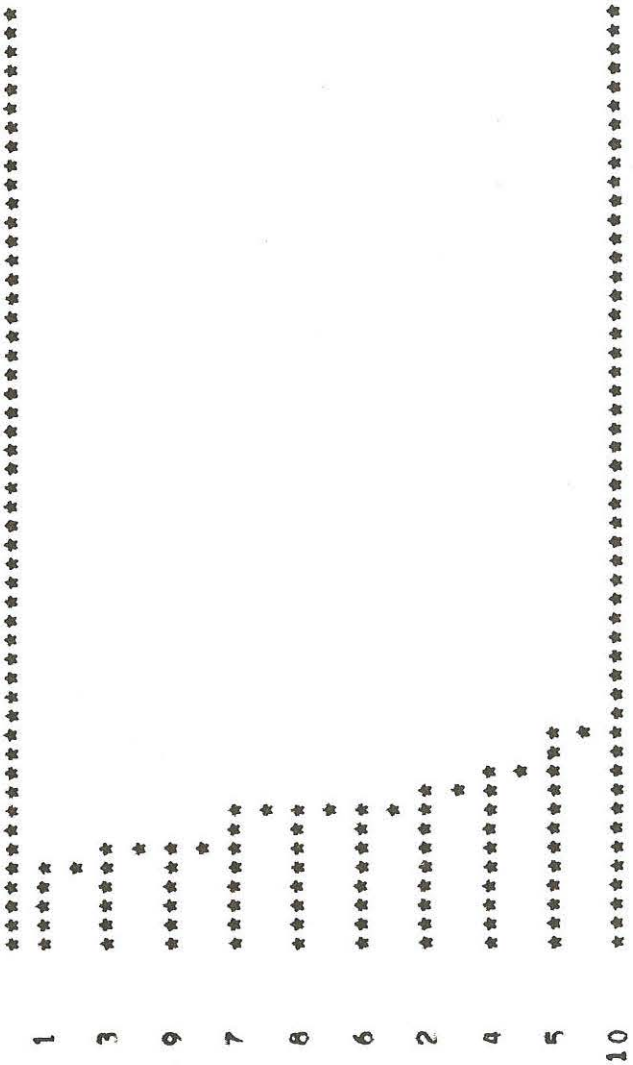
Table 1

Data Set Used to Illustrate Cluster Analysis

	1	2	3	4	5	6	7	8	9	10
1	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
2	0.000	0.000	1.000	0.000	1.000	0.000	1.000	0.000	0.000	0.000
3	0.000	0.000	0.000	1.000	0.000	1.000	0.000	0.000	0.000	1.000
4	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	1.000
5	0.000	0.000	0.000	1.000	0.000	1.000	0.000	1.000	1.000	0.000
6	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
7	0.000	1.000	0.000	1.000	0.000	1.000	0.000	0.000	0.000	1.000
8	0.000	0.000	0.000	0.000	1.000	0.000	0.000	0.000	0.000	0.000
9	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
10	0.000	0.000	0.000	0.000	1.000	0.000	1.000	0.000	0.000	1.000
11	0.000	0.000	0.000	0.000	0.000	0.000	0.000	1.000	0.000	1.000
12	1.000	1.000	0.000	0.000	1.000	0.000	0.000	0.000	0.000	1.000
13	0.000	1.000	0.000	0.000	1.000	0.000	1.000	0.000	0.000	0.000
14	0.000	1.000	0.000	0.000	1.000	0.000	1.000	0.000	0.000	0.000
15	0.000	0.000	0.000	0.000	0.000	1.000	0.000	1.000	0.000	0.000
16	0.000	0.000	0.000	0.000	0.000	1.000	0.000	0.000	0.000	0.000
17	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
18	0.000	0.000	0.000	0.000	0.000	0.000	0.000	1.000	0.000	0.000
19	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
20	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
21	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
22	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
23	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
24	0.000	0.000	0.000	1.000	0.000	0.000	0.000	0.000	0.000	1.000
25	1.000	0.000	1.000	0.000	0.000	0.000	1.000	1.000	0.000	1.000
26	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
27	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	1.000	0.000
28	0.000	0.000	0.000	0.000	0.000	0.000	0.000	1.000	0.000	0.000
29	0.000	0.000	0.000	0.000	0.000	0.000	1.000	0.000	0.000	0.000
30	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
31	0.000	0.000	0.000	1.000	0.000	1.000	0.000	0.000	0.000	1.000
32	0.000	0.000	0.000	0.000	1.000	0.000	0.000	0.000	0.000	0.000
33	1.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	1.000
34	0.000	0.000	0.000	1.000	0.000	0.000	0.000	0.000	0.000	0.000
35	0.000	0.000	0.000	0.000	1.000	0.000	0.000	0.000	0.000	0.000
36	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	1.000	0.000
37	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	1.000
38	0.000	0.000	0.000	0.000	1.000	0.000	0.000	0.000	0.000	0.000
39	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
40	1.000	0.000	0.000	1.000	0.000	0.000	0.000	0.000	0.000	0.000
41	0.000	0.000	0.000	0.000	1.000	0.000	0.000	0.000	0.000	0.000
42	1.000	0.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
43	0.000	0.000	0.000	1.000	0.000	0.000	0.000	0.000	0.000	1.000
44	0.000	0.000	0.000	0.000	0.000	0.000	0.000	1.000	0.000	1.000
45	0.000	0.000	0.000	1.000	0.000	0.000	0.000	0.000	0.000	0.000
46	0.000	1.000	0.000	1.000	0.000	0.000	1.000	0.000	0.000	0.000
47	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000

Figure 9

Dendrograms Drawn by MINT Program



3.2.4 Synagraphic Mapping

The data for synagraphic mapping consists of pairs of observations in the order of: a) quantity measured; b) location of measurement. From these data the measured values can be plotted on a map of the site from which they were obtained. Synagraphic mapping is a technique for projecting the continuous spatial distribution of the measurements based upon the sample measurements. Typical measurements as a function of location are plant cover, numbers, biomass, and microclimatological data.

These maps are produced by the SYMAP* program with the output appearing as a shaded map by contrast printing. The darkest areas denote highest measurement values while the least darkest denote the lowest. There are several user options for generating the continuous spatial surface by interpolation between values at given data points and for drawing surface contour lines.

Figure 10 illustrates the contour map option of SYMAP generated by recording at random locations a measured value, in this case, soil surface temperature. Temperature is recorded as falling into one of four discrete classes numbered from 1 to 4, e.g., 51 - 60 °F, 61 - 70 °F, etc., and the locations and values sampled are superimposed on Figure 10 according to this numbering scheme.

---

* The synagraphic mapping program SYMAP was originally written by Howard T. Fisher at Northwestern Technological University in 1963. The version used here was purchased from the Harvard University Laboratory for Computer Graphics and then made compatible with the Burroughs BL 6700 computer.

Figure 10

Conformal Map Drawn by SYMAP Program (Reduced)

SYMAP

3.2.5 Species Diversity Programs

The SPECDIV program written by the DBDPG, takes as input a list of the number of individuals obtained by sampling for each of two or more species. From this the frequently discussed measures of species diversity are calculated.

The notation used is:

n_i = Number of individuals of the i'th species
N = Number of species in collection
I = Number of individuals in collection ($I = \sum_{i=1}^{N} n_i$)

Several symbols used to represent diversity:

D (subscripted)
H
H' (H-prime)

For each diversity index D the originator's name is parenthetically enclosed. The references from which they were obtained is given at the end of this section.

Species Diversity Measures

D1 (Origin unknown)

$$D1=N$$

That is, the number of species is a measure of diversity without regard to the number of individuals or their distribution.

D2 (Willis, 1922)

$$D2 = \text{(Not implemented)}$$

D3 (Gleason, 1922)

$$D3 = N/\log_e (I)$$

D4 (Preston, 1948)

$$D4 = \text{(Not implemented)}$$

D5 (Margalef, 1957)

$$D5 = (N-1)/\log_e (I)$$

D6 (Mac Arthur, 1957)

$$D6 = -\sum_{i=1}^{N} \frac{n_i}{I} \log_e \frac{n_i}{I}$$

D7 (Menhinick, 1964)

$$D7 = \log_e N/\log_e I$$

D8 (Menhinick, 1964)

$$D8 = N/\sqrt{I}$$

D9 (Monk, 1966)

$$D9 = N/I$$

D10 (Simpson, 1949)

$$D10 = \sum_{i=1}^{N} \left(\frac{n_i}{I}\right)^2$$

D11 (McIntosh, 1967)

$$D11 = \sqrt{\sum_{i=1}^{N} n_i^2}$$

It is possible to compute lower and upper bounds on D11. Denoting these by LB and UB:

$$LB = I/\sqrt{N}$$

$$UB = \sqrt{[I - (N-1)]^2 + (N-1)}$$

Therefore, $LB \leq D11 \leq UB$

D12 (Margalef, 1957)

Margalef partitions the quantity of information into:

1) Count the number of species
2) Distribute the individuals according to species
3) Localize the species
4) Localize the species of equal frequency (hypothetical situation)
5) Localize the individuals (maximum information)

These are denoted by D(1), D(2), D(3), D(4), D(5) and are given by the following formulas:

$$D(1) = 1.443 \log_e I$$

$$D(2) = 1.443 \log_e \frac{(I-1)!}{(I-N)!\ (N-1)!}$$

$$D(3) = 1.443 \log_e \frac{I!}{n_1!\ n_2!\ \ldots\ n_N!}$$

$$D(4) = 1.443 \log_e \frac{I!}{((I/N)!\ )^N}$$

$$D(5) = 1.443 \log_e I!$$

The constant 1.443 converts the information into "bits". The above express the total information. This is given on a per individual basis by dividing each index by the number of individuals I. These are denoted:

D(1)/I, D(2)/I, etc.

The interested user should consult Margalef's paper.

D13 (Fisher, 1943)

D13 is given by α in the following transcendental equation:

$$N = \alpha \log_e \left(1 + \frac{I}{\alpha}\right)$$

D15 (Hurlbert, 1971)

$$D15 = \left[\frac{I}{I-1}\right]\left[1 - \sum_{i=1}^{N} \left(\frac{n_i}{I}\right)^2\right]$$

H, HMAX, J (Pielou, 1969)

H is the uncertainty measure appropriate to a finite population.

$$H = \frac{1}{I} \log_e \frac{I!}{n_1!\ n_2!\ \ldots\ n_N!}$$

HMAX is the maximum diversity a collection of I individuals can have. It is calculated by assuming all individuals are equally distributed in number among all N species.

$$HMAX = \frac{1}{I} \log_e \frac{I!}{\left\{\left[\frac{I}{N}\right]!\right\}^{N-r}\left\{\left(\left[\frac{I}{N}\right]+1\right)!\right\}^{r}}$$

where $I = N\left[I/N\right] + r$

and $[\cdot]$ means the integer part of the argument

$J = \frac{H}{HMAX}$ is a measure of eveness of distribution.

$0 < J \leq 1.0$

0 ↓ uneven; 1.0 ↓ even

HPRIME, HPRIMEMAX, JPRIME (Pielou, 1969)

The index HPRIME corresponds to the diversity of the population from which samples are obtained. It assumes knowledge of the true population proportion of the jth species, p_j. In practice p_j is often estimated by (n_j/I). Use of the estimated value for p_j in the formula results, however, in a biased estimate of HPRIME (Pielou, 1969). HPRIME is included in SPECDIV because incorrect usage persists.

$$\text{HPRIME} = -\sum_{j=1}^{N} \left(\frac{n_j}{I}\right) \log_e \left(\frac{n_j}{I}\right)$$

$$\text{HPRIMEMAX} = \log_e (N)$$

$$\text{JPRIME} = \frac{\text{HPRIME}}{\text{HPRIMEMAX}}$$

The following references were used in the development of SPECDIV:

Auclair and Goff, 1971 - D1, D2, D3, D4, D5, D6, D7, D8, D9, D10, D11

Margolef, 1957 - D12
Fisher, et. al., 1943 - D13
Hurlbert, 1971 - D15

An example problem used with SPECDIV is shown in Figure 11.

Figure 11

Example Problem Using SPECDIV Program

Problem Input

N = 9

$n_1 = 52$ $n_2 = 79$ $n_3 = 24$

$n_4 = 84$ $n_5 = 63$ $n_6 = 76$

$n_7 = 128$ $n_8 = 97$ $n_9 = 41$

Problem Output

```
D1 = 9.00000E+00000                          D2(WILLIS,1922)=
D3(GLEASON,1922)= 1.39153E+00000             D4(PRESTON,1948)=
D5(MARGALEF,1957)= 1.23692E+00000            D7(MENHINICK,1964)=  3.39723E-00001
D8(MENHINICK,1964)=  3.54650E-00001          D9(MONK,1966)=  1.39752E-00002
D13(FISHER,1943)=  1.48091E-00000
D6(MACARTHUR,1957)=  2.11004E+00000          D10(SIMPSON,1949)=  1.29711E-00001
D15(HURLBERT,1971)= 8.71642E-00001
D11(MCINTOSH,1967) =  2.31940E+00002
LOWER BOUND  2.14667E+00002   UPPER BOUND  6.36006E+00002

D12(MARGALEF,1957)
D(1)= 9.33289E+00000                         D(2)= 5.92950E+00001
D(3)= 1.92782E+00003                         D(4)= 2.00819E+00003
D(5)= 5.08708E+00003
D(1) / I = 1.44921E-00002
D(2) / I = 9.20730E-00002
D(3) / I = 2.99351E+00000
D(4) / I = 3.11631E+00000
D(5) / I = 7.89919E+00000
H = 2.07450E+00000                           HMAX = 2.16097E+00000
J = 9.59988E-00001
HPRIME = 2.11004E+00000                      HPRIMEMAX = 2.19722E+00000
JPRIME = 9.60323E-00001
HPRIMELOG2 = 3.04415E+00000                  HPRIMEMAXLOG2 = 3.16993E+00000
JPRIMELOG2 = 9.60323E-00001
```

3.2.6 Random Field Plot Locater

Program RANDPLØ, written by the DBDPG, locates points within a geographical site for random quadrat placement. The site is stratified into subregions with the number of locations by subregion specified by the investigator. Random point locations are subsequently determined by RANDPLØ for sampling within each subregion. These are printed out in the form of a map superimposed upon an x-y coordinate system.

3.2.7 Special Programs

Often a standard data analysis is modified in some way by an investigator for his own needs. Some of the more important of these programs written by the DBDPG are:

Rodent Pouch and Stomach Content Summary
Emlen Census Analysis
Fish Life Table Analysis
Rainfall Probability Analysis

With slight modification these programs can be used for data with similar analysis objectives.

3.3 Statistical Inference

3.3.1 Elementary Statistics

Elementary statistical measures are provided by the STATS program, written by the DBDPG. Given below are descriptive statistics and statistics useful for infering to a larger population on the basis of a sample.

Descriptive Statistics:

The following are printed or computed from the sample:

Number of observations
Mean
Variance
Standard deviation
Standard error
Coefficient of variation

Population Statistics:

Based upon the t-statistic, both 90 and 95 percent confidence intervals on the true population means are given. For cases where the assumption of normality is suspect, 90 and 95 percent confidence intervals based upon the distribution-free Tchebycheff inequality are also computed.

STATS estimates requirements for future sampling based on the sample variance (normality is assumed). For α levels of 0.9 and 0.95 an estimated sample size is computed such that the sample mean differs from the population mean by not more than a given fraction R of the population standard deviation (the sample s.d. is used to estimate the population s.d.). Example STATS output is shown in Figure 12.

Figure 12

Example of STATS Program Output

```
START 12
LENGTH  5
DECIMAL   10

 NUMBER OF OBS.       =          126
 MEAN                 =       43.429
 VARIANCE             =     6385.784
 STANDARD DEVIATION   =       79.786
 STANDARD ERROR       =        7.052
 COEF. OF VARIATION   =      183.716

THE 90 AND 95 % CONFIDENCE INTERVALS BASED ON THE T STATISTIC

ALPHA   LOWER BOUND     UPPER BOUND
 90 %       31.933  =       55.026
 95 %       29.607  =       57.251

THE 90 AND 95 % CONFIDENCE INTERVALS BASED ON
TCHEBYCHEFF INEQUALITY

ALPHA   LOWER BOUND     UPPER BOUND
 90 %       21.124  =       65.730
 95 %       11.891  =       74.967

ESTIMATED SAMPLE SIZE SUCH THAT WITH A GIVEN PROBABILITY
(EXPRESSED BY THE ALPHA LEVEL) THE SAMPLE MEAN DIFFERS
FROM THE POPULATION MEAN BY NOT MORE THAN A GIVEN FRACTION
R OF THE OF THE POPULATION STANDARD DEVIATION.

NORMALITY IS ASSUMED

 90 %     MIN. OBSER.        R
              3852       0.05
               913       0.10
               405       0.15
               228       0.20
               146       0.25
               101       0.30

 95 %     MIN. OBSER.        R
              5106       0.05
              1296       0.10
               576       0.15
               324       0.20
               207       0.25
               144       0.30
```

3.3.2 Multi-way Contingency Table Analysis

Count data are often analyzed in contingency table form using the Chi-Square test statistic. Classification of the observed sample in this case is limited to two criteria. For example, rodents trapped can be classified by age and sex, sex and trap type, age and species, etc. When the data are classified by more than two criteria a multi-way contingency table is created. In the above example the data could be used in a four-way table listing age, sex, trap type, and species criteria.

The proper G test statistic for multi-way contingency analysis is based on information theory as developed by Kullback (1959). (Sokal and Rohlf [1969] give a short discussion and an example illustrating application.) Interpretation of a multi-way analysis requires detailed theoretical understanding, and the user is advised to consult a knowledgeable statistician to ensure proper use.

The MULTITAB program was written by the DBDPG for multi-way contingency analysis. Contingency tables up to five-way can be handled. A series of hypothesis tests are possible; for example, a three-way table with criteria A,B,C will allow the following tests:

AXB	independence
AXC	independence
BXC	independence
AXBXC	independence
AXBXC	interaction

It should be noted that lack of independence among n criteria does not imply independence among (n-1)-way, (n-2)-way, etc. classifications.

Example Program Output:

The following hypothetical data represents classification of

rodents trapped according to 3 species, 2 trap types, 2 ages, and 2 sexes.

Species	Trap Type	Age	Sex	
			M	F
1	Single Trap	J	14	11
		A	16	20
1	Double Trap	J	16	14
		A	9	14
2	Single Trap	J	19	15
		A	27	19
2	Double Trap	J	23	19
		A	71	53
3	Single Trap	J	33	28
		A	36	26
3	Double Trap	J	31	36
		A	38	43

A portion of the computer output is shown in Figure 13. Under "Source" is the criteria being tested, under "DF" is degrees of freedom, and under "G" is the value of the G statistic. The G value is to be compared with a critical chi-square value with appropriate degrees of freedom. A,B,C,D represent species, trap type, age, sex respectively. From the analysis, the hypothesis of four-way independence is rejected. At the three-way level AXBXC independence and AXB and AXC two-way independence are rejected. The four-way interaction term corresponds roughly in meaning to that given by a four-way Analysis of Variance; it is not rejected.

Figure 13

Example of MULTITAB Program Output

ANALYSIS OF INFORMATION

SOURCE	DF	1/2*G	G
A X B	2.	8.346229	16.69246
A X C	2.	8.736795	17.47359
A X D	2.	1.502270	3.004540
B X C	1.	1.819460	3.638921
B X D	1.	0.4215662	0.8431325
C X D	1.	.6129518E-02	.1225904E-01
A X B X C	2.	3.471880	6.943760
A X B X D	2.	0.8186774	1.637355
A X C X D	2.	0.9424595	1.884919
B X C X D	1.	.2773704E-01	.5547407E-01
INTERACTION	2.	0.1208342	0.2416684
FOUR WAY	18.	26.21404	52.42808

3.4 General Statistical Numerical Analysis and Optimization Programs

The DBDPG maintains a library of programs useful in statistical, numerical analysis, and optimization problems. At present, over 150 programs have been acquired, and their use and interpretation is understood by DBDPG personnel. This resource is drawn upon as needed in order to analyze given problems from diverse directions. Statistical routines are useful in analyzing data, numerical analysis is useful for both data and modelling activities, and optimization programs exist for the projected trend of Desert Biome research in desert resource management decision models. A partial list of these programs follows.

Statistical	Numerical Analysis	Optimization
ANOVA	Numerical Integration	Linear Programming
Regression and Correlation Analysis	Polynomial Equation Roots	Goal Programming
Least Squares Fit by Orthogonal Polynomials	Complex and Real Exponential Integral	Quadratic Programming
Curve Fitting with Constraints	Matrix Operations	General Non-Linear Programming
Factor Analysis	Interpolation by Aitken	Critical Path Scheduling
Principal Components Analysis	Numerical Differentiation	Integer Programming
Discriminant Function Analysis	Fibonacci Search	Least Cost Network Flow
Probability Similarity Index		Minimal Spanning Tree
Normal Random Deviates		

4.0 Data Set Abstract Search System

At the end of 1972 the Desert Biome Data Bank contained over 450 data sets. As individual data sets become less manageable with increased size so, too, do collections of data sets. The DBDPG views the data set collection as one large data set under the management control of a computer information retrieval system. This system, named the Data Set Abstract Search System, provides information, in the form of data set abstracts (illustrative abstract shown in Figure 3), describing data covered.

Within the Desert Biome, data sets are available to investigators for use in modelling, ecosystem comparison and syntheses, etc. The retrieval program selects a set of DSCodes of interest to the user. This is accomplished by specifying as input one or more "profile elements," selected from a Data Set Search Profile List, which are relevant to the user's interest. Each DSCode is associated with a "relevant profile element" set. The retrieval program selects DSCodes by comparing the set of "profile elements" specified by the user with the "relevant profile set" of each DSCode (retrieval programs of this principle are commonly known as Keyword Systems).

To use the program the user selects appropriate numbers representing elements from the Profile sheet. For an abstract to qualify for retrieval the "relevant profile set" must contain all of the selected elements. For example, if the user specifies "shrubs" and "Curlew Valley" only abstracts for data sets applying to shrubs in Curlew Valley will be retrieved.

Instructions for abstract retrieval requests should be requested from:

Steve Black
Central Office Assistant
Desert Biome
Utah State University
Logan, Utah 84322

When the user receives these instructions, the request form should be filled out and returned to the same address.

References

Auclair, A.N. and F.G. Goff (1971). Diversity Relations of Upland Forests in the Western Great Lakes Area. Amer. Nat. 105:499-528.

Fisher, R.A., A.S. Corbet, and C.B. Williams (1943). The Relation Between the Number of Species and Number of Individuals in a Random Sample of an Animal Population. J. Animal Ecology 12:42-58.

Goldstein, R.A. and O.F. Grigal (1972). Computer Programs for the Ordination and Classification of Ecosystems. Ecological Sciences Division Publication No. 417, Report Code: ORNL-IBP-71-10. Oak Ridge National Laboratory.

Hurlbert, S.H. (1971). The Nonconcept of Species Diversity: A Critique and Alternative Parameters. Ecology 52:577-586.

Kruskal, J.B. (1964). Multidimensional Scaling by Optimizing Goodness of Fit to a Nonmetric Hypothesis. Psychometrika 29:1-27.

Kullback, S. (1959). Information Theory and Statistics. Wiley, New York.

Margalef, D.R. (1958). Information Theory in Ecology. General Systems 3: 36-71.

Pielou, E.C. (1969). An Introduction to Mathematical Ecology. Wiley-Interscience, New York. pp. 221-235.

Rohlf, F.J. (1972). MINT System Documentation, Dept. of Ecology and Evolution, State Univ. of New York, Stony Brook.

Shannon, C.C. and W. Weaver (1949). The Mathematical Theory of Communication. U. of Illinois Press, Urbana.

Sokal, R.R. and F.J. Rohlf (1969). Biometry. W.H. Freeman; San Francisco.

Sokal, R.R. and P.H.A. Sneath (1963). Principles of Numerical Taxonomy. W.H. Freeman, San Francisco.

Sokal, R.R. and P.H.A. Sneath (1973). Numerical Taxonomy. W.H. Freeman, San Francisco.