SCSUG November 2014

Using SAS to Manage Biological Species Data and Calculate Diversity Indices

Paul A. Montagna, Harte Research Institute, TAMU-CC, Corpus Christi, TX

ABSTRACT

Species level information is necessary for conservation, management, and research into ecological and environmental systems. A common problem is that one doesn't know what the community composition will be, so data management is complex because it has to be managed by species and not by samples. Once the data management problem is solved, there will be a need to summarize information at the sample level, and perhaps calculate a diversity index. This presentation will describe algorithms for commonly used diversity indices, and demonstrate how SAS was used to solve the data management and computational problems.

KEYWORDS

PROC SORT, PROC TRANSPOSE, DATA STEP PROGRAMMING, ARRAYS, DO LOOPS, GOTO

INTRODUCTION

Biodiversity is a common indicator of ecosystem health and often is measured in environmental assessment studies. Both terrestrial and aquatic samples are composed of multiple species and it is useful to summarize the species in a sample by computing a diversity index. Diversity indices intend to describe richness (i.e., how many species are present) and evenness (i.e., how the counts are distributed among the species). For example, it is common for disturbed samples to have dominance (i.e., low evenness regardless of the richness) compared to samples from undisturbed environments.

An important attribute of diversity is that the more samples taken, the more species are found. This concept is referred to as the "species-sample size" principle and it implies that one will always find new species. This attribute implies that when creating biological databases, it is always advisable to use a vector approach where each species from a sample is a record (Table 1). A matrix approach (where each sample is a record and species are in columns) will not work because it would be necessary to continually add new columns when new species are found. Thus, samples are decomposed and must be reassembled in order calculate diversity indices.

Table 1. Biological data.

Sample	Species	Value
1	1	6
1	2	5
2	1	9
2	3	1
2	4	2

In the present study, various diversity algorithms are discussed, and then SAS data base approaches and analytical methods are shown to implement the equations.

MATHEMATICAL METHODS

There are many approaches to calculating diversity indices, and most fall into one of the two classes being: richness indices or evenness indices (Ludwig and Reynolds 1988). The simplest richness index is R, the total number of species (equation 1).

$$R = \sum S_n \quad \text{..... (1)}$$

Richness indices are weak because they vary as a function of sample size. One of the first richness indices to account for sample size (*n*) is R1 (Margalef 1958).

$$R1 = \frac{R-1}{\ln(n)} \quad \text{..... (2)}$$

A similar formulation which is a less severe transformation of s sample size is R2 (Menhinick 1964).

$$R2 = \frac{R}{\sqrt{n}} \quad \text{..... (3)}$$

There are many different forms of diversity indices that try to account for evenness as well as richness. Hill (1973) reviewed many and demonstrated that they were part of a family or series of diversity numbers that are based on the proportional representation of each of the *i* species in a sample. Hill's diversity numbers are N0, N1, and N2. N0 is the total count, which I define above as R, the total number of species. Hill's N1 is the exponential form of the Shannon and Weaver (1948) H' diversity index.

$$N1 = e^{H'} \quad \text{..... (4)}$$

where

$$H' = \sum \left[\left(\frac{n_i}{n} \right) \left(ln \frac{n_i}{n} \right) \right] \quad (5)$$

Hill's N2 is the reciprocal of Simpson (1949) index λ.

$$N2 = \frac{1}{\lambda} \quad (6)$$

where

$$\lambda = \sum \left(\frac{n_i}{n} \right)^2 \quad (7)$$

Evenness indices have also been created. An evenness index should be maximum if all the species are equally abundant, and then then should decrease to zero as relative abundances. Again Hill (1973) shows how a series of proportional ratios of Hill's Numbered indices can be used to generalize the equations and computing. The most common evenness index used by ecologists is Pielou (1975) J' where a perfectly even community has an index value of one because there is one individual on a proportional basis for every species. Hill named J' E1

$$E1 = \frac{H'}{\ln(R)} = \frac{\ln(N1)}{\ln(N0)} \quad (8)$$

Sheldon (1969) proposed the exponential form of E1, and Hill named this E2.

$$E2 = \frac{e^{H'}}{R} = \frac{N1}{N0} \quad (9)$$

Heip (1974) proposed subtracting the minimum from E2, and Hill named this E3.

$$E3 = \frac{e^{H'}-1}{R-1} = \frac{N1-1}{N0-1} \quad (10)$$

Hill (1973) proposed the ration of N2 to N1, and this is named E4. This ratio is useful because is it in units of very abundant species. That is, as N1 and N2 reduce to one species, the ratio reduces to a value of 1.

$$E4 = \frac{1/\lambda}{e^{H'}} = \frac{N2}{N1} \quad (11)$$

Alatalo (1981) proposed a modified Hill's ratio that approaches zero as a single species becomes more dominant.

$$E5 = \frac{\left(\frac{1}{\lambda}\right)-1}{(e^{H'})-1} = \frac{N2-1}{N1-1} \quad (12)$$

DATA METHODS

The method to create biological databases is an important decision. Species data are multivariate responses in a sampling design, meaning the dataset will contain samples as rows and species as columns, and it will be easier to analyze the data once it is in this matrix format. However, it is recommended that biological community data be stored in a vector format (Table 1). The data model should contain every species within every sample as rows and the count (or size) of the species in a column (Table 1). This is important because, we never know which species will be found in any given sample, and new species are found all the time. So, most biological data is stored as a vector, not a matrix. Often samples are collected in a complex experimental design, so you can either include columns for all elements of the design, or include a key variable for each measurement value and then create a second relational table with key variables.

It is important to distinguish your data from your metadata. Metadata is data about the data. Include the data source with contact information, what the weather was like, the exact georeferenced location, the units of the variables, adequate descriptions of variable names and codes used to identify treatments, and any other information that is usually overlooked or assumed to be easy to remember. Just remember this: "You won't remember it." It is very important to write down all the details, and then record it in your data base.

For example, note that in Table 1, species 1 occurs in samples 2 and 3, but in sample 2 there are two new species (3 and 4). This storage approach creates a problem because simply computing sample means with PROC MEANS would return the wrong values. For example the average for Species 2 would be 5, because the zero value for Species 2 was not found in the second sample. Zeros must be in the data set to calculate means. This is easily accomplished by transposing the dataset twice. The first transpose will create the missing cells by using the PREFIX option and the ID statement so that each species is placed in its own column. The second transpose returns the data set to the original format, and then it is easy to replace the missing values with zeros as in the example below.

```
PROC SORT DATA=sp;
    BY sample species;
    RUN;
PROC TRANSPOSE DATA=sp OUT=tsp PREFIX=S;
    BY sample;
    ID species;
    VAR value;
    RUN;
PROC TRANSPOSE DATA=tsp OUT=sp1;
    BY sample;
    RUN;
DATA sp2;
    SET sp1;
    IF value=. THEN value=0;
    RUN;
```

Table 2. Results of first transpose.

Sample	S1	S2	S3	S4
1	6	5	.	.
2	9	.	1	2

The result of the program to put zeros in the dataset for missing species is shown in Table 2. The vector in Table 1 (data "sp") has been converted to a matrix (data "tsp". The second transpose puts the dataset back into a vector format (data "sp1"), and then the data step is used to convert the missing values to zero (Table 3, data "sp2"). The data set with zeros can be used to calculate the average abundance of each species correctly.

Table 3. Results of second transpose.

Sample	Species	Value
1	1	6
1	2	5
1	3	0
1	4	0
2	1	9
2	2	0
2	3	1
2	4	2

While using the ID statement in TRANSPOSE is necessary to create a dataset with zeros, it is not if our goal is to simply convert the species data into an array that can then be used to calculate diversity indices. In fact, to calculate the diversity indices, we only need to TRANSPOSE once without the ID statement. An example of this is listed below:

```
PROC SORT DATA=sp;
    BY sample species;
    RUN;
PROC TRANSPOSE DATA=sp OUT=tsp2 PREFIX=S;
    BY sample;
    VAR value;
    RUN;
```

Table 4. Results of transpose without the ID statement.

Sample	S1	S2	S3
1	6	5	.
2	9	1	2

This data transform creates an array with missing values (Table 4, data "tsp2). Notice that the values of the species are now compressed and that the column names (S1 to S3) no longer represent the individual species, but just the next species found in a sample. In contrast, notice in Table 2, which is a result of using the ID statement, that the column names do represent the individual specific species. Leaving out the ID statement creates an array, which is more useful for creating the programming code to calculate the diversity indices.

The next task is to calculate various measures of biodiversity (Equations 1-13). This is best accomplished by transposing the data and treating each sample as an array without the ID statement as done in Table 4. We can also calculate total abundance of organisms (i.e., to sum values). First, the ARRAY has to be declared, then it is a simple matter to calculate the index contribution for each species, and then sum the individual contributions. Two cautions: 1) include a test to make certain data exists on the line, that is the sample is not empty or no organisms were found, or there are zeros in the row, and 2) make sure there are no "divide by zero" errors, which can happen if there is just one species in a sample. Although not included here, it is simple to add an IF THEN ELSE type statement or GOTO statement and test for the values of denominators in the equations. The SAS DATA step code below calculates the indices. Also, don't forget to FORMAT the calculated variables to display the correct number of significant digits.

```
PROC TRANSPOSE DATA=sp OUT=tsp2 PREFIX=S;
   BY sample;
   VAR value;
   RUN;
DATA diversity;
   SET tsp2;
   ARRAY s s1-s100;                        /* Create temporary species array       */
   DROP s1-s100 _name_;
   DO OVER s;                              /* Loop to convert zero to missing      */
     IF s=0 THEN s=.;
     END;
   maxn=max(of s1-s100);                   /* maxn = highest number of individuals */
   tn=sum(of s1-s100);                     /* tn = total number of individuals     */
   R=n(of s1-s100);                        /* R = Richness, total species EQ-1     */
   R1=(R-1)/log(tn);                       /* R1 = Margalef index EQ-2             */
   R2=R/sqrt(tn);                          /* R2 = Menhinick index EQ-3            */
   Hprime=0; Lambda=0;                     /* Define all indices and set to zero   */
   N0=0; N1=0; N2=0;
   E1=0; E2=0; E3=0; E4=0; E5=0;
   DO OVER s;                              /* Loop to calculate diversity indices  */
     If s=. THEN GOTO msp;                 /* Skip species with missing values     */
     Hprime = Hprime + (-(s/tn) * log(s/tn));   /* Shannon H' index EQ-5           */
     Lambda = Lambda + (s/tn)**2;          /* Simpson λ index EQ-7                 */
     msp: ;
     END;
   N0=R;                                   /* Hill's Number 0, same as R EQ-1      */
   N1=exp(Hprime);                         /* Hill's number 1 EQ-4                 */
   N2=1/Lambda;                            /* Hill's number 2 EQ-6                 */
   E1=log(n1)/log(n0);                     /* Eveness index, J', EQ-8              */
   E2=N1/N0;                               /* Eveness index EQ-9                   */
   E3=(N1-1)/(N0-1);                       /* Eveness index EQ-10                  */
   E4=N2/N1;                               /* Eveness index EQ-11                  */
   E5=(N2-1)/(N1-1);                       /* Eveness index EQ-12                  */
   OUTPUT;                                 /* Add new calculated values to dataset */
   RETURN;                                 /* Start the next data step             */
   FORMAT Hprime Lambda R1 R2 N0 N1 N2 E1 E2 E3 E4 E5 8.2;
   FORMAT tn R 8.0;
   DROP maxn;
RUN;
```

The program above contains two complex action steps necessary to calculate the proportions of each individual species: the ARRAY and DO statements. The ARRAY statement creates a series of temporary variables that are used to identify each individual species. The DO statement allows for repetitive processing in a single data step. For example, the first DO loop tests each species (s) and if it is zero it converts it to a missing value. Combining the ARRAY and DO statements is necessary to calculate the proportional representation of each species to the total abundance. This is implemented in the second DO loop where the individual species (s) is divided by the total number of organisms (tn). One caution about DO loops, you have to be careful which letter used as a name and make sure it is unique to the data set. For example, here "S" is used for each species column, so if we had another numeric variable beginning with an "S" then it could be processed as if it were a species. This would not occur if it was a character variable.

SAS code can be added to test if there are no species present (`If tn=. OR tn=0 THEN …`), and then the indices can be set to either missing or zero. Likewise, it is simple to add a test if the denominator would be zero, and execute the calculation only for non-zero counts.

CONCLUSION

The calculation of diversity indices is a common practice among those who collect, manage, and study biological community samples or data. Biological community data has a unique hierarchical structure because each sample generates measurement values (in this case count data) for a unique number of species within each sample. This structure alone has implications on how to design data models for biological community structure data.

The SAS system provides powerful tools for data management, statistical analysis, and data visualization. In fact, it would very difficult and time consuming to perform the analyses presented here without the ability to easily transpose data and write code for complex equations.

All diversity indices have a weakness in that they are species independent. For example, a sample with 3 species and a second sample with 3 different species have the same richness, but the samples are completely different. Also some indices are very sensitive to sample size, such as R1 and R2. When samples sizes are equal, R is just fine. The most commonly used indices are H' (Equation 5) for diversity and J' (Equation 8) for evenness. While there are many diversity indices, some are more easily interpretable than others. For example Hill's diversity indices N1 and N2 are in the units of number of dominant species, and thus are easy to understand. Likewise, Hill's evenness index E5 is also interpretable because it approaches zero as a single species becomes dominant.

REFERENCES

Alatalo, R.V. 1981. Problems in the measurement of evenness in ecology. *Oikos* 37:199-204.

Heip, C. 1974. A new index measuring evenness. *Journal of the Marine Biological Association* 54:555-557.

Hill, M.O. 1973. Diversity and evenness: a unifying notation and its consequences. *Ecology* 54:427-432.

Ludwig, J.A. and J.F. Reynolds. 1988. Statistical Ecology: A Primer on Methods and Computing, John Wiley & Sons, New York.

Margalef, R. 1958. Information Theory in Ecology. *General Systematics* 3:36-71.

Pielou, E.C. 1975. Ecological Diversity. Wiley, New York.

Shannon, C. E. and W. Weaver. 1949. The Mathematical Theory of Communication. University of Illinois Press. Urbana, IL.

Sheldon, A.L. 1969. Equitability indices: dependence on species count. *Ecology* 50:466-467.

Simpson, E.H. 1949. Measurement of diversity. *Nature* 163:688

ACKNOWLEDGMENTS

This publication was made possible, in part, by the National Oceanic and Atmospheric Administration, Office of Education Educational Partnership Program award (NA11SEC4810001). Its contents are solely the responsibility of the award recipient Paul Montagna and do not necessarily represent the official views of the U.S. Department of Commerce, National Oceanic and Atmospheric Administration.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Paul A. Montagna
Harte Research Institute for Gulf of Mexico Studies
Texas A&M University-Corpus Christi
6300 Ocean Drive, Unit 5869
Corpus Christi, Texas 78712
paul.montagna@tamucc.edu
Office (361) 825-2040
http://harteresearchinstitute.org/

TRADEMARK INFORMATION