THE MEASUREMENT OF SIZE DIVERSITY

This webpage provides a free program for measuring the size diversity of a size distribution. The program "diversity08" was programmed by J.J. Egozcue, February 2008, following the methods presented in Quintana et al. (2008) *Limnology and Oceanography: Methods*, 6:75-86. It is based on a non-parametric kernel estimation, after data standardization by division of sample geometric mean. This method is applicable to any type of size distribution, even if it doesn't fit a parametric function, and allows its immediate comparison with the size diversity of other distributions independently of the method used for size estimation. See the above reference for more details on the method. The webpage was designed by Omar Martínez.

1. THEORETICAL FRAMEWORK

The Shannon-Wiener index is used to measure size diversity. Since size is not a discrete, but a continuous variable, the goal is to estimate a Shannon size diversity index $\mu_2(X)$ corresponding to the probability density function $p_X(x)$ of the size of the individuals, which takes the following integral form:

$$\mu_2(X) = -\int_0^{+\infty} p_X(x)\log_2 p_X(x)\,dx$$

Size diversity is computed by means of a non-parametric kernel estimation. The expression used for the computation of the size diversity is:

$$\hat{\mu}_{\mathrm{kerMC}}(X) = \bar{y} - \frac{1}{n}\sum_{k=1}^{n}\log_2\left[\frac{1}{n\sqrt{2\pi}\,\sigma}\sum_{j=1}^{n}\exp\left(-\frac{1}{2}\frac{(y_k - y_j)^2}{\sigma^2}\right)\right]$$

This equation coincides with equation (9) in Quintana *et al.* (2008), except that the base of the logarithm is binary instead of a natural logarithm (see reference above for further details on notation).

Data are standardized by means of the division by the geometric mean. According to Quintana *et al.* (2008), this standardization has the advantages that 1) the same size diversity value is obtained when using original size or log-transformed data and 2) size measurements with different dimensionality (longitudes, areas, volumes or biomasses) may be immediately compared with the simple addition of ln *k*, where *k* is the dimensionality (1, 2, or 3, respectively). See below for more details (section 3.2 "dimensionality").

2. PROGRAM PROCEDURE

- Download the whole diversity08 folder to your computer.
- Open the input file spreadsheet, diversity08_data, and fill (or change) title, dimensionality and abundances in the three first lines of the input file (see section 3).
- Copy your data into the input file spreadsheet, diversity08_data, immediately below these three first lines (as outlined in section 3). Do not change the name or the extension of this input file. Replace existing data if necessary. It works if two (or three when abundances=1) consecutive columns of an excel sheet are directly copied and pasted in the input file.
- Run the program file diversity08-1. Input and program files must be in the same folder. A normal end of program should be achieved.
- Take the results from the output files (see section 4).

3. INPUT FILE (diversity08_data)

The program requires information to be input into the diversity08_data input file as follows:

Title (first line), for further identification of the output files.

Dimensionality (second line) indicates the dimension in which size is measured. There are only three options:
1- When size measurements are lengths (L)
2- When size measurements are surfaces (L^2)
3- When size measurements are volumes or biomass (L^3)

Abundances (third line) is a binary function and indicates whether the input file includes data on abundance:
0 - no abundance values are included; each value is the size of a single individual (see example 1)
1 - an additional column is included indicating the number of individuals of a determinate size (see example 2).

Data list (fourth and consecutive lines)
Data is entered in columns, the first one is the sample identifier (sample ID). The program computes the size diversity for each one of the sample ID identified in this first column. Size distribution data can be entered in two different ways, depending on whether the input file includes data on abundance, as shown in the following examples. Points, not commas, must be used to separate decimals.

EXAMPLE 1: A phytoplankton sample (Title = "Phytoplankton"), counted by flow cytometry, where the size of each individual is measured. The size of a cell was measured using its equivalent spherical diameter (ESD, µm). The input file will have this form:

```
Phytoplankton.
1 Dimensionality
0 Abundances
1        3
1        3
1       14
1       16
1        6
1        8
27       4
27       3
27       7
27      15
5        6
5        9
5       11
```

The first column of the data list is the sample identifier (sample ID). It must be numerical and have less than 10 digits. The second column is the size of a single individual. In this example, there are three samples (sample 1, sample 27 and sample 5); in sample 1 there are six individuals, two of them with an equivalent spherical diameter of 3 µm, etc.

EXAMPLE 2: A sample of benthic invertebrates (Title = "Benthic invertebrates"). Size is measured in dry weight (µg), by measuring the length of the individuals and converting length in dry weight using allometric relationships found in the bibliography. Units in µgC can also be used. Only a selected number of individuals are measured and the abundance of each size is estimated.

```
Benthic invertebrates.
3 Dimensionality
1 Abundances
1    1.11610E+01     2
1    1.09070E+01     8
1    2.99840E+00    12
1    8.86290E+00    30
1    5.71810E+00    17
1    9.28090E+00    74
2    1.85380E+01     7
2    4.35180E+01     3
2    5.85180E+00    25
2    6.27090E+00    32
3    1.11610E+01    28
3    2.99840E+00     1
```

In this case, there is a new column indicating the abundance of each size. In sample 1, there are 2 individuals with an individual biomass of 1.11610E+01 µg, 8 individuals with an individual biomass of 1.09070E+01 µg, etc.

If you find two (or more) individuals with the same size, DO NOT AGGREGATE them to a single size measurement with double (or more) abundance.

4. OUTPUT FILES

The program generates three different output files: diversity08_summ, diversity08_seq and diversity08_log. These files are created after each run of the program. If information must be saved, files should be saved with another name. The file **diversity08_summ** is the main file, which gives the size diversity values. The file **diversity08_seq** computes additional information on each sample, which may be helpful for users. The file **diversity08_log** only shows the program process. If program does not work correctly, this file is useful for identifying errors. You won't need to use it if you have a normal end of program.

4.1 diversity08_summ output file

This is the summary output file and provides the size diversity of each sample. The arithmetic mean, the geometric mean and the standard deviation of the original data are also included. Data is displayed in a table to facilitate pasting it into a spreadsheet.

An additional warning column indicates whether the data contains some samples with only one value (warning 1) or if there are some negative values (warning 2), for example if logarithms were previously applied to data < 1.

The diversity08_summ output files of examples 1 and 2 are shown in appendix 1. Note that size diversity values may be negative (as in sample 3 of example 2). The model is based on a continuous probability density function, which may take local probability values > 1. Thus, 0 is not the minimum value of size diversity. Negative values may be found when there is a high accumulation of data in a determinate size, meaning extremely low size diversity values.

4.2 Dimensionality

One of the main advantages of the standardization by the geometric mean is that size distribution data measured with different dimensionalities are comparable, when an allometric relationship between them is assumed. For instance, let X be a random length, e.g., a radius of a sphere (or the equivalent spherical diameter, as in example 1), and let V be a scaled power of X, defined by $V = aX^k$, e.g., the volume of a sphere $a = 4\pi/3$, $k = 3$. Assume that both variables, V and X, are divided by the respective geometric means, the relationship of the corresponding diversities is:

$$\mu(V) = \log_2 k + \mu(X)$$

Where k is the dimensionality ($k = 3$ in this example). In practice, it means that data sets which differ in dimensionality may be easily compared by the simple addition (or subtraction) of $\log_2 k$ (see Quintana *et al.* (2008) for details).

By default, results of size diversity which appear in the diversity08_summ output file are normalised to dimensionality 3, that is, when dimensionality chosen in the input file is 1, the program adds $\log_2 3$ to the final result of size diversity; if dimensionality chosen is 3, it doesn't add anything. Using example 1, we can now examine how dimensionality works. For each ESD value, the biovolume may be obtained ($4\pi r^3/3$):

Sample	ESD (µm)	biovolume (μm^3)
1	3	14.13
1	3	14.13
1	14	1436.03
1	16	2143.57
1	6	113.04
1	8	267.95
27	4	33.49
27	3	14.13
27	7	179.50
27	15	1766.25
5	6	113.04
5	9	381.51
5	11	696.56

The resulting diversity08_summ output files for each dataset, using ESD and biovolume values (µm or μm^3 and dimensionalities 1 or 3 respectively), are shown in appendix 2. Note that the output data differ in mean, geometric mean and standard deviation, but they have the same size diversity.

4.3 diversity08_seq output file

The output file diversity08_seq provides additional information for each sample, which may be useful for users (see example in appendix 3). Its use is not strictly necessary, since data on size diversity are summarized in the diversity08_summ output file.

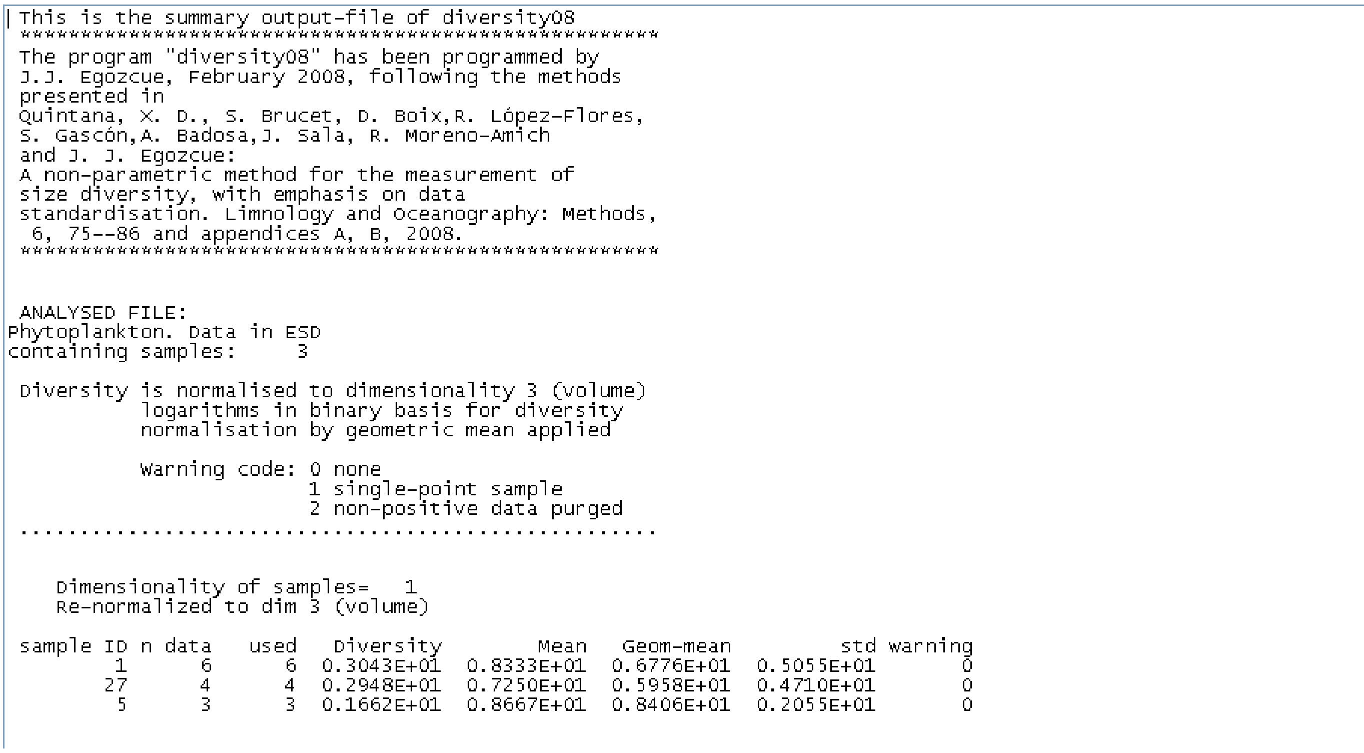
For each sample, a set of descriptors of the original data are listed (sample parameters of the original data). Furthermore, this output file also displays results of different approaches in the measurement of the size diversity (diversity estimation). The first row (MC-kernel) displays the diversity values measured by means of the non-parametric kernel approach, using either a binary logarithm (3 first columns) or a natural logarithm (3 last columns). The second row (Log-normal) displays the diversity values obtained using a parametric approach, by fitting data to a log-normal distribution, again using either a binary logarithm (3 first columns) or a natural logarithm (3 last columns). Results in both rows will be similar if the shape of the original size distribution is similar to a log-normal distribution, but may differ if the shape strongly differs from a log-normal distribution.

The columns dim=1, dim=2 and dim=3 indicate the dimension the data refers to. Examining the first three columns, size diversity values in dim=1 differ from that in dim=2 by $\log_2 2$, and from that in dim=3 by $\log_2 3$ (these differences are ln 2 and ln 3 in the last three columns).

The (MC-kernel diversity, log_2 dim=3) is the default diversity shown in the diversity08_summ output file.

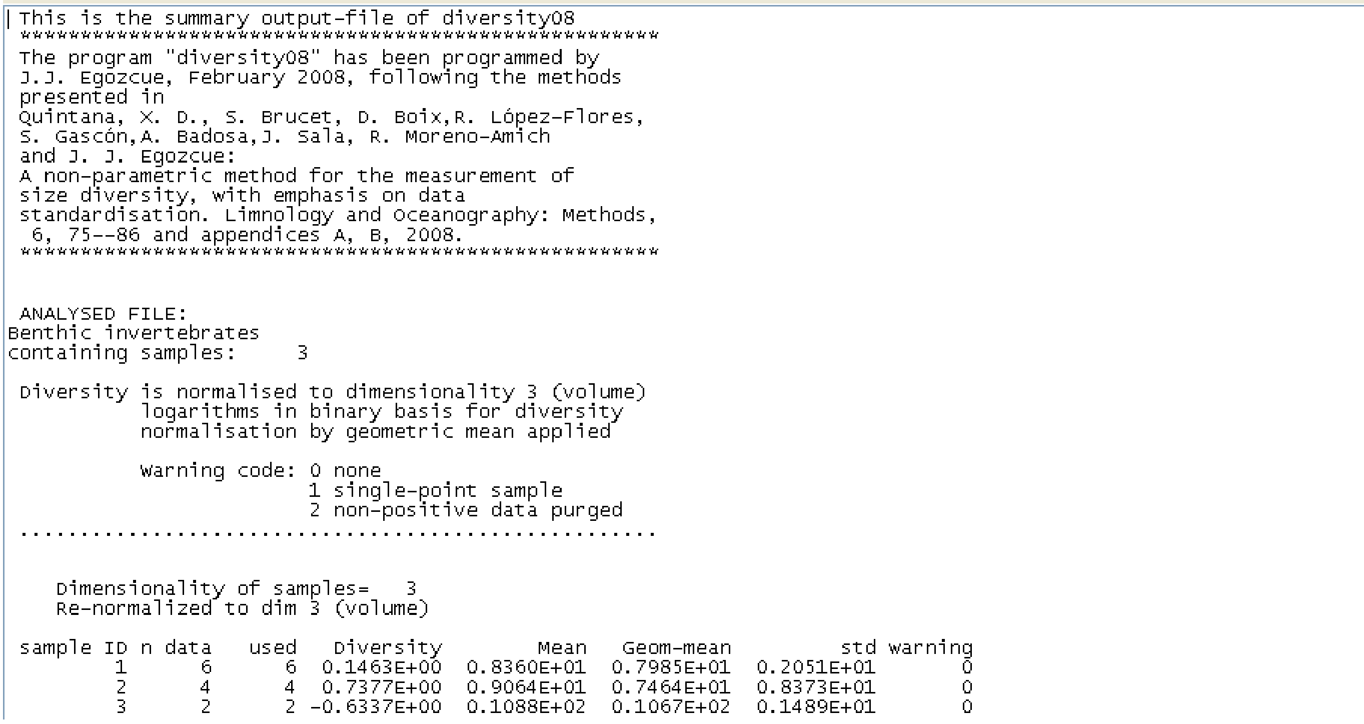
Appendix 1:

diversity08_sum outfile generated in example 1 (Phytoplankton):



```
 This is the summary output-file of diversity08
 *****************************************************
 The program "diversity08" has been programmed by
 J.J. Egozcue, February 2008, following the methods
 presented in
 Quintana, X. D., S. Brucet, D. Boix,R. López-Flores,
 S. Gascón,A. Badosa,J. Sala, R. Moreno-Amich
 and J. J. Egozcue:
 A non-parametric method for the measurement of
 size diversity, with emphasis on data
 standardisation. Limnology and Oceanography: Methods,
  6, 75--86 and appendices A, B, 2008.
 *****************************************************


 ANALYSED FILE:
Phytoplankton. Data in ESD
containing samples:     3

 Diversity is normalised to dimensionality 3 (volume)
           logarithms in binary basis for diversity
           normalisation by geometric mean applied

           Warning code: 0 none
                         1 single-point sample
                         2 non-positive data purged
 .....................................................


    Dimensionality of samples=   1
    Re-normalized to dim 3 (volume)

 sample ID n data   used   Diversity        Mean   Geom-mean         std warning
         1      6      6  0.3043E+01  0.8333E+01  0.6776E+01  0.5055E+01       0
        27      4      4  0.2948E+01  0.7250E+01  0.5958E+01  0.4710E+01       0
         5      3      3  0.1662E+01  0.8667E+01  0.8406E+01  0.2055E+01       0
```

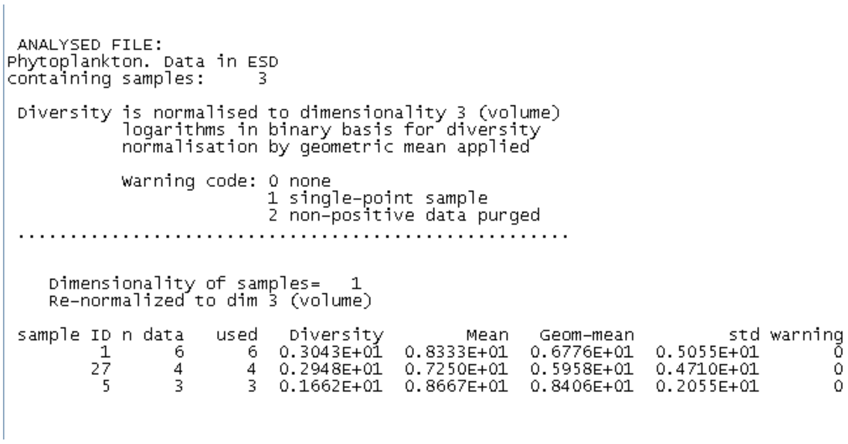
diversity08_sum outfile generated in example 2 (Benthic invertebrates):



```
 This is the summary output-file of diversity08
 *****************************************************
 The program "diversity08" has been programmed by
 J.J. Egozcue, February 2008, following the methods
 presented in
 Quintana, X. D., S. Brucet, D. Boix,R. López-Flores,
 S. Gascón,A. Badosa,J. Sala, R. Moreno-Amich
 and J. J. Egozcue:
 A non-parametric method for the measurement of
 size diversity, with emphasis on data
 standardisation. Limnology and Oceanography: Methods,
  6, 75--86 and appendices A, B, 2008.
 *****************************************************


 ANALYSED FILE:
Benthic invertebrates
containing samples:     3

 Diversity is normalised to dimensionality 3 (volume)
           logarithms in binary basis for diversity
           normalisation by geometric mean applied

           Warning code: 0 none
                         1 single-point sample
                         2 non-positive data purged
 .....................................................


    Dimensionality of samples=   3
    Re-normalized to dim 3 (volume)

 sample ID n data   used   Diversity        Mean   Geom-mean         std warning
         1      6      6  0.1463E+00  0.8360E+01  0.7985E+01  0.2051E+01       0
         2      4      4  0.7377E+00  0.9064E+01  0.7464E+01  0.8373E+01       0
         3      2      2 -0.6337E+00  0.1088E+02  0.1067E+02  0.1489E+01       0
```

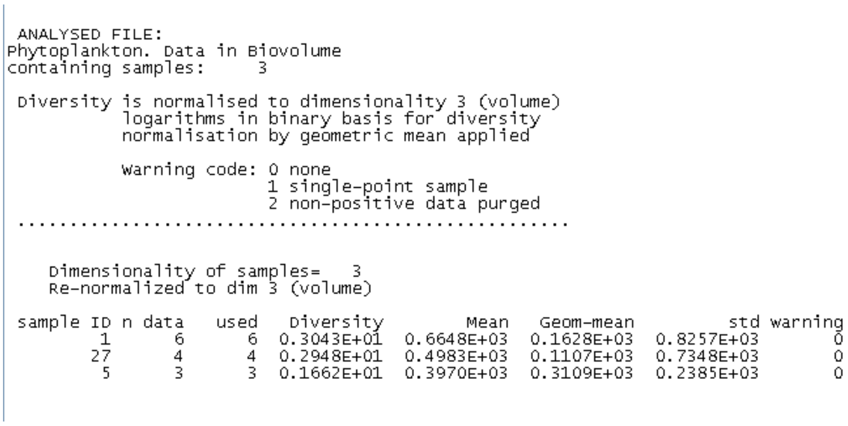
Appendix 2:

diversity08_sum outfile generated in example 1 using ESD data (dimensionality = 1):



```
 ANALYSED FILE:
Phytoplankton. Data in ESD
containing samples:     3

 Diversity is normalised to dimensionality 3 (volume)
           logarithms in binary basis for diversity
           normalisation by geometric mean applied

           Warning code: 0 none
                         1 single-point sample
                         2 non-positive data purged
 ....................................................


    Dimensionality of samples=   1
    Re-normalized to dim 3 (volume)

 sample ID n data   used   Diversity        Mean   Geom-mean         std warning
         1      6      6  0.3043E+01  0.8333E+01  0.6776E+01  0.5055E+01       0
        27      4      4  0.2948E+01  0.7250E+01  0.5958E+01  0.4710E+01       0
         5      3      3  0.1662E+01  0.8667E+01  0.8406E+01  0.2055E+01       0
```

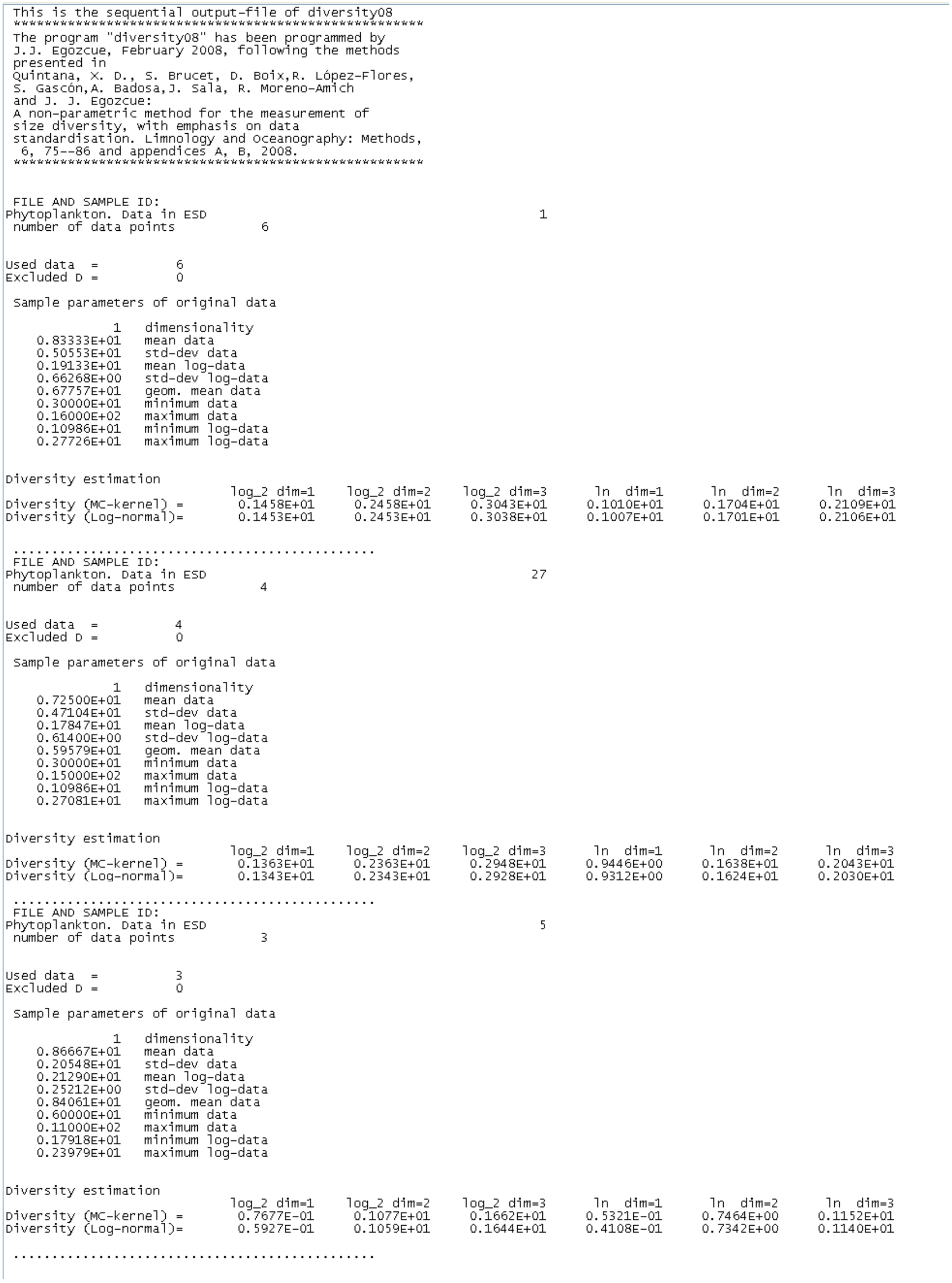
diversity08_sum outfile generated in example 1 using biovolume data (dimensionality = 3):



```
 ANALYSED FILE:
Phytoplankton. Data in Biovolume
containing samples:     3

 Diversity is normalised to dimensionality 3 (volume)
           logarithms in binary basis for diversity
           normalisation by geometric mean applied

           Warning code: 0 none
                         1 single-point sample
                         2 non-positive data purged
 ....................................................


    Dimensionality of samples=   3
    Re-normalized to dim 3 (volume)

 sample ID n data   used   Diversity        Mean   Geom-mean         std warning
         1      6      6  0.3043E+01  0.6648E+03  0.1628E+03  0.8257E+03       0
        27      4      4  0.2948E+01  0.4983E+03  0.1107E+03  0.7348E+03       0
         5      3      3  0.1662E+01  0.3970E+03  0.3109E+03  0.2385E+03       0
```

Appendix 3:

diversity08_seq outfile generated in example 1:



```
 This is the sequential output-file of diversity08
 *****************************************************
 The program "diversity08" has been programmed by
 J.J. Egozcue, February 2008, following the methods
 presented in
 Quintana, X. D., S. Brucet, D. Boix,R. López-Flores,
 S. Gascón,A. Badosa,J. Sala, R. Moreno-Amich
 and J. J. Egozcue:
 A non-parametric method for the measurement of
 size diversity, with emphasis on data
 standardisation. Limnology and Oceanography: Methods,
  6, 75--86 and appendices A, B, 2008.
 *****************************************************


 FILE AND SAMPLE ID:
Phytoplankton. Data in ESD                                           1
 number of data points           6


Used data  =          6
Excluded D =          0

 Sample parameters of original data

              1   dimensionality
    0.83333E+01   mean data
    0.50553E+01   std-dev data
    0.19133E+01   mean log-data
    0.66268E+00   std-dev log-data
    0.67757E+01   geom. mean data
    0.30000E+01   minimum data
    0.16000E+02   maximum data
    0.10986E+01   minimum log-data
    0.27726E+01   maximum log-data


Diversity estimation
                             log_2 dim=1    log_2 dim=2    log_2 dim=3      ln  dim=1      ln  dim=2      ln  dim=3
Diversity (MC-kernel) =       0.1458E+01     0.2458E+01     0.3043E+01     0.1010E+01     0.1704E+01     0.2109E+01
Diversity (Log-normal)=       0.1453E+01     0.2453E+01     0.3038E+01     0.1007E+01     0.1701E+01     0.2106E+01

 ..............................................
 FILE AND SAMPLE ID:
Phytoplankton. Data in ESD                                          27
 number of data points           4


Used data  =          4
Excluded D =          0

 Sample parameters of original data

              1   dimensionality
    0.72500E+01   mean data
    0.47104E+01   std-dev data
    0.17847E+01   mean log-data
    0.61400E+00   std-dev log-data
    0.59579E+01   geom. mean data
    0.30000E+01   minimum data
    0.15000E+02   maximum data
    0.10986E+01   minimum log-data
    0.27081E+01   maximum log-data


Diversity estimation
                             log_2 dim=1    log_2 dim=2    log_2 dim=3      ln  dim=1      ln  dim=2      ln  dim=3
Diversity (MC-kernel) =       0.1363E+01     0.2363E+01     0.2948E+01     0.9446E+00     0.1638E+01     0.2043E+01
Diversity (Log-normal)=       0.1343E+01     0.2343E+01     0.2928E+01     0.9312E+00     0.1624E+01     0.2030E+01

 ..............................................
 FILE AND SAMPLE ID:
Phytoplankton. Data in ESD                                           5
 number of data points           3


Used data  =          3
Excluded D =          0

 Sample parameters of original data

              1   dimensionality
    0.86667E+01   mean data
    0.20548E+01   std-dev data
    0.21290E+01   mean log-data
    0.25212E+00   std-dev log-data
    0.84061E+01   geom. mean data
    0.60000E+01   minimum data
    0.11000E+02   maximum data
    0.17918E+01   minimum log-data
    0.23979E+01   maximum log-data


Diversity estimation
                             log_2 dim=1    log_2 dim=2    log_2 dim=3      ln  dim=1      ln  dim=2      ln  dim=3
Diversity (MC-kernel) =       0.7677E-01     0.1077E+01     0.1662E+01     0.5321E-01     0.7464E+00     0.1152E+01
Diversity (Log-normal)=       0.5927E-01     0.1059E+01     0.1644E+01     0.4108E-01     0.7342E+00     0.1140E+01

 ..............................................
```