

Diversity characterizations in R

Béla Tóthmérész

Ecological Institute

Debrecen University

P. O. Box 71

Debrecen, H-4010

Hungary

E-mail: tothmerb@delfin.klte.hu

Abstract

Diversity measuring methods are presented in the R program language with special emphasis on the modern techniques that provide scalable diversity comparisons.

Key words: Diversity indices, scalable one-parametric diversity index families, species accumulation plots

Introduction

R is a programming environment for data analysis and graphics (Ihaka & Gentleman, 1996). It is a free-ware version of the program package S-Plus. A wide range of statistical methods are implemented in R, and new applications can easily be developed (Crawley, 2002; Venables and Ripley, 2002). Knowledge of the R and/or S language is useful, although not absolutely necessary to use the procedures in the 'diversity' package. As an introductory book, Krause and Olsen (2002) may be mentioned.

This paper provides a description and/or explanation of the diversity add-on package for R through examples. It is best used while sitting at a console with R running. This draft of the Guide should be considered a work in progress. It is not a commercial product. Effort is put into maintaining the documentation and make it accessible through the web.

How can R be installed

R can be installed through the web (<http://www.r-project.org/>) from a CRAN site. It contains binaries for a base distribution and a large number of add-on packages to run on Windows 95, 98, ME, NT4, 2000, and XP on Intel and clones.

Executing commands and saving output to a file

Commands can be executed via the command line of R. Longer sequences of commands or implementations of new functions can be written in a separate file. When the commands are stored in a file, named 'diversity.R' that has to reside in the working (default) directory, then the command line statement

```
> source("diversity.R")
```

will execute the commands in 'diversity.R' and load function code contained in 'diversity.R' file.

If you started the R program package and you choose in the pull down menu the 'File' and then 'Source R code' options, you can load the source code of the R procedures and the demo datasets. You can use the usual "Browse" button or you make a copy of the 'diversity.R' file in the default directory of your R version. R is usually installed into the 'C:\Program Files\R' directory and the working directory is identical with the actual version of the program; at the time of writing the paper it is 'rw2001'. If you make a copy into the 'C:\Program Files\R\rw2001' directory you can load the program without browsing the directories.

You can also open 'diversity.R' in your favourite text editor and use copy/paste to paste commands or sequences of commands into the command line. The output normally will appear on the screen only. You can change this by using the function `sink`. The command:

```
> sink("DivOutput.txt")
```

will divert all subsequent output from the console to an external file, "DivOutput.txt". The command

```
> sink()
```

restores it to the console once again. Alternatively you may copy the results from the R desktop through the clipboard in the usual way. To save the figures, right-click with the mouse button when the pointer is within the figure, and chose a suitable option from the appearing menu.

What is the "diversity" package?

The "diversity" package is an add-on package for the R statistical computing system. Several data sets are also included with this package and will be used in examples in this

guide. The examples do not include a detailed theoretical and/or technical explanation. See Patil and Tailie (1982), Tóthmérész (1993, 1995, 1997, 1998) and Tóthmérész & Magura (2005) for an introduction to the statistics of diversity.

Sample data sets

There are six sample data frames included in the package. Data frames are matrix-like structures (see e.g. Krause & Olsen, 2002). The simplest data set is named "trichotom", originally published by Tóthmérész (2002). It may be used to check and/or demonstrate the calculations. Tóthmérész (1993) used to illustrate the diversity ordering by the "ABC" data set (see also Southwood & Henderson, 2000). The three assemblages included in the "dens" data frame is to demonstrate the density dependent and density independent representations. This was originally published and discussed in Tóthmérész (1998), as well as the data sets of the "oAB" data frame. This latter is a pooled data set, by pooling the 10 separate samples contained in the "oA" and "oB" data frames.

Methods available in the package

The methods included in the package are presented in the paper of Tóthmérész & Magura (2005). There are a few slightly different methods that are equally useful from a biological point of view. These methods are reviewed by Patial & Tallie (1983), and Tóthmérész (1995, 1997).

Examples of the usage of the procedures

The data set `trichotom` contains the number of individuals of the species in three assemblages. Entering

```
> trichotom
```

displays the entire data set.

```
> trichotom
      nC1 nC2
species1  40  60
species2  30  20
species3  30  10
species4   0  10
```

It contains two assemblages (`nC1` and `nC2`) and 4 species (`species1`, ..., `species4`).

Entering

```
> t(trichotom)
```

displays the data set in the transposed form:

	species1	species2	species3	species4
nC1	40	30	30	0
nC2	60	20	10	10

Using the

```
> trichotom$nC1
```

commands, only the data vector of the nC1 assemblage is displayed:

```
[1] 40 30 30 0
```

Traditional diversities are obtained as a special case of the one-parametric Rényi diversity index family. Shannon diversity is obtained for $\alpha=1$ for the nC1 and nC2 assemblages:

```
> div.Renyi(trichotom$nC1, alpha=1)
```

```
[1] 1.0889
```

```
> div.Renyi(trichotom$nC2, alpha=1)
```

```
[1] 1.0889
```

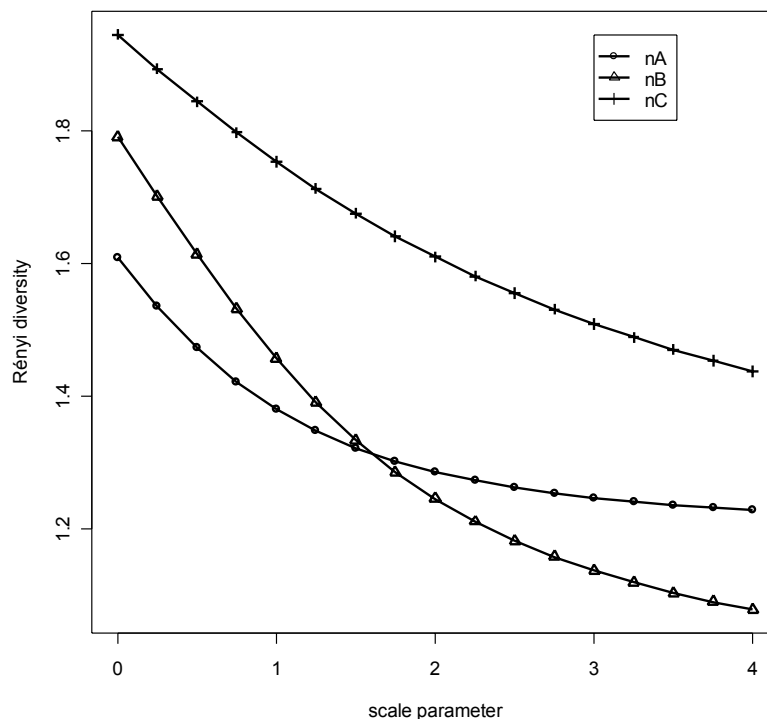


Figure 1. Diversity profiles of the Rényi's one-parametric diversity index family for the ABC data set. This resulted from the command line: (plot.div.Renyi(ABC, from=0, to=4, step=0.25)).

The default option is the natural logarithm; choosing `base=2` you receive the Shannon diversity with the logarithm of base 2:

```
> div.Renyi(trichotom$nC1, alpha=1, base=2)
[1] 1.570951
```

When `alpha=0`, the logarithm of the number of species is obtained:

```
> div.Renyi(trichotom$nC1, alpha=0)
[1] 1.098612
```

Using the exponential function (`exp`) we get the number of species of the `nC1` assemblage:

```
> exp(div.Renyi(trichotom$nC1, alpha=0))
[1] 3
```

The three variants of the Simpson (or quadratic) diversity is received in the following way:

```
> div.Renyi(trichotom$nC1, alpha=2)
[1] 1.078810
> exp(div.Renyi(trichotom$nC1, alpha=2))
[1] 2.941176
> 1-1/exp(div.Renyi(trichotom$nC1, alpha=2))
[1] 0.66
```

The first one is $-\log \sum_{i=1}^S p_i^2$, the second one is the effective number of species for the quadratic diversity, i.e. $1/\sum_{i=1}^S p_i^2$. The last one is the $1 - \sum_{i=1}^S p_i^2$ for of the quadratic diversity.

A particular value of the right-tail sum (*RTS*) diversity is received in the following way (`i=2`):

```
> div.RTS(trichotom$nC2, 2)
[1] 0.2
```

If you would like to know each value of the *RTS* diversity for $i=1, \dots, S$, you should enter:

```
> div.RTS.all(trichotom$nC2)
[1] 0.4 0.2 0.1 0.0
```

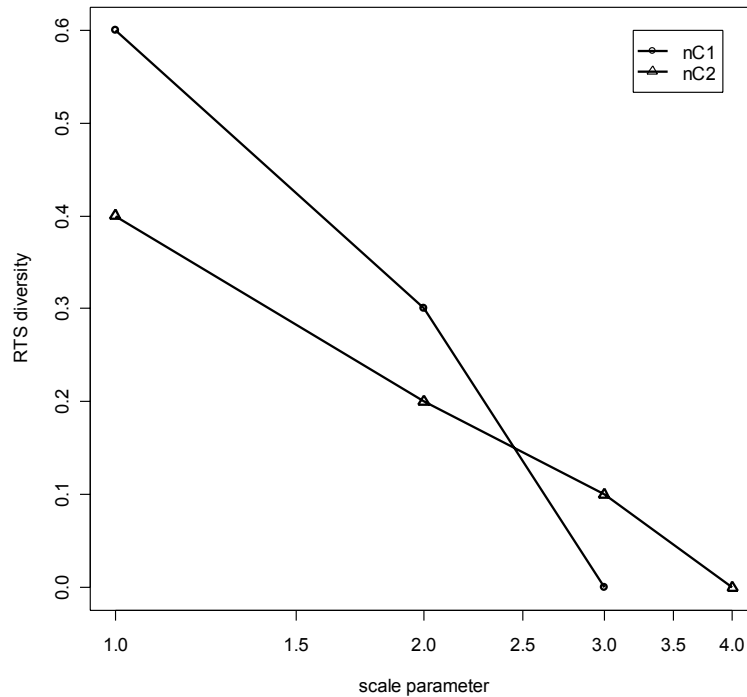


Figure 2. *RTS* diversity profiles for the `trichotom` data set. Output of the command `line: (plot.div.RTS(trichotom, xLabel=0, log="x"))`.

The species accumulation plots have more options to choose from than for the other one-parametric diversity index families. For the `dens` data set the expected number of species in a sub-sample containing 25 individuals is the following (for an infinitely large assemblage):

```
> div.ESm.Infinite(dens$nE, 25)
[1] 7.070675
> div.ESm.Infinite(dens$nF, 25)
[1] 6.181578
> div.ESm.Infinite(dens$nG, 25)
[1] 7.070675
```

For the finite case the result is slightly different:

```
> div.ESm.Finite(dens$nE, 25)
[1] 7.082133
> div.ESm.Finite(dens$nF, 25)
[1] 6.192088
> div.ESm.Finite(dens$nG, 25)
[1] 7.099477
```

There are commands to draw the diversity profiles of the assemblages. The simplest way to draw the Rényi diversity profiles of the assemblages of the ABC data set is the following (using the default option):

```
> plot.div.Renyi(ABC)
```

You can change the parameters as described in the documentation of the procedures (Appendix):

```
> plot.div.Renyi(ABC, from=0, to=4, step=0.25)
> plot.div.Renyi(ABC, from=0, to=4, step=0.25, main="")
```

RTS diversity profiles of the assemblages in the `trichotom` data set with the default option are produced by the command

```
> plot.div.RTS(trichotom)
```

Using the `log="x"` option, the *x*-axis scale is not logarithmic:

```
> plot.div.RTS(trichotom, log="x")
```

This may be useful for a species-poor assemblage (fewer than 15 species). In the case of species-rich assemblages, the dominance plot using logarithmic scale usually produces a much nicer graph (Tóthmérész, 1997).

To draw a species accumulation plot is slightly more sophisticated than to draw other diversity profiles. The default options still works. On entering

```
> plot.div.ESm(dens)
```

you get a density independent representation for an infinitely large assemblage and default starting and ending values of the number of individuals along the *x*-axis. You can use the options in a longer (and safer) way

```
> plot.div.ESm(dens, type="infinite", representation="individual",
from=2, to=-1, nSteps=10)
```

or just shortly:

```
> plot.div.ESm(dens, "infinite", "individual", from=2, to=-1, nSteps=10)
```

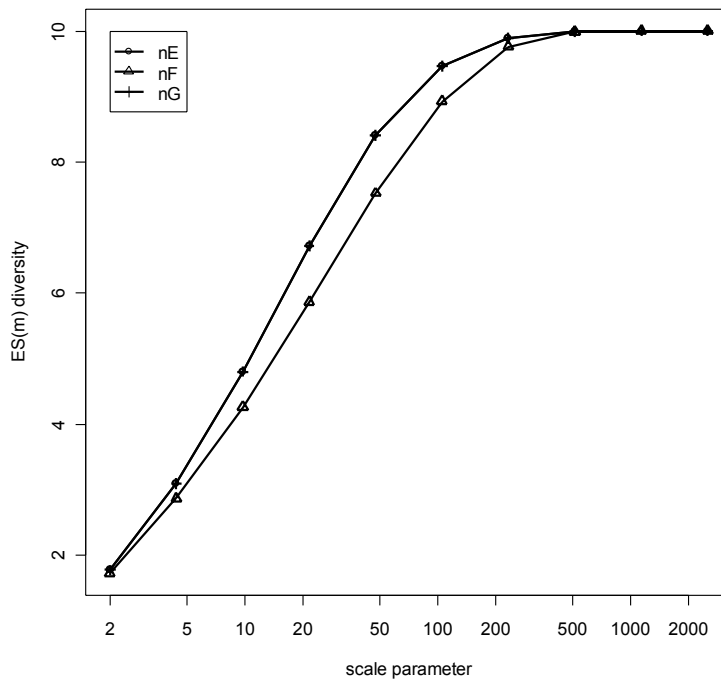


Figure 3. Species accumulation curves for the `dens` data set; density independent representation and infinitely large assemblage is assumed, resulting from: `(plot.div.ESm(dens, "infinite", "individual", from=2, to=-1, nSteps=10))`. Please note that $nE \equiv nG$.

Let us assume that you would like to draw the expected number of individuals starting with 5 individuals and increase the sub-sample size up to 400 individuals. You have to type the following command line:

```
> plot.div.ESm(dens, "infinite", "individual", from=5, to=400, nSteps=15)
```

The following two commands produce a figure demonstrating that finite and infinite options may produce really different outcomes. This occurs because the infinite option is based on the relative frequencies of the species:

```
> plot.div.ESm(dens, "finite", "individual", from=60, to=100, nSteps=5)
> plot.div.ESm(dens, "infinite", "individual", from=60, to=100, nSteps=5)
```

There are further commands to help working with the package. You can print the values of the Rényi diversity in the following way:

```
> print.div.Renyi(trichotom, from=0, to=2, step=0.5)
```

This will result in the following output:

```
scale      nC1      nC2
```

1	0.0	1.098612	1.3862944
2	0.5	1.093814	1.2349776
3	1.0	1.088900	1.0889000
4	1.5	1.083891	0.9643264
5	2.0	1.078810	0.8675006

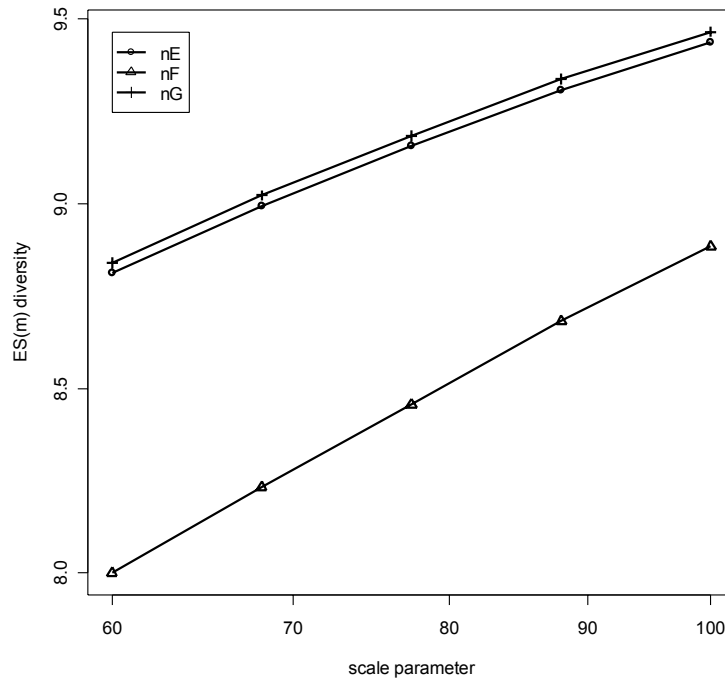


Figure 4. Species accumulation curves for the dens data set; density independent representation and finite assemblage is assumed. Command line: `(plot.div.ESm(dens, "finite", "individual", from=60, to=100, nSteps=5))`. Please note that $nE \neq nG$, while using the same representation with infinite option they were identical.

If you would like to know the effective number of species, which is defined as $\exp(\text{Renyi})$, simply enter:

```
> exp(print.div.Renyi(trichotom, from=0, to=2, step=0.5))
      scale      nC1      nC2
1  1.000000  3.000000  4.000000
2  1.648721  2.985641  3.438302
3  2.718282  2.971004  2.971004
4  4.481689  2.956158  2.623020
5  7.389056  2.941176  2.380952
```

Note that the scale parameter values are also printed as $\exp(\alpha)$ instead of α .

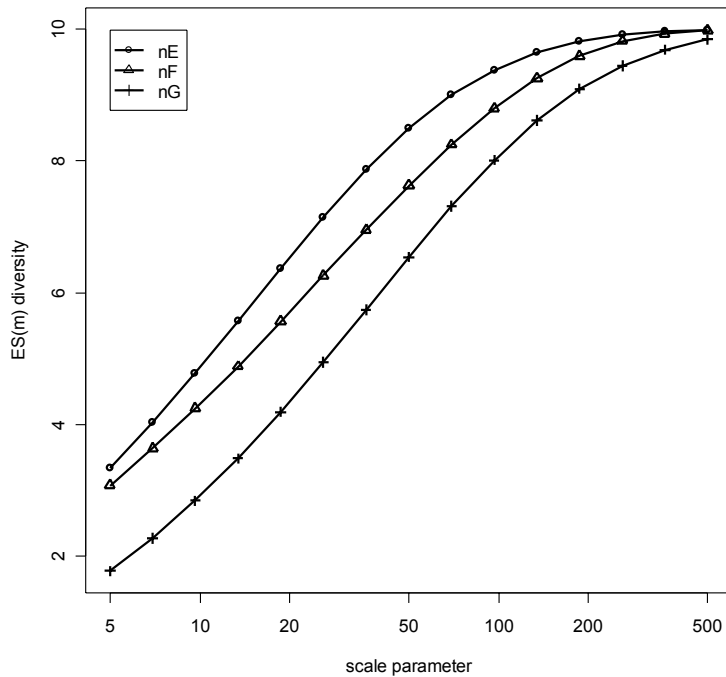


Figure 5. Species accumulation curves for the dens data set; density dependent representation and infinitely large assemblage is assumed. Command line: (plot.div.ESm(dens, type="infinite", representation="area", from=5, to=500, nSteps=15)).

The total number of species of the oA data set, which contains 10 assemblages of 10 samples of an assemblage is produced by

```
> S.total(oA)
[1] 20
```

and similarly, the median of the number of species and the average number of species of the assemblages (or samples) is received as

```
> Si.average(oA)
[1] 5.0 4.7
```

The number of species for each assemblage (samples of an assemblage) is provided by the following command:

```
> Si(oA)
a01 a02 a03 a04 a05 a06 a07 a08 a09 a10
  5   5   4   5   4   5   5   5   4   5
```

The total number of individuals for the 20 species present in the oA date set is provided by

```
> ni.total(oA)
  1    2    3    4    5    6    7    8    9   10   11   12   13   14   15   16
6140 4060 2190 1600  900  752  576  45   15  42  43  12  51  119  16  62
 17  18  19  20
 72  34  97  84
```

The number of individuals for each assemblage (samples of an assemblage) is produced by entering

```
> Ni.total(oA)
 a01  a02  a03  a04  a05  a06  a07  a08  a09  a10
1688 1674 2147 1433 1421 1561 1893 2176 1457 1460
```

Appendix. List of the Procedures

1. `div.Renyi(xv, alpha=1, base=exp(1))`
2. `div.RTS(xv, parameter=1)`
3. `div.ESm.Infinite(xv, subsample=2)`
4. `div.ESm.Finite(xv, subsample=2)`
5. `div.RTS.all(xv)`
6. `plot.div.Renyi(xm, from=0, to=4, step=0.25, xLabel=4, xlab="scale parameter", ylab="Rényi diversity", main="Rényi diversity profiles", lty=1, lwd=2, ...)`
7. `plot.div.RTS(xm, xLabel=0, log="x", xlab="scale parameter", ylab="RTS diversity", main="RTS diversity profiles", lty=1, lwd=2, ...)`
8. `plot.div.ESm(xm, type="infinite", representation="individual", from=2, to=-1, nSteps=10, log="x", xlab="scale parameter", ylab="ES(m) diversity", main="ES(m) diversity profiles", lty=1, lwd=2, ...)`
9. `print.div.Renyi(xm, from=0, to=3, step=0.5)`
10. `S.total(xm)`
11. `Si.average(xm)`
12. `Si(xm)`
13. `ni.total(xm)`
14. `Ni.total(xm)`

References

Crawley M. 2002. Statistical Computing. An Introduction to Data Analysis using S-Plus. Wiley, Chichester.

- Ihaka R & Gentleman R. 1996. R: A Language for Data Analysis and Graphics, *Journal of Computational and Graphical Statistics* 5, 299-314.
- Krause A & Olsen M. 2002. *The Basics of S-PLUS*. Third Edition. Springer, Berlin.
- Patil GP & Taillie C. 1982. Diversity as a concept and its measurement. *Journal of the American Statistical Association* 77, 548-567.
- Southwood TRE & Henderson PA. 2000. *Ecological Methods*. 3rd ed. Blackwell Science, Oxford.
- Tóthmérész B. 1993. DivOrd 1.50: A Program for Diversity Ordering. *Tiscia* 27, 33-44.
- Tóthmérész B. 1995. Comparison of different methods for diversity ordering. *Journal of Vegetation Science* 6, 283-290.
- Tóthmérész B. 1997. *Diversity Orderings*. Scientia Kiadó, Budapest. (Hungarian).
- Tóthmérész B. 1998. On the characterization of scale-dependent diversity. *Abstracta Botanica* 22, 149-156.
- Tóthmérész B. 2002. A diverzitás jellemzésére szolgáló módszerek evolúciója. In: Magyar botanikai kutatások az ezredfordulón. (Salamon-Albert É. ed.), JPTE Növénytani Tanszék és MTA Pécsi Akadémiai Bizottság, Pécs, pp. 607-638. (Evolution of the methods of diversity characterizations.).
- Tóthmérész B & Magura T. 2005. Diversity and scalable diversity characterizations. *European Carabidology 2003. Proceedings of the 11th European Carabidologists' Meeting*. DIAS report, No. 114, Flakkebjerg, Denmark. 353-368.
- Venables WN & Ripley BD. 2002. *Modern Applied Statistics with S*. Fourth Edition. Springer, Berlin.