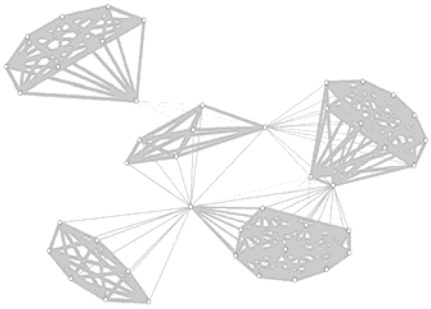
DIETA1

October 21 2007

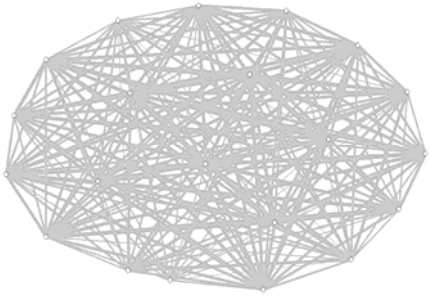
Paulo R. Guimarães

Daitan Labs, Galleria Office, Bloco 4, cj 444, Campinas, SP, Brazil



Paulo R. Guimarães Jr.
and
Márcio S. Araújo

Pós-Graduação em Ecologia, CP 6109
Universidade Estadual de Campinas
Campinas, SP, Brazil 13083-970
+55-19-3521-6279



DIETA1 is a PC-compatible program that calculates two indices of intrapopulation variation in resource use, based on complex network theory.

Downloading this program will place a single .exe file in the folder you save it to. It is recommended you create a folder exclusively for **DIETA1**, save the .exe file to this folder, and place all data text files you want to read in that folder.

Overview of the program:

Double-clicking the icon for the program will open an MS-DOS window, and the program will start. You will go through the following steps (discussed in detail further on in this manual):

1. You will be prompted to enter a data file name. See below for details.
2. You will be asked what type of data is in the file. There are three possible types: (i) proportions; (ii) integers; and (iii) decimal numbers (see below). If the data are not already converted into proportions, the computer will convert the data matrix into a matrix of proportions, calculating the proportion of each individual's diet that falls into a given resource category.
3. If the data are integers, the program goes to step 4, otherwise step 4 is skipped and the program goes directly to step 5.
4. You will be asked if Monte Carlo bootstrap simulations will be run.
    a. In case of a positive answer, the program will need to calculate the proportion of the population diet that falls into each resource category for the simulations. There are two ways of doing this (see below). You will be prompted for which way you prefer. Then you will be asked how many bootstrap replicates (1 to 10,000) you would like to run.
5. You will be asked to enter a value for the 'weight factor' (see below).
6. If Monte Carlo bootstraps will be run (a 'yes' answer in step 4), you will be asked if you want to print the file 'Boot[*filename*].txt'

Program details:

Heading numbers follow the preceding outline of the program.

1) Data entry

DIETA1 reads text files. These can be easily produced in Microsoft Excel by choosing "Save As", and changing the type option to Text (.txt). The program can also read files with other extension, such as .net, but the data must be in ASCII format and the columns must be separated using <space> or <tab>. The end of the line must be a <new line> character. Save the file into the same folder that holds DIETA1.

1.1) Data file format

The program assumes that the data is in the following format: each row represents the diet data for a given individual; each column represents a distinct class of resource, such as food taxon.

Each cell can be: 1. the proportion in the diet (all entries in the row must sum to 1); 2. counts of individual diet items, as in the following hypothetical example – these will be integers; 3. total mass of each food category in an individual's gut (floating point numbers will work here). In the following example, the table is of diet item counts for 4 food types on 5 individuals. Each cell is the number n_{ij} of items of resource j that individual i consumed.

Individual	Food type A	Food type B	Food type C	Food type D
1	88	7	2	3
2	152	3	0	0
3	0	7	3	8
4	0	1	5	10
5	0	0	4	12

IMPORTANT:

1. Do not leave blank spaces where there are no diet items, please fill empty cells with zeros.

2. Do not include the header row naming the resource categories, but **DO** include the first column that identifies individuals. The file for the above data will look like this prior to analysis:

1	88	7	2	3
2	152	3	0	0
3	0	7	3	8
4	0	1	5	10
5	0	0	4	12

2) Data type

The program will prompt you to tell it whether the data is:

1. Already converted into proportions
2. Integers, such as counts of diet items
3. Data with decimal places such as prey mass within each category.

2.1) Converting to proportions

The measures of intrapopulation diet variation rely on mathematical operations on diet proportions, so the first step is transforming the data matrix (**X)** with elements n_{ij} into a matrix of proportions (**P)** with elements p_{ij}.

$$p_{ij} = \frac{n_{ij}}{\sum_{j} n_{ij}}$$

The matrix for the above data will look like this after the conversion to proportions:

1	0.88	0.07	0.02	0.03
2	0.98	0.02	0.00	0.00
3	0.00	0.39	0.17	0.44
4	0.00	0.06	0.31	0.63
5	0.00	0.00	0.25	0.75

This operation is skipped if the data are already in proportion format. The bootstrapping routine (see below) requires the number of prey items, and cannot be calculated with data already in proportion format or data with decimal places (e.g. prey mass).

4) Data analysis

4.1) Population diet proportions

The Monte Carlo bootstrapping routine (see below) depends on the calculation of the population's diet proportions q_j, the proportion of resource j in the population's diet. There are two ways of doing this.

4.1.1) The most straightforward way of calculating q_j is to sum up all prey items falling into category j (sum all i individuals) and then convert it into a proportion by dividing it by the total number of prey items of the total population diet:

$$q_j = \frac{\sum_{i} n_{ij}}{\sum_{i}\sum_{j} n_{ij}}$$

The shortcoming of this approach is that individuals consuming large numbers of prey items will bias the population to look like them. Let's take the above diet matrix as an example. This matrix could represent the food consumption of frogs feeding on terrestrial arthropods. Ants are small, clumped prey that are consumed in large numbers, whereas the other prey categories are large, mobile prey consumed in small numbers:

Individual	Ants	Crickets	Spiders	Beetles
1	88	7	2	3
2	152	3	0	0
3	0	7	3	8
4	0	1	5	10
5	0	0	4	12

The population's proportions would be:

Ants	Crickets	Spiders	Beetles
0.78	0.06	0.05	0.11

which is strongly biased towards ants. In this case, it would maybe be preferable to use prey mass instead of prey number, since there is a correlation between prey size and number in the frogs' diets. Taking the same dataset with dry mass (mg) factored in would yield:

Individual	Ants	Crickets	Spiders	Beetles
1	30.8	58.0	4.1	4.9
2	53.2	24.9	0.0	0.0
3	0.0	58.0	6.1	13.0
4	0.0	8.3	10.2	16.3
5	0.0	0.0	8.2	19.6

and the population's proportions would be:

Ants	Crickets	Spiders	Beetles
0.27	0.47	0.09	0.17

Now, crickets, which are the larger prey consumed would be the most important prey category in the population diet. There is no best way to calculate proportions and the choice between prey number or mass depends on careful considerations on the features of the system being studied, such as the presence of prey number × prey mass correlations. However, bootstrap-based hypothesis testing is restricted to data on the number of prey items, since bootstrapping prey mass data is substantially more complex.

4.1.2) An alternative approach to calculating the population's proportions is to average across each individual's proportions:

$$q_j = \frac{1}{N}\sum_i p_{ij}$$

This will be equivalent to the previous measure when all individuals consume the same number or mass of items. The advantage of this approach is that it weights all individuals equally. Whereas the previous approach to population diet proportions is a measure of resource utilization, this second measure is more a measure of electivity: the proportion of decision-making events that resulted in capturing resource type *j*.

4.2) Calculating the measures of intrapopulation diet variation

We refer readers to Araújo *et al.* (in press) for details on the two proposed indices. Below we give a brief account on them.

4.2.1) The index *E* of interindividual variation

First, we define *O*, a measure of the network overall degree of pairwise overlap:

$$O = \sum w_{ij} \qquad (1)$$

where

$$w_{ij} = 1 - 0.5\sum_{k=1}^{K}\left|p_{ik} - p_{jk}\right| \qquad (2)$$

is a measure of niche pairwise overlap between individuals *i* and *j* (adapted from Schoener 1968); p_{ik} is the frequency of category *k* in individual *i*'s diet, and p_{jk} is the frequency of category *k* in individual *j*'s diet. The pairwise niche overlap ranges from close to 0 (very little overlap) to 1 (total overlap).

Our measure of the degree of interindividual niche variation in the network is defined as:

$$\breve{O} = \frac{O}{n(n-1)/2} = 2O/n(n-1), \qquad (3)$$

in which *n* is the number of nodes in the network. $\breve{O}$ will be 1 if there is no interindividual niche variation and will tend to 0 as variation increases. Our measure $\breve{O}$ is not intuitive, however, as one tends to think that an index to measure interindividual variation will increase its value with increasing interindividual variation and decrease its value otherwise. Therefore, we go a step further and define an index of individual specialization, *E*, as:

$$E = 1 - \breve{O} \qquad (4)$$

Now, in the absence of interindividual niche variation, E will be zero, and will increase towards 1 with the increase of interindividual variation.

4.2.1.1) Variance

A Jackknife estimation of the variance of *E* can be derived using the formalism of *U*-statistics (Arversen 1969). The variance in turn can be used to compare two populations as follows. Given two populations A and B with E_A and E_B measures of interindividual variation, it follows that E_A is approximately Normal with mean θ_A and variance $\sigma^2{}_A$, i.e. $E_A \sim N(\theta_A, \sigma^2{}_A)$ and $E_B \sim N(\theta_B, \sigma^2{}_B)$. Therefore, $(E_A - E_B) \sim N(\theta, \sigma^2{}_A + \sigma^2{}_B)$. Now, we want to test the null hypothesis H_0: $\theta = 0$ vs. H_1: $\theta \neq 0$. One can perform a simple test using the Normal distribution by calculating

$$\frac{|E_A - E_B|}{\sqrt{\hat{\sigma}_A^2 + \hat{\sigma}_B^2}} \qquad (5)$$

and looking up a Normal distribution table for the *P*-value. If the *P*-value is smaller than α (usually 0.05), one rejects H_0; otherwise there is not enough statistical evidence against the null hypothesis.

4.2.2) The index C_{ws} of clustering

We propose a measure of the relative degree of clustering in the niche overlap network:

$$C_{ws} = \frac{(C_w - \breve{O})}{(C_w + \breve{O})} \qquad (6)$$

where C_w is the network weighted clustering coefficient.

In a totally random network (in our case, a network consisting of individuals that sample randomly from the population niche), $C_{ws} \sim 0$, indicating no clustering. If individuals form discrete groups specialized on distinct sets of resources, $C_w > \breve{O}$, $C_{ws} > 0$, and the network is clustered. If $C_w < \breve{O}$, $C_{ws} < 0$, the network degree of clustering is actually lower than what would be expected solely on the overall network density of connections, indicating that the individuals' diets are overdispersed.

4.2.3) Visualizing networks

DIETA1 outputs a matrix with all w_{ij} values, and a binary (zeros and ones) matrix in which ones represent edges whose weights are higher then the average network weight $\breve{O}$ (strong edges). Both matrices can be used to visualize networks, and the binary matrix can be used to identify *w-cliques*, which are defined as groups of nodes in which all nodes are connected to each other by the so-defined strong connections (Araújo et al. in press). These *w-cliques* in turn can be used as a way to visualize clusters in the niche overlap network and determine the affiliation of individuals to the different clusters. Another approach that can be taken following the identification of clusters is to map resources onto those clusters. This allows the identification of the resources associated to each cluster, which can be a useful tool in identifying the resources underlying the resource polymorphism.

The visualization of the networks can be done in commonly used software of network analyses (e.g. Pajek). The matrices generated by DIETA1 can be imported into Pajek, one of the most popular of such programs (Batagelj and Mrvar 1998) that can be downloaded for free at http://vlado.fmf.uni-lj.si/pub/networks/pajek. Users interested in visualizing their networks are encouraged to read carefully the Pajek manual to get details on how to import the files generated by DIETA1 and use them to do the above-mentioned analyses.

4.2.4) Bootstrapping

In case the data are integers, you will be asked if Monte Carlo bootstrap simulations will be run to test the null hypothesis that any observed diet variation arose from individuals sampling stochastically from a shared distribution. In the simulations, each individual is assigned a number of prey items equal to the number of items it was observed eating, and then prey items are randomly assigned to the individual's diet via multinomial sampling from the observed population resource distribution. Next, both *E* and C_{ws} are recalculated for the resulting simulated population. The program can generate up to 10,000 such populations. In the case of *E*, the null hypothesis can be rejected if the empirical value is higher than 95% of the null *E* values. In the case of C_{ws}, which assumes both positive and negative values, the hypothesis test is two-tailed, so that the null hypothesis of $C_w \sim \breve{O}$ can be rejected if C_{ws} is higher than 97.5% of the null positive C_{ws} values or lower than 97.5% of the null negative C_{ws} values. This Monte Carlo procedure assumes that every prey item observed in an individual's diet represents an independent feeding event. We acknowledge, however, that this assumption may be violated for prey that are found in tightly clumped groups.

5) The 'weight factor'

The output matrix containing the all the w_{ij} values or the binary matrix containing zeros and ones can be imported by network-visualization programs, such as Pajek, and used to draw networks. If the matrix of w_{ij} values is used, Pajek will use these values to determine the width of the edges, so that users can have a visual representation of the strength of connections between nodes. Pajek reads real numbers varying from zero to

infinite, working well with values between 5 and 12. If the raw w_{ij} values ranging from 0 to 1 are used, differences in the edge widths are too subtle to be visualized. The function of the weight factor, which multiplies all w_{ij} values in the matrix, is to circumvent this problem, by scaling the w_{ij} values to values more tractable to Pajek. You will be prompted to enter a weight factor varying from 1 to 100; values between 5 and 12 will generally allow a proper network visualization.

6) Output

The program will output five files as follows (*filename* corresponds to the name of the data file entered by the user):

1) one text file named 'Indices[*filename*].txt' containing the empirical *E* and its variance, var(*E*), as well as the empirical C_{ws};

2) one optional text file named '*P*-values[*filename*].txt' containing the non-parametric *P*-values of *E* and C_{ws} generated by the Monte Carlo bootstrap procedure;

3) one optional text file named 'Boot[*filename*].txt' containing the calculated values of $\breve{O}$, *E*, and C_w for every bootstrap simulation;

4) one *.mat file that can be imported into programs of network analysis named '*filename*.mat' containing the matrix of w_{ij} values multiplied by the 'weight factor' previously entered by the user; *.mat files can be easily opened in Microsoft Notepad; the file for the above hypothetical example will look like this:

*Vertices 5

1 "1"

2 "2"

3 "3"

4 "4"

5 "5"

*Matrix

0.000	0.899	0.120	0.113	0.050
0.899	0.000	0.019	0.019	0.000
0.120	0.019	0.000	0.674	0.611
0.113	0.019	0.674	0.000	0.875
0.050	0.000	0.611	0.875	0.000

where the number 5 indicates the number of vertices (nodes) in the network; the numbers below '*Vertices' index the vertices and the characters inside quotes (e.g. "1") are the labels identifying individuals in the first column of the data file; '*Matrix' indicates the square matrix with *i* rows and *j* columns of the pairwise niche overlaps between individuals *i* and *j* (w_{ij}) multiplied by the 'weight factor' entered by the user; in the above example the weight factor was set to 1.0; diagonals are arbitrarily set to zeroes, because pairwise overlaps are only calculated between different individuals.

5) and finally one *.mat file named 'B[*filename*].mat' representing the binary matrix of strong edges, in which cells correspond to either 1 (strong edge present) or zero (strong edge absent). The binary matrix for the above example will look like this:

*Vertices 5

1 "1"

2 "2"

3 "3"

4 "4"

5 "5"

*Matrix

0	1	0	0	0
1	0	0	0	0
0	0	0	1	1
0	0	1	0	1
0	0	1	1	0

Note that in the binary network only the connections with $w_{ij} > \breve{O}$ (0.338 in the example) are maintained.

7) Troubleshooting

This program is brand new and has not been extensively tested. If you have trouble, please write to either prguima.pm@gmail.com, prguima@gmail.com, or msaraujo@gmail.com, giving us as much detail as possible on the problem. If you wish, you might also send the data files so that we can run the program on them and a have a better idea of where the problem is.

Acknowledgements

Daniel I. Bolnick, Eduardo G. Martins, Aluisio Pinheiro, and Sérgio F. dos Reis helped in different parts of the development of the indices proposed here. Pedro Jordano helped with fruitful discussions on network metrics.

References

Araújo, M. S., P. R. Guimarães Jr., R. Svanbäck, A. Pinheiro, P. Guimarães, S. F. dos Reis, and D. I. Bolnick. In press. Network analysis reveals contrasting effects of intraspecific competition on individual versus population diets. Ecology.

Arversen, J. N. 1969. Jackknifing *U*-statistics. Annals of Mathematical Statistics **40**:2076-2100.

Batagelj, V., and A. Mrvar. 1998. Pajek - Program for Large Network Analysis. Connections **21**:47-57.

Schoener, T. W. 1968. The *Anolis* lizards of Bimini: resource partitioning in a complex fauna. Ecology **49**:704-726.