CoOccurrence – a FORTRAN program for species co-occurrence analysis
Version 2.0

Werner Ulrich

Nicolaus Copernicus University in Toruń

Department of Animal Ecology

Gagarina 9, 87-100 Toruń; Poland

e-mail: ulrichw @ uni.torun.pl

Latest update: 10.06.2008

1. Introduction

The study of species co-occurrences has a long tradition in ecology (Weiher and Keddy 1999). Particularly the long lasting and still ongoing discussion around community assembly rules (Diamond 1975, Diamond and Gilpin 1982, Gilpin and Diamond 1982, Connor and Simberloff 1979, 1983, 1984, Gotelli and McCabe, 2002, Ulrich 2004) has inspired the development of statistical tools to infer non random patterns in community assembly (Gotelli and Graves 1996, Gotelli 2000, 2001).

Community assembly is often studied in terms of nestedness (Atmar and Patterson 1986, Wright et al. 1998, Ulrich and Gotelli 2007) and non-random patterns of species co-occurrences (Diamond 1975, Gotelli 2000, 2001) and the respective standard software is widely used: The *Nestedness Temperature Calculator* (Atmar and Patterson 1995), *EcoSim* (Gotelli and Entsminger 2002, and *Nestedness* (Ulrich 2006).

The present program *CoOccurrence* implements two nestedness and three common co-occurrence metrics :The discrepancy metric of Brualdi and Sanderson (1999), NODF (Almeida-Neto et al. (2008), the species combinations score (Pielou and Pielou 1968), the C-score (Stone and Roberts 1990), and the checkerboard score (Gotelli 2000). While most of these metrics are also implemented in *Nestedness* and *EcoSim*, the present program is particularly designed for the study of multiple matrices for instance in null model analysis, for the analysis of the statistical behaviour of certain metrics, and in studies of neutral models.

2. Metrics

Nestedness

CoOccurrence contains the discrepancy metric *BR* and *NODF*. The first counts the number of discrepancies (absences or presence) that must be erased to produce a perfectly nested matrix (Brualdi and Sanderson 1999). This is done for tows and columns separately and the lower value is taken. NODF is a matrix wide count of species associations for rows and columns and separates the importance of columns (NODFcolumns) and rows (NODFrows) for the total degree of nestedness (Almeida-Neto et al. 2008).

Co-occurrence

The present program implements five measures of co-occurrences:

1. The species combinations score (COMBO) screens the columns of the presence absence or abundance matrix for unique species combinations (Pilou and Pielou 1968). Hence, it counts the number of species that always co-occur.

2. The checkerboard score (Gotelli 2000) screens the matrix for checkerboards. These are 2*2 submatrices of the structure $\begin{matrix} 1 & 0 \\ 0 & 1 \end{matrix}$ or $\begin{matrix} 0 & 1 \\ 1 & 0 \end{matrix}$. The score is a simple count of the number of such submatrices.

3. The C-score (Stone and Roberts 1990) is the average number of checkerboards for two species i and j. The score is calculated from

$$C S = \sum_{i,j} (N_i - N_{ij})(N_j - N_{ij})$$

where N_i and N_j are the row totals (numbers of occurrences) of species i and j and N_{ij} is the number of co-occurrences of both species.

4. A correlation test calculates the average Spearman rank order correlation between all species pairs and compares this with the null expectation.

5. The variance test of Schluter (1984) compares the variance of row totals (s^2_{row}) with the mean variance of *N* columns (s^2_{column}). For a random matrix the quotient $s^2_{row}/s^2_{columns}$ should be χ^2 distributed with *N* degrees of freedom.

For comparing matrices of different size and shape the metrics have to be standardized. This is done by dividing the effect size through the expected value.

$$\text{Metric}_{Standardized} = \frac{\text{Obs} - \text{Exp}}{\text{Exp}}$$

```
S  1  2  3  4  5  6  7  8  9  10 11 12
a 0.00 0.00 1.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 1.00 0.00
b 1.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 1.00 1.00 1.00
c 0.00 0.00 1.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00
d 1.00 0.00 1.00 0.00 0.00 0.00 0.00 1.00 1.00 0.00 0.00 0.00
e 1.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 1.00 0.00
f 0.00 1.00 0.00 0.00 0.00 0.00 0.00 1.00 1.00 0.00 1.00 1.00
g 1.00 1.00 1.00 0.00 1.00 0.00 1.00 1.00 0.00 1.00 1.00 0.00
h 0.00 0.00 1.00 1.00 1.00 1.00 0.00 1.00 0.00 0.00 0.00 0.00
i 1.00 0.00 1.00 0.00 0.00 0.00 0.00 1.00 1.00 0.00 0.00 0.00
j 0.00 0.00 0.00 0.00 0.00 1.00 1.00 1.00 0.00 1.00 1.00 0.00
k 0.00 0.00 1.00 1.00 1.00 1.00 1.00 1.00 0.00 0.00 0.00 0.00
l 1.00 1.00 0.00 1.00 1.00 1.00 0.00 0.00 1.00 0.00 1.00 1.00
m 0.00 1.00 1.00 1.00 1.00 1.00 0.00 1.00 1.00 0.00 0.00 0.00
n 0.00 1.00 1.00 0.00 0.00 1.00 1.00 1.00 0.00 1.00 0.00 0.00
o 1.00 1.00 1.00 1.00 1.00 1.00 1.00 1.00 0.00 0.00 1.00 0.00
p 1.00 1.00 1.00 1.00 0.00 1.00 1.00 1.00 1.00 0.00 1.00 0.00
q 0.00 0.00 0.00 1.00 0.00 0.00 1.00 1.00 0.00 0.00 1.00 0.00
r 1.00 1.00 1.00 0.00 1.00 1.00 0.00 1.00 0.00 0.00 0.00 0.00
s 0.00 1.00 0.00 1.00 1.00 1.00 1.00 1.00 1.00 0.00 0.00 0.00
t 0.00 1.00 0.00 1.00 0.00 1.00 1.00 1.00 1.00 0.00 1.00 0.00
```

3. Data structure

CoOccurrence needs one main plain text data file of the following structure. The columns of the matrix are sites, the rows species. Hence the matrix below contains 20 species distributed over 12 sites. The data file has to be a simple ASCII file with data delimitated by one or more spaces. Accepted are either abundance or presences absence data of the integer (In) or real format (Fn.k) The first row contains site names, the first column species names. The file has therefore the same format that is needed for *EcoSim* (Gotelli and Entsminger 2002). The number of species is not limited, the maximum number of sites is about 150.

Alternatively, you may run *CoOccurrence* in a batch. In this case the input files must have a structure as shown above and the driver file contains only the names of the data files. The first line has to be a comment line.

```
* Comment
Test1.txt
Test2.txt
Test3.txt
Test4.txt
Test5.txt
```

4. Program run

First, the program asks for the files names. The default output file names are *CoOccurrence.txt* and *Matrix.txt*. You get the default values after returning *enter*. If you don't give the name of the data file and return *enter* the program expects a batch run and a file name with the data files.

Next, the program asks for the model for randomization. You have seven possibilities: A null model with fixed row and column constraints (input: s) using the sequential swap algorithm (Gotelli 2000, 2001), no constraints (equiprobable row and columns, input: e), or fixed row (input: f) or fixed column (input: c) constraints only. For details of the null models used see Gotelli (2000, 2001). The sequential swap model uses ten times the matrix size (10*rows*columns) single swaps to generate a randomized matrix.

The fifth null model (o) assigns species with a probability according to the number of site occurrences. This model is therefore identical to the Random 1 model of Patterson and Atmar (1986) and Wright et al. (1998).

The sixth null model is a sampling model, where the sites are filled with species using a random sampling of individuals from a common species pool that is structured according to a lognormal species abundance distribution. In this case the program asks for the shape generating parameter *a* of the lognormal model. This has the typical form $[S=S_0Exp(-a(R-R_0)^2)]$ and is computed using a normally distributed random number on a log scale. Preston's canonical lognormal has the parameter value a = 0.2 (May 1975). In the case of the lognormal null model column (site) species numbers are fixed to the observed values (fixed column constraint).

The seventh null model resamples rows according to the observed species abundance distribution calculated from row totals of abundance. This last null model, of course, needs abundance data as input.

Next the program asks for the number of randomizations to compute the null model means and standard deviations, as well as upper and lower 95% confidence limits. In most cases 100 such randomizations will be enough.

5. The output files

CoOccurrence produces two output files. The first file (*Cooc.txt*) contains basic information about the matrix and the measurements. First it gives species

```
Cooc.txt - Notatnik

Plik  Edycja  Format  Widok  Pomoc
> File:  test.txt        Species:    20 Sites:    24 MatFill: 0.82 Occ:        394 Model: a
Index         File                In    SimIn   StdDevIn       Z-In    StIndex     SkewIn      L95%CIn     U95%CIn
Checkerboard test.txt              0     0.00       0.00       0.00     0.0000       0.00        0.00        0.00
Combinations test.txt              3     1.63       1.35       1.01     0.0050       1.04        0.00        6.00
C-Score      test.txt          10.56    11.47       1.41      -0.64     0.0000       0.36        9.18       14.43
BR           test.txt             50    52.18       4.66      -0.47    -0.0055       0.16       44.00       62.00
Correlation  test.txt           0.21     0.26       0.03      -1.65    -0.0665       0.20        0.20        0.32
Variancetest test.txt           2.77     2.38       0.21       1.84     0.0103       0.34        2.00        2.85
NODF         test.txt          77.25    77.09       2.08       0.07     0.0067      -0.72       73.02       80.38
NODFcolumns  test.txt          75.64    74.21       3.00       0.48     0.0554      -0.79       66.38       79.32
NODFrows     test.txt          79.58    81.28       2.52      -0.67    -0.0906      -2.63       75.93       84.87
```

and site numbers, matrix fill, the total number of occurrences, and the null model algorithm. Then observed metric values, simulated values, the respective standard deviations, Z-scores, standardized values, skews of the null model distribution, and upper and lower 95% confidence limits of this distribution are provided. The second file *Matrix.txt* contains the packed original data matrix and the last randomized packed matrix. The examples above show both files. Due to the fixed output format (I5) the matrix file may

have problems to proper print the values in the case of relative abundances as input values.

6. Null model properties

Different null models point to different degrees of randomness in the data. Matrix size and fill might also influence the results. The fixed row - fixed column model (sequential swap) appears to be the most conservative model while retaining the basic matrix

```
Matrix.txt - Notatnik

Plik  Edycja  Format  Widok  Pomoc
Packed matrix: test.txt

              1.0 2.0 3.0 4.0 5.0 6.0 7.0

1               1   1   1   1   1   1   1           7
2               1   1   1   1   1   0   0           5
3               1   1   1   1   1   0   0           5
4               1   1   1   1   0   1   0           5
5               1   1   1   0   0   0   0           3
6               1   1   0   0   1   0   0           3
7               1   1   0   1   0   0   0           3
8               1   0   1   0   0   0   0           2
9               0   1   0   0   0   0   0           1
10              1   0   0   0   0   0   0           1

                9   8   6   5   4   2   1

Last packed randomized matrix: test.txt          No. of Swaps:   25000

              1.0 2.0 3.0 4.0 5.0 6.0 7.0

1               1   1   1   1   1   1   1           7
2               1   1   1   1   1   0   0           5
3               1   1   1   1   1   0   0           5
4               1   1   1   1   1   0   0           5
5               1   1   1   0   0   0   0           3
6               1   1   0   1   0   0   0           3
7               1   0   1   0   0   1   0           3
8               1   1   0   0   0   0   0           2
9               1   0   0   0   0   0   0           1
10              0   1   0   0   0   0   0           1

                9   8   6   5   4   2   1
```

CoOccurrence

properties during simulation. However, standard deviations of this model are often smaller than those of the other models increasing thus the probability of pointing to non-randomness. The equiprobable model, in turn, is the most liberal. Fixed row equiprobable column and fixed column equiprobable row constraints behave quite similar, which is not surprising due to the theoretical equivalence of matrix rows and columns. The lognormal sampling model appears to be robust against matrix fill but not against matrix size. Theoretically this model should behave similar to the fixed row - fixed column model.

There is no consensus about the question which null model is best suited to account for passive sampling. Because the lognormal and the resample models are passive sampling models they should theoretically be preferred. However, we seldom know the underlying abundance distribution from which the individuals are drawn. Recent evidence (Ulrich and Gotelli 2007a, b) points to the fixed row - fixed column model to account best for passive sampling. However, there is still no comparative study. In my view the lognormal model and the fixed row - fixed column should be preferred.

7. Citing *CoOccurrence*

CoOccurrence is freeware but nevertheless if you use *CoOccurrence* in scientific work you should cite *CoOccurrence* as follows:
Ulrich W. 2006. CoOccurrence – a FORTRAN program for species co-occurrence analysis.
www.uni.torun.pl/~ulrichw

8. System requirements

CoOccurrence is written in FORTRAN 95 and runs under Windows 9.x, XP, and Vista. Computation abilities are only limited by the computer's memory.

10. References

Almeida-Neto M., Guimarães P., Guimarães Jr P. R., Loyola R. D., Ulrich W. 2008. A consistent metric for nestedness analysis in ecological systems: reconciling concept and quantification. Oikos: in press.

Atmar W., Patterson B. D. 1993. The measure of order and disorder in the distribution of species in fragmented habitat. Oecologia 96: 373-382.

Atmar W., Patterson B. D. 1995. The nestedness temperature calculator: a visual basic program, including 294 presence absence matrices. AICS Research Incorporate and The Field Museum.

Brualdi R. A., Sanderson J. G. 1999. Nested species subsets, gaps, and discrepancy. Oecologia 119: 256-264.

Connor E. F., Simberloff D. 1979. The assembly of species communities: chance or competition. Ecology 60: 1132-1140.

Connor E. F., Simberloff D. 1983. Interspecific competition and species co-occurrence patterns on islands: null models and the evaluation of evidence. Oikos 41: 455-465.

Connor E. F., Simberloff D. 1984. Neutral models of species co-occurrence' patterns. In: Strong D. R., Simberloff D., Abele L. G., Thistle A. B., (eds.), Ecological Communities: Conceptual Issues and the Evidence, Princeton Univ. Press, pp. 316-331.

Diamond J. M. 1975. Assembly of species communities. In: Cody, M. L., Diamond, J. M. (eds.), Ecology and Evolution of Communities, Harvard Press, Cambridge, pp. 342-444.

Diamond J. M., Gilpin M. E. 1982. Examination of the 'null' model of Connor and Simberloff for species co-occurrences on islands. Oecologia 52: 64-72.

Gilpin M. E., Diamond J. M. 1982. Factors contributing to non-randomness in species co-occurrences on islands. Oecologia 52: 75-84.

Gotelli N. J. 2000. Null model analysis of species co-occurrence patterns. Ecology 81: 2606-2621.

Gotelli N. J. 2001. Research frontiers in null model analysis. Global Ecology and Biogeography Letters 10: 337-343.

Gotelli N. J., McCabe D. 2002. Species co-occurrence: a meta-analysis of J. M. Diamond's assembly rules model. Ecology 83: 2091-2096.

Gotelli N.J., Entsminger G.L. 2002. EcoSim: Null models software for ecology. Version 7. - Acquired Intelligence Inc. & Kesey-Bear. Burlington, VT 05465. http://homepages.together.net/~gentsmin/ecosim.htm.

Gotelli N. J., Graves G. R. 1996. Null Models in Ecology. Smithsonian Institution. Press, Washington D.C.

McAbendroth L., Foggo A., Rundle D., Bilton D. T. 2005. Unravelling nestedness and spatial pattern in pond assemblages. J. Anim. Ecol. 74: 41-49.

Patterson B. D., Atmar W. . 1986. Nested subsets and the structure of insular mammalian faunas and archipelagos. In: Heaney, L. R., Patterson B. D. (eds) Island biogeography of mammals, Academic Press, London, pp. 65-82.

Pielou D. P., Pielou E. C. 1968. Association among species of infrequent occurrence: the insect and spider fauna of *Polyporus betulinus* (Bulliard) Fries. - Journal of Theoretical Biology 21: 202-216.

Schluter D. 1984 A variance test for detecting species associations, with some example applications. Ecology 65: 998-1005.

Stone L. and Roberts A. 1990. The checkerboard score and species distributions. - Oecologia 85: 74-79.

Ulrich W. 2004. Species co-occurrences and neutral models: reassessing J. M. Diamond's assembly rules. Oikos 107: 603-609.

Ulrich W. 2006. Nestedness—a Fortran progam for program for calculating ecological matrix temperatures. Online available at http//www.uni.torun.pl/~ulrich.

Ulrich W., Gotelli N. J. 2007a. Null model analysis of species nestedness patterns. Ecology: 88: 1824-1831.

Ulrich W., Gotelli N. J. 2007b. Disentangling community patterns of nestedness and species co-occurrence. Oikos 116: 2053-2061.

Weiher E., Keddy P. A. (eds.) 1999. Ecological Assembly Rules: Perspectives, Advances, Retreats. Cambridge Univ. Press, New York.

Wright D. H., Patterson B. D., Mikkelson G. M. , Cutler A., Atmar. W. 1998. A comparative analysis of nested subset patterns of species composition. Oecologia 113: 1-20.