TRGRPS—An interactive algorithm for group recognition with an example from Spartinetea

1 AUTHOR:

Laszlo Orloci
Western University, London, Canada

150 PUBLICATIONS   **1,911** CITATIONS

TRGRPS – AN INTERACTIVE ALGORITHM FOR GROUP RECOGNITON WITH AN EXAMPLE FROM SPARTINETEA*

László ORLÓCI**

Department of Plant Sciences, University of Western Ontario, London, Canada N6A 5B7

Keywords: Algorithm, Clustering, Groups, Hypotheses, Numerical phytosociology, Testing

Introduction

I know that group recognition is a subject of great interest to this group. I realize, of course, that there are preferences for different methods, according to different objectives.

The algorithms used by phytosociologists appear to be of two generic types:

1. The first focuses on homogeneity within the groups. Groups are sought in which the member objects are as similar as possible in terms of given characters. The clustering algorithm, which I described in an earlier paper (Orlóci 1967) with group homogeneity measured in terms of the sum of squares, represents an example in this connection.

2. The second type shifts focus from homogeneity within the groups to discontinuities in the entire sample. The clustering algorithm TRGRPS is of this type. It is formulated around the definition that one group of relevés is discontinuous with another group if none of its neighbourhoods, with given radius, overlap with any of the neighbourhoods with the same radius in the other. The definition and related concepts are explained in the text.

The algorithm is programed in BASIC under file name TRGRPS.BAS. When it is hypothesized that discontinuities subdivide the sample into a given number of groups and the hypothesis is presented to the program with data, TRGRPS. BAS performs the necessary tests automatically and, if groups are found, it produces group descriptors.

Relevant Concepts

Sample space

It will be assumed that there are N relevés in the sample labelled by symbols A, B, C, etc. It will be assumed also that in the associated sample space the relevés are represented by points in the line,

The relative placements of the points to one another are important. The assumption of such a simple unidimensional arrangement is not necessary, but convenient, because it can simplify matters when the concepts are explained without detracting from the generality of the conclusions.

Nearest neighbour relevés

The length of the line segment, connecting any two points in sample space, is proportional to differences in the species composition of the relevés. The line segments represent distances in p-dimensions. Each dimension in p-space corresponds to a different species in the sample. It is noted that differences need not be measured in terms of a distance function. Any symmetric measure of dissimilarity is permissible. The point nearest to another point in terms of any such distance or dissimilarity is said to be that point's nearest neighbour. In the example shown, A is nearest neighbour to B, E to F, etc.

The neighbourhood of a relevé

Each relevé may or may not have other relevés in its neighbourhood within a given radius (r) for which the neighbourhood is defined. At $r = d(B, C)$, where $d(B, C)$ is the distance of point B from C, the following neighbourhoods of relevés exist:

{AB}   for relevé A

{ABC}   for relevé B

{BC}   for relevé C

{DE}   for relevé D and E

{F}   for relevé F

It is noted that {AB} and {BC} are overlapping neighbourhoods; *i.e.* relevé B is possessed in common by both. {ABC}, {DE} and {F} are non-overlapping. Neighbourhoods that are non-overlapping are said to be discontinuous.

Clustering by Neighbourhoods

Given $r = 0$ as the initial state of the neighbourhood radius in a clustering procedure, as many discontinuous neughbourhoods can be recognized in the sample A, B, C, D, E, F as there are non-identical relevés,

  {A}, {B}, {C}, {D}, {E}, {F}

Now, let r be incremented by a constant, small quantity in steps until the condition $d(A, B) \leqslant r < d(B, C)$ is satisfied. Here, $d(A, B)$ represents the smallest distance other than zero in the sample and $d(B, C)$ the second smallest distance. At such a radius,

  {AB}, {C}, {DE}, {F}

are recognized as discontinuous neighbourhoods. If r is further incremented, so that it will satisfy the condition $d(B, C) \leqslant r < d(E, F)$, where $d(E, F)$ is the third smallest distance in the sample, the discontinuous neighbourhoods are

  {ABC}, {DE}, {F}

Further increments in r, until $d(E, F) \leqslant r < d(C, D)$ is satisfied, yield the discontinuous neighbourhoods,

  {ABC}, {DEF}

Finally, when $r \geqslant d(C, D)$, only a single group

  {ABCDEF}

is recognized. All neighbourhoods at a given r under the given constraints satisfy the original definition of discontinuous groups stated in the introduction. In other words, none of the neighbourhoods in any one group overlaps with any neighbourhood in the other. It is evident from what has been said that the procedure of clustering by neighbourhoods uses single link fusions similarly to Sneath's (1957) method of single linkage clustering.

The Algorithm TRGRPS

TRGRPS incorporates clustering by neighbourhoods and various other devices which enable the user to test hypotheses about the existence of discontinuous groups of revelés. The algorithm regards a group of relevés as discontinuous if none of its neighbourhoods overlap with any neighbourhood in other groups. The flow pattern of operations and decisions is described below:

a. Present data; specify a hypothesis for the number of groups R; compute distances between relevés.

b. Specify initial value of neighbourhood radius and its increments.

c. Cluster relevés and count the number of groups. Designate this number by k.

d. If $k = R$ then accept the hypothesis.
   If $k < R$ then modify the hypothesis to less than R groups and continue with step b.
   If $k > R$ then increment neighbourhood radius and continue at c.

I realize that based on sampling considerations tests of the kind as in this algorithm should preferably be probalistic. However, the algorithm as I present it here does not incorporate probabilistic considerations. I plan to discuss these aspects in a separate eassay presently in the preparatory stage.

A Related Algorithm

TRGRPS evolved through modifications of the original Jancey (1974) algorithm which I abstracted (Orlóci 1975) under the name TRGRUP. I stressed a nearest neighbour property of the groups in the original descriptions of TRGRUP, although I incorporated the idea of clustering by neighbourhoods in the program TRGRUP.BAS. The present paper corrects this anomaly by recognizing that although the condition that every member in a discontinuous group must have its nearest neighbour within the group is necessary it is by no means sufficient because when this condition is satisfied neighbourhoods ean overlap between the groups. The groups with contents

  {AB}, {C}, {DEF}

represent an eclatant example in this connection. In these, the nearest neighbour property is satisfied; every member in each group has its nearest neighbour within its own group. Yet two groups, {AB} and {C}, appear continuous because the neighbourhoods of relevés A, B and C overlap at the implied radius $d(E, F)$.

TRGRPS.BAS differs from TRGRUP.BAS by the speci-

fic constraint that none of the relevés are allowed to remain unclustered when the hypothesis is true. This means that for the hypothesis of R groups to be accepted in TRGRPS, the sample must divide into exactly R groups, and not just R noda plus a number of points that float unclustered among the groups. Programs TRGRUP.BAS and TRGRPS.BAS yield identical results when the group structure is strong in the sample.

Running the Program TRGRPS.BAS

The program uses the BASIC programing language, written in conversational mode for processing from a teletype console. After the RUN command, the program requests values for the following parameters:

Number of species (P)
Number of relevés (N)
Minimum group size (S)
Number of groups (R)
Initial value of neighbourhood radius (C)
Increment in neighbourhood radius (D)

Values for P and N are requested first, and after P and N, values for S and $R.S$ can be set to one if no hypothesis specifies *a priori* minimum group size. An initial value for R derives from the hypothesis. Alternate values, small or large, may be assigned to R in successive runs until a solution is accepted. It is noted that running time and R are inversely related.

Data input is from DISK file RAWD which contains P sets of N numbers. The data in computer core are stored in array **D** which has P rows and N columns. Data input is followed by computation of relevé distances. The program uses the Euclidean distance. However, if so desired, any symmetric function of a relevé dissimilarity may replace the Euclidean distance. The distances are stored in array **A** which has N^2 elements.

After the elements of **A** are generated, a search begins for the smallest value. The message initial value of linkage parameter is printed followed by display of the smallest distance value. A request to confirm this value is displayed next. The user responds by typing either the same value, or another value, which then serves as the initial value for C. The program determines a maximum of 50 values, each representing a difference between pairs of the smallest distance values in **A**, ordered according to size, to assist in the choice of a value for the increment D of C. Any value can be chosen. However,

it must be kept in mind that a large value of D shortens the computation time required, while a small value increases precision. A zero is not permissible for D.

Once the program gets underway, the sample is searched for discontinuous groups. When exactly R groups are found, the group descriptors are printed and the execution of the program stops. If R groups are not found a new search begins for $R-1$ groups.

Printout of Results

Examples are given in the next section. When a solution is found under a current hypothesis for R, as modified by parameters S, C, D, the following descriptors are printed:
a. A value of C representing the current size of the neighbourhood radius r at the last fusion. This can be used as a global measure for coherence in the groups.
b. A label for each group.
c. Relevé labels in each group.
Results of repeated runs at different R can be used to trace a hierarchical group structure of relevés in the sample.

Results from Spartinetea

The results presented here derive from the data in Table 1 representing saltmarsh relevé section f (Van der Maarel, Orlóci & Pignatti 1976, Lausi, Beeftink & Kortekaas 1976). Inspection of the raw data reveals that an initial hypothesis of 5 relevé groups may be tenable with S set equal to 5. When TRGRPS is run a solution is indeed found at this value of R (see Table 2).

Table 1. Sample of 50 relevés from Spartinetea. The Braun-Blanquet scale is replaced by the following scores in the table: $1 = +$, $2 = +$, $3 = 1$, $5 = 2$, $7 = 3$, $8 = 4$, $9 = 5$. Relevé order follows syntaxonomic affiliation.

Species	Relevé
Triglochin maritima	000006020000000000000000000000000000000000000002000000
Agrostis stolonifera	000006000000002000000000000006000000000000000000000000
Spartina alterniflora	000000000000599900000000000000000000000000090000000000
Spartina maritima	7988953331100000000000000000000000000000000000000000000
Spartina townsendii	000000000000000000000078958553111133353135525525555555
Spartina patens	0000000000000000999900000000000000000000000000000000000
Phragmites australis	00000000000000002025000000000000000000000000000000000
Puccinellia maritima	000008225300023000000000088775005352202103005000000
Scirpus maritimus	00000000000050000000000000000000000000000000500000
Juncus gerardi	000000000000000000200000000000000000000000000000000
Juncus maritimus	0000000000000655200000000000000000000000000000000
Atriplex hastata	000000000000000000000000000000000000000000008797900000
Halimione portulacoides	0000023332000000000220000220220020022530010035500
Salicornia ramosa	000007768900000000000000000500000004000000000000000000
Salicornia fruticosa	000000000000000000620000000000000000000000000000000
Arthrocnemum glaucum	00000000000000000200000000000000000000000000000000
Salicornia europaea	00002050220000000000000000001555388773702523000500033533
Suaeda maritima	0002655200000000000000200022212232227788739002505555555
Spergularia media	00000000000000000000000000300000000000000000000000
Limonium bellidifolium	00000000000000000000001020000000000000000000000000000
Limonium vulgare	003100000000020015220000000000002000000000000000
Plantago crassifolia	0000000000000000200000000000000000000000000000000
Plantago maritima	0000000000000000000000000000000000000000000002000000
Aster squamatus	000000000000000200000000000000000000000000000000000
Aster tripolium	0030200323002250000200200032233232303523033233053353
Inula crithmoides	00000000000000000220100000000000000000000000000000
Sonchus arvensis	000000000000000022000000000000000000000000000000000

```
TRGRPS          16:15       22-Nov-75
PROGRAM NAME -- TRGRPS
????????????????????    ........
NUMBER OF SPECIES ?27
NUMBER OF QUADRATS ?50
MINIMUM GROUP SIZE REQUIRED ?5
NUMBER OF GROUPS REQUIRED ?5
INITIAL VALUE OF NEIGHBOURHOOD RADIUS : 1
CONFIRM BY TYPING  1 ELSE TYPE DESIRED VALUE: ?1
SELECT A VALUE FROM THE LIST BELOW
TO SERVE AS THE INCREMENT IN LINKAGE PARAMETER.
  1  2
TYPE SELECTED VALUE:
  ?2
  5 --GROUPS
MAXIMUM NEIGHBOURHOOD RADIUS 67
GROUP 1
  1   2   3   5   4
GROUP 2
  6   7   8  10   9
GROUP 3
 11  12  14  15  13
GROUP 4
 16  19  20  17  18
GROUP 5
 21  22  25  24  23  27  28  29  30  34  31  32  35  33  36  37  49
 50  46  47  48  38  40  39  26  41  42  43  45  44

TIME:  177.74 SECS.
```

The correspondence of the groups at the 5—group level in
TRGRPS and the units in conventional syntaxonomy is most
remarkable. The first group, and the second, are subassocia-
tions (*typicum* and *salicornietosum*) of *Spartinetum mariti-
mae*. The third group represents *Spartinetum alterniflorae*,
the fourth *Spartinetum patensis*, and the fifth *Spartinetum
townsendii*.

A hierarchical pattern of fusions, derived by repeated runs
of TRGRPS. BAS with different values of *R*, is traced in Fig.
1. Interestingly, the first fusion does not occur between the
subassociations of *Spartinetum maritimae*, notwithstanding
expectations, but between *Spartinetum maritimae salicor-
nietosum* and *Spartinetum townsendii*. With these the other
subassociation, *typicum*, fuses only in the second step, fol-
lowed by fusions with *Spartinetum alterniflorae*, and lastly,
Spartinetum patensis.

Discussion

It is evident intuitively that the relevés in a discrete group
must exhibit a density pattern that permits their recognition as
members in a coherent group. Coherence, the discreteness of
groups, is associated with criteria of non-overlapping neigh-
bourhoods in algorithm TRGRPS. When applied to a sample
of *Spartinetea*, the results of a comparison of the groups
suggest that the method of group recognition in TRGRPS
resembles the method of group recognition in traditional syn-
taxonomy.

The algorithm TRGRPS should be a useful tool in the
hands of phytosociologists, interested in the construction of
syntaxonomic systems and in the identification of a parsime-
nious path through neighbouring groups for which hierarchi-
cal relationships are to be shown.

Summary

TRGRPS can detect groups or signal if discrete groups can-
not be found in a sample. The present paper elaborates on the
concepts, describes the algorithm and provides illustrations
from syntaxonomy. A computer program (TRGRPS.BAS) is
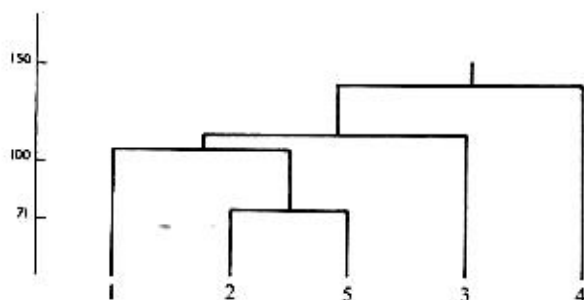available from the author upon request.

Figure 1. Hierarchical structure in Spartinetea (Table 1) at the level
of 5 groups. Vertical scale indicates values of the neighbourhood
radius. Symbols at the base designate syntaxonomic units recogni-
zed in TRGRPS. Legend to symbols: 1 - *Spartinetum maritimae typi-
cum*, 2 - *Spartinetum maritimae salicornietosum*, 5 - *Spartinetum
townsendii*, 3 - *Spartinetum alternifoliae*, 4 - *Spartinetum batensis*.
Numbers 1 to 5 identify groups in the computer printouts at the
5 groups level in Table 2.

References

Jancey, R.C. 1974. Algorithm for detection of
   discontinuities in data sets. Vegetatio 29: 131-133.
Lausi, D., W.G. Beeftink & W.M. Kortekaas, 1976. Survey
   of salt marsh relevés selected for the working group for
   Data Processing in Phytosociology, to be published
   in Vegetatio.
Maarel, E. van der, L. Orlóci & S. Pignatti 1976. Data
   Processing in Phytosociology: retrospect and
   anticipation. Vegetatio 32: 65-72.
Orlóci, L. 1967. An agglomerative method to classify plant
   communities. J. Ecol. 55: 193-205.
Orlóci, L. 1975. Multivariate Analysis in Vegetation
   Research. Dr. W. Junk b.v. Publishers, The Hague, 276
   pp.
Sneath, P.H.A. 1957. The application of computers to
   taxonomy. J. Gen. Microbiol. 17: 201-226.