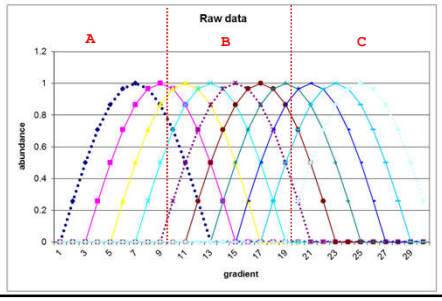
Indicator Species Analysis

- **What it does:** Analysis aims to identify what species are "indicators" of groups of samples or experimental treatment.
- Typical application involves a priori grouping of samples.



What is an indicator species?

- **Indicator species** – a species that is characteristic of a group of samples.
  - A perfect indicator species will only occur in one group.
  - Generalist species will occur across multiple groups and therefore not a good indicator.
  - Rare species are not sampled often, also not a good indicator.
  - Indicator values range from 0 to1 but may also be expressed as a percentage.
  - Each species is assigned an indicator value for each group.
  - Significance of values assessed through permutations.

Indicator Species Analysis

- **Appropriate questions**
  - A priori grouping of samples
    - Cluster analysis → find indicators for clusters
    - Disturbance or other discrete environmental variable → what species are indicators for pre/post disturbance?
    - Distinct habitat types → what species are indicators of each habitat?
    - Diet analysis → what diet items are indicators for a species or group from a particular habitat?

Indicator Species Analysis

- **Indicator Values**
  - Calculated for each species-group combination
    - X_{kj}=mean abundance of species j in group k
    - n_k=number of samples in group k
    - a_{ijk}=abundance of species j in sample i of group k
    - RA_{kj}=relative abundance of species j in group k

$$X_{kj} = \frac{\sum_{i=1}^{n_k} a_{ijk}}{n_k}$$

$$RA_{kj} = \frac{x_{kj}}{\sum_{k=1}^{g} x_{kj}}$$

Indicator Species Analysis

- RF_{kj}= Relative frequency of species j in group k
- B_{ijk}=presence/absence of species j in sample i of group k

$$RF_{kj} = \frac{\sum_{i=1}^{n_k} b_{ijk}}{n_k}$$

- **Indicator Value** – product of the relative frequency and relative abundance.
  - A measure of exclusiveness for a species in a group.
  - Typically, only the highest indicator value for a species across groups is used.

$$IV_{kj} = 100(RA_{kj} * RF_{kj})$$

Significance of Indicator Values - Monte Carlo Tests

- Test the significance of indicator values through permutation test
  - randomly assign samples to groups, generate a distribution of indicator values.
  - Null – observed IV will not differ from those generated through permutations.

Indicator Species Analysis

- Some other uses
  - Determine stopping point for clustering
    - If the goal is to have clusters that describe distinct communities, groups should contain indicator species
    - One could define optimal clustering as when you maximize indicator values
- Has a few things in common with Polar (Bray-Curtis) ordination. Can be thought of as ordinating species along a categorical environmental gradient or among experimental treatments.

Indicator species analysis functions

- indval function in labdsv package
  - Provide community matrix and grouping variable
- multipatt function in indicspecies package
  - Provide community matrix and grouping variable
  - Variety of other options
  - To do the traditional indicator species analysis
    - Options: duleg=TRUE and func="IndVal.g"

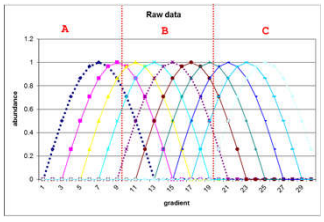
Indicator Species Analysis

- Input – raw (not transformed) community data and grouping variable. Analysis assumes count data.
- Code
  - `Indicator_species<-multipatt(community,clust,duleg=TRUE,func="IndVal.g")`
- Technique originally described in:
  - Dufrene, M. and P. Legendre. 1997. Species assemblages and indicator species: The need for a flexible asymmetrical approach. *Ecological Monographs* 67: 345-366.

Output

- Relative frequency of occurrence for each species in each cluster (RF_{jk})
- Relative abundance of each species in each cluster (RA_{jk})
- Indicator values 100 (RF_{jk}* RA_{jk})
  - Note that multipatt function returns the square roots, square them if you are reporting them as traditional indicator values.
- Significance of indicator values from permutations
- Group that each species is associated with (highest value)

```
ind_species<-multipatt(community,envdata$spatial)
$B
              a   b   c
sp1  1.0000000 0.2 0.0
sp2  0.7777778 0.4 0.0
sp3  0.5555556 0.6 0.0
sp4  0.3333333 0.8 0.0
sp5  0.1111111 1.0 0.0
sp6  0.0000000 0.9 0.2
sp7  0.0000000 0.7 0.4
sp8  0.0000000 0.5 0.6
sp9  0.0000000 0.3 0.8
sp10 0.0000000 0.1 1.0

$A
               a           b           c
sp1  0.90918235 0.09081765 0.00000000
sp2  0.71494203 0.28505797 0.00000000
sp3  0.46021386 0.53978614 0.00000000
sp4  0.20993468 0.79006532 0.00000000
sp5  0.03771739 0.96228261 0.00000000
sp6  0.00000000 0.90009958 0.09990042
sp7  0.00000000 0.69299280 0.30700720
sp8  0.00000000 0.43417375 0.56582625
sp9  0.00000000 0.19299280 0.80700720
sp10 0.00000000 0.03407417 0.96592583
```




```
$sign
      s.a s.b s.c index      stat    p.value
sp1     1   0   0     1 0.9535105 0.00019996
sp2     1   0   0     1 0.7456983 0.00319936
sp3     0   1   0     2 0.5690973 0.13957209
sp4     0   1   0     2 0.7950171 0.00079984
sp5     0   1   0     2 0.9809600 0.00019996
sp6     0   1   0     2 0.9000498 0.00019996
sp7     0   1   0     2 0.6964876 0.00859828
sp8     0   0   1     3 0.5826626 0.11377724
sp9     0   0   1     3 0.8034960 0.00119976
sp10    0   0   1     3 0.9828153 0.00019996
```

multipatt options

- max.order - calculates indicator values for combinations of groups. Set to 2 for pairs of groups where the default is to use all pairwise (can be modified)
- min.order – set the minimum order for groups (set to 2 to look at combinations only)

```
$A
               a           b           c         a+b         a+c         b+c
sp1  0.90918235 0.09081765 0.00000000 1.00000000 0.90918235 0.09081765
sp2  0.71494203 0.28505797 0.00000000 1.00000000 0.71494203 0.28505797
sp3  0.46021386 0.53978614 0.00000000 1.00000000 0.46021386 0.53978614
...

$sign
      s.a s.b s.c index      stat    p.value
sp1     1   0   0     1 0.9535105 0.00019996
sp2     1   1   0     4 0.7608859 0.00539892
sp3     1   1   0     4 0.7608859 0.00539892
sp4     0   1   0     2 0.7950171 0.00159968
sp5     0   1   0     2 0.9809600 0.00019996
sp6     0   1   0     2 0.9000498 0.00019996
sp7     0   1   1     6 0.7416198 0.00979804
sp8     0   1   1     6 0.7416198 0.01139772
sp9     0   0   1     3 0.8034960 0.00059988
sp10    0   0   1     3 0.9828153 0.00019996
```

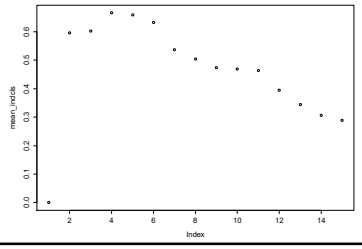
multipatt options

- control option lets you change how permutations are done.
  - Number of permutations
  - `control=how(nperm=5000)`
  - Perform permutations within blocks
  - `control=how(nperm=5000,blocks=envdata$spatial)`

```
$sign
      s.a s.b s.c index      stat p.value
sp1     1   0   0     1 0.9535105       1
sp2     1   0   0     1 0.7456983       1
sp3     0   1   0     2 0.5690973       1
sp4     0   1   0     2 0.7950171       1
sp5     0   1   0     2 0.9809600       1
sp6     0   1   0     2 0.9000498       1
sp7     0   1   0     2 0.6964876       1
sp8     0   0   1     3 0.5826626       1
sp9     0   0   1     3 0.8034960       1
sp10    0   0   1     3 0.9828153       1
```

Use indicator species analysis to determine optimal grouping of cluster analysis

```
distance<-vegdist(community,method="bray")
cluster<-agnes(distance, diss=TRUE, method="average", keep.diss=TRUE)
mean_indcls <- numeric(15)
for (k_cuts in 2:15)
      {
      cut_tree<-cutree(cluster,k=k_cuts)
      ind_species<-multipatt(community,cut_tree,duleg=TRUE)
      mean_indcls[k_cuts]<-mean(ind_species$sign$stat)
      }
# plot the number of clusters (x) vs. the mean indicator value for all species (y)
plot(mean_indcls)
```



Two Way Indicator Species (TWINSPAN)

- Hill, M. O. 1979. TWINSPAN—A FORTRAN program for arranging multivariate data in an ordered two-way table by classification of the individuals and attributes. Ecology and Systematics. Cornell University Press, Ithaca, NY.
- Seeks to define groups and the dominant species within.
- Not intended to find indicator species in pre-defined groups.
- Input community data, the analysis performs it's own clustering based on the assumption that indicator species are present
- Analysis is based on an initial CA
  - Inherits issues associated with CA (arch)
  - Assumption is there is one strong gradient.
  - Samples are divided along CA axes. As with DCA, cut points are arbitrary and criteria vague.

Two Way Indicator Species (TWINSPAN)

- Method for making divisions is complex, not well described in literature describing the technique.
  - At each iteration
    - Division of CA ordination is made
    - Minimal subset of species necessary to reproduce the ordination = indicator species
    - The division is refined by adjusting the weighting of species/samples
- There is no R code for doing this, the author of the package that does regular indicator species analysis considers TWINSPAN to be fundamentally "broken".
- TWINSPAN program:
  - http://www.canodraw.com/wintwins.htm

Sample Unit #

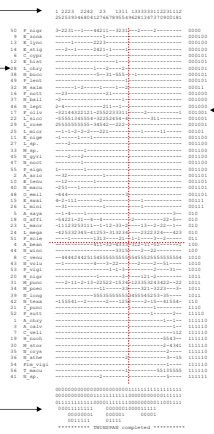
Species name and number

Interior numbers on the analysis are abundance “classes” for each species in each sample.

Red lines represent the first division of samples and species.

Cluster division for species

Cluster divisions for samples

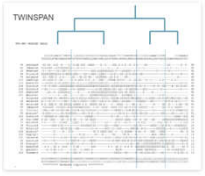


- If you want to run TWINSPAN…there is a twinspanR package but it has to be installed manually…

TWINSPAN in R

15 Replies

TWINSPAN is perhaps one of the most popular clustering methods (at least among vegetation ecologists), which is not implemented in R. R-sig-eco forum has several posts (mostly from Jari Oksanen and Dave Roberts) on the topic of TWINSPAN in R, where they described difficulties with importing the original TWINSPAN code (written in FORTRAN) into R. Seems that both Jari and Dave spent considerable effort trying to implement the method into R, but seems like there is some problem which is not easy to crack.

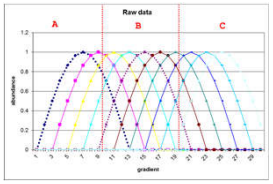


Indicator Power of Species

- Question – “is species X an indicator of species Y”, or “do species X and Y co-occur more than expected at random”.
- Applications – facilitation, competitive interactions etc.
- First approach
  - Categorize samples based on the presence of species X, then perform traditional indicator species analysis
  - Abundances of species X not included, values would represent the relative frequency and abundance of species Y where X occurred.
  - See also signassoc function in indicspecies package.

Indicator Power of Species

- Say we are interested in co-occurrence with species 3 (yellow)




```
classification<-community$sp3>0

ind_species<-multipatt(community,classification,duleg=TRUE)

ind_species
```

```
$sign
      s.FALSE s.TRUE index      stat p.value
sp1         0      1     2 0.7076616   0.020
sp2         0      1     2 0.8753357   0.005
sp3         0      1     2 1.0000000   0.005
sp4         0      1     2 0.8753357   0.005
sp5         0      1     2 0.7076616   0.010
sp6         0      1     2 0.5030200   0.595
sp7         1      0     1 0.5651889   0.310
sp8         1      0     1 0.7247320   0.020
sp9         1      0     1 0.7817360   0.015
sp10        1      0     1 0.7817360   0.010
```

Indicator Power of Species

- Function indpower in vegan package
  - Follows method of Halme et al (2009) and calculates pairwise indicator power.

Target Species

Indicator Species

```
> indpower(community)
           t.sp1     t.sp2     t.sp3     t.sp4     t.sp5     t.sp6     t.sp7     t.sp8     t.sp9    t.sp10
i.sp1  1.0000000 0.8528029 0.7035265 0.5504819 0.3892495 0.2010076 0.0000000 0.0000000 0.0000000 0.0000000
i.sp2  0.8528029 1.0000000 0.8528029 0.7035265 0.5504819 0.3892495 0.2010076 0.0000000 0.0000000 0.0000000
i.sp3  0.7035265 0.8528029 1.0000000 0.8528029 0.7035265 0.5504819 0.3892495 0.2010076 0.0000000 0.0000000
i.sp4  0.5504819 0.7035265 0.8528029 1.0000000 0.8528029 0.7035265 0.5504819 0.3892495 0.2010076 0.0000000
i.sp5  0.3892495 0.5504819 0.7035265 0.8528029 1.0000000 0.8528029 0.7035265 0.5504819 0.3892495 0.2010076
i.sp6  0.2010076 0.3892495 0.5504819 0.7035265 0.8528029 1.0000000 0.8528029 0.7035265 0.5504819 0.3892495
i.sp7  0.0000000 0.2010076 0.3892495 0.5504819 0.7035265 0.8528029 1.0000000 0.8528029 0.7035265 0.5504819
i.sp8  0.0000000 0.0000000 0.2010076 0.3892495 0.5504819 0.7035265 0.8528029 1.0000000 0.8528029 0.7035265
i.sp9  0.0000000 0.0000000 0.0000000 0.2010076 0.3892495 0.5504819 0.7035265 0.8528029 1.0000000 0.8528029
i.sp10 0.0000000 0.0000000 0.0000000 0.0000000 0.2010076 0.3892495 0.5504819 0.7035265 0.8528029 1.0000000
```

Halme, P., Mönkkönen, M., Kotiaho, J. S, Ylisirniö, A-L. 2009. Quantifying the indicator power of an indicator species. Conservation Biology 23: 1008–1016.

Indicator Species

- **Characteristic species in discrete groups of samples**
  - Indicator species
  - TWINSPAN
- **Species associations**
  - Indicator power of species
- **Identification of species most responsible for differences among groups of samples**
  - SIMPER (similarity percentage)
    - Based on abundance, does not weigh occurrence frequency as indicator species analysis does.

SIMPER (similarity Percentage)

- Attempt to assess each species contribution to Bray-Curtis similarity among groups. For a given sample, one species contribution is:

$$d_{ijk} = \frac{\left|(x_{ij} - x_{ik})\right|}{x_{ij} + x_{ik}}$$

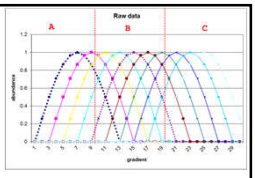
- Where x is the abundance of species *i* in sample *j* and *k*
- Average contribution for that species is the average across samples:

$$d_{jk} = \sum_{i=1}^{S} d_{ijk}$$

- Function simper(community, grouping)
  - limited to use with Bray Curtis index
  - Vegan package

Clarke, K.R. 1993. Non-parametric multivariate analyses of changes in community structure. Australian Journal of Ecology, 18, 117–143

```
sim<-simper(community,envdata$spatial)

> sim$a_b$overall
[1] 0.8269227

Contrast: a_b

      average      sd  ratio    ava    avb   cumsum
sp5  0.130366 0.05589 2.3325 0.02876 0.73369 0.1577
sp6  0.124990 0.06518 1.9176 0.00000 0.68369 0.3088
sp1  0.119654 0.04263 2.8069 0.75966 0.07588 0.4535
sp4  0.101245 0.07584 1.3349 0.16288 0.61298 0.5759
sp7  0.096174 0.07910 1.2159 0.00000 0.52638 0.6922
sp2  0.083789 0.05719 1.4651 0.58487 0.23320 0.7936
sp3  0.079117 0.07034 1.1248 0.36643 0.42979 0.8892
sp8  0.060202 0.07152 0.8418 0.00000 0.32979 0.9620
sp9  0.026711 0.04619 0.5783 0.00000 0.14659 0.9943
sp10 0.004675 0.01443 0.3240 0.00000 0.02588 1.0000
```



=.8269/.1303

Species in rank order by contribution.

d_{ijk} | mean abundance | cumulative d_{ijk}

Overall (mean of all pairwise) Bray-Curtis dissimilarity for a vs. b

```
> sim$a_b$overall
[1] 0.8269227

Contrast: a_b

      average      sd  ratio     ava     avb   cumsum
sp5  0.130366 0.05589 2.3325 0.02876 0.73369 0.1577
sp6  0.124990 0.06518 1.9176 0.00000 0.68369 0.3088
sp1  0.119654 0.04263 2.8069 0.75966 0.07588 0.4535
sp4  0.101245 0.07584 1.3349 0.16288 0.61298 0.5759
sp7  0.096174 0.07910 1.2159 0.00000 0.52638 0.6922
sp2  0.083789 0.05719 1.4651 0.58487 0.23320 0.7936
sp3  0.079117 0.07034 1.1248 0.36643 0.42979 0.8892
sp8  0.060202 0.07152 0.8418 0.00000 0.32979 0.9620
sp9  0.026711 0.04619 0.5783 0.00000 0.14659 0.9943
sp10 0.004675 0.01443 0.3240 0.00000 0.02588 1.0000
```

Sum d_{ijk} for all paired samples, each species contribution to overall Bray Curtis

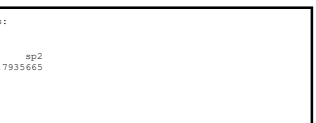
Variability in contribution and ratio to overall abundance..is the species a strong and consistent contributor to differences?

Species in rank order by contribution.

Sum = overall Bray-Curtis dissimilarity

cumulative % d_{jk} =1.0

```
cumulative contributions of most influential species:

$a_b
      sp5       sp6       sp1       sp4       sp7       sp2
0.1576526 0.3088029 0.4535014 0.5759371 0.6922408 0.7935665

$a_c
      sp1      sp10       sp9       sp2
0.2261496 0.4481796 0.5952704 0.7356632

$b_c
      sp5      sp10       sp6       sp4       sp9       sp7
0.1531625 0.2952327 0.4290845 0.5571064 0.6584465 0.7595939

> summary(sim)

Contrast: a_b

       contr      sd  ratio    av.a    av.b cumsum
sp5  0.130366 0.05589 2.3325 0.02876 0.73369 0.1577
sp6  0.124990 0.06518 1.9176 0.00000 0.68369 0.3088
sp1  0.119654 0.04263 2.8069 0.75966 0.07588 0.4535
…
Contrast: a_c

       contr      sd  ratio    av.a    av.b cumsum
sp1  0.226150 0.12048 1.8770 0.75966 0.00000 0.2261
sp10 0.222030 0.12394 1.7915 0.00000 0.73369 0.4482
sp9  0.147091 0.09365 1.5707 0.00000 0.61298 0.5953
…
Contrast: b_c

       contr      sd  ratio    av.a    av.b cumsum
sp5  0.13057 0.05169 2.5262 0.73369 0.00000 0.1532
sp10 0.12112 0.04058 2.9845 0.02588 0.73369 0.2952
sp6  0.11411 0.06768 1.6860 0.68369 0.07588 0.4291
…
```


Methods in Ecology and Evolution 2012, 3, 89–101 doi: 10.1111/j.2041-210X.2011.00127.x

Distance-based multivariate analyses confound location and dispersion effects

David I. Warton[1*], Stephen T. Wright[2] and Yi Wang[1,2]

[1]School of Mathematics and Statistics and Evolution & Ecology Research Centre; and [2]School of Computer Science and Engineering, The University of New South Wales, NSW 2052, Australia

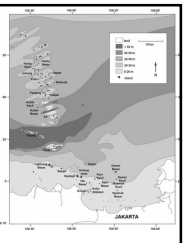
Caveat: Warton et al. document some potentially serious issues with SIMPER based on the positive relationship between mean and variance in abundance data.

Assignment

- Text: Section 4.10
- Reading:
  - Taylor, C.M. and M. E. Roberts. 2008. Using community-level analyses to identify dietary patterns for species in space and time. *Journal of Freshwater Ecology 23:519-528.*
  - Heino, J. et al. 2003. Defining macroinvertebrate assemblage types of headwater streams: implications for bioassessment and conservation. *Ecological Applications* 13:842-852.

Assignment

Warwick, R. M., & Clarke, K. R. (1990). A statistical analysis of coral community responses to the 1982–83 El Niño in the Thousand Islands, Indonesia. Coral Reefs, 8(4), 171-179.



- Tikus island coral reef dataset
- Code to load data from package
  - `library(mvabund)`
  - `data(tikus)`
- 10 transects sampled 6 years (81, 83-88). Coral bleaching event in 82-83.
  - `tikus$abund`: Abundance of 75 coral species
  - `tikus$x$time`: Years pre (81) and post (all others). Code to define pre/post variable:
    - `pre_post<-tikus$x$time!=81`
- Do an indicator species analysis. What species are indicators for pre vs. post?
- Visualize with an NMDS
  - Square-root transform data, k=2, Bray-Curtis distance, symbols representing years or pre/post
- Use simper to identify species most responsible for differences between pre/post.
- Synthesis should explain similarities/differences between approaches.