A bunch of R functions to assist phytosociological tabulation

[1,2] Tiago Monteiro-Henriques
[1] Centro de Estudos Florestais, Instituto Superior de Agronomia, Universidade de Lisboa, Tapada da Ajuda 1349-017 Lisboa, Portugal.
[2] Centro de Investigação e de Tecnologias Agro-Ambientais e Biológicas, UTAD, Quinta de Prados, Apartado 1013, 5000-801 Vila Real, Portugal.
tmh@isa.ulisboa.pt

Key-words Tabulation, Classification, Optimization, Hill climbing, R Statistical Software

1 Question

The main purpose of table sorting is clustering: the researcher seeks to partition the relevé set in groups (subsets), with ecologic or biogeographic significance, easily and floristically differentiable. A plethora of methods have been proposed aiming at assisting phytosociological tabulation. Although direct optimization is a natural way to approach the issue, the huge number of possible combinations of relevé groups hinders any such attempt (Peet and Roberts 2013). Clustering methods based on (dis)similarity measures are one practical alternative, although rarely producing optimal outputs and can be far from the phytosociologist's objectives. After all: **is it possible, with today computers, to directly optimize the search for differential species patterns in a phytosociological table?**

2 Definitions

k-partition If we have *N* relevés that we want to classify in *k* groups, a *k-partition* is any set of nonempty *k* subsets of the *N* relevés, such that every relevé is in exactly one of these subsets.

Partition A, N=8, k=2

rel. 1 rel. 2	rel. 4 rel. 6
rel. 3 rel. 5	rel. 7 rel. 8

Partition B, N=8, k=2

rel. 1	rel. 2 rel. 4 rel. 6
rel. 3	rel. 5 rel. 7 rel. 8

n-neighbour We define a *n-neighbour* of a partition as a second partition in which *n* relevés are placed in a different group. E.g. partition A and B in the above figure are *2-neighbours* of each other.

optimization Mathematical optimization aims at selecting the best value from a set of alternatives. In this work we use hill-climbing algorithm to search for a partition of the relevés that maximizes an index.

3 What to optimize?

After the presentation of the $DiffVal$ index (Monteiro-Henriques & Bellu 2014), I've come to notice that I could improve the optimization procedure and that the proposed index needed a minor correction (to maintain it between 0 and 1). Therefore, a better performing method is now ready for disclosure, optimizing $TotDiffVal1$ directly. I illustrate the use of the technique on real and virtual datasets.

$DiffVal1$ and $TotDiffVal1$ can be calculated in the following way:

$$DiffVal1_{s,P} = \sum_{g \ni s} \frac{\frac{a}{b}}{e} \times \frac{\frac{c}{d}}{e}$$

$DiffVal1_{s,P}$	is the differential value of taxon *s*, given the partition *P*
$g \ni s$	the sum is obtained for each group *g* containing taxon *s*
a	is the total no. of relevés of the groups that do not contain *s*
b	is the total no. of relevés of all groups except *g*
c	is the no. of relevés of *g* containing taxon *s*
d	is the total no. of relevés in group *g*
e	is the total no. of groups containing taxon *s*

$$TotDiffVal1_{T,P} = \frac{1}{n} \sum_{i=1}^{n} DiffVal1_{i,P}$$

$TotDiffVal1_{P,T}$ is the total differential value of partition P for phytosociological table T, where n is the total number of taxa present in T, and $DiffVal1_{i,P}$ is the differential value of taxon i, given the partition P.

4 R Functions

HC.optim.tdv accepts a phytosociological table and searches for a *k-partition* (*k* defined by the user) optimizing $TotDiffVal1$ index, i.e. searches, using a hill-climbing algorithm, for patterns of differential species by rearranging the relevés into *k* groups. Optimization can start from a random partition, or from a given partition (e.g. produced by any clustering method, or even a manual classification). Each iteration searches for a $TotDiffVal1$ improvement screening all 1-neighbours, until the given number of maximum iterations (*maxit*) is reached. Optionally, a faster search (stochastic hill-climbing) can be performed firstly (defining *random.first*=TRUE), consisting on searching for $TotDiffVal1$ improvements, by randomly selecting *n-neighbours* (*n* defined by the user), until a given number of maximum iterations (*rf.maxit*) is reached.

tdv accepts a phytosociological table and a partition, returning the respective $TotDiffVal1$ index.

tabulate accepts a phytosociological table, a partition and taxa names, returning an ordered table based, firstly, on the number of groups a species occurs in, and secondly, on the within-group relative frequency; optionally an image of the ordered table can be produced.

explore.tabulation accepts an object returned by the *tabulate* function, plotting a condensed image of the table, permitting the user to click on the coloured blocks and receive the respective list of taxa names on the console.

identical.p checks if two *k-partitions* are identical, i.e. if their groupings are the same (useful in the case that the partitions have divergent group numberings).

select.taxa accepts a phytosociological table, and minimum and maximum 'constancy' values. It returns a trimmed phytosociological table, removing species outside the given 'constancy' range and removing relevés that become empty. Optionally, a minimum number of presences (in relevés) might be used instead of the 'constancy' range.

5 Example 1 - Virtual dataset

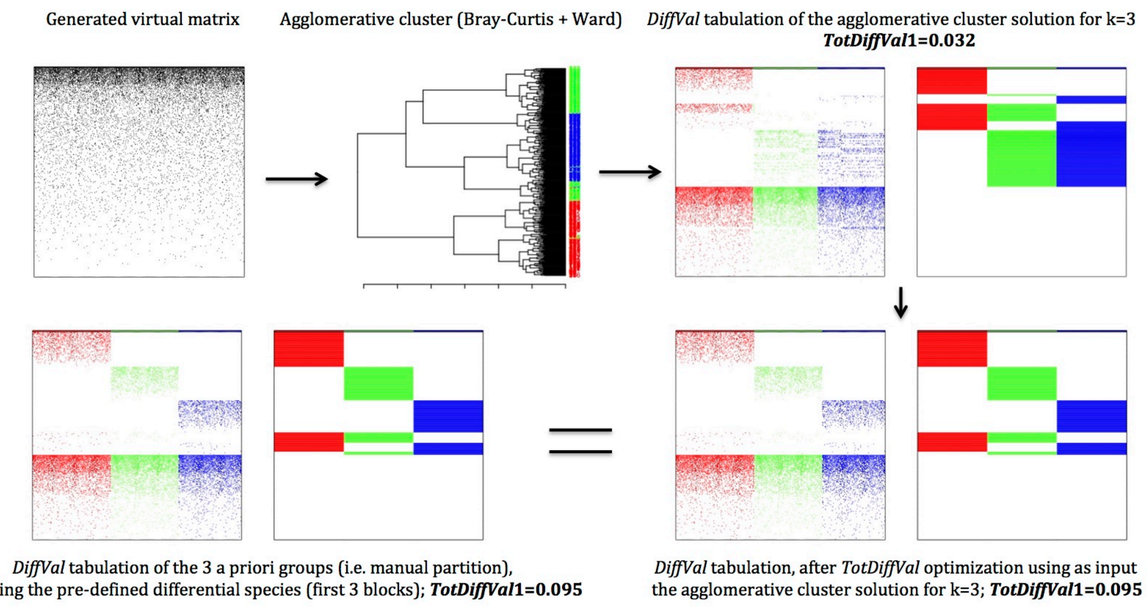
This example was thought purposely to illustrate the robustness of this methodology, and to show a verisimilar case where the method outperforms agglomerative clustering.

I've built a matrix of 400 relevés and 270 taxa, firstly generating uniformly random constancy values for each taxa within pre-defined intervals and, after, generating random dominance levels for each taxa, using the following probabilities of occurrence:

No. of taxa		Probability of occurrence of each dominance level. In grey, relative values, as final probability depended on uniformly selected constancy value in the referred interval.						Constancy (C) (%)	Observations	
		0	1	2	3	4	5			
1		-	-	5%	40%	40%	15%	100	Dominant taxa	
5	(13%)	100 - C	15%	30%	40%	15%	-	[20, 90]	Co-dominant taxa	
30		100 - C	55%	25%	15%	4%	1%	[20, 95]	Higher rank and other ±faithful taxa	
20		100 - C	65%	25%	10%	-	-	[5.25, 25]	Other less faithful randomly distributed taxa	Present in ≈ 21 to 100 relevés
25		100 - C	70%	25%	5%	-	-	[2.75, 5]		Present in ≈ 11 to 20 relevés
30	(45%)	100 - C	70%	29%	1%	-	-	[1, 2.5]		Present in ≈ 4 to 10 relevés
35		100 - C	70%	30%	-	-	-	[0.5, 0.75]		Present in ≈ 2 to 3 relevés
10		99.8%	0.25%	-	-	-	-	0.25		Present in ≈ 1 relevé
40		100 - C	35%	35%	20%	10%	-	[0.25, 40]	Kept only in a predefined 1st group of 150 relevés	
38	(42%)	100 - C	35%	35%	20%	10%	-	[0.25, 40]	Kept only in a predefined 2nd group of 130 relevés	
36		100 - C	35%	35%	20%	10%	-	[0.25, 40]	Kept only in a predefined 3rd group of 120 relevés	

≈13% of dominant, co-dominant and ±faithful taxa (>20% constancy, i.e. present in >80 relevés)
≈45% of randomly distributed taxa, e.g. other companion, transgressive, cosmopolitan, erratic, and other lower constancy taxa
≈42% of taxa with low to high differential value (simulating e.g. a (bio)geographical or an ecological distribution constraint)

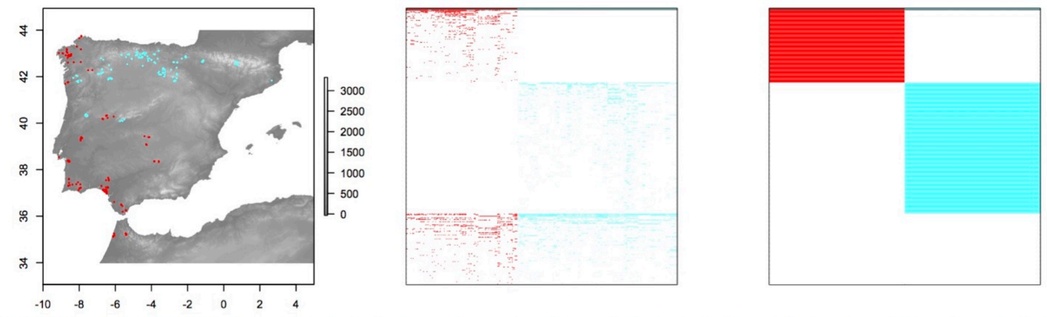
5 Example 1 - Virtual dataset (colours correspond to the a priori groups)



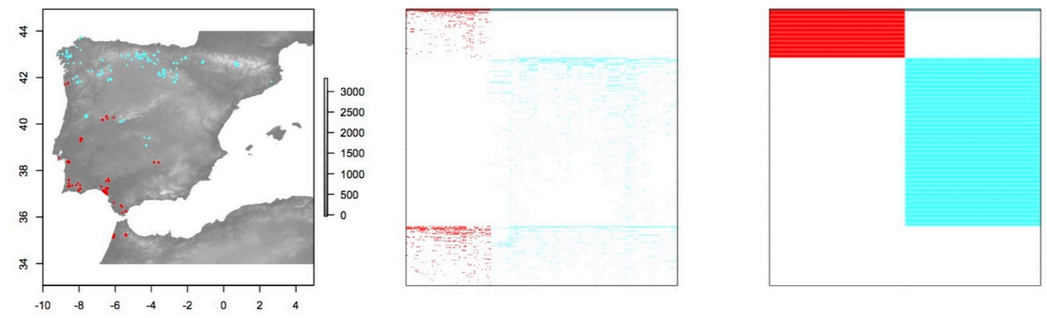
5 Example 2 - Real dataset

In this example I show an application to a real-world dataset on Iberian damp heathlands, analysing matrix M with 410 relevés x 352 species.

TotDiffVal1$_{M,P1}$**=0.0370**; P1 is the solution for k=2 of an agglomerative hierarchical clustering (Bray-Curtis + Ward); 261 (93+168) taxa are exclusive to one of the groups; 66 with relative frequency greater than 5%.

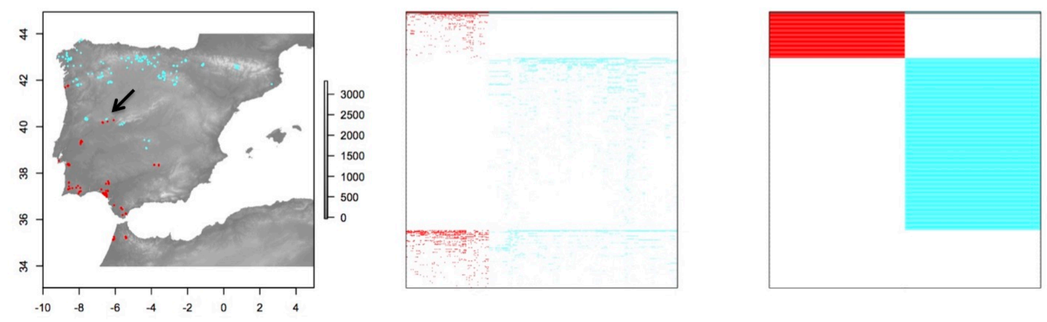


TotDiffVal1$_{M,P2}$**=0.0409**; P2 is the solution for k=2 of an agglomerative hierarchical clustering (Bray-Curtis + Complete), using only the taxa with 'constancies' from 5 to 70%; 276 (60+216) taxa are exclusive of one of the groups; 69 with relative frequency greater than 5%.



5 Example 2 - Real dataset

TotDiffVal1$_{M,P3}$**=0.0411**; P3 is the result of hill-climbing optimization (on M matrix) of the solution for k=2 of an agglomerative hierarchical clustering (Bray-Curtis + Complete), using only the taxa with 'constancies' from 10 to 70%; 278 (57+221) taxa are exclusive of one of the groups; 70 with relative frequency greater than 5%.



Functions *tdv* and *tabulate* enabled the comparison of the two clustering procedures for k=2 solutions (P1 and P2), showing the differences between $TotDiffVal1$ and showing which species and relevés contributed to that change. I have compared two cluster strategies ("Bray-Curtis + Ward" and "Bray-Curtis + Complete linkage") in three different matrices: M, i.e. the full matrix; and two other, obtained by trimming M using two 'constancy' ranges: 5 to 70% (M2) and 10 to 70% (M3). Generally, *HC.optim.tdv* was able to find partitions presenting higher $TotDiffVal1$ than the clustering algorithms. All the obtained partitions (from clusters and optimizations) were tested on matrix M checking which produced the highest $TotDiffVal1$ (regardless of the matrix which originated the partitions). Best solution found (P3), until the present moment, is the result of hill-climbing optimization (on M matrix) of the solution for k=2 coming from an agglomerative clustering (Bray-Curtis + Complete) on M3, which changed the cluster solution only very slightly. **Notice that for k=2, as in the presented example, optimizing $TotDiffVal1$ is equivalent to optimize the mean relative frequency of all taxa occurring exclusively in one of the two groups.**

6 Download

Functions and indices are available at http://home.isa.utl.pt/~tmh/.

7 Conclusions

■Answering the first question: **yes**, it's possible to directly optimize the search for differential species patterns in a phytosociological table, with nowadays computers. However, the process is time consuming, worsening on bigger tables. Depending on data structure and size, sometimes, conversion to a known best solution might be very difficult. But don't despair! Results can be surprising. ■**Yet, and importantly, you can always compare the result of any clustering procedure (including manual classifications!).** ■It is also possible to check if a partition is a 1-neighbour local maximum. ■Theoretically, other indices (e.g. IndVal derived) can be optimized in this way too. ■It is important to bear in mind, however, that it is not mathematically feasible to optimize two or more criteria at once expecting a single best value. I.e., for a nontrivial problem, finding a unique solution that simultaneously shows a maximum value for a certain first criteria and a maximum value for a certain second criteria it is simply not possible(!). ■The virtual example was purposely thought to converge to the a priori groups, to illustrate that *HC.optim.tdv* can move closer to optimality (considering $TotDiffVal1$ as the objective function) than other clustering strategies. Of course, such clustering strategies were not developed for $TotDiffVal1$ optimization, but some are fair heuristics to it! ■Using only the taxa that presents intermediate levels of constancy as input, as proposed in Mueller-Dombois and Ellenberg (1974), can improve the method's performance, as well as using the output of other cluster strategies as starting partition. ■The used hill-climbing optimization technique cannot assure the finding of the global maximum, however, multiple starts increase the probability of finding it. ■The expert must always validate the obtained differential taxa patterns (the best solutions should be plotted on maps or on the environmental space). ■**I invite you to try these functions on your data, and would be happy in case you give me some feedback about it** (tmh@isa.ulisboa.pt)**.**

8 References

Monteiro-Henriques T. & Bellu A. (2014). An optimization approach to the production of differentiated tables based on new differentiability measures. 23rd Internaational Workshop of the European Vegetation Survey - Book of Abstracts. Ljubljana. p.43-44.
Mueller-Dombois D, Ellenberg H (1974) Aims and Methods of Vegetation Ecology. John Wiley & Sons, New York
Peet RK, Roberts DW (2013) Classification of Natural and Semi-natural Vegetation. In: Maarel E van der, Franklin J (eds) Vegetation ecology, 2nd ed. John Wiley & Sons, Chichester, West Sussex, England ; Hoboken, NJ, pp 28–70