Use of mean Ellenberg indicator values in vegetation analysis

Theoretical background

Ellenberg indicator values (EIVs; Ellenberg et al. 1992) are estimates of species ecological optima along seven main ecological gradients - light, temperature, continentality, moisture, nutrients, soil reaction and salinity (with salinity having no meaning outside the oceanic region). Mean of species EIVs for species occurring at given stand (Fig. 1a) is often used as an estimate of local ecological conditions and a surrogate for measured environmental variables.

When analysing vegetation data, one should bear in mind that unlike real measured environmental factors, mean EIVs are variables *calculated* from species-plot vegetation matrix, and this feature has important consequences for their further use. While calculated, mean EIVs inherit two types of information, one derived from external information about species ecological behaviour (i.e. tabulated species indicator values), and another derived from species composition data itself.

How does the composition inherit into mean EIVs will be clear from the following example. If two vegetation samples have exactly the same species composition, they will have the same mean EIVs (if mean EIVs is calculated as mean weighted by species abundances, species in the samples would have to have also the same abundances). There is no other option, given the way how the values are calculated, although it's not probable that real measured environmental factors would be also exactly the same. If the samples differ in one or few species, their mean EIVs will not differ too much - if you count average from let's say 20 numbers and you change one or few of them, mean won't change much. Although it may seem unlikely that these situations will occur in real dataset, they illustrate how influenced are mean EIVs by compositional information.

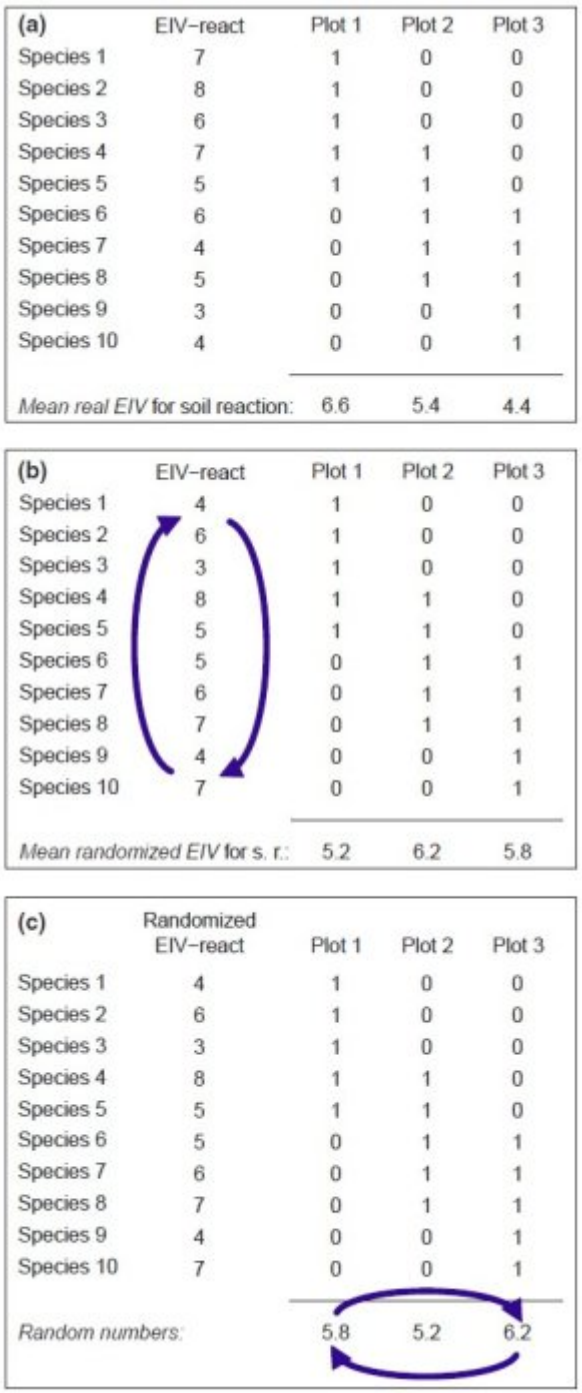
While interpreting importance of mean EIVs, we usually acknowledge only the first type of information and interpret mean EIVs in terms of the underlying ecological gradients, silently ignoring the information derived from compositional data itself. This basically doesn't matter if the mean EIVs are used for purely descriptive purpose, e.g. when evaluating which stand is dryer and which wetter. The problem occurs when we analyse the relationship between mean EIVs and other variables derived from the same compositional data and we attempt to test the significance of this relationship, which is often the case.

An example could be the correlation of mean EIVs with ordination axes in an unconstrained ordination analysis (e.g. DCA, PCA, CA, NMDS), which is done whenever the mean EIVs are passively projected onto ordination diagram. Ordination axes (or the positions of samples along these axes, respectively) directly reflect the similarity in species composition between samples - the more similar the samples are, the more close they will be. Being derived from the same vegetation dataset, both mean EIVs and ordination axes are indeed not independent, and standard statistical test of their relationship (which always require independence of variables) will return biased results. The bias is not just an theoretical issue, but it's rather remarkable - the relationships tend to be more often and highly significant, which may lead to wrong impression that mean EIVs are actually better environmental factors then real measured variables.

Other types of potentially biased analyses include testing the differences (e.g. by ANOVA) in mean EIVs between groups of samples, which have been assembled together according to their similarity (e.g. by cluster analysis, or applying the same experimental treatment on vegetation leading to similar species composition). Yet other type is correlation (or regression) between mean EIVs and species richness (number of species in the sample), or using mean EIVs as explanatory in classification or regression trees (CART), if the response variable is also derived from vegetation (e.g. species richness in regression trees or classification into vegetation types in classification trees).

There are several ways how to deal with this issue. The simplest one is not to use mean EIVs in vegetation analysis at all - if there are relevant measured environmental variables at hand, there is no need to use these

calculated surrogates, just because it is so easy to calculate them. But it often happens that measured factors are not available, and in that case the informational potential offered by mean EIVs may be useful. Mean EIVs may be simply used in descriptive analysis without further statistical inference (without testing for their significance). In case you need statistical inference, the following lines will briefly introduce the modified test of significance, which can be used for analysing the relationship of mean EIVs with other variables derived from the same compositional data, including the practical way how these could be calculated in a stand-alone application based on R script.



(a)

	EIV–react	Plot 1	Plot 2	Plot 3
Species 1	7	1	0	0
Species 2	8	1	0	0
Species 3	6	1	0	0
Species 4	7	1	1	0
Species 5	5	1	1	0
Species 6	6	0	1	1
Species 7	4	0	1	1
Species 8	5	0	1	1
Species 9	3	0	0	1
Species 10	4	0	0	1
Mean real EIV for soil reaction:		6.6	5.4	4.4

(b)

	EIV–react	Plot 1	Plot 2	Plot 3
Species 1	4	1	0	0
Species 2	6	1	0	0
Species 3	3	1	0	0
Species 4	8	1	1	0
Species 5	5	1	1	0
Species 6	5	0	1	1
Species 7	6	0	1	1
Species 8	7	0	1	1
Species 9	4	0	0	1
Species 10	7	0	0	1
Mean randomized EIV for s. r.:		5.2	6.2	5.8

(c)

	Randomized EIV–react	Plot 1	Plot 2	Plot 3
Species 1	4	1	0	0
Species 2	6	1	0	0
Species 3	3	1	0	0
Species 4	8	1	1	0
Species 5	5	1	1	0
Species 6	5	0	1	1
Species 7	6	0	1	1
Species 8	7	0	1	1
Species 9	4	0	0	1
Species 10	7	0	0	1
Random numbers:		5.8	5.2	6.2

Modified permutation test

What is modified is the null hypothesis, which is going to be tested. The original null hypothesis reads as “there is no relationship between variable X and mean EIVs”. However, as was said above, principally there *is* a relationship between both variables, because they are both derived from the same vegetation dataset. To test such hypothesis is therefore not too informative - it's easy to reject it, but this rejection doesn't tell us if the

relationship is caused by differences in external information from species indicator values or the simply by differences in species composition itself. We can, however, modify the null hypothesis to be more informative: “there is no relationship between variable X and the part of information in mean EIVs derived from external information about species ecological behavior (i.e. the list of species EIVs)”. If rejected, we know that the relationship is significant because of differences in external indicator values, e.g. about species demands for soil moisture.

The original null hypothesis can be tested using parametric or permutation test of significance. In case of correlation, the parametric test of significance of the correlation coefficient *r* is based on a t-test. Permutation test, on the other hand, compares the test statistic (e.g. correlation coefficient *r*) of the real variables with the distribution of the test statistics occurring if the relationship is random (one of the variables, e.g. mean EIVs, is randomized prior to analysis, Fig. 1c). If the real *r* is higher (or lower) then let's say 95% of values generated by random permutation, we consider it as being significantly different. Modification of this permutation test is based on changing the permutation schema; instead of randomizing calculated mean EIVs, we randomize species indicator values between the species in the vegetation matrix, and using these randomized values we calculate new mean EIVs (Fig. 1b) and correlate them with the other variable to create the null distribution. In other words the modified permutation test is asking if the real assignment of species indicator values to the species brings more information than if the indicator values are assigned to species randomly, without relationship to knowledge about real environmental behaviour of these species. See Zelený & Schaffers (2012) for further details of this analysis.

Practical guide how to calculate modified permutation test

Modified permutation test can be calculated for two types of analysis with mean EIVs:

1. correlation or regression of mean EIVs with other variables derived from species composition (e.g. ordination axes from unconstrained ordination, or species richness), and
2. ANOVA of mean EIVs between groups of samples (providing that the samples are clustered into groups according to their similarity, e.g. using cluster analysis or applying the same experimental treatment resulting in similar species composition of samples within groups).

Modified permutation test can be calculated

1. using the R functions, namely `envfit.iv` (for correlation with ordination axes) and `summary.aov.iv` (for ANOVA), which are part of online Supplementary materials of the paper Zelený & Schaffers (2012), namely Appendix S2 and Appendix S3;
2. using the stand-alone application, based on the algorithms mentioned above; the application is essentially a wrapper to both functions (containing additional functionality) and doesn't require knowledge of R programming, although you still need to have R program installed on your computer.

References

- Zelený D. & Schaffers A.P. (2012): Too good to be true: pitfalls of using mean Ellenberg indicator values in vegetation analyses. *Journal of Vegetation Science*, 23: 419-431