To test or not to test, that is the question: relating mean Ellenberg indicator values to ordination axes is an example of spurious correlation

David Zelený

Institute of Ecology and Evolutionary Biology
National Taiwan University
Taipei, Taiwan
Data for community weighted mean approach (CWM approach)

- **species**
  - matrix of species composition $L$
- **sample attributes**
  - characteristics of individual samples (plots, localities)
  - environmental variables,
  - experimental treatment,
  - sample classification,
  - species richness
  - diversity index

**species attributes** – characteristics of individual species
- traits, Ellenberg-type species indicator values, species phylogenetic relationship, species specialization values etc.
CWM

community weighted mean of species attributes (t)

- CWM of traits
- mean Ellenberg indicator values (mEIV)
- diatom index in hydrobiology
- transfer function in paleoecology

\[
\text{CWM} = f_1 (L, t)
\]
CWM

*community weighted mean of species attributes* \( (t) \)
- CWM of traits
- mean Ellenberg indicator values \((mEIV)\)
- diatom index in hydrobiology
- transfer function in paleoecology

\[
CWM = f_1 (L, t)
\]

**extrinsic sample attributes**

*measured/estimated environmental variables, experimental treatments*
- soil pH (measured),
- canopy cover (estimated)
- temperature (GIS derived)
- fertilized vs control (treatment)

\[
SA_{extr} = e
\]
CWM

*community weighted mean of species attributes* \((t)\)
- CWM of traits
- mean Ellenberg indicator values (mEIV)
- diatom index in hydrobiology
- transfer function in paleoecology

\[
CWM = f_1 (L, t)
\]

**Extrinsic sample attributes**

*measured/estimated environmental variables, experimental treatments*
- soil pH (measured),
- canopy cover (estimated)
- temperature (GIS derived)
- fertilized vs control (treatment)

\[
SA_{extr} = e
\]

**Intrinsic sample attributes**

*calculated from species composition*
- sample ordination scores,
- species richness,
- diversity indices,
- assignment of sample into a cluster by numerical classification

\[
SA_{intr} = f_2 (L)
\]
Testing CWM approach with extrinsic sample attributes

\[ \text{CWM} \sim \text{SA}_{\text{extr}} \]

\[ f_1(L, t) \sim e \]


Testing CWM approach with extrinsic sample attributes

(a) Community-weighted mean approach

L – species composition
t – species attributes (traits, EIV)
e – sample attributes (env. vars)
ep – randomized sample attr.
c – CWM of species attributes
pr – significance of row-based test
pc – significance of column-based test
pmax – significance of max test

David Zelený: CWM approach with intrinsic species attributes - 6
Testing CWM approach with intrinsic sample attributes

\[
\text{CWM} \sim \text{SA}_{\text{intr}}
\]

\[
f_1 (L, t) \sim f_2 (L)
\]


“Mathematical Contributions to the Theory of Evolution.—On a Form of Spurious Correlation which may arise when Indices are used in the Measurement of Organs.” By Karl Pearson, F.R.S., University College, London. Received December 29, 1896,—Read February 18, 1897.

(1) If the ratio of two absolute measurements on the same or different organs be taken it is convenient to term this ratio an index.

If \( u = f_1(x, y) \) and \( v = f_2(z, y) \) be two functions of the three variables \( x, y, z \), and these variables be selected at random so that there exists no correlation between \( x, y, z \), or \( z, x \), there will still be found to exist correlation between \( u \) and \( v \). Thus a real danger arises when a statistical biologist attributes the correlation between two functions like \( u \) and \( v \) to organic relationship. The particular case that is likely to occur is when \( u \) and \( v \) are indices with the same denominator for the correlation of indices seems at first sight a very plausible measure of organic correlation.
Examples of spurious correlations (✓) in trait-trait relationships (by Shen Yen-Cheng)

David Zelený: CWM approach with intrinsic species attributes - example of spurious correlation
Examples of spurious correlations (×) in trait-trait relationships (by Shen Yen-Cheng)

LDMC vs SLA

LDMC = LDW*1000/LFW,
SLA = LA/LDW,

LDMC ~ SLA
LDW*1000/LFW ~ LA/LDW

LDMC – leaf dry matter content
SLA – specific leaf area
LA – leaf area
LDW – leaf dry weight
LFW – leaf fresh weight
What to do with spurious correlation?

Alternative opinions:
- do not test
- spurious correlation is not a problem
- test by Monte Carlo permutation test, which permutes the parent variable present only at one side of the equation (Jackson & Somers 1991, Brett 2004)

\[ CWM \sim SA_{\text{intr}} \]
\[ f_1 (L, t) \sim f_2 (L) \]
Too good to be true: pitfalls of using mean Ellenberg indicator values in vegetation analyses

David Zelený & André P. Schaffers

Abstract

Question: Mean Ellenberg indicator values (EIVs) inherit information about compositional similarity, because during their calculation species abundances (or presence–absences) are used as weights. Can this similarity issue actually be demonstrated, does it bias results of vegetation analyses correlating mean EIVs with other aspects of species composition and how often are biased studies published?

Methods: In order to separate information on compositional similarity possibly present in mean EIVs, a new variable was introduced, calculated as a weighted average of randomized species EIVs. The performance of these mean randomized EIVs was compared with that of the mean real EIVs on the one hand and random values (randomized mean EIVs) on the other. To demonstrate the similarity issue, differences between samples were correlated with dissimilarity matrices based on various indices. Next, the three mean EIV variables were tested in canonical correspondence analysis (CCA), detrended correspondence analysis (DCA), analysis of variance (ANOVA) between vegetation clusters, and in regression on species richness. Subsequently, a modified permutation test of significance was proposed, taking the similarity issue into account. In addition, an inventory was made of studies published in the Journal of Vegetation Science and Applied Vegetation Science between 2000 and 2010 likely reporting biased results due to the similarity issue.

Results: Using mean randomized EIVs, it is shown that compositional similarity is inherited into mean EIVs and most resembles the inter-sample distances in correspondence analysis, which itself is based on iterative weighted averaging. The use of mean EIVs produced biased results in all four analysis types examined.

Use modified (column-based) permutation test!
Relating mEIV to ordination axes in DCA: standard (row-based) test

Sample-based ordination space

<table>
<thead>
<tr>
<th>Attribute</th>
<th>$r^2$</th>
<th>Pr($&gt;r$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>light</td>
<td>0.7157</td>
<td>0.001</td>
</tr>
<tr>
<td>temp</td>
<td>0.4449</td>
<td>0.001</td>
</tr>
<tr>
<td>cont</td>
<td>0.1170</td>
<td>0.006</td>
</tr>
<tr>
<td>moist</td>
<td>0.9236</td>
<td>0.001</td>
</tr>
<tr>
<td>react</td>
<td>0.4176</td>
<td>0.001</td>
</tr>
<tr>
<td>nutr</td>
<td>0.8706</td>
<td>0.001</td>
</tr>
</tbody>
</table>
Relating mEIV to ordination axes in DCA: modified (col-based) permutation test

Sample-based ordination space

<table>
<thead>
<tr>
<th>Variable</th>
<th>r²</th>
<th>Pr(&gt;r)</th>
</tr>
</thead>
<tbody>
<tr>
<td>light</td>
<td>0.7157</td>
<td>0.0002 ***</td>
</tr>
<tr>
<td>temp</td>
<td>0.4449</td>
<td>0.0224 *</td>
</tr>
<tr>
<td>cont</td>
<td>0.1170</td>
<td>0.5322</td>
</tr>
<tr>
<td>moist</td>
<td>0.9236</td>
<td>0.0002 ***</td>
</tr>
<tr>
<td>react</td>
<td>0.4176</td>
<td>0.0374 *</td>
</tr>
<tr>
<td>nutr</td>
<td>0.8706</td>
<td>0.0002 ***</td>
</tr>
</tbody>
</table>

David Zelený: CWM approach with intrinsic species attributes - example of spurious correlation
Alternative solution: standard permutation test in species-based ordination space

David Zelený: CWM approach with intrinsic species attributes - example of spurious correlation
Conclusions

• Sample attributes are either *extrinsic* or *intrinsic*, and each type requires different treatment

• CWM with *extrinsic* variables – test depends on the hypothesis tested (max test, row- or column-based tests)

• CWM with *intrinsic* variables – no test, or modified (column-based) permutation test
New Results

Which results of the standard test in community weighted mean approach are too optimistic?

David Zeleny

doi: https://doi.org/10.1101/349589

This article is a preprint and has not been peer-reviewed [what does this mean?].
R package *weimea*

[https://github.com/zdealveindy/weimea](https://github.com/zdealveindy/weimea)

David Zelený: CWM approach with intrinsic species attributes - example of spurious correlation
Thank you for your attention!

Resources:
Zeleny (submitted): Which results of standard test in community weighted mean approach are too optimistic?
   biorXiv, https://doi.org/10.1101/349589

weimea: R package for analysis species and sample attributes,
   GitHub, https://github.com/zdealveindy/weimea

This presentation: https://www.davidzeleny.net/doku.php/conf

This study has been supported by Ministry of Science and Technology, Taiwan (MOST, 106-2621-B-002-003-MY3)