Fusion de données spectroscopiques appliquée aux sols

Lauric Cécillon
Outline

Intro – Data fusion: why?

1 – Outer product analysis: OP-PLS / OP-PCA

2 – Principal component transform
Conclusion of our SPIRSOL paper

« Coupling NIR spectral libraries with other diffuse reflectance measurements of soils, such as mid-infrared reflectance spectra, will probably be the next step towards spectral sensing of soil quality worldwide »

Soil spectroscopy
- VIS-NIR vs MIR
- Different analyzers, techniques, labs, teams

Spectroscopic data fusion
- 2D correlation spectroscopy (I. Noda et al., since 1986)
- Outer product analysis (A. Barros, D. Rutledge et al., since 1997)
Data fusion: an example with outer product analysis (OPA)

First: compute OP matrices for each sample

Mutual weighting of each signal by the other:
- if intensities simultaneously high in the two domains, the product is higher;
- if intensities simultaneously low in the two domains, the product is lower;
- if one intensity high and the other low, the product tends to an intermediate value

(Figure from D. Rutledge, Pers. Com.)
Detail of outer product computing

![Diagram showing the calculation of outer products](image-url)

1. **OPA**

Figure from Jaillais et al., 2007 (ChemoLabS)
Example of the NIR-MIR OP matrix for 1 soil sample
NIR-NMR OP matrix for the same soil sample
MIR-NMR OP matrix for the same soil sample
Data fusion: an example with outer product analysis (OPA)
Unfold OP matrices, perform analysis on OP vectors, then fold-back result vector

Figure from Jaillais et al., 2006 (Vibr. Spec.)
Predicting soil organic matter composition with IR-OPA

Outer product – partial least squares: OP-PLS

N soil samples → \(^{13}\)C NMR spectra → MMM → Inferred SOM composition

NIR Spectra → Infrared-OPA

MIR Spectra → PLS model Infrared-OPA

Unknown soil sample → Predicted SOM composition

Figure from Cécillon et al., in revision
Predicting soil organic matter composition with IR-OPA

Study site

- Test on soil samples from Storgama catchment (Norway)

Photo from Live Semb Vestgarden
Predicting soil organic matter composition with IR-OPA

Study site: soils and vegetation

- Test on soil samples from Storgama catchment (Norway)

Photos from Strand et al., Ambio (2008)
Predicting soil organic matter composition with IR-OPA

$^{13}$C CPMAS NMR Spectrum of a soil sample under Calluna

<table>
<thead>
<tr>
<th></th>
<th>Alchile (0-45ppm)</th>
<th>O-alchile (45-110ppm)</th>
<th>Aromatico (110-160ppm)</th>
<th>Carbonile (160-220ppm)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Classical interpretation</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>37</td>
<td>45</td>
<td>13</td>
<td>5</td>
</tr>
</tbody>
</table>

NMR – Molecular Mixing Model
Nelson & Baldock, 2005

Data from Cécillon et al., in revision
Predicting soil organic matter composition with IR-OPA
Predicting soil organic matter composition with IR-OPA

- Cross-validated PLS model with 21 soil samples

{Graphs showing IR fingerprints predicted values versus NMR-MMM reference values for Carbohydrate and Lipid with respective Q², RMSECV, and RPD values.}

Figure from Cécillon et al., in revision
Predicting soil organic matter composition with IR-OPA

- Cross-validated PLS model with 21 soil samples

![Graphs showing predicted vs. reference values for Protein and Black carbon](Image)
Predicting soil organic matter composition with IR-OPA

- Cross-validated PLS model with 21 soil samples

![Graph showing Lignin predictions](image)
OPA-PLS with variable selection (VIP)

Identification of wavebands associated with each biochemical component

- Phenol group (C-C-O or C-O stretching) and/or Aryl group (C-O-C or C-C-C stretching)
- Aliphatic CH₂ & CH₃ groups (CH stretching)
- Tertiary Alcohol (C-O stretching) and/or Ether group (C-O-C)
- Primary, secondary, tertiary Amides (C=O stretching) and/or primary Amine and Amide groups (NH₂ scissoring)
### Predicting soil organic matter composition with IR-OPA

<table>
<thead>
<tr>
<th>Biochemical component</th>
<th>NIR</th>
<th>Soon RPIQ!</th>
<th>MIR</th>
<th>IR fingerprints (NIR-MIR)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Q²</td>
<td>RMSECV</td>
<td>RPD</td>
<td>Q²</td>
</tr>
<tr>
<td>Carbohydrate</td>
<td>0.95</td>
<td>18</td>
<td>4.8</td>
<td>0.88</td>
</tr>
<tr>
<td>Protein</td>
<td>0.92</td>
<td>17</td>
<td>3.7</td>
<td>0.98</td>
</tr>
<tr>
<td>Lignin</td>
<td>0.01</td>
<td>161</td>
<td>1.0</td>
<td>0.06</td>
</tr>
<tr>
<td>Lipid</td>
<td>0.79</td>
<td>31</td>
<td>2.3</td>
<td>0.78</td>
</tr>
<tr>
<td>Black carbon</td>
<td>0.55</td>
<td>51</td>
<td>1.5</td>
<td>0.70</td>
</tr>
</tbody>
</table>

Table from Cécillon et al., in revision
OP-PLS (prediction of lipid content with NIR-MIR OP matrices) → Map of B-coefficients

Model Stat. (CV-LOO):
4LV / Q² = 0.77 / RPIQ = 4.1

Raw spectra
OP-PLS (prediction of lipid content with MIR-NMR OP matrices) → Map of B-coefficients

Model Stat. (CV-LOO):
5LV / Q² = 0.92 / RPIQ = 7.0

Raw spectra
11 months decomposition experiment: crop residues in soil

→ Principal component analysis of NIR-NIR OP matrices: OP-PCA

IR fingerprints obtained from the outer product of two NIR spectral domains:

4,220–4,835 cm\(^{-1}\) (160 wavenumbers) & 5,685–6,100 cm\(^{-1}\) (109 wavenumbers)

=17,440 outer product variables per sample

These two NIR regions show numerous absorbance peaks related to carbohydrates, lipids, proteins, or aromatic structures

<table>
<thead>
<tr>
<th></th>
<th>Wheat</th>
<th>Rape</th>
<th>Alfalfa</th>
</tr>
</thead>
<tbody>
<tr>
<td>C/N ratio</td>
<td>74</td>
<td>60</td>
<td>27</td>
</tr>
</tbody>
</table>

W : Wheat
R : Rape
A : Alfalfa
11 months decomposition experiment: crop residues in soil

→ Principal component analysis of NIR-NIR OP matrices: OP-PCA
11 months decomposition experiment: crop residues in soil

→ Principal component analysis of NIR-NIR OP matrices: OP-PCA

(c) PC1 loadings (wheat)  
(e) PC1 loadings (rapeseed)  
(g) PC1 loadings (alfalfa)
### 11 months decomposition experiment: crop residues in soil

Principal component analysis of NIR-NIR OP matrices: OP-PCA

#### Table 1  NIR wavebands explanatory for PC1 and PC2 of principal component analysis of NIR fingerprints for all samples and for each residue and their corresponding chemical bonds and biochemical components

<table>
<thead>
<tr>
<th>PC1</th>
<th>All residues</th>
<th>Wheat</th>
<th>Rapeseed</th>
<th>Alfalfa</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wavebands  (cm⁻¹)</td>
<td>4800–4825&lt;sup&gt;a&lt;/sup&gt; 4710–4725&lt;sup&gt;a&lt;/sup&gt; 4620–4640&lt;sup&gt;a&lt;/sup&gt; 4485–4505&lt;sup&gt;a&lt;/sup&gt;</td>
<td>4790–4835&lt;sup&gt;a&lt;/sup&gt; 4515–4550&lt;sup&gt;a&lt;/sup&gt; 5785–5820&lt;sup&gt;b&lt;/sup&gt; 4620–4635&lt;sup&gt;b&lt;/sup&gt; 4485–4505&lt;sup&gt;b&lt;/sup&gt; 4355–4370&lt;sup&gt;b&lt;/sup&gt;</td>
<td>4320–4350&lt;sup&gt;a&lt;/sup&gt; 4280–4305&lt;sup&gt;a&lt;/sup&gt; 4245–4270&lt;sup&gt;a&lt;/sup&gt; 5825–5865&lt;sup&gt;b&lt;/sup&gt; 5770–5795&lt;sup&gt;b&lt;/sup&gt; 5730–5750&lt;sup&gt;b&lt;/sup&gt; 4575–4605&lt;sup&gt;b&lt;/sup&gt; 4365–4380&lt;sup&gt;b&lt;/sup&gt;</td>
<td>4425–4475&lt;sup&gt;a&lt;/sup&gt; 4315–4355&lt;sup&gt;a&lt;/sup&gt; 4275–4295&lt;sup&gt;a&lt;/sup&gt; 4230–4265&lt;sup&gt;a&lt;/sup&gt; 5765–5790&lt;sup&gt;b&lt;/sup&gt; 4705–4765&lt;sup&gt;b&lt;/sup&gt; 4560–4610&lt;sup&gt;b&lt;/sup&gt;</td>
</tr>
<tr>
<td>Chemical bonds&lt;sup&gt;e&lt;/sup&gt;</td>
<td>NH, OH, CO, CN, aromatic CH&lt;sup&gt;a&lt;/sup&gt;</td>
<td>NH, OH, aromatic CH; CHO&lt;sup&gt;a&lt;/sup&gt;; aliphatic CH; aromatic CH; NH; CONH&lt;sub&gt;2&lt;/sub&gt;&lt;sup&gt;b&lt;/sup&gt;</td>
<td>Aliphatic CH, aromatic CH, and CH&lt;sup&gt;a&lt;/sup&gt;; aliphatic CH, CH, aromatic CH, NH, CONH&lt;sub&gt;2&lt;/sub&gt;&lt;sup&gt;b&lt;/sup&gt;</td>
<td>CH, aliphatic CH, and aromatic CH&lt;sup&gt;a&lt;/sup&gt;; CH, CO, and NH&lt;sup&gt;b&lt;/sup&gt;</td>
</tr>
<tr>
<td>Biochemical components&lt;sup&gt;b&lt;/sup&gt;</td>
<td>Protein, alcohol, water, and aryl&lt;sup&gt;a&lt;/sup&gt;</td>
<td>Protein, water, alcohol, aryl, and carbohydrate&lt;sup&gt;a&lt;/sup&gt;; lipid&lt;sup&gt;b&lt;/sup&gt;</td>
<td>Lipid, carbohydrate&lt;sup&gt;a&lt;/sup&gt;; aryl, protein, and carbohydrate&lt;sup&gt;b&lt;/sup&gt;</td>
<td>Lipid and carbohydrate&lt;sup&gt;a&lt;/sup&gt;; aryl, protein, carbohydrate&lt;sup&gt;b&lt;/sup&gt;</td>
</tr>
</tbody>
</table>
Principal component transform
Speeding-up processing without losing any information

R script available!
### Principal component transform

Speeding-up processing without loosing any information

#### High-level description for the OP-PCT-PCA algorithm

<table>
<thead>
<tr>
<th>Step</th>
<th>Computation</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>X,Y</td>
<td>Input of X and Y matrices</td>
</tr>
<tr>
<td>2</td>
<td>[T&lt;sub&gt;X&lt;/sub&gt;, P&lt;sub&gt;X&lt;/sub&gt;] ← PCA(X)</td>
<td>Full rank PCA of X</td>
</tr>
<tr>
<td>3</td>
<td>[T&lt;sub&gt;Y&lt;/sub&gt;, P&lt;sub&gt;Y&lt;/sub&gt;] ← PCA(Y)</td>
<td>Full rank PCA of Y</td>
</tr>
<tr>
<td>4</td>
<td>K=OP(T&lt;sub&gt;X&lt;/sub&gt;, T&lt;sub&gt;Y&lt;/sub&gt;)</td>
<td>Outer product (OP) between T&lt;sub&gt;X&lt;/sub&gt; and T&lt;sub&gt;Y&lt;/sub&gt;</td>
</tr>
<tr>
<td>5</td>
<td>[T, P&lt;sub&gt;PCT&lt;/sub&gt;] = PCA(K)</td>
<td>Full rank PCA of K. T represents the scores of the original space. PPCT represents the PCT loadings (compressed space)</td>
</tr>
<tr>
<td>6</td>
<td>for a=1:PC unfold P&lt;sub&gt;PCTa&lt;/sub&gt; P&lt;sub&gt;a&lt;/sub&gt;=P&lt;sub&gt;Y&lt;/sub&gt; P&lt;sub&gt;PCTa&lt;/sub&gt; P&lt;sub&gt;X&lt;/sub&gt;&lt;sup&gt;T&lt;/sup&gt;</td>
<td>Rebuild each Principal Component's loadings (a) — Eq. (13)</td>
</tr>
</tbody>
</table>

R script available!
Principal component transform
Speeding-up processing without losing any information

R script available!
OP-PCT-PLS implemented in Cemagref DB (MOLTER initiative)
Merci !

Remerciements particuliers :

B. Barthès

L.T. Strand, G. Certini, C. Forte, H. Lange

N. Pascault, P.A. Maron

D. Rutledge, C. Cordella, A. Barros

F. Bray, S. De Danieli, E. Ancelet, A. Saenger, J.J. Brun

D. Rasse

Contact :

http://lauric.cecillon.free.fr/    cecillonlauric@yahoo.fr