Nonlinear data mining techniques or clustering to improve predictions of a large Brazilian vis-NIR library

Johanna Wetterlind, Bo Stenberg
Suzana Romero Aarújo, José Alexandre Demattê
Objectives

- Improve the prediction performance of SOM and clay content of a large heterogeneous spectroscopic library.
- Non-linear methods
- Reduce the variation by dividing the library into subsets
Vis-NIR library

- 7172 soil samples from four Brazilian states, *J. Demattê*
- SOM: 0.1% - 6.0%
- Clay content: 0.2% - 99%
- 350 – 2500 nm
3 strategies

1. Global PLSR model
2. Global nonlinear models:
   - Boosted regression tress (BT), Support vector machines (SVM)
3. Clustering – cluster-wise PLSR models
Clustering strategy

1. Calibration:
   Clustering of the calibration samples and build PLSR models for each cluster

2. Validation:
   Assign the validation samples to the clusters (discriminant analysis) and do the predictions using the cluster-wise PLSR-models
Clustering

- Based on the spectral features
  - $k$-mean clustering
- 3 transformations:
  1. 1$^{\text{st}}$ derivative - 4 clusters
  2. Mean normalized - 5 clusters
  3. Continuum removal - 4 clusters
Results

ultureom 
Measured Clay, % Measured OM, % 
Predicted Clay, % 
(a) 
(b) 
(c)

RMSEP reduced with 3 % compared with PLSR

RMSEP reduced with 10 % compared with PLSR
Results

**Clay**

**Predicted**

**Measured**

**Predicted OM, %**
**Measured Clay, %**
**Predicted Clay, %**

(a) 
(b) 
(c)

**SOM**

**Predicted**

**Measured**

RMSEP reduced with 17 % compared with PLSR

RMSEP reduced with 30 % compared with PLSR

**RMSEP**

R2 = 0.79
RMSE = 11.6
RPD = 2.27

R2 = 0.81
RMSE = 11.00
RPD = 2.30

R2 = 0.83
RMSE = 10.8
RPD = 2.35

R2 = 0.53
RMSE = 0.60
RPD = 1.45

R2 = 0.61
RMSE = 0.54
RPD = 1.60

R2 = 0.55
RMSE = 0.62
RPD = 1.44
# Prediction results

<table>
<thead>
<tr>
<th>Strategy</th>
<th>Clay</th>
<th></th>
<th>SOM</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>R²</td>
<td>RMSEP</td>
<td>R²</td>
<td>RMSEP</td>
</tr>
<tr>
<td>Clustering (normalized)</td>
<td>0.87</td>
<td>9.3</td>
<td>0.60</td>
<td>0.42</td>
</tr>
<tr>
<td>Clustering (1st derivative)</td>
<td>0.76</td>
<td>12.8</td>
<td>0.30</td>
<td>0.59</td>
</tr>
<tr>
<td>Clustering (continuum removal)</td>
<td>0.81</td>
<td>11.0</td>
<td>0.56</td>
<td>0.41</td>
</tr>
<tr>
<td>BT</td>
<td>0.83</td>
<td>10.8</td>
<td>0.61</td>
<td>0.54</td>
</tr>
<tr>
<td>SVM</td>
<td>0.81</td>
<td>11.0</td>
<td>0.55</td>
<td>0.62</td>
</tr>
<tr>
<td>PLSR</td>
<td>0.79</td>
<td>11.16</td>
<td>0.52</td>
<td>0.60</td>
</tr>
</tbody>
</table>
## Mean normalized spectra

<table>
<thead>
<tr>
<th>Cluster</th>
<th>n</th>
<th>Clay</th>
<th>SOM</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>R²</td>
<td>RMSEP</td>
</tr>
<tr>
<td>1</td>
<td>512</td>
<td>0.75</td>
<td>9.6</td>
</tr>
<tr>
<td>2</td>
<td>238</td>
<td>0.83</td>
<td>10.4</td>
</tr>
<tr>
<td>3</td>
<td>347</td>
<td>0.40</td>
<td>11.1</td>
</tr>
<tr>
<td>4</td>
<td>175</td>
<td>0.68</td>
<td>9.2</td>
</tr>
<tr>
<td>5</td>
<td>726</td>
<td>0.77</td>
<td>7.8</td>
</tr>
</tbody>
</table>
Discussion

Goethite

Hematite

Wavelength/nm

Continuum removed spectra

Cluster 1
Cluster 2
Cluster 3
Cluster 4
Cluster 5
Results and Discussion

Indicates smaller proportions of 1:1 minerals like kaolinite
• Reducing the variation in a very dominant feature such as soil mineralogy within the clusters enhanced the prediction performance of SOM content.

• It also improved the predictions of clay content but not to the same extent.

• Reduce the variation of the soil property of interest?
Conclusions

• Division of the large library into smaller subsets based on variation in mean normalized spectra was the best alternative for both clay and SOM.
• Clustering divided the library into more mineralogically uniform clusters.
• The additional step of assigning the validation samples to the right cluster did not add substantial to the prediction error.