New approaches of soil similarity analysis using manifold-based metric learning from proximal vis–NIR sensing data

Leonardo Ramirez-Lopez
Thorsten Behrens
Karsten Schmidt
Raphael Viscarra Rossel
Thomas Scholten

Second global workshop on Proximal Soil Sensing
May 17, 2011
Montreal

Soil vis–NIR and distance metrics

• Clustering
• Outlier detection
• k-Nearest neighbors
• Locally weighted regression
• Support vector machines
• Similarity search - assessing unknown soil samples, composition elucidation, finding proper calibration sets

Soil vis–NIR distances

• Clustering
• Outlier detection
• k-Nearest neighbours
• Locally weighted regression
• Support vector machines
• Similarity search - assessing unknown soil samples, composition elucidation

Soil vis–NIR similarity search

Motivation

Are really Xu samples similar to Xr samples in terms of soil attributes (e.g. clay content, mineralogy)?

The soil vis–NIR similarity should reflect the soil compositional similarity (at least those attributes that have strong influence on the soil vis–NIR spectra)

Which distance metric strategy should we use?

Distance metrics in proximal soil vis–NIR sensing

Usual methods
• Euclidean distance (ED) in the original variable space
• Mahalanobis distance (MD) in the original variable space
• Mahalanobis distance in the principal component space (PC distance):

Optimal compression depends on the explained variance of each component

Proposed methods
• Surface difference spectrum (SDS): works in the original feature space
• o-Linearly embedded (o-ALLE): works in the projected space and uses a SDS distance matrix to derive a new distance matrix
• Optimized PC distance (o-PC): works in the PC space. The selection of the number of PC features to retain is based on the soil compositional similarity
The Surface Difference Spectrum method (SDS)

The SDS method involves a multi-resolution (contextual) analysis of the Euclidean distance spectra (f) between two given soil spectrums \( X_r \) and \( X_u \)

\[
f(X_r, X_u) = \frac{1}{n} \sum_{i=1}^{n} (X_r(i) - X_u(i))^2
\]

from this spectral surface a new distance metric is formulated as:

\[
P = \frac{1}{n} \sum_{i=1}^{n} f(X_r, X_u)
\]

\( \alpha \)-Locally linear embedding (\( \alpha \)LLE)

The main goal is to discover geometric structures of high dimensional manifolds finding low dimensional and less complex representations of them

1. Select the neighbors of \( X_i \) by using the distance matrix information
2. Compute the weights for \( X_i \) reconstruction based on the differences
3. Map to embedded coordinates or low dimensional space
4. Re-compute the distance matrix

Optimized principal components (\( \alpha \)–PC) distance

• The goal is to optimize the number of PC features to retain
• In the \( \alpha \)–PC method a successive number of PCs is used to search for the most similar samples in the SSL.
• The correspondent compositional information of \( X_i \) (\( Y_i \)) is compared with the compositional information of the most similar samples found (\( Y_{\text{SSL}} \)). The RMSD is calculated. In this way we can find the optimum number of PC features that minimizes the RMSD (root mean square of compositional differences).

Results and conclusions

Australian soil vis–NIR dataset

• Unknown set: 278 samples (randomly selected)
• Reference set: 837 samples
• Compositional attribute: Clay content
• Algorithms:
  • ED
  • MD
  • SDS

<table>
<thead>
<tr>
<th>Method</th>
<th>Parameter</th>
<th>( R^2 )</th>
<th>RMSD</th>
</tr>
</thead>
<tbody>
<tr>
<td>ED</td>
<td></td>
<td>0.74</td>
<td>11.49</td>
</tr>
<tr>
<td>MD</td>
<td></td>
<td>0.74</td>
<td>21.29</td>
</tr>
<tr>
<td>SDS</td>
<td>10</td>
<td>0.74</td>
<td>8.71</td>
</tr>
</tbody>
</table>

• SDS returned the best search results in the spectral space outperforming the ED and MD methods and even the standard PC distance.

Mahalanobis distance returns poor results in the spectral space.

The \( \alpha \)–PC distance method reflects better the compositional similarity between samples than the standard PC method.

• By using the conventional selection of PC features important soil compositional information contained in the vis–NIR spectra can be lost.

• The best results of soil compositional similarity search were obtained by using the distance metric learning approach (\( \alpha \)LLE).
Conclusion remarks

- ED (in the spectral space): Very simple, easy to implement, assumes differences of each wavelength to have equal weighting.

- MD (in the spectral space): Very simple, easy to implement, does not reflect the compositional similarity. In some cases, multicollinearity in the vis-NIR data leads to a singular or nearly singular variance-covariance matrix that cannot be inverted and therefore the MD cannot be calculated.

- SDS: Takes into account the order of the wavelengths, easy to implement, works in the spectral space, returns vis-NIR distances which reflects the soil compositional similarity, good candidate for distance metric learning algorithms.

- PC: Low computational cost, well known technique, works in a projected space, could return non-reliable distances because PC features (containing important soil information) with low significance are ignored.

- o–PC: Low computational cost, better reflects the soil compositional similarity between spectra than the standard PC method.

- o–LLE: High computational cost, reliable distances, handles non-linearity.

Thank you for your attention

Questions?