
  
**ERSTHERR FRIEDRICH**  
**UNIVERSITÄT**  
**TÜBINGEN**

## 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


leonardo.ramirez-lopez@uni-tuebingen.de      Institute of geography, chair of physical geography and soil science

  
**ERSTHERR FRIEDRICH**  
**UNIVERSITÄT**  
**TÜBINGEN**

### 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


Leonardo Ramirez-Lopez      Institute of geography, chair of physical geography and soil science

  
**ERSTHERR FRIEDRICH**  
**UNIVERSITÄT**  
**TÜBINGEN**

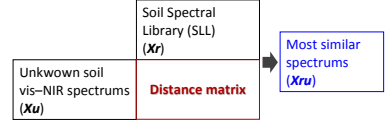
### Soil vis-NIR distances

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

Leonardo Ramirez-Lopez      Institute of geography, chair of physical geography and soil science

  
**ERSTHERR FRIEDRICH**  
**UNIVERSITÄT**  
**TÜBINGEN**

### Soil vis-NIR similarity search




**Motivation**

Are really  $Xu$  samples similar to  $Xru$  in terms of soil attributes (eg. 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?

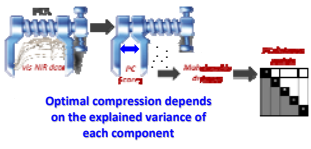
Leonardo Ramirez-Lopez      Institute of geography, chair of physical geography and soil science

  
**ERSTHERR FRIEDRICH**  
**UNIVERSITÄT**  
**TÜBINGEN**


### 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):**



Leonardo Ramirez-Lopez      Institute of geography, chair of physical geography and soil science

  
**ERSTHERR FRIEDRICH**  
**UNIVERSITÄT**  
**TÜBINGEN**

### Distance metrics in proximal soil vis-NIR sensing

**Proposed methods**

- **Surface difference spectrum (SDS):** works in the original feature space
- **$\sigma$ -Locally linear embedding ( $\sigma$ LLE):** works in an projected space and uses a SDS distance matrix to derive a new distance matrix
- **Optimized PC distance ( $\sigma$ -PC):** works in the PC space. The selection of the number of PC features to retain is based on the soil compositional similarity

Leonardo Ramirez-Lopez      Institute of geography, chair of physical geography and soil science

**The Surface Difference Spectrum method (SDS)**

The SDS method involves a multi-resolution (contextual) analysis of the Euclidean distance spectra ( $E$ ) between two given soil spectrums  $Xr$  and  $Xu$

$$SDS(\sigma) = \begin{cases} E(Xr, Xu) = (Xr_1 - Xu_1)(Xr_1 - Xu_1)^T & \text{if } \sigma = 0 \\ A(E, E) = \frac{1}{\sigma} (E_i - E_{i+\sigma})(E_i - E_{i+\sigma})^T & \text{otherwise} \end{cases}$$

Wavelengths  
Frequency delay

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

$$D = \sum_{i=1}^n \sum_{j=1}^n W_{ij} SDS$$

optimal  $\sigma$   
Number of wavelengths  
Weight obtained from a robust correlation analysis

Leonardo Ramirez-Lopez Institute of geography, chair of physical geography and soil science

**An example...**

Euclidean D. Spectrum  
Euclidean D. Spectrum  
Euclidean D. Spectrum

Leonardo Ramirez-Lopez Institute of geography, chair of physical geography and soil science

**$\sigma$ -Locally linear embedding ( $\sigma$ 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

Roweis and Saul (2000)

Leonardo Ramirez-Lopez Institute of geography, chair of physical geography and soil science

**Optimized principal components ( $\sigma$ -PC) distance**

- The goal is to optimize the number of PC features to retain
- In the  $\sigma$ -PC method a successive number of PCs is used to search for the most similar samples in the SSL.
- The correspondent compositional information of  $Xr$  ( $Yr$ ) is compared with the compositional information of the most similar samples found ( $Yrr$ ). 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).

Leonardo Ramirez-Lopez Institute of geography, chair of physical geography and soil science

**The experiment...**

*Australian soil vis-NIR dataset*

- Unknown set: **278** samples (randomly selected)
- Reference set: **837** samples
- Compositional attribute: **Clay content**
- Algorithms:
  - ED
  - MD
  - SDS
  - PC
  - $\sigma$ -PC
  - $\sigma$ LLE

Leonardo Ramirez-Lopez Institute of geography, chair of physical geography and soil science

**Results and conclusions**

| Method                                 | Param. value | R <sup>2</sup> | RMSD  |
|--|--------------|----------------|-------|
| <i>Searching in the spectral space</i> |              |                |       |
| ED                                     | -            | 0.65           | 11.49 |
| MD                                     | -            | 0.24           | 21.29 |
| SDS                                    | 10           | 0.74           | 9.71  |
| <i>Searching in a projected space</i>  |              |                |       |
| PC                                     | 3            | 0.64           | 11.67 |
| $\sigma$ -PC                           | 8            | 0.75           | 9.73  |
| $\sigma$ LLE                           | 35           | 0.76           | 9.57  |

- 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  $\sigma$ -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 ( $\sigma$ LLE).

Leonardo Ramirez-Lopez Institute of geography, chair of physical geography and soil science

UNIVERSITÄT TUBINGEN

### 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 can not 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
- **$\sigma$ -PC:** low computational cost, better reflects the soil compositional similarity between spectra than the standard PC method
- **$\sigma$ LLE:** High computational cost, reliable distances, handles non-linearity.

Leonardo Ramirez-Lopez      Institute of geography, chair of physical geography and soil science

UNIVERSITÄT TUBINGEN

# Thank you for your attention

---

## Questions?

Leonardo Ramirez-Lopez      Institute of geography, chair of physical geography and soil science

UNIVERSITÄT TUBINGEN

Leonardo Ramirez-Lopez      Institute of geography, chair of physical geography and soil science

UNIVERSITÄT TUBINGEN

Leonardo Ramirez-Lopez      Institute of geography, chair of physical geography and soil science

UNIVERSITÄT TUBINGEN

### Results

*Parameter optimizations of our proposed approaches*

**SDS**      Max. Frequency delay  
 **$\sigma$ LLE**      Number of neighbors (k)  
 **$\sigma$ -PC**      PC features

Leonardo Ramirez-Lopez      Institute of geography, chair of physical geography and soil science