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

L. Ramírez–López[1]*, T. Behrens[1], K. Schmidt[1], R. Viscarra Rossel[2], T. Scholten[1]
[1]*Institute of Geography, Physical Geography, University of Tübingen, Rümelinstraße 19–23, 72074, Tübingen, Germany.*
[2]*CSIRO Land and Water, Bruce E. Butler Laboratory, GPO Box 1666 Canberra ACT 2601, Australia*
leonardo.ramirez-lopez@geographie.uni-tuebingen.de

Abstract

In this presentation we introduce two new methods of similarity search and evaluate the performance of commonly used distance metrics compared to our approaches. The first method uses a surface difference spectrum (SDS) and works in the spectral space. The second one works in a projected space and is based on the SDS–locally linear embedding algorithm. We also propose a parameter optimization for a principle components distance. To test our approaches we used an Australian soil vis–NIR spectral library. The performance of the methods was evaluated by their ability to identify spectrums with similar clay content.

Keywords: Manifold learning, distance metric learning, dimensionality reduction, spectral similarity, soil spectroscopy.

Introduction

In vis–NIR spectroscopy based proximal soil sensing (PSS) distance metrics and similarity search play key roles in assessing unknown soil samples, composition elucidation, finding proper calibration sets, and outlier detection. Choosing an appropriate distance metric is an important step for the success of many applications of PSS. Much of the soil vis–NIR research efforts have been focused in improving soil predictions. Despite this, few studies have been carried out to establish adequate distance metric algorithms for similarity search in soil vis–NIR data sets. In this work, we introduce two new distance metric approaches for similarity search in vis–NIR data. The first one, called spectral difference surface (SDS), is based on a multi-resolution analysis in the spectral space of differences between samples. The second one is built on manifold–based distance metric learning, which learns the SDS distance by using the locally linear embedding (σLLE) algorithm. Distance metric learning and manifold learning are recent and fast growing research areas in data mining (Weinberger et al., 2010; Qiao et al, 2011). In distance metric learning the underlying metric is itself adapted to improve the results of classification and pattern recognition (Weinberger et al., 2010). Several works in machine learning have demonstrated that distance metrics learning leads to substantial improvements over the commonly used metrics (Mordohai and Medioni, 2010). Regarding the principle components distance method, we will show empirically that the amount of the total variance explained by PC features is not a good indicator of the optimal number of PC features to retain in the case of soil vis–NIR data. In this respect we also propose a simple framework to identify the adequate number of PC features to retain based on a optimization method.

Algorithms

Surface Difference Spectrum – SDS:
The SDS method involves a multi-resolution analysis of the Euclidean distance spectra (E) between two soil spectrums X and Y. An auto–distance function (A) is applied on E as a

function of frequency or wavelength delay (σ) returning a 3D spectrum of differences. In SDS the only parameter that needs to be set is σ. Outputs of A may be interpreted as multiple derivative energy spectrums of the spectral difference between X and Y.
SDS–Locally Linear Embedding – σLLE: The standard LLE was introduced first by Roweis and Saul (2000). Basically LLE is an unsupervised metric learning algorithm, which learns the global manifold structure from local neighborhoods. Unsupervised metric learning is usually referred as manifold learning (ML). The ML concept was introduced simultaneously by Seung and Lee (2000), Roweis and Saul (2000) and Tenenbaum et al. (2000). In ML is assumed that high–dimensional data lie on or close to a low–dimensional smooth manifold (Qiao et al., 2010). The main goal of ML algorithms is to discover geometric structures of high dimensional manifolds finding low dimensional and less complex representations of them. In this respect, LLE works as a non–linear dimensionality reduction (or projection) method and it is carried out in three main steps : 1. Selecting neighbors: In this step the Euclidean distance is used to find the k–nearest neighbors X_j of each data point X_i; 2. Computing a weight matrix: Here a weight matrix (W) is computed in order to find the optimal reconstruction of X_i by its neighbors (X_j); 3. Computing low–dimensional coordinates. In LLE there is only one free parameter that needs to be set: the number of neighbors (k). Here we have used the SDS distance for neighbor search instead the conventional Eucidean distance (ED) used in the LLE algorithm. Here, our version the standard LLE is called σLLE method.

Material and methods

We used an Australian soil spectral library (SSL) which includes 1115 samples. The soils were diverse and represented by various Australian Soil Classification orders. We randomly selected 278 samples from the SSL as unknown set (Xu). The remaining samples (837) were used as reference set (Xr). Distances between samples were computed by our proposed approaches: the SDS and the σLLE. We compared these to the commonly used methods in PSS: Euclidean Distance (ED), Mahalanobis Distance (MD) and PC distance. In the ED and MD methods, distances are calculated directly in the spectral space. In the PC method, distances are computed on the projected PC scores space. In this projected space the Mahalanobis distance is used. In the PC distance method the choosing of PC features to retain is based on the explained variance of components. We also proposed and used an optimized PC distance (o–PC distance) method. The o–PC distance is based on a very simple parameter optimization framework to determining the number of PC features to retain before the distance computations. For samples in *Xu*, we searched their most similar samples in *Xr*. We compared the clay content of the Xu samples and the clay content of the most similar samples found. We have used this soil attribute because it has a strong effect on the vis–NIR reflectance intensity (Demattê et al., 2004). The root mean square of differences (RMSD) of cross validations was used as parameter to evaluate the performance of the methods tested. To find the optimal parameters of SDS (number of wavelengths delays, *σ*), o–PC (number of PC features, PCs) and σLLE (number of nearest neighbors, *k*) we propose a framework based on the minimization of the RMSD. Let *h* be the parameter to optimize in each method (either *σ,* PCs or *k*). A successive number of *h* is used to search for the most similar samples in the reference set. So that, for each sample in the reference set its correspondent similar sample in the same set is found as function of *h*. In this way we can find the optimum *h* that minimizes the RMSD.

Results and discussion

Using the SDS method we found that the distance metric for similarity searching can be improved gradually by increasing the number of wavelengths delays (σ) involved in the distance computation. Figure 1a shows that around 12 frequency delays (σ=12) were necessary to

reduce the RMSD of searching samples with similar clay content. This shows that new important information about the spectral similarity between samples is emerging when the context in neighborhood wavelengths is taken into account. We selected 12 as optimum σ to run the distance computations of for similarity searching. For the o–PC method we found that 8 was the optimum number of features to compute the distance matrix. By using the σLLE algorithm we observed that the RMSD can be reduced gradually by increasing the SDS–neighbors (Figure 1c). We used 35 as optimum number of k.

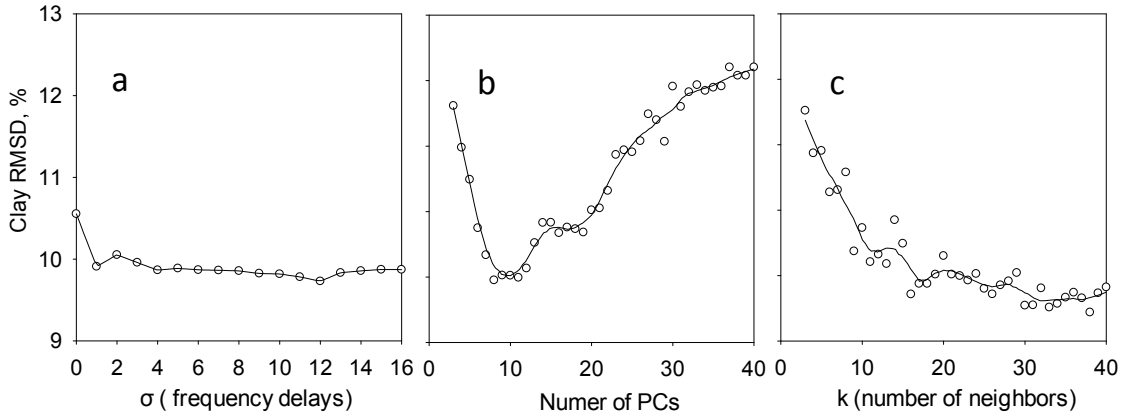

Figure 1. Root mean square of differences (RMSD) of the different methods which required parameter optimization. a. SDS distance; b. o–PC distance and c. σLLE distance

Once the parameters of the SDS, o–PC and σLLE methods were optimized we searched for the most similar samples of unknown samples (Xu) in the reference set (Xr). Results are presentenced in Table 2.

Table 2. Results of the most similar sample search.

Method	Param. value	R^2	RMSD
	Searching in the spectral space		
ED	-	0.65	11.49
MD	-	0.24	21.29
SDS	10	0.74	9.71
	Searching in a projected space		
PC	3	0.64	11.67
o–PC	8	0.75	9.73
σLLE	35	0.76	9.57

The soil similarity search approach that returned the best results was σLLE (RMSD = 10.83%). For searching methods in the original spectral space the best results were obtained by the SDS method. On the other hand, the SDS method returned also better results than those obtained with the standard PC method and similar to those returned by the o–PC method. In the projected space we found that the standard PC distance method can be improved by using our proposed parameter optimization. The RMSD of the o–PC search was lower than the RMSD obtained with the standard PC method. This confirms that important soil information is also contained in the first PC features with low explained variance. For the unsupervised distance metric learning method (σLLE), we found that it is a reliable method for similarity search.

Conclusions

The results show that the SDS returned the best soil compositional search results in the spectral space outperforming the ED and MD methods and even the standard PC distance. The SDS method returns similar results to the o–PC method. Except the standard PC, the projection methods (o–PC and σLLE) have a good similarity search performance. The reason for this is that the projected geometric structures are less complex representations of the original spectral space. The standard PC approach is not an adequate method since some PC features containing important information on soil compositional are ignored. In light of our results, we suggest that the o–PC method for distance computations should be preferred over the standard PC method. However, the best results of soil compositional similarity search were obtained by using the σLLE approach. In general the unsupervised distance metric learning approach (σLLE) can learn the SDS distance calculated in the spectral space and return better results because the reduction of the original spectral complexity. On the other hand, the SDS distance can be used to reduce the complexity of the vis–NIR data by using the LLE algorithm.

References

Demattê, J.A.M., Campos, R.C., Alves, M.C., Fiorio, P.R., Nanni, M.R., 2004. Visible–NIR reflectance: a new approach on soil evaluation. Geoderma, 121: 95–112.

Mordohai , P. and Medioni, G. 2010. Dimensionality estimation, manifold learning and function approximation using tensor voting. Journal of Machine Learning Research. v 11. pp 411–450.

Qiao, H., Zhang, P., Zhang, B., Zheng, S. 2011. Tracking feature extraction based on manifold learning framework. Journal of Experimental & Theoretical Artificial Intelligence, 23: 23–38.

Roweis, S.T., Saul, L.K. 2000. Nonlinear dimensionality reduction by locally linear embedding. Science, 290: 2323–2326.

Seung, H.S. & Lee, D.D. 2000. The manifold ways of perception. Science 290:2268-2269.

Tenenbaum, J.B., de Silva, V., Langford, J.C. 2000. A global geometric framework for nonlinear dimensionality reduction. Science, 290: 2319–2323.

Weinberger, K. Q., Sha, F., Saul, K. 2010. Convex optimizations for distance metric learning and pattern classification. IEEE Signal Processing Magazine, 27(3):146–158.