Improved spectral estimation of multiple soil properties by stratification on ancillary and spectral data

H.M. Bartholomeus[1]*, C.R. Sabajo[1], A. van Groenestijn[2], and L. Kooistra[1]
[1]*Wageningen UR – Centre for Geo-Information, Wageningen, The Netherlands*
[2]*Adviestalent.nl, Amersfoort, The Netherlands*
Harm.Bartholomeus@wur.nl

Abstract

Visible and Near Infrared spectroscopy has proven to be a valuable tool for rapid estimation of multiple soil properties, but wide implementation is limited due to low robustness of developed models. We investigated if a large and highly variable spectral dataset in combination with stratification methods can be used to get good estimates of within field variation of soil properties. The results show that stratification by ancillary or spectral data can be used to improve the estimates of a large number of soil properties and result in models that are robust enough to be used for minimally pre-processed samples.

Keywords: spectral database, stratification, soil properties

Introduction

Visible and Near Infrared spectroscopy (VNIRS) has proven to be a valuable tool for rapid estimation of multiple soil properties, both in the laboratory (Sørensen and Dalsgaard 2005) and directly in the field (Stevens et al. 2008). In recent years, chemometric techniques like partial least squares regression (PLSR) and Multiple linear regression (MLR) have frequently been used in VNIR soil spectroscopy (Stevens et al. 2008; Viscarra Rossel et al. 2006). These models perform well under controlled conditions, and yield good predictions on local scale. However, true implementation of VNIRS in proximal soil sensing is limited because of the low robustness of chemometric techniques and consequently, the use of many local calibrations. The development of large (regional to global) soil spectral libraries has been proposed as a possible solution for this problem. But on datasets with a large range in soil properties the prediction accuracy drops dramatically, because of non-linearity in the relation between reflectance and the soil properties. This limits the use of large spectral databases for proximal soil sensing for local applications. However, for smaller ranges of soil properties linear functions can describe the relation well (Bartholomeus et al. 2008).
In this paper we discuss the use of stratification by means of thematic and spectral criteria to improve the estimates of multiple soil properties with soil spectral measurements. This is first examined by means of a soil spectral database of the Netherlands, covering a wide variety in soil types, mineralogical backgrounds and soil properties. Finally, to examine if the developed methods can improve the estimates in a proximal sensing setup, we tested if the stratification procedures improve the within field estimation of soil organic matter (SOM) of an experimental field, which includes several soil types. .

Methodology

For this study, we used two datasets. Dataset 1 consists of a large number of soil samples (N=422) spread over The Netherlands, covering a large range in soil types and properties. Samples were collected in natural environments, taking 4 samples with a gouge to a depth of 10 cm which were mixed. From this mixture a subsample was taken for chemical analysis and spectral measurements. Samples were dried and sieved (2 µm) before spectral reflectance was measured with an ASD Fieldspec Pro FR in combination with an ASD contact probe. The spectral range from 450-2500 nm was used for analysis. Soil samples were chemically analyzed Nitrogen (Nt), total Phosphorus (Pt), potassium (K), Magnesium (Mg), Ammonium-Nitrogen (N-NH4), Nitrate-Nitrogen (N-NO3/NO2), Soluble Nitrogen (Nts), Aluminum (Al), Calcium (Ca), SOM and pH. Furthermore, to test the applicability of the stratification methods for proximal soil sensing, we used spectra of 51 locations collected from a single field close to Oud-Annerveen in the North of the Netherlands (Dataset 2), which were analyzed for SOM.

Dataset 1 was divided in a calibration (N=371) and validation set (N=51), both roughly covering the same range between minimum and maximum SOM content. Since many other soil properties are correlated with SOM this implies that the ranges for many other soil properties are comparable for both sets as well. First, to decrease the dimensionality of the dataset, the spectral data were rotated using a principal component analysis (PCA). The first 20 components were kept and used further for clustering and spectral analysis. Next, a reference model was made, based on a Stepwise Multiple Regression (SMLR), using all the samples in the calibration set and using 20 principal components. Finally, the dataset was stratified in a number of subclasses, according to three different criteria. First, we used the texture classes (sand, clay, peat), derived from the soil type map of the Netherlands. Second, we used a two-step approach, in which the values predicted by the reference model were subdivided in a variable number of clusters by means of k-means clustering. The third method was based on k-means clustering of the 20 principal components. PCA, k-Means clustering and SMLR were done in the software package R, . First, the samples in the training set were clustered. For the samples of the reference dataset, the Euclidian distance to all cluster centers was calculated and the samples were assigned to the class with the minimum Euclidian distance. The soil properties were related to the PCA transformed spectral properties by means of SMLR. The initial model included 20 principal components, which was cut-back in forward and backward direction. For each cluster and each soil property a separate model was fitted, validation samples were assigned to the clusters and values for the soil properties were estimated. Performance of the stratification methods was evaluated by means of the Ratio of Performance to Deviation (RPD), where the levels defined by Chang et al. (2001) were used to evaluate the models.

To show that large spectral databases can serve as a calibration for field studies and that the use of stratification can improve the estimate of the within field variation of SOM, we applied the models stratification methods to Dataset 2, which includes both peat and sandy soils.

Results and Discussion

The RPD of the stratified models is compared to the RPD of the model with no stratification (1 cluster). Practically all clustering methods show an increase in RPD during the calibration phase (results not shown). In Table 1, the RPD's of the validation set are given, which show that stratification can increase the RPD for all measured soil properties, although for some soil properties the increase in accuracy is limited (K, Mg, N) and the RPD remains lower than 2. For SOM and pH the increase in prediction accuracy is the strongest and all investigated clustering methods do improve the estimation of the soil variable. Overall, using a 2-step approach results in the best estimates. The improvement of the estimated pH and SOM values for the validation set is graphically shown in figure 1.

Predictions of the SOM content for Dataset 2 show that stratification strongly improves the estimated SOM, compared to the use of a single model. The single model overestimates the samples with low SOM content and underestimates the samples with a high SOM content and is therefore not well able to determine the within parcel variation. Stratified models (figure 2) are much better able to determine the within field variation in SOM ($RPD_{\text{single cluster}} = 1.16$, $RPD_{\text{clustering by soiltype}} = 2.00$, $RPD_{\text{2step, 2 clusters}} = 1.75$). Since these samples were only air-dried and measured under different illumination conditions than the calibration data, this gives a good impression of how well VNIRS in combination with large spectral databases for calibration can be used for proximal soil sensing.

Table 1: Ratio of Performance to Deviation (RPD) values for stratified models by means of different clustering procedures and cluster sizes. The green cells indicate an improved performance, compared to the prediction by means of the non-stratified reference models.

Attribute	-	Soil Type	2-Step	2-Step	PC1-20	PC1-20
# clusters	1	3	2	3	2	3
Nt	1.52	1.45	1.69	**1.82**	1.72	1.72
Pt	2.22	1.92	2.50	2.51	**2.55**	2.10
K	1.29	**1.36**	1.13	1.06	1.01	0.91
Mg	1.33	**1.45**	1.35	1.31	1.31	1.32
N-NH4	1.35	1.28	1.31	**1.46**	1.30	1.39
N-NO3/NO2	1.16	1.13	1.17	**1.18**	1.16	1.10
Nts	1.38	1.46	1.41	1.43	**1.47**	1.34
Al	1.67	1.65	2.19	2.26	2.20	**2.28**
Ca	1.76	1.92	**1.99**	1.89	1.89	1.65
SOM	1.62	1.93	**2.75**	1.71	1.79	1.96
pH	2.46	2.72	**3.11**	2.81	3.00	2.95

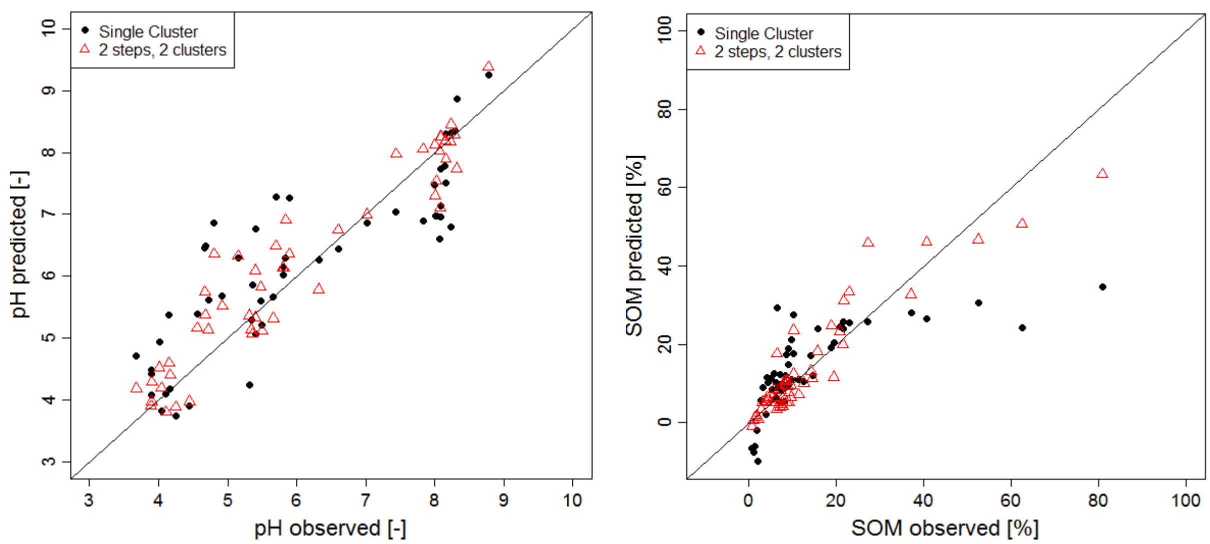

Figure 1: Scatterplots of observed vs. predicted values of the Dataset 1 validation samples for pH and SOM, displaying the estimates by stratified and non-stratified models.

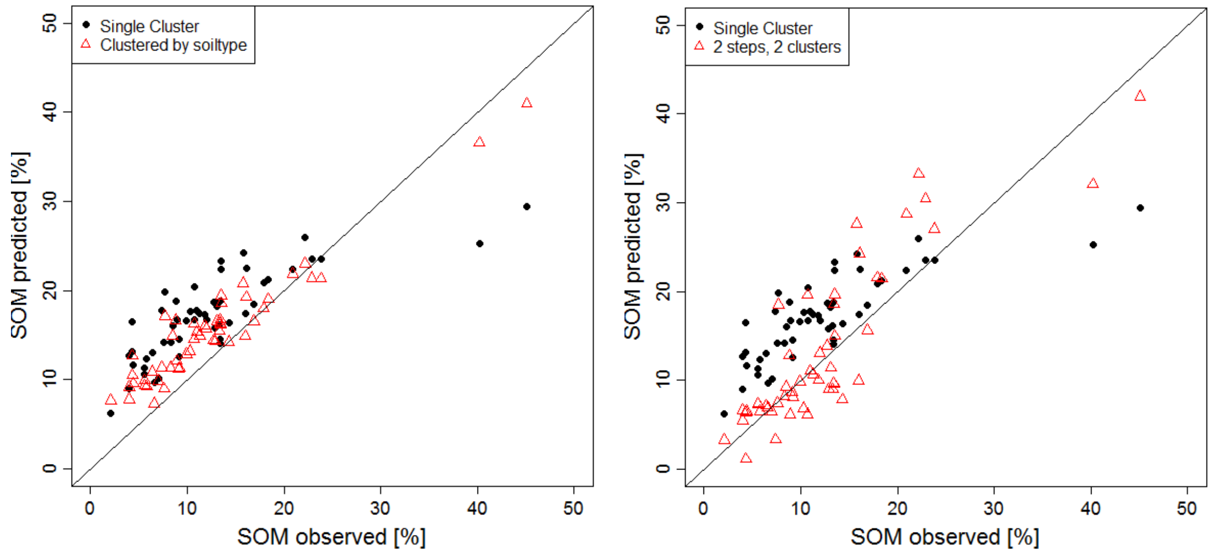

Figure 2: Observed vs. predicted SOM content of 51 soil samples within the test field (Dataset 2), determined by stratification based on soil type (left) and stratification based on the two-step approach with 2 clusters (right).

Conclusions

Our research shows that the stratification of VNIRS models, in combination with SMLR can strongly improve the estimate of multiple soil properties, compared to the use of a single model. In general, a two-step approach yields the best results. With appropriate stratification methods, the full potential of large spectral databases for proximal soil sensing can be used.

References

Bartholomeus, H.M., Schaepman, M.E., Kooistra, L., Stevens, A., Hoogmoed, W.B., & Spaargaren, O.S.P. (2008). Spectral reflectance based indices for soil organic carbon quantification. Geoderma, 145 (1-2), 28-36

Chang, C.W., Laird, D.A., Mausbach, M.J., & Hurburgh Jr., C.R. (2001). Near-infrared reflectance spectroscopy - Principal components regression analysis of soil properties. Soil Science Society America Journal, 65, 480-490

Sørensen, L.K., & Dalsgaard, S. (2005). Determination of clay and other soil properties by Near Infrared Spectroscopy. Soil Science Society America Journal, 69, 159-167

Stevens, A., Wesemael, B.v., Bartholomeus, H., Rosillon, D., Tychon, B., & Ben-Dor, E. (2008). Laboratory, field and airborne spectroscopy for monitoring organic carbon content in agricultural soils Geoderma, 144 (1-2), 395-404

Viscarra Rossel, R.A., Walvoort, D.J.J., McBratney, A.B., Janik, L.J., & Skjemstad, J.O. (2006). Visible, near infrared, mid infrared or combined diffuse reflectance spectroscopy for simultaneous assessment of various soil properties. Geoderma, 131, 59-75.