Using soil spectral libraries in support of proximal soil sensing

Y. Ge[1], C.L.S. Morgan[2]*, J.A. Thomasson[1]
[1]*Biological & Agricultural Engineering, Texas A&M University, TX 77843-2117, USA*
[2]*Soil & Crop Sciences, Texas A&M University, TX 77843-2474, USA*
cmorgan@ag.tamu.edu

Abstract

The effectiveness of using the Texas Soil Spectral Library (TSSL) to predict soil sample constituents from local Texas fields was investigated in the context of proximal soil sensing. Two sample selection methods (Kennard-Stone algorithm and spectral angle mapper method) and boosting the calibration model with local samples were also investigated in an attempt to improve prediction performance. Results showed that the calibration model from the complete TSSL could predict soil organic carbon in local samples satisfactorily, with relative percent difference greater than 1.40. The Kennard-Stone algorithm consistently improved the accuracy for all local fields, while the spectral angle mapper method and boosting samples had mixed results, with prediction greatly improved for some fields but reduced for others. It is concluded an existing soil spectral library can be helpful in proximal soil sensing, particularly when the application is qualitative or semi-quantitative.

Keywords: multivariate, organic carbon, partial least squares, spectral similarity.

Introduction

One of the essential components of a successful optical proximal soil-sensing (PSS) system is a reliable and accurate calibration model that relates soil spectral information to various soil constituents. The common method of spectral model calibration (as in VisNIR spectroscopy) requires samples with known constituents and spectra in advance, a situation neither practical nor suitable for use with PSS. Alternatively, an existing spectral library can be used for in-field model calibration. However, the concern is that poor performance would result from apparent discrepancies between library and field samples in terms of parental material, pedological features, clay mineralogy, and other spectral characteristics, not to mention the effect that variations in moisture content and aggregate size can have.

“Boosting” a generalized spectral library with local samples has been studied by several researchers (Brown, 2007; Sankey et al, 2008). Local samples are usually relatively similar to the target samples in many aspects, so including them should improve prediction accuracy by making the calibration dataset more representative. On the other hand, some researchers (particularly in agricultural products such as grain) have investigated selecting a subset of a spectral library for model calibration (Bouveresse and Massart, 1996). The idea in this case is to remove extraneous data in the spectral library and thereby improve the representativeness of the calibration set.

The overall goal of this study was to investigate the usefulness of a soil spectral library in support of PSS and to consider specific methods of use to improve prediction accuracy. Specific objectives were (1) to use the Texas Soil Spectral Library (TSSL) to predict soil sample constituents from local fields, and (2) to compare methods to improve prediction performance of TSSL.

Materials and methods

TSSL consists of ~2300 soil samples from various regions of Texas. Six field-scale local datasets (50 samples each) from Erath (3 fields), Comanche (2), and McLennan (1) counties were equally split for boosting and model validation. All soil samples were scanned with an ASD (Boulder, CO) AgriSpec spectrometer from 350 to 2500 nm at a 1-nm interval. Each soil spectrum was preprocessed with a custom algorithm (Brown et al., 2005) to obtain the first derivative spectrum at 10-nm interval. In this study the focus was given to soil organic carbon (OC) measured by subtracting Inorganic C (modified pressure calcimeter) from total C (dry combustion). Though the soils in the TSSL and the six local datasets were scans of dried and ground samples, this type of exercise could also be performed on samples collected in situ. However we have not collected a large enough in situ Texas Spectral Library for this type of exercise.

The methods used to improve prediction performance of TSSL involved subsetting the data and boosting the data. Subsetting the data was done with two techniques, the Kennard-Stone algorithm (K-S, Kennard and Stone, 1969) and the spectral angle mapper (SAM) method. First, the K-S algorithm was implemented while retaining 10, 20, …, and 90% of samples in the library, and the best retention percentage was identified by cross-validation and tested on the local samples. Second, the spectral angle mapper (SAM) method was implemented to select 75 library samples having the smallest SAM distance to the test samples to form a calibration set. Boosting the library with local samples involved including additional local samples in with the K-S or SAM selected samples for calibration. In total, five calibration models were developed and compared: library only, K-S, K-S+Boost, SAM, and SAM+Boost. The RMSE (root mean squared error) and RPD (relative performance deviation) statistics were used for model assessment. Partial least squares regression was used for model calibration, and data analyses were performed in the R computing environment.

Results and discussion

TSSL had greater variation in OC (in terms of both range and CV) but a smaller mean than all local fields (table 1). The RPD of the OC models with different K-S retention percentages in the calibration set (figure 1) indicates that when the percentage is small (≤ 30%, or 690 samples), the performance of TSSL is very poor. Performance improves greatly when the retention percentage increases from 40 to 70% and drops when more than 80% of the samples are retained. This behavior suggests redundancy exists in TSSL, and the redundant samples reduce prediction accuracy. The optimal size of TSSL appears to be 70% of its original size (or 1600 samples).

Table 1. Summary statistics of soil organic carbon (g kg^{-1}) in Texas Soil Spectral Library and six local fields.

Field	N	Min.	Max.	Mean	SD	CV (%)
McLennan	50	0.0	27.0	8.5	6.8	80
Erath1	50	0.3	55.9	14.5	11.2	77
Erath2	50	0.6	47.7	8.6	9.0	105
Comanche1	50	0.0	48.2	7.9	9.1	115
Comanche2	50	0.1	39.7	10.2	8.7	85
Erath3	50	1.4	49.5	12.6	11.6	93
TSSL	2298	0.0	88.5	6.4	8.8	139

Min. = Minimum value; Max. = Maximum value; SD = Standard deviation; CV = Coefficient of variation.

The RMSE and RPD statistics of the OC calibration models indicate that the five different methods behave differently in the different fields (table 2). With only the TSSL data included, all fields except for Comanche1 had an RPD greater than 1.4, the threshold suggested by Chang et al. (2001) for useful calibration models in soil VisNIR applications. This level of accuracy is generally adequate for detecting differences in OC across a landscape. These results therefore provide evidence that a soil spectral library can be helpful for supporting PSS applications at the landscape and watershed scale.

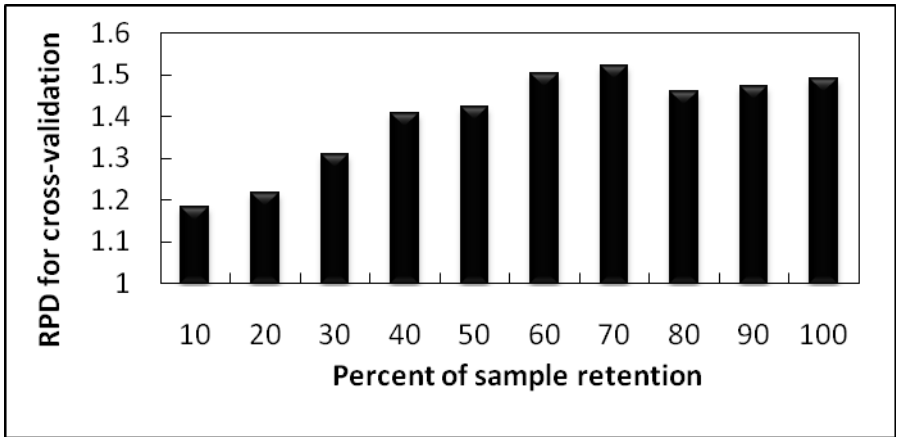

Figure 1. The relative percent difference (RPD) of the soil organic carbon model with different retention percents using the Kennard-Stone algorithm.

Table 2. Root mean squared error (g kg^{-1}) and relative percent difference (in parentheses) of five soil organic carbon calibration models for six local fields.

Field	Library only	K-S	K-S+Boost	SAM	SAM+Boost
McLennan	3.9 (1.61)	3.9 (1.62)	4.1 (1.55)	4.7 (1.34)	4.3 (1.47)
Erath1	9.2 (1.44)	8.9 (1.48)	8.8 (1.50)	6.2 (2.11)	6.6 (1.99)
Erath2	5.0 (1.57)	5.0 (1.57)	4.5 (1.71)	5.0 (1.54)	4.5 (1.71)
Comanche1	6.9 (1.61)	6.8 (1.63)	6.8 (1.64)	5.9 (1.86)	5.9 (1.87)
Comanche2	4.7 (1.95)	4.5 (2.04)	4.6 (1.99)	5.4 (1.68)	5.8 (1.57)
Erath3	5.7 (1.45)	5.4 (1.53)	5.3 (1.55)	5.8 (1.42)	5.0 (1.65)

K-S = the Kennard-Stone algorithm (70% retention); SAM = Spectral Angle Mapper.

Comparing the other methods to using the entire TSSL, slight but consistent improvements in prediction accuracy were seen with the K-S algorithm (70% retention) in all test fields. The SAM method improved prediction performance substantially for Erath1 and Comanche1, but RMSE increased substantially for Comanche2. It is possible that an analysis of the spectral structure of each validation set relative to that of TSSL might shed light on the inconsistency of the SAM method. Boosting samples were not very effective when used with the K-S method; only minor improvements in a few fields were found. On the other hand, noticeable improvements were obtained when boosting samples were used with the SAM method for Erath2 and Erath3 fields. When considering all five different methods, SAM+Boost gave the best overall performance for local prediction, but the results were not consistent from field to field.

If one is conducting a study with PSS, local boost samples with known constituents may not be available. If they are available, they are usually in small quantities compared to the number of samples in a spectral library. Therefore, it is advisable to use a spectral selection method to select a smaller subset of samples and then incorporate boost samples (as was done with SAM

+ Boost). An alternative is to implement weighted regression with more weight on the boost samples. It is noteworthy that K-S and SAM are two calibration selection methods based solely on soil spectral information. Other methods allow selection of an appropriate calibration set for a specific group of target samples (e.g., soil series or Major Land Resource Area). These methods have potential in PSS applications but require a more sophisticated library design to include such information.

Conclusions

The major conclusions drawn from this study are as follows.

1. Using TSSL to predict soil OC in local fields yielded satisfactory (RPD > 1.4) results. The TSSL would be useful in support of PSS applications that are qualitative a/o semi-quantitative in nature and at the landscape or watershed scale.
2. The K-S algorithm maintained or improved prediction accuracy for all validation fields.
3. Inconsistent results were obtained with the SAM method.
4. Boosting with local samples resulted in significant improvements in prediction for three fields when used with SAM methods. However, boosting efficacy was less apparent when used with K-S selection methods.

References

Bouveresse, E., D.L. Massart. 1996. Improvement of the piecewise direct standardization procedure for the transfer of NIR spectra for multivariate calibration. Chemometrics and Intelligent Laboratory Systems, **32** 201-213.

Brown, D.J. 2007. Using a global VNIR soil-spectral library for local soil characterization and landscape modeling in a 2nd-order Uganda watershed. Geoderma, **140** 444-453.

Brown, D.J., Shepherd, K.D., Walsh, M.G., Mays, M.D., Reinsch, T.G. 2006. Global soil characterization with VNIR diffuse reflectance spectroscopy. Geoderma, **132** 273-290.

Chang, C., 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 of American Journal, **65** 480-490.

Kennard, R.W., and Stone, L.A., 1969. Computer aided design of experiments. Technometrics, **11** 137-148.

Sankey J.B., Brown, D.J., Bernard, M.L., Lawrence, R.L. 2008. Comparing local vs. global visible and near-infrared (VisNIR) diffuse reflectance spectroscopy (DRS) calibrations for the prediction of soil clay, organic C and inorganic C. Geoderma **148** 149-158.