2010

Evaluation of spectral pretreatments, partial least squares, least squares support vector machines and locally weighted regression for quantitative spectroscopic analysis of soils

Benoit Igne
Iowa State University, benoitigne@yahoo.fr

James B. Reeves
United States Department of Agriculture

Gregory McCarty
United States Department of Agriculture

W. Dean Hively
United States Department of Agriculture

Eric Lund
Vis Technologies

See next page for additional authors

Evaluation of spectral pretreatments, partial least squares, least squares support vector machines and locally weighted regression for quantitative spectroscopic analysis of soils

Abstract
Soil testing requires the analysis of large numbers of samples in the laboratory that is often time consuming and expensive. Mid-infrared spectroscopy (mid-IR) and near infrared (NIR) spectroscopy are fast, non-destructive and inexpensive analytical methods that have been used for soil analysis, in the laboratory and in the field, to reduce the need for measurements using complex chemical/physical analyses. A comparison of the use of spectral pretreatment as well as the implementation of linear and non-linear regression methods was performed. This study presents an overview of the use of infrared spectroscopy for the prediction of five physical (sand, silt and clay) and chemical (total carbon and total nitrogen) soil parameters with near and mid-infrared units in bench top and field set-ups. Even though no significant differences existed among pretreatment methods, models using second derivatives performed better. The implementation of partial least squares (PLS), least squares support vector machines (LS-SVM) and locally weighted regression (LWR) for the development of the calibration models showed that the LS-SVM did not out- perform linear methods for most components while LWR that creates simpler models performed well. The present results tend to show that soil models are quite sensitive to the complexity of the model. The ability of LWR to select only the appropriate samples did help in the development of robust models. Results also proved that field units performed as well as bench-top instruments. This was true for both near infrared and mid-infrared technology. Finally, analysis of field moist samples was not as satisfactory as using dried-ground samples regardless of the chemometrics methods applied.

Keywords
near infrared spectroscopy, NIR, mid-infrared spectroscopy, mid-IR, Fourier transform infrared spectroscopy, FT-IR, partial least squares regression, PLS, least squares support vector machines, LS-SVM, locally weighted regression, LWR, soil physical–chemical properties

Disciplines
Agriculture | Bioresource and Agricultural Engineering

Comments
This article is from *Journal of Near Infrared Spectoscopry* 18 (2010): 167–176, doi:10.1255/jnirs.883.

Authors
Benoit Igne, James B. Reeves, Gregory McCarty, W. Dean Hively, Eric Lund, and Charles R. Hurburgh Jr.

Received: 13 January 2010 ■ *Revised: 18 June 2010* ■ *Accepted: 18 June 2010* ■ *Publication: 15 July 2010*

JNIRS JOURNAL OF NEAR INFRARED SPECTROSCOPY

Evaluation of spectral pretreatments, partial least squares, least squares support vector machines and locally weighted regression for quantitative spectroscopic analysis of soils

Benoit Igne,[a] James B. Reeves, III,[b,*] Gregory McCarty,[b] W. Dean Hively,[b] Eric Lund[c] and Charles R. Hurburgh, Jr[a]

[a]Department of Agricultural and Biosystems Engineering, Iowa State University, Ames, Iowa, USA

[b]Hydrology & Remote Sensing Laboratory, BARC West, Beltsville, Maryland, USA. E-mail: james.reeves@ars.usda.gov

[c]Veris Technologies, 601 Broadway, Salina, Kansas, USA

Soil testing requires the analysis of large numbers of samples in the laboratory that is often time consuming and expensive. Mid-infrared spectroscopy (mid-IR) and near infrared (NIR) spectroscopy are fast, non-destructive and inexpensive analytical methods that have been used for soil analysis, in the laboratory and in the field, to reduce the need for measurements using complex chemical/physical analyses. A comparison of the use of spectral pretreatment as well as the implementation of linear and non-linear regression methods was performed. This study presents an overview of the use of infrared spectroscopy for the prediction of five physical (sand, silt and clay) and chemical (total carbon and total nitrogen) soil parameters with near and mid-infrared units in bench top and field set-ups. Even though no significant differences existed among pretreatment methods, models using second derivatives performed better. The implementation of partial least squares (PLS), least squares support vector machines (LS-SVM) and locally weighted regression (LWR) for the development of the calibration models showed that the LS-SVM did not out-perform linear methods for most components while LWR that creates simpler models performed well. The present results tend to show that soil models are quite sensitive to the complexity of the model. The ability of LWR to select only the appropriate samples did help in the development of robust models. Results also proved that field units performed as well as bench-top instruments. This was true for both near infrared and mid-infrared technology. Finally, analysis of field moist samples was not as satisfactory as using dried-ground samples regardless of the chemometrics methods applied.

Keywords: near infrared spectroscopy, NIR, mid-infrared spectroscopy, mid-IR, Fourier transform infrared spectroscopy, FT-IR, partial least squares regression, PLS, least squares support vector machines, LS-SVM, locally weighted regression, LWR, soil physical–chemical properties

Introduction

The measurement of physical and chemical parameters of soil is an important step toward sustainable farming practices, landscaping management and, more generally, the understanding of terrestrial ecosystem processes. Standard soil analytical procedures are often complex, time-consuming, and expensive for many applications. Research sampling

ISSN: 0967-0335
doi: 10.1255/jnirs.883

strategies such as grid soil sampling require collecting large numbers of samples and their individual analysis in the laboratory is often a tedious process. The challenge of modern soil science is the development of technologies able to perform rapid, accurate, and precise analyses. Mid-infrared (mid-IR) spectroscopy and near infrared (NIR) spectroscopy are fast and non-destructive secondary analytical methods. NIR spectroscopy has been used for more than 30 years in agriculture and the pharmaceutical and petroleum industries—amongst the best known applications—for screening and quality control. In recent years, mid-IR has seen increasing interest, especially for soil analysis.[1]

Infrared (IR) spectroscopy is based on the absorption of infrared radiation by molecules proportionally to their concentration in the sample of interest. The physical structure of the sample is also responsible for differences in absorption patterns due to the scattering of light in the sample related to particle size and porosity. Since the first applications of IR spectroscopy for the determination of soil properties,[2] many authors have published bench-top and field research results on the determination of chemical and physical properties.[1,3–5] Successful applications have been reported from both mid-IR and NIR spectroscopy with an apparent advantage to mid-IR measurements.[6]

Nevertheless, the technology involved in NIR instruments allows a better customisation of applications for remote sensing. NIR spectroscopy is also better suited for field trials due to the limited impact of CO_2 and water vapour that make the treatment of mid-IR spectra potentially difficult and perhaps impossible with moist field samples.[7]

The development of models using data from mid-IR and NIR spectra involves the use of chemometric tools. There are two main steps in the development of a prediction model: (i) the treatment of the spectral data and (ii) the development of the models. Pretreatment of the spectral data before model development is seen as a critical step since it aims to increase signal-to-noise ratio or enhance variations in the signal, but a fine threshold has to be found between removing noise and removing information. Model development is also of importance. Partial least square regression (PLS) has largely been used, but a recent publication introduced neural networks to improve the relationship between spectral data and soil properties.[8] Authors have reported satisfactory results, but the number of samples used seemed low compared to standard practices reported in the NIR literature.[9]

In the present study, we provide an overview of the chemometric methods available to soil scientists. We experimented with the use of various spectral pretreatment methods as well as linear and non-linear regression methods for the prediction of five physical (sand, silt and clay) and chemical (total carbon and total nitrogen) soil parameters based on spectral data collected on four infrared units: a Fourier transform near infrared (FT-NIR) spectrometer and a Fourier transform infrared (FT-IR) spectrometer (bench-top units) and two field portable proximal sensing systems (an FT-IR spectrometer and an NIR system consisting of CCD and InGaAs linear array-based spectrometers).

Materials and methods

Soil samples

A set of 315 ultisol samples collected in April 2007 from a plough layer (0–20 cm) of five bare-soil fields (recently plough-tilled and generally vegetation free) located on the eastern shore of Maryland, USA, were used in this study. Sampling occurred in transects across the fields and corresponded to transects of measurement produced by the tillage-based NIR spectrometer (Veris On-The-Go; Veris Technology, Salina, KA, USA). For the bench-top units, soil samples were dried at 50°C for two days and crushed using a hammer mill to pass through a 2 mm screen. The crushed material was further ground in scintillation vials containing two stainless steel rods and placed on a roller mill overnight. For the portable FT-IR unit, samples were scanned field moist in the laboratory.

Spectral and reference data collection

Bench-top units

Spectra were collected on dried and sieved samples in NIR and mid-IR regions by a Digilab Fourier transform spectrometer (FTS7000 FTS; Varian, Inc., Palo Alto, CA, USA) equipped with DTGS (deuterated triglycine sulfate) and InSb detectors using a Pike (Pike Technologies, Madison, WI, USA) AutoDiff autosampler (sample cups ~1 cm in diameter). Each sample was scanned 64 times, at a resolution of four wavenumbers and scans co-added to give a final spectrum. These two instruments will be called bench-FT-NIR and bench-FT-MIR throughout the article.

A field portable FTIR spectrophotometer (Surface Optics Corp, San Diego, CA, USA) was used to collect FTIR spectra on field moist samples. This mid-IR instrument collects spectra from 4000 cm^{-1} to 400 cm^{-1}. A rotating sample cup was used to increase the area scanned and 64 scans were co-added to produce a spectrum. The beam splitter was KBr and the resolution was 8 cm^{-1}. Note that even though this unit was used in the laboratory, samples were not dried before scanning to simulate field experiment conditions.

Field unit

Transects of NIR spectral data were collected using a tractor-mounted Veris On-The-Go spectrophotometer built into a shank mounted on a toolbar and pulled behind a tractor. Spacing between transects was 20 m on average. Spectral measurements were acquired through a sapphire window mounted on the bottom of the shank. The spectrophotometer used a tungsten–halogen bulb to illuminate the soil and the reflected light was collected into a fibre-optic cable for transmission to the spectrometer. Two spectrometers were used to collect spectral data in the visible and near infrared (350–2225 nm) range with an average resolution of 8 nm. The sensors used

Table 1. Statistical parameters of the samples used in the calibration models.

Parameter	*n*	Average concentration (%)	Range (%)	Standard deviation (%)
Calibration set				
Carbon	209	1.26	0.55–2.04	0.25
Nitrogen	209	0.06	0.02–0.12	0.02
Sand	209	46.67	19.01–87.69	13.23
Silt	209	43.09	9.50–67.87	11.23
Clay	209	10.23	1.34–19.72	3.01
Validation set				
Carbon	106	1.25	0.59–1.88	0.25
Nitrogen	106	0.06	0.02–0.11	0.02
Sand	106	46.94	21.32–86.84	12.81
Silt	106	43.15	11.07–63.91	11.11
Clay	106	9.91	2.81–19.65	2.80

were a 3648-element TCD1304AP Linear CCD Array (Toshiba, Tokyo, Japan) and a 256-element InGaAs linear image sensor (Hamamatsu, Shizuoka, Japan). Shutters in the shank were manipulated to acquire dark and reference spectra at approximately 10 min intervals. The shank was pulled through the soil at approximately 10 cm depth at 6 km h^{-1}, acquiring approximately 20 spectra per second. Tillage-based reflectance data associated with soil sampling locations were extracted by Gaussian elliptical weighting with approximately five spectra averaged for each location. Field sampling and field-based spectral measurements were collected on the same day.

Reference measurements

Five physical and chemical parameters were measured; physical parameters were sand, silt and clay; chemical parameters were total carbon and total nitrogen.

Soil organic carbon and total nitrogen content were determined by dry combustion using a TruSpec CN analyser (Leco Corp., St Joseph, MI, USA). None of the samples contained significant inorganic C. Sand, silt and clay fractions in soil were determined by the hydrometer method described by Gee and Bauder.[10]

From the 315 samples, every third sample based on spatial distribution in the fields was used to create an independent validation set. Table 1 presents summary statistics of both calibration and validation sets.

Chemometrics methods

Since most of the work done on soil used classical chemometric methods, this study presents alternative approaches to spectral pretreatment and calibration development.

Spectral pretreatment

Two types of spectral pretreatment were used. The first type consisted of common derivative and scatter correction methods (called spatial pretreatment methods—in relation to frequency-based pretreatment methods presented later). Nineteen different spectral pretreatments and/or combinations of treatments were used and compared. Vasques *et al.*[11] presented an exhaustive study on the use of derivative methods for soil analysis. The present study reused some of the methods and proposed alternative techniques as well as combinations of methods. They were variations of the Savitzky–Golay derivative,[12] normalisation (unit area under curve) and scatter correction methods [standard normal variate (SNV),[13] multiplicative scatter correction (MSC),[14] extended multiplicative scatter correction (EMSC)[15] and loopy MSC and EMSC].[16] Table 2 presents the different preprocessing methods used.

The second type of spectral pretreatment used information present in the Fourier and wavelet decompositions of raw spectra. They are novel techniques to spectral pretreatment and have shown their ability to develop robust models for calibration transfer situations.[17] This property could be useful in the development of soil calibrations that are often affected by a large variability from sample to sample and from field to field. There are two approaches to process frequency information. The first involves filtering the high frequency components of Fourier coefficients or wavelet detail coefficients with a smoothing filter applied to a range of Fourier coefficients representing the highest frequency components or to the entire detail component of a wavelet transform. Spectra are (i) transformed in periodigrams (fast Fourier transform) or detail and approximation coefficients (wavelet transform), (ii) their Fourier coefficients representing the highest frequency components or detail coefficients are smoothed using a Savitzky–Golay smoothing filter, and (iii) converted back to spectra using an inverse Fourier or wavelet transform. These methods are, respectively, called Fourier smoothing and wavelet smoothing. The tuning of the algorithms is done by choosing the smoothing filter (window size and polynomial order) for both Fourier and wavelet filtering and the range

Table 2. Preprocessing methods.

ID	Combination	Parameter
1	—	—
2	Second derivative	25-point window, third-order polynomial
3	Second derivative	15-point window, third-order polynomial
4	Second derivative	35-point window, third-order polynomial
5	First derivative	25-point window, third-order polynomial
6	Smoothing	25-point window, third-order polynomial
7	Normalisation	Unit area under curve
8	Normalisation + second derivative	25-point window, third-order polynomial
9	Normalisation + second derivative	15-point window, third-order polynomial
10	Normalisation + second derivative	5-point window, third-order polynomial
11	Normalisation + second derivative	35-point window, third-order polynomial
12	SNV	—
13	Second derivative + SNV	25-point window, third-order polynomial
14	EMSC	—
15	MSC	—
16	Loopy MSC	5 cycles
17	Loopy MSC	10 cycles
18	Loopy EMSC	5 cycles
19	Loopy EMSC	10 cycles

of Fourier coefficients to filter (detail component of wavelet transform was entirely filtered since wavelet transform keep the localisation information). The second approach consists in correcting the high-frequency components of a fast Fourier transform with a slope and a bias. As with Fourier smoothing, spectra are first transformed into periodigrams, then, similar to multiplicative scatter correction performed in the wavelength domain,[14] a slope and a bias are calculated and applied to each periodigram of the calibration set. Coefficients are obtained by regressing the magnitudes of the calibration set on the magnitude component of the average spectrum of 20 samples selected from the calibration set covering the range of contents. Filtered periodigrams are finally transformed back into spectra. The tuning of the algorithm involves finding the appropriate range of high frequency Fourier coefficients to correct. This third method was called Fourier signal correction. For more information about the methods, refer to the flowchart in Figure 1 and to Igne and Hurburgh.[17]

In the present study, Fourier filtering, wavelet filtering and Fourier signal correction were tuned by iterative processes to determine the best smoothing filters and range of frequency components to filter (for Fourier based methods only). A

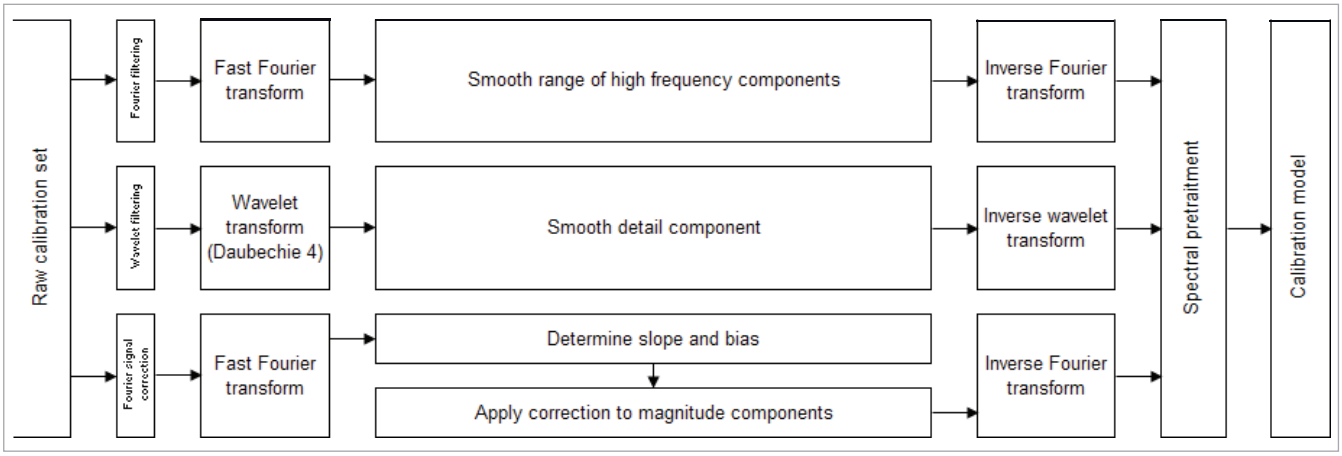

Figure 1. Flow-chart of frequency-based preprocessing methods.

Daubechie 4[18] wavelet was used to decompose the spectra in the wavelet domain.

Frequency-filtered spectra were then preprocessed in the wavelength domain before calibration. The best performing preprocessing methods (or combination of preprocessing methods) found when developing PLS models were used to complement frequency-based pretreatment methods.

Calibration techniques

Partial least squares,[19] a linear method, least squares support vector machines (LS-SVM),[20–22] a non-linear regression technique, and locally weighted regression (LWR), a local linear method,[23–25] were used to develop calibration models for the five parameters. Least squares support vector machines were preferred to neural networks for their ability to perform on small datasets and by the fact that only one global minimum exists in the optimisation plane.

Partial least squares regression

Partial least squares regression extracts from the spectral data (**X**_matrix) the information that is related to the reference value of interest (**Y**_matrix). This extraction is performed by calculating principal components or latent variables that maximise the covariance between the **Y**_matrix and all possible linear functions of the **X**_matrix.[19] The choice of latent variable to include in the model is usually determined by minimising the standard error of cross-validation and limiting overfitting, the result of creating a model only valid on the calibration set. Model parameters (preprocessing methods and number of latent variables to include) were determined by comparing performances in cross-validation leave-one-out. The validation set was applied to the tuned model.

Least squares support vector machines

Least squares support vector machines has been developed to perform on data presenting non-linear relationships with a limited number of observations. Similar to support vector machines classification that looks for the maximum margin between clusters, LS-SVM tries to minimise the prediction error relative to an error rate determined by the user. The main advantage of LS-SVM is that only two parameters need to be determined: the error rate and the parameter of the kernel function (to correct for non-linearity). The error plane presents only one minimum. However, its main drawback is the computation time; it is exponentially proportional to the size of the dataset and can take several hours to perform on a set of several hundred samples. Cogdill and Dardenne[20] provided a good overview of LS-SVM. Shawe-Taylor *et al.*[21] and Suykens *et al.*[22] are references for theoretical aspects of support vector machines. An exhaustive search of the error rate and kernel parameter was performed to optimise the model. The validation set was

Table 3. PLS and LWR model parameters.

Instrument	Parameter	PLS		LWR	
		Number of samples	Number of latent variables	Number of samples	Number of latent variables
Portable FT-IR	Total carbon	209	7	175	7
	Total nitrogen	209	6	200	6
	Sand	209	9	205	9
	Silt	209	9	205	9
	Clay	209	12	150	9
Bench-FT-NIR	Total carbon	209	12	65	8
	Total nitrogen	209	9	209	9
	Sand	209	14	95	10
	Silt	209	14	95	10
	Clay	209	11	145	9
Veris	Total carbon	209	12	205	12
	Total nitrogen	209	6	110	5
	Sand	209	12	90	10
	Silt	209	13	90	10
	Clay	209	12	125	11
Bench-FT-MIR	Total carbon	209	13	195	13
	Total nitrogen	209	12	170	12
	Sand	209	11	110	10
	Silt	209	11	110	10
	Clay	209	9	160	8

applied to the tuned model. Fernández Pierna *et al*.[23] first applied LS-SVM to soil analysis on a chemometrics contest dataset. The present study aimed at showing the method performances on another dataset with various instrumental settings and different parameters.

Locally weighted regression

Local models look for the closest samples in the calibration set to the sample to predict. A sub-calibration set is then created to only include in the model samples that are relevant. This approach allows the development of linear models in situations where, over the range of concentrations, the relationship between NIR response and chemical concentration are non-linear. It also reduces the complexity of models by including less noise, but is subject to overfitting. The locally weighted regression (LWR) algorithm[24,25] uses a principal component analysis to determine the closest samples by calculating score Mahalanobis distances in principal component analysis space between the new sample and the database using sample scores. An exhaustive search of the best parameters was performed. The size of the calibration set was increased by five samples every generation, starting with an initial set of 20 samples. Partial least squares model performances at each generation were evaluated by cross-validation leave-one-out. The number of principal components included in the PLS models varied from five to 15. The model presenting the best results was applied to the validation set.

Calibration and validation procedures

MATLAB R2007b (The MathWorks, Natick, MA, USA) was used for all calculations. The pretreatment of the data and the development of PLS and LWR models were performed with the PLS_toolbox 4.2.1 and the EMSC_Toolbox 1.2 (both toolboxes are from Eigenvector Research, Wenatchee, WA, USA); LS-SVM calibrations were developed with the LS-SVMlab toolbox v. 1.5 for MATLAB by Suykens *et al*.[22] Table 3 presents the number of latent variables and samples used for PLS and LWR models.

Spatial preprocessing methods used for LS-SVM and LWR were the best performers found with PLS. All models were validated on the same validation sets. Autoscaling (mean zero and unit variance) was used to scale spectra, after pretreatment methods, before developing all regression models.

Model evaluation and comparison

Standard error of prediction (*SEP*) was used to evaluate the precision of each model. Standard error of prediction is the standard deviation of differences not corrected for bias between the **Y**_matrix of validation and the prediction matrix $\hat{y}$. The models' fit was evaluated using the coefficient of determination (r^2) that represents the percentage of variability explained by the model.

Results and discussion

Effect of spectral preprocessing methods

Figure 2 presents the effect of all spatial preprocessing methods on the precision of the PLS predictions of total carbon and silt content (see Table 2 for a list of settings used for all methods). No significant effect of preprocessing methods was observed (α = 0.05). For total carbon, when considering only FT-NIR and the field unit, normalisation to unit area gave significantly higher *SEP*s. This was not true for silt. The difference between the 18 methods (except normalisation) was not large and the method that gave the lowest errors for one instrument was often different for others. The development of models without preprocessing methods was a

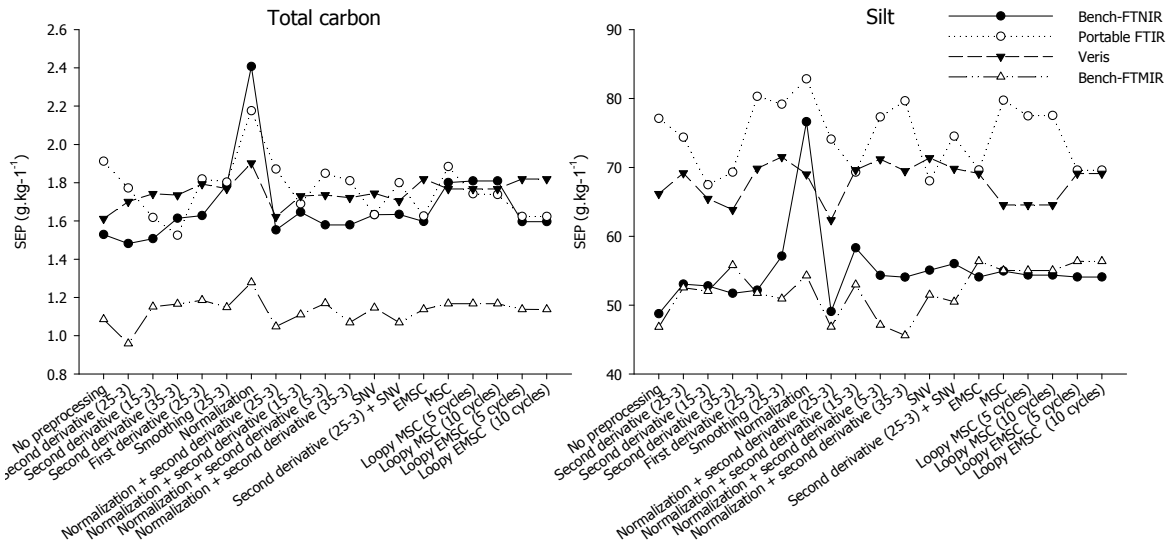

Figure 2. Effect of preprocessing methods on the prediction of total carbon and silt.

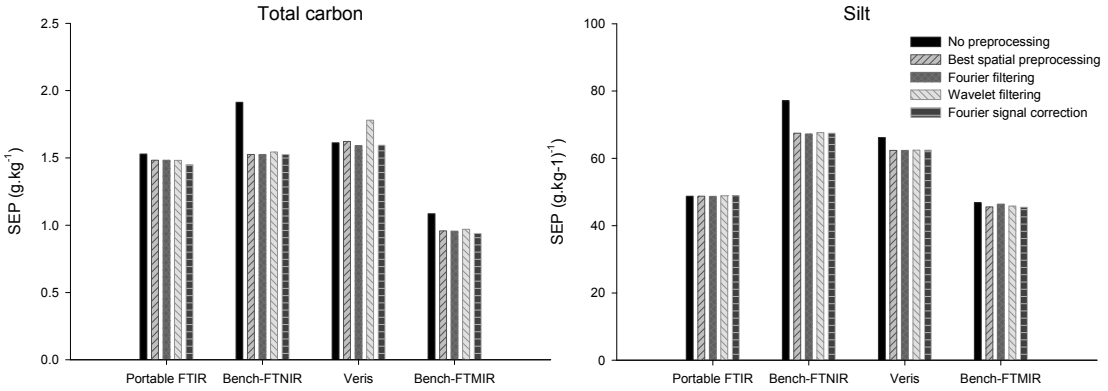

Figure 3. Comparison on spatial and frequency-based preprocessing methods.

good option for the Veris, bench-FT-NIR, and bench-FT-MIR. It had the advantage of simplifying the model and providing as precise results as with pretreatment methods. For silt, the difference between methods was larger than for total carbon and the use of preprocessing methods was really beneficial for instruments such as the bench-FT-NIR and the Veris units, although not using preprocessing methods was not the worst option. This situation may come from the fact that silt is much harder to predict by NIR spectroscopy than carbon and that an enhancement of the signal-to-noise ratio can improve the overall performances of the models.

The present results showed that the use of spectral preprocessing methods was instrument- and parameter dependent. These results are consistent with the literature.[26] This situation makes the interpretation of the effect of each method difficult, since a good preprocessing method [i.e. second derivative (35-point window, third-order polynomial)] worked for a specific instrument (i.e. the portable-FT-IR)

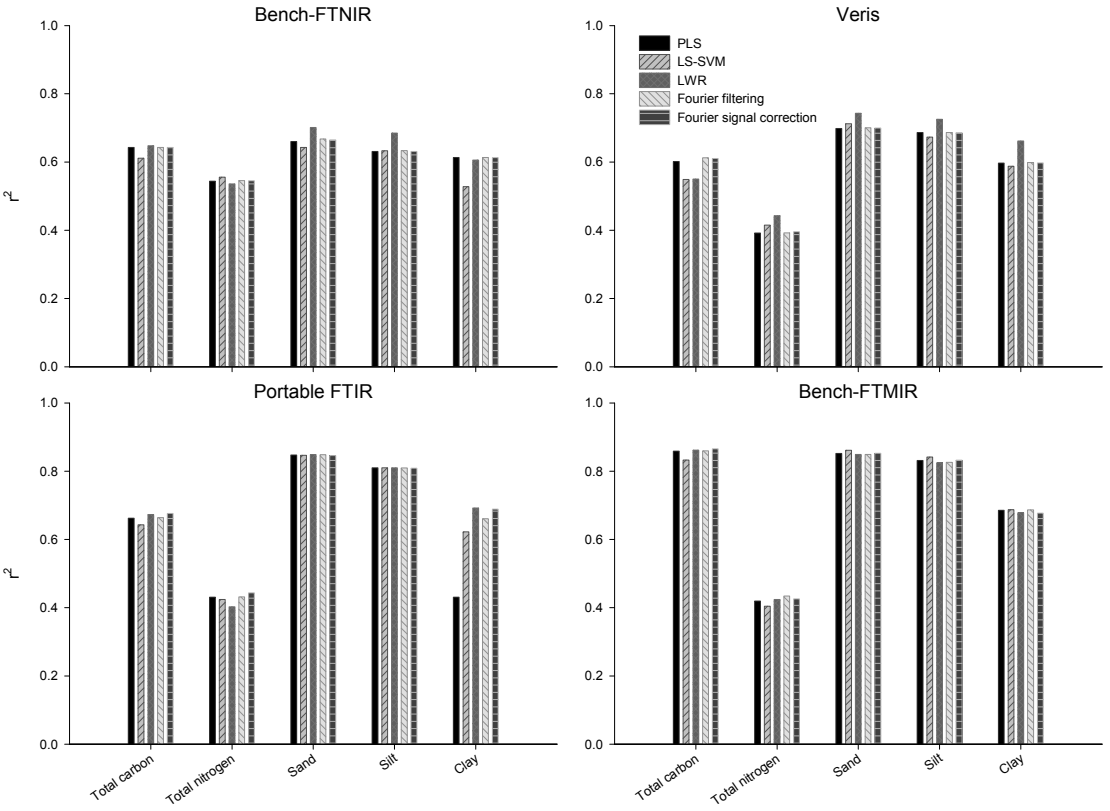

Figure 4. Comparison of coefficients of determination of models developed with partial least squares (PLS), least squares support vector machines (LS-SVM), and locally weighted regression (LWR).

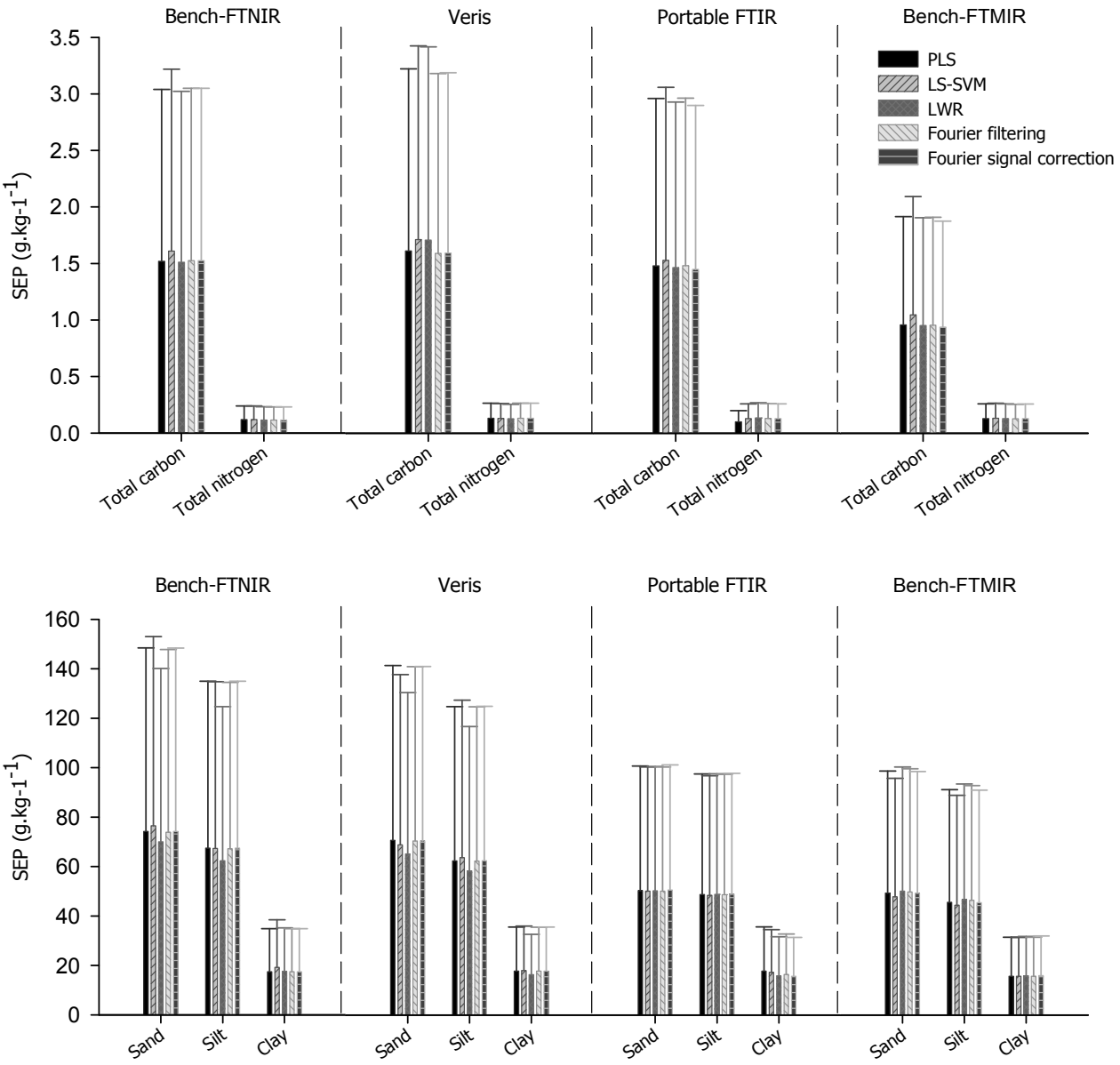

Figure 5. Comparison of standard errors of predictions of chemical and physical soil properties models developed with partial least squares (PLS), least squares support vector machines (LS-SVM), and locally weighted regression (LWR).

and both parameters, but did not perform well for others (i.e. bench-FT-MIR). Differences in technologies, interpolation, sample presentation, number of subsamples (signal-to-noise ratio), parameter, etc. have a large impact on the final results and only an iterative process can help in the development of the best models. However, the second derivative was always in the combination of preprocessing methods that gave the lowest error (in agreement with Vasques *et al.*[11]) and MSC-based methods did not perform well, especially with the bench-FT-MIR instrument.

For total carbon and silt, Figure 3 presents the best spatial preprocessing method: (*for carbon*: second derivative 25-point window—third-order polynomial, second derivative 35-point window—third-order polynomial, normalisation + second derivative 25-point window—third-order polynomial, second derivative 25-point window—third-order polynomial for portable FT-IR, bench-FT-NIR, Veris, and bench-FT-MIR, respectively; *for silt*: none, second derivative 15-point window—third-order polynomial, normalisation + second derivative 25-point window—third-order polynomial, normalisation + second derivative 25-point window—third-order polynomial for portable FT-IR, bench-FT-NIR, Veris and bench-FT-MIR, respectively) as well as the frequency-based filtering methods. Similar to what was observed with spatial preprocessing, frequency-based preprocessing gave similar precision to other methods. Fourier signal correction appeared to be the best method, even

though the improvement was not significant. Improvements were observed when compared with the use of only spatial preprocessing methods. Again, the use of these preprocessing methods was parameter- and instrument dependent.

Effects of regression methods

The comparison of preprocessing methods showed that it is often beneficial to perform an iterative search of the best algorithms to enhance the performances of a calibration model, but as demonstrated in Figures 4 and 5, the choice of the regression method is also of importance. Figures 4 and 5 present the coefficient of determination and the standard error of prediction for the five parameters of interest when models were developed with PLS (spatial preprocessing), LS-SVM, LWR, PLS + Fourier filtering and PLS + Fourier signal correction. Wavelet filtering was left out, as it gave some low results for silt and total carbon. A bootstrap method[27] was implemented to estimate the uncertainty around the *SEP*.

Even though no significant differences existed between results, by using PLS results as the benchmark, we observed that LS-SVM did not always perform well (clay) and did not outperform more traditional methods when predicting other parameters. Figure 4 shows that LWR was the method of choice for NIR instruments while other regression techniques performed similarly for mid-IR instruments. It is interesting to notice that for clay, PLS did not perform very well for mid-IR instruments while PLS + frequency based methods gave good results. The noise removed by modifying the frequency information was beneficial for the prediction of these parameters.

These results tend to show that soil calibration performances are very sensitive to the complexity of the models. With autoscaling only or a single preprocessing method and the selection of samples to include in the calibration set (LWR), results are as good as, or better than, when combining many preprocessing methods and employing more advanced regression methods. Note that spectral preprocessing methods have been optimised for PLS models and used for LWR and LS-SVM. An optimisation of the spectral preprocessing methods might improve the validation results.

For all parameters except total nitrogen, MIR instruments gave significantly better results than NIR-based units. The bench-FT-MIR unit outperformed the field unit for total carbon. Those results are consistent with the literature, but show that field units can be good alternatives to bench-top instruments. It is nevertheless important to note that the portable FT-IR unit had a larger interpolation than the bench MIR unit (8 cm^{-1} versus 4 cm^{-1}). Calibrations were thus based on spectra with half as many data points. This may have impacted final results in addition to the fact that the portable FT-IR unit may be noisier. Also, it is necessary to insist on the fact that the Veris and the portable FTIR spectra were collected on moist samples and this was most likely responsible for the poorer results of these units.

Compared with previous publications,[4,28] the present results for the prediction of total carbon and nitrogen were rather low. This was most likely due to the lower standard deviations of calibration sets and readers should not infer on the predictive ability of the chemometrics techniques since sand, silt and clay results were in agreement with reported results.

Conclusions

The present study compared 22 different preprocessing methods and three different regression algorithms for the prediction of five physical and chemical parameters in soil samples. While no statistical differences existed among the preprocessing methods, derivative-based models gave the best precisions and the use of Fourier-based methods allowed the development of more robust models. The comparison of regression methods showed that the simplest model (locally weighted regression) provided the overall best performances. Finally, MIR units gave the best results. Differences between bench-top and field units were instrument- and parameter dependent.

This article provides soil scientists with an example of calibration development with different instrumentations and approaches to model the data. The results of the locally weighted regressions suggested it is a good alternative to classic PLS models and reinforced the need for samples in a calibration set to have analyses which are uniformly distributed over the range for the parameter of interest. Locally weighted regression, being easily subject to overfitting due to the limited number of samples, may also require a solid validation strategy for field applications.

Advances in chemometrics that constitute LS-SVM, advanced spatial preprocessing methods and frequency-based filtering methods did not give different or better results than others, but the situation might be different in other cases, with different samples, different instruments and different concentration ranges in the various parameters. They should be tested through an iterative process and compared to PLS that remains the best benchmark method for soil analysis.

References

1. J.B. Reeves, III, G.W. McCarty and J.J. Meisinger, "Near infrared reflectance spectroscopy for the determination of biological activity in agricultural soils", *J. Near Infrared Spectrosc.* **8,** 161 (2000). doi: 10.1255/jnirs.275
2. S.A. Bower and R.J. Hanks, "Reflection of radiant energy from soils", *Soil Sci.* **100,** 130 (1965). doi: 10.1097/00010694-196508000-00009
3. C.W. Chang, D.A. Laird and C.R. Hurburgh, "Influence of soil moisture on near-infrared reflectance spectroscopic measurement of soil properties", *Soil Sci.* **170,** 44 (2005). doi: 10.1097/00010694-200504000-00003
4. R. Zornoza, C. Guerrero, J. Mataix-Solera, K.M. Scow, V. Arcenegui and J. Mataix-Beneyto, "Near infrared

spectroscopy for determination of various physical, chemical and biochemical properties in Mediterranean soils", *Soil Biol. Biochem*. **40,** 1923, 1930 (2008).

5. B. Stenberg, R.A. Viscarra Rossel, A. Mounem Mouazen and J. Wetterlind, "Visible and near infrared spectroscopy in soil science", in *Advances in Agronomy,* Volume 107**,** Ed by D.L. Sparks. Academic Press, not yet published (2010).

6. G.W. McCarty and J.B. Reeves, III, "Comparison of near infrared and mid infrared diffuse reflectance spectroscopy for field scale measurement of soil fertility parameters", *Soil Sci*. **171,** 94 (2006). doi: 10.1097/01.ss.0000187377.84391.54

7. J. Wang, J.C. Gille, P.L. Bailey, J.R. Drummond and L. Pan, "Instrument sensitivity and error analysis for the remote sensing of tropospheric carbon monoxide by MOPITT", *J. Atmos. Oceanic Technol*. **16,** 465, 474 (1999).

8. L.H. Zheng, M.Z. Li, L. Pan, J.Y. Sun and N.Tang, "Estimation of soil organic matter and soil total nitrogen based on NIR spectroscopy and BP neural network", *Spectrosc. Spectral Anal*. **28,** 1160 (2008).

9. C. Borggaard, "Neural networks in near-infrared spectroscopy", in *Near-Infrared Technology in Agricultural and Food Industries*, 2nd Edn, Ed by P. Williams and K. Norris. American Association of Cereal Chemists, St Paul, Minnesota, USA, pp. 101–108 (2001).

10. G.W. Gee, and J.M. Bauder, "Particle-size analysis", in *Methods of Soil Analysis, Part 1, Physical and Mineralogical Methods*, Agronomy Monograph No. 9, 2nd Edn. American Society of Agronomy, Madison, WI, USA, pp. 383–411 (1986).

11. G.M.Vasques, S. Grunwald and J.O. Sickman, "Comparison of multivariate methods for inferential modeling of soil carbon using visible/near-infrared spectra", *Geoderma* **146,** 14 (2008). doi: 10.1016/j.geoderma.2008.04.007

12. A. Savitzky and M.J.E. Golay, "Smoothing and differentiation of data by simplified least squares procedures", *Anal. Chem*. **36,** 1627 (1964). doi: 10.1021/ac60214a047

13. R.J. Barnes, M.S. Dhanoa and S.J. Lister, "Standard normal variate transformation and de-trending of near infrared diffuse reflectance spectra", *Appl. Spectrosc*. **43,** 772 (1989). doi: 10.1366/0003702894202201

14. P. Geladi, D. MacDouglas and H. Martens, "Linearization and scatter-correction for near-infrared reflectance spectra of meat", *Appl. Spectros*. **39,** 491 (1985). doi: 10.1366/0003702854248656

15. H. Martens, J.P. Nielsen and S.B. Engelsen, "Light scattering and light absorbance separated by extended multiplicative signal correction. Application to near-infrared transmission analysis of powder mixtures", *Anal. Chem*. **75,** 394 (2003). doi: 10.1021/ac020194w

16. W. Windig, J. Shaver and R. Bro, "Loopy MSC: A simple way to improve multiplicative scatter correction", *Appl. Spectrosc*. **62,** 1153 (2008). doi: 10.1366/000370208786049097

17. B. Igne and C.R. Hurburgh, "Using the frequency components of near infrared spectra: optimising calibration and standardisation processes", *J Near Infrared Spectrosc*. **18,** 39 (2010). doi: 10.1255/jnirs.865

18. I. Daubechies, "Ten lectures on wavelets", *CBMS-NSF Lecture Notes*, nr. 61, SIAM (1992).

19. T. Næs, T. Isakson, T. Fearn and T. Davies, *Multivariate Calibration and Classification*, NIR Publications, Chichester, UK (2002).

20. R.P. Cogdill and P. Dardenne, "Least-squares support vector machines for chemometrics: an introduction and evaluation", *J. Near Infrared Spectrosc*. **12,** 93 (2004). doi: 10.1255/jnirs.412

21. J. Shawe-Taylor and N. Cristianini, *Support Vector Machines and Other Kernel-based Learning Methods*. Cambridge University Press, Cambridge, UK (2000).

22. J.A.K. Suykens, T. Van Gestel, J. De Brabanter, B. De Moor and J. Vandewalle, *Least Squares Support Vector Machines*. World Scientific, River Edge, NJ, USA (2002). doi: 10.1142/9789812776655

23. J.A. Fernández Pierna and P. Dardenne, "Soil parameter quantification by NIRS as a chemometric challenge at 'Chimiométrie 2006'", *Chemometr. Intell. Lab. Syst*. **91,** 94 (2008). doi: 10.1016/j.chemolab.2007.06.007

24. W.S. Cleveland and S.J. Devlin, "Locally weighted regression: an approach to regression analysis by local fitting", *J. Am. Stat. Assn* **83,** 596 (1988). doi: 10.2307/2289282

25. T. Næs, T. Isaksson and B. Kowalski, "Locally weighted regression and scatter correction for near-infrared reflectance data", *Anal. Chem*. **62,** 664 (1990). doi: 10.1021/ac00206a003

26. V.M. Fernández-Cabanás, A.Garrido-Varo, D. Pérez-Marín and P. Dardenne, "Evaluation of pretreatment strategies for near-infrared spectroscopy calibration development of unground and ground compound feedingstuffs", *Appl. Spectrosc*. **60,** 17 (2006). doi: 10.1366/000370206775382839

27. NM. Faber, "Improved computation of the standard error in the regression coefficient estimates of a multivariate calibration model", *Anal. Chem*. **72,** 4675 (2000). doi: 10.1021/ac0001479

28. D.F. Malley, P.D. Martin and E. Ben-Dor, "Application in analysis of soils", in *Near-Infrared Spectroscopy in Agriculture*, Agronomy 44, Ed by C.A. Roberts, J. Workman, Jr and J.B. Reeves, III. American Society of Agronomy, Crop Science Society of America, and Soil Science Society of America, Madison, WI, USA, pp. 729–784 (2004).