Soil Organic Carbon Prediction Using Vis-NIR Spectroscopy with a Large Dataset

Yang Shi[1,2(✉)], Rujing Wang[1], and Yubing Wang[1]

[1] Institute of Intelligent Machines, Chinese Academy of Science, Hefei, China
yshi@mail.ustc.edu.cn
[2] Department of Automation, University of Science and Technology of China, Hefei, China

Abstract. Visible and near-infrared reflectance spectroscopy based soil properties estimation is an alternative to traditional laboratory analysis. The calibration model is a main factor influencing predictive performance. In this study, a large scale soil database, which contains 19,036 soil samples, was compared to its subsets to validate the effect of sample size on predictive performance. Four regression techniques based on linear model, namely, multiple linear regression (MLR), principal components regression (PCR), partial least squares regression (PLSR), and stepwise regression (SR) were compared to identify suitable models to predict the content of organic carbon in soil samples. The impact of derivatives or the raw spectra as predictor variables, and the interval of spectra were also studied. The best predictions were obtained using SR and MLR on raw spectra, yielding root mean square of error of cross validation (*RMSECV*) and coefficient of determination (R^2) values of 25.3912, 25.4254 and 0.9227, 0.9225, indicating excellent models.

Keywords: Soil organic carbon · MLR · PCR · PLSR · Stepwise regression

1 Introduction

The processes, mechanisms, and variability of soil is complex and difficult to fully comprehend [1]. Although soil is a heterogeneous system, the management of soil resource needs quantitative and sustainable analysis of soil quality, especialy under agricultural systems [2]. Visible and Near-InfraRed (Vis-NIR) reflectance spectroscopy, which may replace traditional laboratory analysis, can be used to rapidly and precisely characterise the chemical and physical of the soil at a low cost [3]. The analytical ability of Vis-NIR spectroscopy, depending on absorptivity feature of Vis-NIR light determined by overtones and combinations of C-H, O-H and N-H bonds, make it possible for quantitive analysis of carbon, nitrogen and water forms [4]. As Vis-NIR spectroscopy is an indirect solution to obtain properties of soil samples, the estalishment of reliable calibration models is essential to describe the relation of soil property and its spectroscopy.

In order to tackle the task of accurate prediction of soil properties, there are fundamentally two classes of approaches: memory-based methods and model-based methods. The main goal of memory-based methods is to develop a reasonable metric of

Published by Springer Nature Switzerland AG 2019
D. Li and C. Zhao (Eds.): CCTA 2017, IFIP AICT 545, pp. 76–86, 2019.
https://doi.org/10.1007/978-3-030-06137-1_8

the distance between samples. The hypothesis is proposed that when two samples are similar in terms of soil compositional characteristics, the distance of them are relatively small in spectra space [5, 6]. The model-based methods, which treat the content of soil samples as regression results from the spectra as predictor variables, have gained wide acceptance of researchers. Traditional approach include Multiple Linear Regression (MLR), Principal Components Regression (PCR), Partial Least Squares Regression (PLSR). Among these methods, PLSR is the most commonly used method and performs well in many applications [7]. Recently machine learning based methods are becoming more popular. For example, Artificial Neural Networks (ANNs), Support Vector Regression (SVR), Cubist are employed to train a non-linear regression model and outperform traditional linear methods even in industrial applications [8, 9].

In recent studies, many chemometrics experiments are carried on a small dataset, which contains hundreds of calibration samples even less than one hundred. The problem of overfitting often occurs especially when the dimension is greater than the sample size as explained in [10]. Thus, many excellent solutions are proposed on small dataset problems. One solution is that a linear transformation on the spectra are usually performed to compress the high dimension data to several representative orthogonal components in PCR and PLSR methods. Another way to achieve a better prediction is variable selection. These algorithms can select relevent features and discard noisy or unreliable ones. Key wavelengths selection or Uninformative Variables Elimination (UVE), is taken to reduce predictor variables in the final model in UVE-PLS [11], competitive adaptive reweighted sampling [12], iteratively variable subset optimization [13] and stepwise regression (SR) [14]. Meanwhile, for most spectra data, values at neighboring wavelengths are usually highly correlated [10]. It is possible to use a large dataset with a suitable wavelength interval avoiding spectra redundancy and overfitting. However, the effects of wavelength interval selection and the number of calibration samples on predictive performance are lack of attention and less discussed.

In this paper, we compare the predictive accuracy of Soil Organic Carbon (SOC) using the Vis-NIR spectroscopy based on different linear models, namely, MLR, PCR, PLSR and SR. Especially, the effect of sample size of the calibration set, the wavelength interval, input variables include raw reflectance, first-order and second-order derivatives on predictive performance were analyzed and compared. The evaluation was based on Root Mean Square of Error of Cross Validation (*RMSECV*) and coefficient of determination (R^2) in the 5-fold cross validation.

2 Materials and Methods

2.1 The LUCAS Soil Database

A large scale soil database was compiled in the framework of the European Land Use/Cover Area frame Statistical Survey (LUCAS). LUCAS samples covered all the major soil types in 23 countries in Europe (EU). According to environmental conditions and the criteria of accessibility, about 19 thousand survay points were finally selected from original 250 thousand points. Physical testing and chemical analysis on seleted soil samples were taken in a central laboratory [15].

The soil samples were air-dried and sieved before spectra measuring. The Vis-NIR spectra were measured at wavelengths with a range of 400 nm to 2,500 nm and an interval of 0.5 nm, yielding a vector of 4,200 dimensions. By using ISO standard methods, the properties of soil samples were analyzed include coarse fragments percentage, particle size distribution, as well as Soil Organic Carbon (SOC, g/kg) et al. In this paper, only the content of SOC was concerned, and the statistics of which is shown in Table 1.

Table 1. Statistical data of soil organic carbon measured by chemical methods in LUCAS.

	Soil Organic Carbon (g/kg)
Minimum value	0.0
Maximum value	586.8
Mean value	50.0

The LUCAS topsoil dataset used in this work was made available by the European Commission through the European Soil Data Centre, which is managed by the Joint Research Centre (JRC), http://esdac.jrc.ec.europa.eu/ [16, 17].

2.2 The Pretreatment of Spectrum

Suitable spectrum pretreatment produces positive impact on model performance, either lower prediction error or reduction of computation. Because values at neighboring wavelengths are usually highly correlated for most spectra data, a down sampling step is often taken to simplify the calculation process. As mentioned in [18], derivatives are used in analytical spectroscopy for decades due to the capability to reduce additive or multiplicative effects and eliminate noises of raw spectra. Derivatives can make the spectra smoother and more characteristic. The averaged spectrum of all 19,036 samples is displayed in Fig. 1(a) with solid line, compared to spectra of the samples with the maximal and minimal SOC with dash line. The first and second derivatives of these spectra are also displayed in Fig. 1(b) and (c).

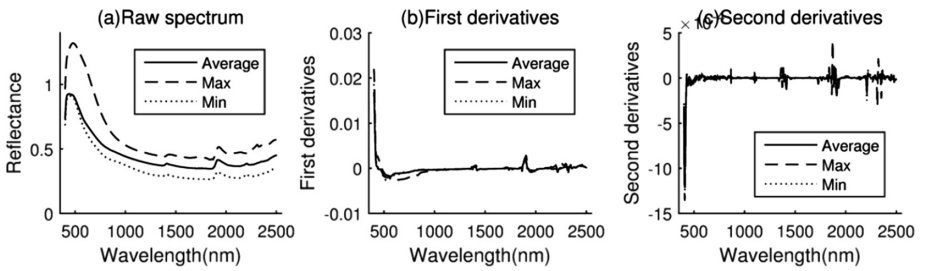

Fig. 1. The Vis-NIR spectra of soil samples from LUCAS dataset with different pretreatment methods. (a) raw spectra, (b) first derivatives, (c) second derivatives.

2.3 Linear Calibrate Models

Multiple Linear Regression. We assume the value of interested soil property is y, and the spectral data vector used as independent variables into the model is X, where $X = [x_1, x_2, \ldots, x_n]$, and x_i is the reflectance value at a specific wavelength. Suppose the relationship existed between the specific soil property and the spectra is linear, the prediction of soil property is a linear regression problem, and the simplest multiple linear model can be summarized by the following equation:

$$y = X \cdot W \tag{1}$$

where W is weight vector of corresponding variable. The linear least square method should be selected when the dimension of X is smaller than the rank of the matrix which is combined by all samples. However, modern instruments offer more dimension but still strongly correlated. A dimension reduction or variable selection step based on domain knowledge or statistics is necessary and often taken before modelling unless the sample size is significantly large.

Principal Components Regression. PCR is a linear regression method based on principal components calculated by Principal Component Analysis (PCA) algorithm. PCA is a powerful dimension reduction method based on statistics. After PCA of spectral data, original X is compressed to several components. The eigenvalue of each principle component show the percentage accounts for the variation of dataset, and the eigenvector is a weight vector of variables at all wavelengths to calculate principle components. In theory, the advantage of PCR is the elimination of correlation and noise existed in original spectra compared to MLR method. Finally, the chosen number of principal components are input to a MLR model instead of raw data [7].

Partial Least Squares Regression. PLSR, as proposed in [19], is another effective model based on dimension reduction. PLSR has a same main structure as PCR, as both employ linear transformation to overcome the problems of high-dimensionality and multi-collinearity. What is different, the calculation of principal components of PLSR takes y variables into consideration [20]. It performs a linear transition from raw spectrum to much smaller number of orthogonal principle components called latent variables, and the covariance between X and target soil property y is maximized resulting in a high predictive capability.

Stepwise Regression. Stepwise Regression (SR) is a variable selection based MLR method. Unlike PCR and PLSR, it applies a forward or backward method to gradually select several important variables based on entrance/exit tolerances for the p-values of F-statistics. At each step, the explanatory power of larger/smaller models, which is tested by the p-value of an F-statistic, is compared to current model. An in/out action is taken when modified model performs better. The final result is regressed from the selected subset of raw full spectrum.

2.4 Assessing Model Performance

The performance of regression models should be evaluated by the use of an independent validation set before application. A *k*-fold cross validation with *k* set to 5 was adopted in this paper. We created five random partitions of the data and then predicted SOC values of samples in one partition with the model calculated using the rest four partitions. Therefore, all the samples were predicted once in a round. The quality of models was assessed by following statistics:

$$RMSECV = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(y_{i,p} - y_{i,m}\right)^2} \tag{2}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}\left(y_{i,p} - y_{i,m}\right)^2}{\sum_{i=1}^{n}\left(\overline{y_m} - y_{i,m}\right)^2} \tag{3}$$

where n is the number of samples, $y_{i,m}$ is target content measured by chemical method, $y_{i,p}$ is the predicted target content value of ith sample, $\overline{y_m}$ is the average content value of all samples. The *RMSECV* is the Root Mean Square of Error of Cross Validation and the R^2 is coefficient of determination.

The *RMSECV* and R^2 are important statistical parameters to evaluate the estimation accuracies of calibration models. A smaller *RMSECV* means better accuracy of prediction. Meanwhile, $R^2 < 0.5$ indicate unsuccessful models, which are not recommended; $0.50 < R^2 < 0.65$ indicate poor models; a model with R^2 between 0.65 and 0.81 is a fair model can be used for approximate quantitative estimation; a model with R^2 between 0.81 and 0.9 is a good quantitative model; and $R^2 > 0.9$ indicate excellent models [21].

In this paper, we focused on the relationship between SOC content and Vis-NIR spectrum of soil samples. Three datasets contain different number of samples were used for comparison. The largest dataset contains all 19,036 samples of LUCAS, called full dataset. The other two datasets, namely subset 4 k and subset 1 k, contain 4,000 and 1,000 samples randomly chosen from LUCAS, respectively. The geographic information, the notes of many stones, classification of soil and any other information in reference records were all ignored in modeling.

The raw spectrum from the Vis-NIR spectrometer offers 4,200 wavelengths in the region with an interval of 0.5 nm. A down-sampling step were performed on the raw data to intervals of 2 nm, 4 nm, 6 nm, …, 30 nm to evaluate the effect of different intervals. Neither mathematical pre-treatments like standard normal variate, multiplicative scatter correlation, nor useless variables removal based on the knowledge of the instrumental artifacts or the cause of the spectrum were employed different from the usual practices [22]. In addition, normalization of data was accomplished to enhance features and curve shape. Because of the random partition for cross validation, all these methods were calculated 20 times for each setting of datasets to obtain a fair result using averaged evaluation.

The chemometric and statistical experiments were self-written using the R 3.3.2 software [23] with package prospectr 0.1.3 [24] and MATLAB R2014b (The MathWorks Inc., Natick, MA, USA) with statistics toolbox.

3 Results and Discussion

3.1 Raw Spectrum or Derivatives

The estimation of derivatives by Savitzky-Golay algorithm [25]. The first derivatives used a 3-point window and a first-order polynomial, and the second derivatives used a 3-point window and a second-order polynomial. Three different variables input to MLR, PCR, PLSR and SR models using full dataset and subset 1 k, and the result of *RMSECV* are shown in Fig. 2. The interval of spectra shown were 4 nm.

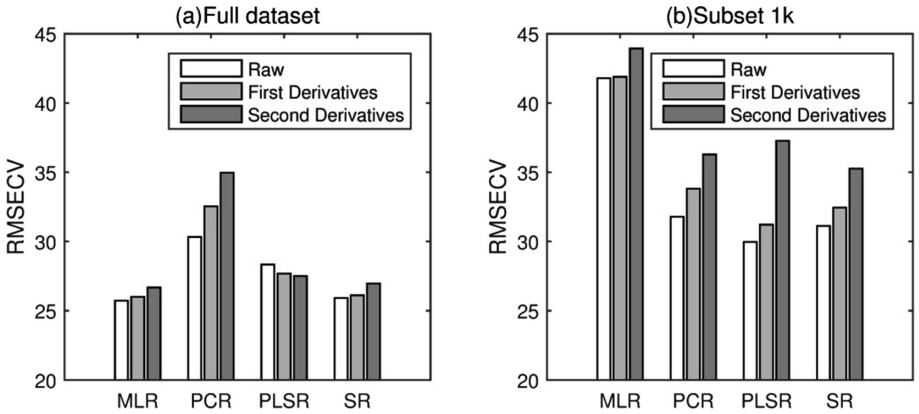

Fig. 2. The evaluation of raw spectrum and derivatives for the four methods. (a) comparison of *RMSECV* using full dataset; (b) comparison of *RMSECV* using subset 1 k.

The *RMSECV* of raw spectra used as input of MLR, PCR and SR models with full dataset was smaller than that of both first derivatives and second derivatives. The PLSR model with input of second derivatives achieved *RMSECV* = 27.5059, better than *RMSECV* = 28.3380 of raw spectrum and *RMSECV* = 27.6787 of first derivatives. When subset 1 k was applied, the best performance of all models were achieved by input of raw spectrum.

3.2 Component Number in PCR and PLSR

The key step of both PLSR and PCR are the compression of original predictor variables to several representative components. Figure 3 displays the *RMSECV* results of different numbers of components, range from 1 to 40, which were evaluated using full dataset with 4 nm interval. Using PCR method, the *RMSECV* dropped rapidly from 81.6008 to 38.5681 when the number of principle components increased from 1 to 12,

and decreased to 30.3287 gradually. The result of PLSR method showed the same trend, with the *RMSECV* decreased sharply from 80.2717 to 36.3023 when PLS component number increased from 1 to 8, and slowly achieved 28.3380 with the components continue to rise to 40. It should be noted that 40 components were used both in PCR and PLSR methods when compared to other methods in this paper.

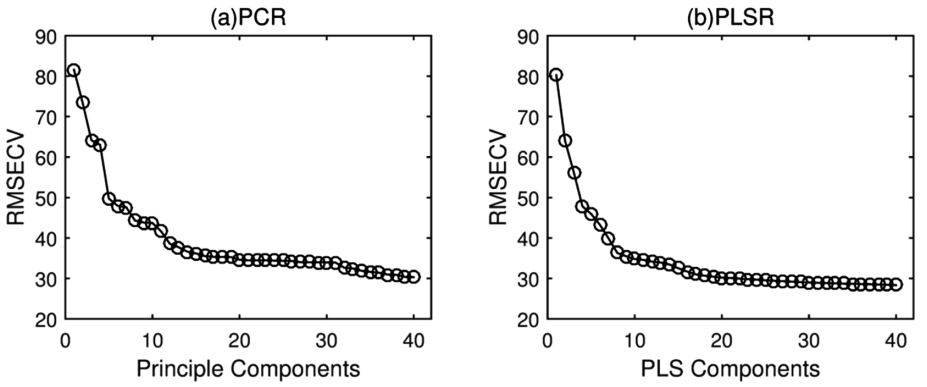

Fig. 3. The effect of the number of components. (a) the comparison of principle components in PCR method; (b) the comparison of PLS components in PLSR method.

3.3 Interval and Sample Number

The developed calibration models were validated with three different sizes of datasets with intervals range 2 nm to 30 nm. Figure 4 shows the comparison of *RMSECV*, and Table 2 lists *RMSECV* and R^2 in some parts of experimental results.

Among all these experiments, the best results were achieved by SR (*RMSECV* = 25.3912, R^2 = 0.9227) and MLR (*RMSECV* = 25.4254, R^2 = 0.9225), both using full dataset with 2 nm interval.

The accuracies of different models were compared under different conditions. For full dataset with intervals of 2 nm and 16 nm, the performance of MLR (*RMSECV* = 25.4254 and 27.3204), were very close to that of stepwise (*RMSECV* = 25.3912 and 27.3638), both outperformed PLSR (*RMSECV* = 28.3407 and 28.5977) and PCR (*RMSECV* = 30.3544 and 30.2450).

The repeated accuracies of these models were concerned so that these validation were calculated for 20 times. Table 2 lists the standard deviation (STD) of repeated attempts in the last column. The STDs using subsets were several times larger than that using full dataset.

3.4 Discussion

As stated above, the prediction of SOC based on Vis-NIR spectroscopy were studied. The performance of four linear calibration models, namely MLR, PCR, PLSR and SR

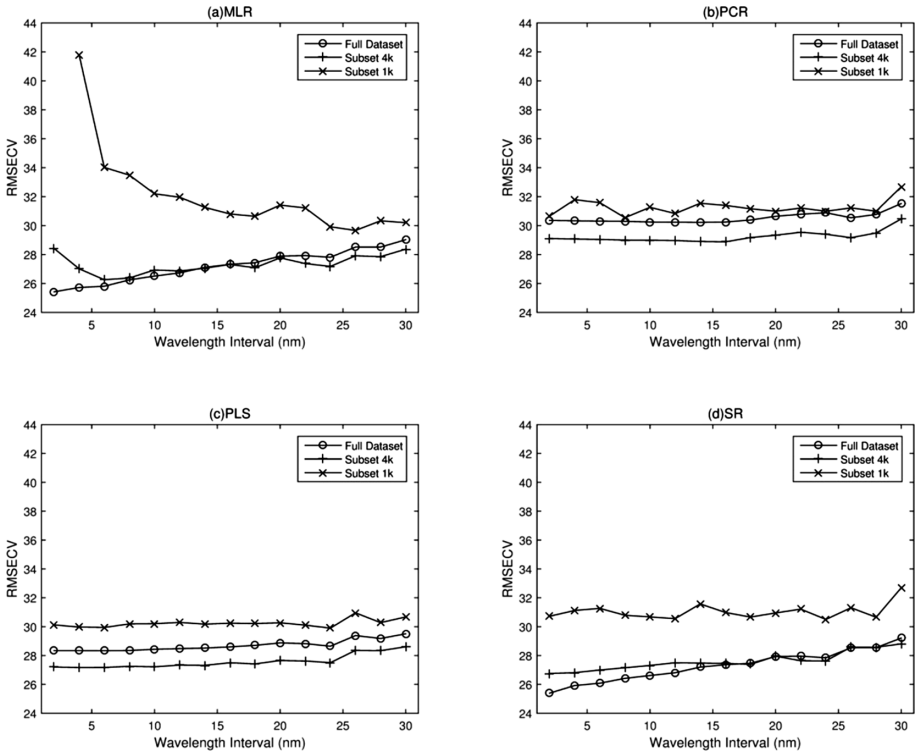

Fig. 4. The effect of wavelength interval with three datasets of different sample size.

were compared under different conditions, differed in model input, spectral interval and number of samples.

Raw spectra are more suitable than the derivatives as the input of a linear calibration model. Although derivatives make the spectra smooth, reduce the noise and enhance the feature in theory, raw spectra used as input of models offered better prediction.

The results for PCR and PLSR are slightly affected by the influence of wavelength interval. Because the predictions were based on several orthogonal components compressed from spectra which were highly collinear, and the down sampling step brought little information loss unless the interval was bigger than 24 nm. In PCR and PLSR methods, the performance was gradually increasing with more component number, however the promotion with more than 20 was limited.

If the sample number is relatively large, MLR and SR methods are better choice, and offer more accurate and stable prediction, even outperform PCR and PLSR. The downsampling step leads to some loss in accuracy, but trades off reduction of computation.

Table 2. Comparison of prediction accuracy using different datasets.

	Samples	Interval (nm)	Dims	Model	R^2	*RMSECV* (g/kg)
Full dataset	19036	2	1050	MLR	0.9225	25.4254 ± 0.0507
				PCR	0.8895	30.3544 ± 0.0177
				PLSR	0.9037	28.3407 ± 0.0269
				SR	0.9227	25.3912 ± 0.1549
		16	132	MLR	0.9105	27.3204 ± 0.0419
				PCR	0.8903	30.2450 ± 0.0168
				PLSR	0.9019	28.5977 ± 0.0277
				SR	0.9102	27.3638 ± 0.0408
Subset 4 k	4000	2	1050	MLR	0.9010	28.4219 ± 0.2745
				PCR	0.8962	29.1095 ± 0.0872
				PLSR	0.9094	27.2023 ± 0.1438
				SR	0.9123	26.7577 ± 0.3615
		16	132	MLR	0.9086	27.3131 ± 0.1985
				PCR	0.8979	28.8758 ± 0.0760
				PLSR	0.9074	27.4993 ± 0.1567
				SR	0.9077	27.4538 ± 0.1518
Subset 1 k	1000	2	1050	MLR	–	–
				PCR	0.9048	30.6532 ± 0.3338
				PLSR	0.9081	30.1207 ± 0.5210
				SR	0.9044	30.7190 ± 0.6910
		16	132	MLR	0.9039	30.8037 ± 0.7147
				PCR	0.8995	31.3939 ± 2.7045
				PLSR	0.9074	30.2397 ± 0.4229
				SR	0.9026	30.9834 ± 1.4537

4 Conclusion

In this study, the performance of four regression techniques include MLR, PCR, PLSR and SR were compared to identify a best model to predict the content of organic carbon in soil samples. Although derivatives make the spectra smooth, reduce the noise and enhance the feature theoretically, the prediction accuracies are not as good as raw spectra when used as predictor variables. The full spectrum MLR and SR outperforms PCR and PLSR when the dataset is quite large. PCR, PLSR and SR are candidate models for either large or small dateset, however MLR is only feasible when the number of samples are far bigger than the dimension of predictor variables. The performance of PCR and PLSR has little effect on the interval of the spectra due to the highly collinearity. In addition, more samples make a model more stable and eliminate the randomness of the dataset.

Acknowledgments. This work was supported by National Science Foundation of China under Grant No. 31671586.

References

1. Rossel, R.V., Walvoort, D., McBratney, A., Janik, L.J., Skjemstad, J.: Visible, near infrared, mid infrared or combined diffuse reflectance spectroscopy for simultaneous assessment of various soil properties. Geoderma **131**(1), 59–75 (2006)
2. Askari, M.S., Cui, J., O'Rourke, S.M., Holden, N.M.: Evaluation of soil structural quality using VIS–NIR spectra. Soil Tillage Res. **146**, 108–117 (2015)
3. Rossela, R.A.V., et al.: Guest editorial: near infrared spectroscopy for a better understanding of soil (2016)
4. Li, X., He, Y., Wu, C.: Non-destructive discrimination of paddy seeds of different storage age based on Vis/NIR spectroscopy. J. Stored Prod. Res. **44**(3), 264–268 (2008)
5. Ramirez-Lopez, L., Behrens, T., Schmidt, K., Stevens, A., Demattê, J.A.M., Scholten, T.: The spectrum-based learner: A new local approach for modeling soil vis–NIR spectra of complex datasets. Geoderma **195**, 268–279 (2013)
6. Ramirez-Lopez, L., Behrens, T., Schmidt, K., Rossel, R.V., Demattê, J., Scholten, T.: Distance and similarity-search metrics for use with soil vis–NIR spectra. Geoderma **199**, 43–53 (2013)
7. Ladoni, M., Bahrami, H.A., Alavipanah, S.K., Norouzi, A.A.: Estimating soil organic carbon from soil reflectance: a review. Precis. Agric. **11**(1), 82–99 (2010)
8. Balabin, R.M., Lomakina, E.I.: Support vector machine regression (SVR/LS-SVM)an alternative to neural networks (ANN) for analytical chemistry? comparison of non-linear methods on near infrared (NIR) spectroscopy data. Analyst **136**(8), 1703–1712 (2011)
9. Morellos, A., et al.: Machine learning based prediction of soil total nitrogen, organic carbon and moisture content by using vis-nir spectroscopy. Biosyst. Eng. **152**, 104–116 (2016)
10. Andersen, C.M., Bro, R.: Variable selection in regressiona tutorial. J. Chemom. **24**(11–12), 728–737 (2010)
11. Centner, V., Massart, D.L., de Noord, O.E., de Jong, S., Vandeginste, B.M., Sterna, C.: Elimination of uninformative variables for multivariate calibration. Anal. Chem. **68**(21), 3851–3858 (1996)
12. Li, H., Liang, Y., Xu, Q., Cao, D.: Key wavelengths screening using competitive adaptive reweighted sampling method for multivariate calibration. Anal. Chim. Acta **648**(1), 77–84 (2009)
13. Wang, W., Yun, Y., Deng, B., Fan, W., Liang, Y.: Iteratively variable subset optimization for multivariate calibration. RSC Adv. **5**(116), 95771–95780 (2015)
14. Hummel, J., Sudduth, K., Hollinger, S.: Soil moisture and organic matter prediction of surface and subsurface soils using an nir soil sensor. Comput. Electron. Agric. **32**(2), 149–165 (2001)
15. Montanarella, L., Tóth, G., Jones, A.: Soil component in the 2009 lucas survey. Land quality and land use information in the European Union. JRC, Office for Official Publications of the European Communities, Luxembourg, pp. 209–220 (2011)
16. Panagos, P., Van Liedekerke, M., Jones, A., Montanarella, L.: European soil data centre: response to european policy support and public data requirements. Land Use Policy **29**(2), 329–338 (2012)
17. Tóth, G., Jones, A., Montanarella, L.: LUCAS Topsoil Survey: Methodology, Data and Results. Publications Office (2013)

18. Rinnan, Å., van den Berg, F., Engelsen, S.B.: Review of the most common preprocessing techniques for near-infrared spectra. TrAC Trends Anal. Chem. **28**(10), 1201–1222 (2009)
19. Wold, S., Sjöström, M., Eriksson, L.: PLS-regression: a basic tool of chemometrics. Chemom. Intell. Lab. Syst. **58**(2), 109–130 (2001)
20. Vohland, M., Ludwig, M., Thiele-Bruhn, S., Ludwig, B.: Determination of soil properties with visible to near-and mid-infrared spectroscopy: effects of spectral variable selection. Geoderma **223**, 88–96 (2014)
21. Vohland, M., Besold, J., Hill, J., Fründ, H.C.: Comparing different multivariate calibration methods for the determination of soil organic carbon pools with visible to near infrared spectroscopy. Geoderma **166**(1), 198–205 (2011)
22. Wang, Y., et al.: Soil pH value, organic matter and macronutrients contents prediction using optical diffuse reflectance spectroscopy. Comput. Electron. Agric. **111**, 69–77 (2015)
23. R Core Team: R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria (2016). https://www.R-project.org/
24. Stevens, A., Ramirez-Lopez, L.: An introduction to the prospectr package (2013), r package version 0.1.3
25. Savitzky, A., Golay, M.J.: Smoothing and differentiation of data by simplified least squares procedures. Anal. Chem. **36**(8), 1627–1639 (1964)