Indian Journal of Agricultural Sciences **91** (9): 1328–32, September 2021/Article
https://doi.org/10.56093/ijas.v91i9.116080

Rapid prediction of soil available sulphur using visible near-infrared reflectance spectroscopy

BHABANI PRASAD MONDAL[1], RABI NARAYAN SAHOO[1]*, NAYAN AHMED[1], RAJIV KUMAR SINGH[1], BAPPA DAS[2], NILIMESH MRIDHA[3] and SHALINI GAKHAR[1]

ICAR-Indian Agricultural Research Institute, New Delhi 110 012, India

Received: 24 December 2020; Accepted: 25 February 2021

ABSTRACT

Rapid and accurate prediction of soil available S, an important secondary nutrient, is crucial for its site-specific management in a cultivated region. Although traditional chemical analysis of any nutrient is an accurate method, but often costly, time-consuming and destructive in nature. Recently visible near-infrared (VIS-NIR) reflectance spectroscopic technique has gained its popularity for rapid, non-destructive and cost-effective assessment of soil nutrients. Hence, a study was carried out in an intensively cultivated region of Katol block of Nagpur, Maharashtra, during 2018–20 for rapid prediction of soil available S using spectroscopic technique. Both spectroscopic and chemical analyses were carried out using 132 georeferenced surface soil samples (0-15 cm depth). The descriptive statistical analysis showed that the available S content varied from 1.09 to 47.88 mg/kg. Multivariate models namely partial least square regression (PLSR) and random forest (RF) were applied to develop spectral models for S prediction from spectral dataset. Several statistical diagnostics like coefficient of determination (R^2), root mean square error (RMSE), ratio of performance deviation (RPD) and ratio of performance to interquartile distance (RPIQ) were used to evaluate the performances of two models. The best prediction of S was achieved from nonlinear RF model (R^2 = 0.71, RMSE = 8.86, RPD =1.18, RPIQ = 1.69) as compared to linear PLSR model (R^2 = 0.53, RMSE = 9.04, RPD = 1.16, RPIQ = 1.66) datasets. Therefore, the result suggested applying non-linear multivariate model (RF) for obtaining best predictability for S from spectroscopic technique.

Keywords: Available sulphur, Multivariate models, PLSR, Reflectance spectroscopy, RF

Traditional laboratory-based soil chemical analysis for a large number of soil samples is very much time-intensive and thus making the precise management of soil nutrients very difficult at a large scale (Takele and Iticha 2020). Soil available sulphur (S) is considered as an important secondary nutrient for crop production along with other major nutrients because it is essential for enhancing oil content, nitrogen-fixation and protein synthesis. Therefore, precise management of S is necessary for getting satisfactory crop yield. However, chemical analysis of this nutrient is often neglected especially when large samples have to be analysed for addressing the spatial variability issues due to time limiting constraints and cost. Therefore, soil analysis techniques should be very fast and accurate for quantifying a large number of samples in a very short period of time (Shepherd and Walsh 2002). Visible-Near Infrared (VIS-NIR) reflectance spectroscopic technique coupled with statistical multivariate modelling is a suitable alternative which has obtained its popularity because of its rapid predictability for several nutrients like soil organic carbon, nitrogen, phosphorus, potassium and other nutrients (Peng *et al*. 2014, Xu *et al*. 2018, Mondal *et al*. 2019, Mondal and Sekhon 2019, Mondal *et al*. 2020). However, very limited studies have been conducted for the prediction of soil available S by using VIS-NIR spectroscopy (Chodak *et al*. 2001). Prediction of S especially organic form of S using VIS-NIR spectroscopy mainly depends on the characteristic absorption of spectra by the functional groups (C-S, S-H, C-C) present in organic form of S (Chodak *et al*. 2001). Multivariate statistical techniques like partial least square regression (PLSR), random forest (RF), multivariate adaptive regression splines (MARS) etc. are required for developing spectral models for the prediction of any soil available nutrient (Viscarra Rossel and Behrens 2010, Wold *et al*. 2012).

As the research area regarding soil available S prediction employing VIS-NIR technique is very much limited, the present study has been designed to assess the potential

Present address: [1]ICAR-Indian Agricultural Research Institute, New Delhi; [2] ICAR-Central Coastal Agricultural Research Institute, Goa; [3]ICAR-National Institute of Natural Fibre Engineering and Technology, Kolkata. *Corresponding author e-mail: rnsahoo.iari@gmail.com.

of VIS-NIR spectroscopy in soil S prediction using two important multivariate techniques namely PLSR and RF.

MATERIALS AND METHODS

The present study was carried out during 2018–20 in an intensively cultivated region of Katol block, Nagpur district of Maharashtra, India. Geographically it is situated at 21.27° N latitude and longitude of 78.58°E and the approximate elevation of the study area is 417 meters above mean sea level. The study site receives annual rainfall of around 1000 mm and the dominant cultivable crops are rice, wheat, mustard etc. Citrus and orange are the major plantation crops of the study area.

Nearly 132 georeferenced surface (0-15 cm depth) soil samples were collected using a portable Global Positioning System (GPS) device (Trimble). After grinding and passing through 2 mm sieve, part of the soil sample was subjected to S wet chemistry analysis following the standard methodology, i.e. $CaCl_2$ extractable sulphur (Williams and Steinbergs 1969).

Another part of the sample was used for recording the VIS-NIR spectra using a contact probe of Field Spectroradiometer (Analytical Spectral Devices Inc., Boulder, Colorado, USA, ASD) in the spectral range of 350-2500 nm at 1 nm sampling interval. The sensor was calibrated by measuring reference spectra of white spectralon before spectral measurements. After acquiring the spectra, several spectral preprocessing techniques like splice corrections, noise removal, smoothening of spectra using Savitzky-Golay (Savitzky and Golay 1964) filtering techniques etc. were applied to transform the raw spectra and to increase the signal to noise ratio. Here, the noise containing bands lied at the starting portion of the spectrum ranging from 350-359 nm and at the edges of the spectrum ranging from 2451-2500 nm were removed before model development. Further, the Savitzky-Golay smoothening technique with second order polynomial with a window size of 15 was employed to enhance the spectral features. Spectral data points were resampled at 5 nm interval due to existence of spectral collinearity. Outliers were carefully removed from the dataset and then the standard normal variate (SNV) method was used to normalize the dataset for effective modeling purposes. Several descriptive statistical parameters like mean, median, minimum, maximum, standard deviation (SD), coefficient of variation percentage (%CV), skewness and kurtosis were computed to know the nature of distribution of soil available S in the study area.

In next phase, whole dataset was randomly divided into calibration and validation datasets using 70% and 30% data respectively in order to develop spectral model. In this study, two popular multivariate statistical techniques namely PLSR and RF were employed for spectral modeling of soil S. PLSR is the most commonly used regression technique that has the capability to reduce the multicollinearity of the dataset and simultaneously, it can also increase the predictive performance of the spectral model by enhancing the covariance between predictor and response variables (Thissen *et al*. 2004). Important wavebands associated with available S prediction were identified using a statistical technique, i.e. the variable importance for projection (VIP) given by Wold *et al*. (2001).

The model RF is based on the concept of ensemble learning algorithms and it mainly comprises of numerous classification and regression trees (Breiman 2001). It provides better prediction results by avoiding the under or overfitting of the data.

A couple of statistical indices like coefficient of determination (R^2), root mean square error (RMSE), ratio of performance deviation (RPD) and ratio of performance to interquartile distance (RPIQ) were applied to test the efficacy of the newly developed spectral models using PLSR and RF. Higher R^2, RPD and RPIQ values and lower RMSE values indicate good performance of the prediction model. The RPD values lower than 1.4, 1.4-2.0 and more than 2.0 indicates poor, good and very good predictions respectively by the model (Chang *et al*. 2001); whereas, the RPIQ values in the range of lower than 1.5, 1.5-2.5 and greater than 2.5 approximately denotes poor, good and very good performances of the model correspondingly (Salazar *et al*. 2020).

RESULTS AND DISCUSSION

Descriptive statistical analysis for soil available S: The data (Table 1) describes the classical or descriptive statistical parameters of soil available S for the whole, calibration and validation datasets separately.

In case of entire dataset, the available S content varied from 1.09 to 47.88 mg/kg with a coefficient of variation (CV) of 81.74%. For calibration dataset, the S content varied in the same range like entire dataset. But the variability of calibration dataset in terms of CV (CV=80.88%) was little bit lower than the entire dataset. However, S content showed the largest variability (CV =83.55%) in the validation dataset. In the validation dataset, the S content ranged from 1.83 to 39.84 mg/kg with a mean value of 12.52 mg/kg. The positive skewness values for all three datasets denoted that the distribution was mostly concentrated around lower values coupled with a few higher values. The overall summary statistics represented a good variability of all these three datasets which signified that the model would be calibrated

Table 1 Descriptive statistical parameters of soil available sulphur (mg/kg) in the study area

Soil property	Minimum	Maximum	Mean	Median	SD	% CV	Skewness	Kurtosis
Whole dataset	1.09	47.88	13.42	9.14	10.97	81.74	1.16	0.43
Calibration dataset	1.09	47.88	12.82	8.41	10.37	80.88	1.34	1.16
Validation dataset	1.83	39.84	12.52	8.41	10.46	83.55	1.13	0.39

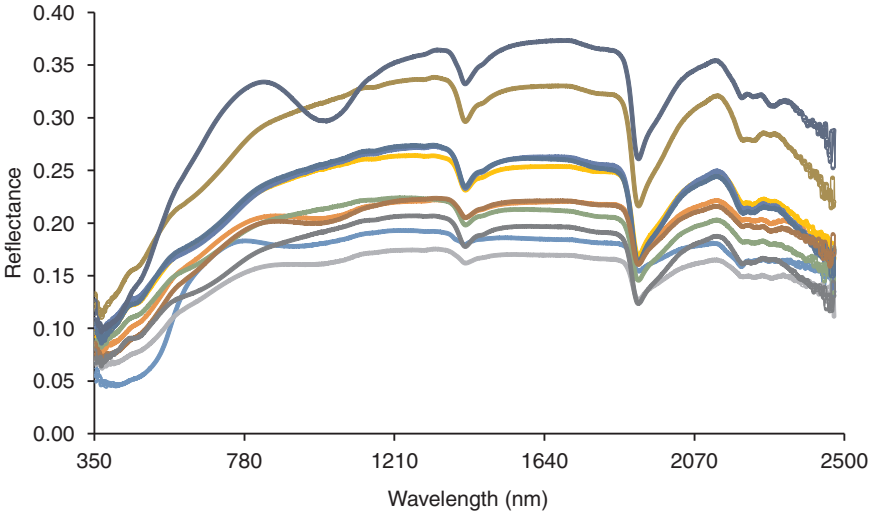

Fig 1 Reflectance spectra of few soil samples of the study area in the spectral range of 350-2500 nm.

well for the prediction of studied soil property (available S) using multivariate models.

Spectral reflectance characteristics of soil samples: Spectrally active constituents called chromophores present in soil samples are mainly responsible for obtaining typical spectral response curve of soil samples, represented in Fig 1.

The presence of organic matter, texture, moisture content and so many other factors modify the characteristics of soil spectral response curve. Some typical absorption features were noticed at certain wavebands in the spectral response curve of soil samples like around 1400 and 1900 nm due to presence of soil moisture and near 2200 nm due to presence of hydroxyl group (-OH) of moisture or clay minerals like kaolinite present in soil.

Performance of spectral model in calibration dataset: Fig 2 illustrated the performance of the spectral model for the prediction of soil available S in the calibration dataset through constructing two scatter plot diagrams. Two important and popular multivariate methods namely PLSR and RF were employed to model the spectroscopic data in order to predict soil S.

For calibration dataset, the best laboratory-based spectroscopic prediction of available S was obtained from

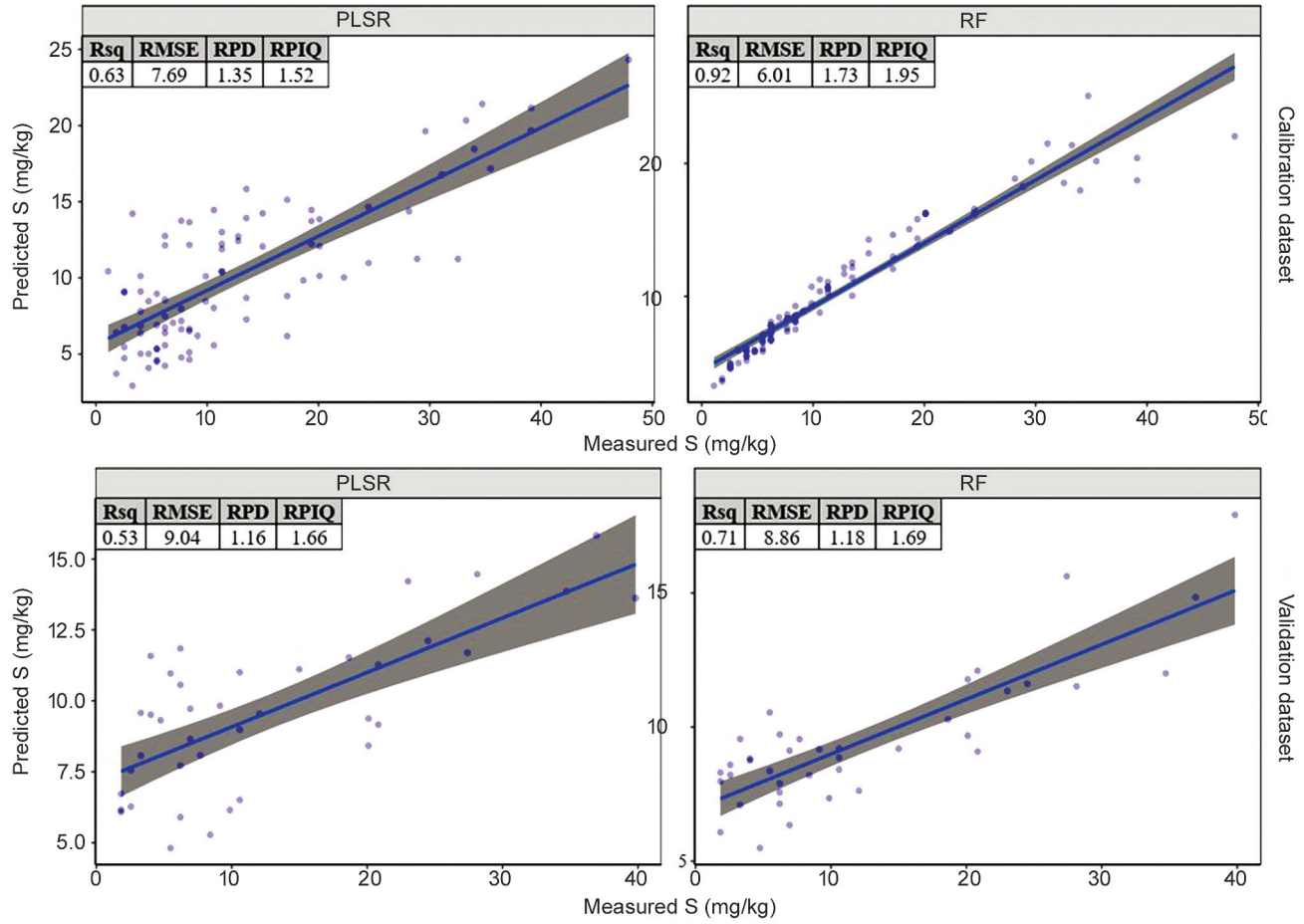

Fig 2 Scatter plots of measured vs spectral model predicted values of soil available S in calibration (upper row) and validation datasets (lower row).

RF model (R^2 = 0.92, RMSE = 6.01, RPD = 1.73 and RPIQ = 1.95), whereas the least prediction performance was achieved from linear PLSR based model (R^2 = 0.63, RMSE = 7.69, RPD = 1.35 and RPIQ = 1.52). Therefore, the results during calibration showed that the RF model outperformed the PLSR model for available S prediction in terms of statistical diagnostics like R^2, RMSE, RPD and RPIQ using VIS-NIR spectroscopy. The results were supported by the findings of Chodak *et al*. (2001), who stated that near infrared reflectance spectroscopy (NIRS) technique was better in soil S prediction of the top organic soil layer. Previous research areas were mainly confined to the use of PLSR model. However, in this present study, the performance of machine learning based nonlinear model namely RF was tested and compared with linear model PLSR in S prediction. Nawar and Mouazen (2019) reported that RF model could outperform several other machine learning based models in soil property prediction. Our findings also supported that RF was better than PLSR in available S prediction.

Performance of spectral model in validation dataset: The scatterplot diagrams of the validation datasets for available S prediction were also portrayed in the lower row (Fig 2). The results followed a similar trend in validation dataset also. For the validation dataset, the best prediction performance was attained from the RF model (R^2 = 0.71, RMSE = 8.86, RPD = 1.18 and RPIQ =1.69), whereas the lowest prediction performance was gained from PLSR model (R^2 = 0.53, RMSE = 9.04, RPD = 1.16 and RPIQ = 1.66). The higher R^2, RPD and RPIQ values and relatively lower RMSE values denoted that the RF model predicted the studied soil parameter (available S) well. Chodak *et al*. (2001) obtained better values of correlation coefficient (r = 0.86) in the validation dataset of soil available S prediction. Although they used PLSR model in their study, their prediction performance was relatively better than the present study using PLSR model. However, the present study demonstrated the better potential of RF model over PLSR model in available S prediction. The better predictive performance of RF model could be explained by the non-linear nature of the model, which made it capable to capture the non-linearity existing in the spectral datasets (Nawar and Mouazen 2019). Our results differed with some previous findings, reporting superior performance of PLSR over RF (Das *et al*. 2020).

The significant functional groups, i.e. C-S, S-H, C-C etc. present in the organic form of S and their overtones and combinations around NIR region influence the prediction of S using VIS-NIR spectroscopy. The present study demonstrated relatively lower performances of PLSR model than the previously reported studies because this present study was confined to the prediction of mainly inorganic form of S, which was present in a very little amount in the soil system to influence the spectral reflectance curve of the soil samples. Most of the soil S exists (90-98%) in the form of organic S, having higher capabilities to influence the spectral response patterns of soil. The better results may be obtained if the spectral measurements of organic form of S or total S were recorded rather than inorganic form of S or available S due to excessive influence of organic or total S in soil spectral characteristics (Chodak *et al*. 2001). The predictions of available S can also be improved through selecting the important wavelengths associated with S by employing statistical techniques. Therefore, significant wavebands associated with the prediction of available S were selected using feature selection technique namely VIP (Wold *et al*. 2001). This technique is mainly based on PLSR models, which help to find significant spectral features or wavebands from the spectra based on VIP threshold value. If the VIP threshold value was greater than one for a particular waveband and also the peak maxima of reflectance was observed for that point, then that waveband was considered as important waveband (Chong and Jun 2005). The most important waveband regions identified for available S prediction were 400-621 nm, 814-1206 nm, 1380-1386 nm and 2153-2207 nm in this study.

To conclude, the present study emphasized the potential of reflectance-based VIS-NIR spectroscopy for rapid prediction of soil available S with the help of two important multivariate techniques namely PLSR and RF. The study demonstrated that the non-linear model RF outperformed the linear PLSR model due to better capacity of handling non-linear data. Therefore, the present study recommends to apply nonlinear technique especially RF for fast prediction of S using spectroscopic technique. Moderate performance of two multivariate models compared to previously conducted studies might be attributed to the form of S studied (available S) and relatively smaller sample size. It is expected that better result may be obtained for organic form of S along with large sample size. Therefore, the study could be further extended to observe the potential of VIS-NIR spectroscopy with organic form of S for a large sample size. The other non-linear techniques may be exploited in future studies to achieve better prediction accuracies for all forms of S.

ACKNOWLEDGEMENTS

The first author would like to thank the Department of Science and Technology (DST) for providing financial assistance in the form of Junior Research Fellowship (JRF) for conducting this Ph D research work.

REFERENCES

Breiman L. 2001. Random forests. *Machine Learning* **45**: 5–32.

Chang C W, Laird D A, Mausbach M J and Hurburg C R J. 2001. Near-infrared reflectance spectroscopy – principal component regression analysis of soil properties. *Soil Science Society of America Journal* **65**: 480–90.

Chodak M, Ludwig B, Khanna P and Beese F. 2001. Use of near infrared spectroscopy to determine biological and chemical characteristics of organic layers under spruce and beech stands. *Journal of Plant Nutrition and Soil Science* **165**: 27–33.

Chong I G and Jun C H. 2005. Performance of some variable selection methods when multicollinearity is present. *Chemometrics and Intelligent Laboratory Systems* **78**: 103–12.

Das B, Sahoo R N, Pargal S, Krishna G, Verma R, Chinnusamy V, Sehgal V K and Gupta V K. 2020a. Comparative analysis of index and chemometric techniques-based assessment of leaf area index (LAI) in wheat through field spectroradiometer, Landsat-8, Sentinel-2 and Hyperion bands. *Geocarto International* **35**: 1415–32.

Mondal B P and Sekhon B S. 2019. Using diffuse reflectance spectroscopy for assessment of soil phosphorus status of an intensively cropped region. *Agricultural Research Journal* **56**: 657–61.

Mondal B P, Sekhon B S, Sahoo R N and Paul P. 2019. VIS-NIR reflectance spectroscopy for assessment of soil organic carbon in a rice-wheat field of Ludhiana district of Punjab. *International Archives of the Photogrammetry, Remote Sensing & Spatial Information Sciences* XLII-3/W6: 417–22.

Mondal B P, Sekhon B S, Paul P, Barman A, Chattopadhyay A and Mridha N. 2020. VIS-NIR reflectance spectroscopy as an alternative method for rapid estimation of soil available potassium. *Journal of the Indian Society of Soil Science* **68**: 323–30.

Nawar S and Mouazen A M. 2019. On-line vis-NIR spectroscopy prediction of soil organic carbon using machine learning. *Soil and Tillage Research* **190**: 120–27.

Peng X, Shi T, Song A, Chen Y and Gao W. 2014. Estimating soil organic carbon using VIS/NIR spectroscopy with SVMR and SPA methods. *Remote Sensing* **6**(4): 2699–2717.

Salazar D F U, Dematte J A M, Vicente L E, Guimaraes C C B, Sayao V M, Cerri C E P, Padilha C de M C and Mendes W D S. 2019. Emissivity of agricultural soil attributes in south eastern Brazil via terrestrial and satellite sensors. *Geoderma* 114038.

Savitzky A and Golay M J E. 1964. Smoothing and differentiation of data by simplified least squares procedure. *Analytical Chemistry* **36**: 1627–39.

Shepherd K D and Walsh M G. 2002. Development of reflectance spectral libraries for characterization of soil properties. *Soil Science Society of America Journal* **66**: 988–98.

Takele C and Iticha B. 2020. Use of infrared spectroscopy and geospatial techniques for measurement and spatial prediction of soil properties. *Heliyon* **6**(10): e05269.

Thissen U, Pepers M, Ustun B, Melssen W J and Buydens L M C. 2004. Comparing support vector machines to PLS for spectral regression applications. *Chemometrics and Intelligent Laboratory Systems* **73**: 169–79.

Viscarra Rossel R A and Behrens T. 2010. Using data mining to model and interpret soil diffuse reflectance spectra. *Geoderma* **158**(1-2): 46–54.

Williams C H and Steinbergs A. 1969. Soil sulphur fractions as chemical indices of available sulphur in some Australian soils. *Australian Journal of Agricultural Research* **10**: 340–52.

Wold S, Sjöström M and Eriksson L. 2001. PLS-regression: a basic tool of chemometrics. *Chemometrics and Intelligent Laboratory Systems* **58**: 109–30.

Xu S, Zhao Y, Wang M and Shi X. 2018. Comparison of multivariate methods for estimating selected soil properties from intact soil cores of paddy fields by Vis–NIR spectroscopy. *Geoderma* **310**: 29–43.