Evaluation of Multivariate Regression Models to Predict Electrical conductivity Using Vis-NIR and MIR Spectra

Seema Chahar[1], Amlan Kumar Ghosh[1], B S Das[2], Nagarjuna Reddy[2], K.M Hati[3], Nishant Sinha[3], and Nilimesh Mridha[4]

[1]Banaras Hindu University
[2]Indian Institute of Technology Kharagpur
[3]Indian Institute of Soil Science
[4]ICAR -National Institute of Natural Fibre Engineering and Technology

August 10, 2020

Abstract

Salts in the root zone have high spatial variability, changes rapidly and adversely affects soil quality and crop productivity. Rapid detection of electrical conductivity (EC) using visible-near infrared (Vis-NIR) and midinfrared (MIR) spectroscopy can alleviate the adverse effects on soil and plant, which through conventional method is time consuming. Soils were collected from the Indo-Gangetic plains and analyzed for EC using conventional, Vis-NIR, MIR spectroscopy and there was wide variation in EC measured by the conventional method. The spectral regions in 460-500 and 1890-1906 nm in the Vis-NIR region and 4200-4310, 5275-5280, 6660-6670, 7305-7310 and 8290-8300 nm in the MIR region were sensitive to detection of EC. Partial least square regression (PLSR) outperformed random forest regression (RF), support vector regression (SVR), and multivariate adaptive regression splines (MARS) both in Vis-NIR and MIR region during calibration. The ratio of performance deviation (RPD), coefficient of determination (R2) and root mean square error (RMSE) of the validation dataset were used to assess the prediction accuracy and the predictive performance of PLSR (2.44, 0.84, 0.21), RF (1.95, 0.81, 0.20), SVR (2.09, 0.78, 0.22) and MARS (1.81, 0.73, 0.27) models. PLSR model performed very well in the Vis-NIR range; however, in the MIR range, RF (1.43, 0.52, 0.20), followed by PLSR (1.40, 0.55, 0.35), performed better than SVR (1.39, 0.53, 0.35) and MARS (1.29, 0.44, 0.37). Vis-NIR spectroscopy with PLSR algorithm predicted EC better than MIR spectroscopy and would be the method of choice for rapid estimation and prediction of EC in the study region.

1. Introduction

Salt affected soils are caused by excess accumulation of salt which are pronounced at the soil surface. Salt is often derived from geological formations featuring shale, marl, limestone, sylvite, gypsum, and halite, but variability of soil salinity is mostly due to parent material, soil type, and landscape position (Clay *et al.,* 2001). Moreover, salts can be transported to the soil surface by capillary action from brackish water tables and can accumulate due to evaporation; they can also accumulate as a result of anthropogenic activities such as fertilization or oil production. Soil salinization is a universal problem and current estimations of the proportion of salt-affected soils in irrigated lands for several countries were 27 % in India, 20 % in Australia, 28 % in Pakistan, 50 % in Iraq and 30 % in Egypt (Stockle, 2013). The accumulation of soluble salts in the root zone greatly affect plant growth, resulting in lower crop yields and adversely affecting the soil fertility (Li *et al.* 2013).

Soil salinity is typically assessed by measuring the soil electrical conductivity in saturated paste extracts (ECe) or by using extracts with different soil-to-water ratios (Sonmez *et al.* , 2008). Developed in the mid-1950s, ECe is one of the most widely reported soil quality assessment parameters (Karlen *et al.,* 2008), regular

monitoring of which is essential for efficient soil and water management and sustainability of agricultural lands (Bilgili *et al.,* 2011). Electrical conductivity can act as an indirect indicator of important soil physical properties (Rhoades *et al.,* 1999) and provides important information about the impact that farm practices, such as irrigation and soil and crop management, have at both the field and regional scales. Therefore, reliable information on the nature and spatial extent of soil salinity is a prerequisite for restoring fertility and preventing further degradation. Thus, timely detection of the extent and magnitude of soil salinity is important for agriculture practices.

It is difficult to obtain up-to-date soil salinity information by using conventional techniques, to identify and monitor soil salinity because these techniques are time consuming and expensive and require high sampling densities and frequencies; hence efforts are being made to obtain more cost-effective methods for mapping soil salinity. During the last two decades, visible and near-infrared spectroscopy has been used as a rapid, cost-effective and relatively accurate method for analyzing conventional soil properties (Nocita *et al.,* 2015). Previously, several studies indicated that pure sodium chloride is featureless in Vis-NIR regions because salt is not a strong or direct chromophore (Metternicht *et al.,* 1997). However, the presence of salts in soils may result in subtle spectral responses when combined with—OH, which is common in soils. Therefore, soil salinity can be characterized by soil spectral reflectance or salinity spectral indices using partial least squares regression, artificial neural network, and stepwise multiple linear regression methods (Zhang *et al.,* 2011) and can be detected using high-resolution spectroscopy (Jin *et al.,*2015). Accordingly, interest in using reflectance spectroscopy as a rapid and effective tool for mapping soil salinity has recently grown and several studies have estimated salt contents of air-dried soils with reasonable accuracy using reflectance spectroscopy (Yong-Ling *et al.,* 2010).

Hyperspectral visible and near-infrared reflectance spectroscopy displays promise as a result of its performance, accuracy and cost effectiveness in the determination of most soil properties (Shepherd and Walsh 2002). Various statistical modeling techniques help to correlate a single reflectance spectrum of soil to a host of physical, chemical, mineralogical and microbiological attributes of that soil after proper calibration and validation of models. Principal component regression (PCR), partial least squares regression (PLSR), multivariate adaptive regression splines (MARS), artificial neural networks (ANN) are some of the commonly used diagnostics for calibration and validation of hyperspectral models (Bilgili *et al.* , 2010). The reliability of calibration of spectral data with chemical analysis data needs to be enhanced by factoring in variations on account of land use and choice of scale. The calibration process also needs to be made indubitable by using optimal sample size and sampling strategy. Once the calibration models between soil reflectance spectra and soil variables have been established, they can be used to predict unidentified parameters. Several regression methods based on visible near IR have been used to estimate soil salinity, and partial least-squares regression is the most common (Farifteh et al. 2007; Bilgili et al. 2011). The PLSR approach has inference capabilities that are useful for modelling a probable linear relationship between the measured reflectance spectra and salt content in soils (Farifteh et al. 2007). The MARS method is considered a nonparametric method that estimates complex nonlinear relationships among independent and dependent variables (Friedman 1991), and it has been effectively applied in different fields (Bilgili *et al.* , 2010; Felicísimo *et al.,* 2012) and generally exhibits high performance results compared with other linear and non-parametric regression models, such as principal component regressions, classification and regression trees and artificial neural networks.

This study was conducted to evaluate multivariate regression models to predict electrical conductivity using Vis-NIR and MIR Spectra as a substitute to conventional soil analysis. The specific goals of this study were to find out sensitive regions of the spectrum for modelling electrical conductivity and compare the performance of multivariate regression models PLSR, RF, SVR, and MARS for predicting EC, both in the Vis-NIR and MIR spectral region.

2. Materials and Methods

2.1. Study Area and Soil Sampling

The study was carried out in the middle Indo-Gangetic plain zone, India situated in the state of Uttar

Pradesh covering the administrative districts of Varanasi, Chanduali, Sant Ravidas Nagar and Mirzapur, localized between 82°30' and 83deg30' East and between 24deg30' and 25deg30' North covering a total area of approximately 9604 km^2(Fig 1). A total of 280 geo-referenced composite soil samples were collected from surface (0–15 cm) layers after crop harvest. Stainless steel soil auger was used for collection of soil samples and the coordinates of the sample points were recorded via hand held Global Positioning System (Model Garmin *etrex*). Samples were air dried and ground in a wooden pestle and mortar to pass through a 2 mm sieve and electrical conductivity of soil was measured in 1:2.5 soil water suspensions using electrical conductivity meter and expressed as dS m^{-1} (Ghosh *et al.,* 2012).

The processed soil samples were further homogenized in Retsch Mortar Grinder RM 200 for 3 minutes wherein samples were grounded up to 0.5 mm size. The RM 200 is suitable for the homogenous and reproducible sample preparation for precision of analysis. These ground soil samples were subsequently used for recording spectral signature in alpha- MIR spectrometer and the processed soil samples that passed through 2 mm sieve was used for obtaining spectral signature in NIR spectroradiometer.

2.2. Spectroscopic Measurement and Pre-Processing of Spectra

Collection of soil spectral data in Vis-NIR range

A portable spectroradiometer (Model: FieldSpec3 FR; Analytical Spectral Devices Inc., USA) equipped with a contact probe (10 mm spot size) was used for spectral reflectance acquisition across the wavelength range of 350 to 2500 nm, covering the visible (VIS), near-infrared (NIR) and shortwave-infrared (SWIR) regions. About 50 g of soil was placed in a special container (10 cm diameter), and the soil surface was leveled with a rubber cork used as a mallet (Mouazen *et al.,* 2010). A spectrum from each quadrant of the container was acquired by keeping the contact probe at the respective positions so as to have four reflectance spectra per soil sample. For each soil sample, a reference spectrum was also collected using a 9.2-cm diameter Spectralon white reference panel (Labsphere).

Collection of soil spectral data in MIR range

Air-dried, crushed and 0.5 mm ground samples, mentioned earlier, were filled in the cups meant of the Bruker alpha Fourier Transformed MIR Spectrometer for recording spectral signatures. The FT-MIR was stabilized for two hours to increase the amplitude count to more that 9000 and corrections were made for the background during instrument calibration, before recording spectra in the near and middle MIR range. For the Vis-NIR range, a spectroradiometer (Model: FieldSpec3 FR; Analytical Spectral Devices Inc., USA). The raw spectra collected from the Vis-NIR range ASD FieldSpec(r) and OPUS file from and MIR-spectrometer respectively were subsequently pre-processed using R software (version 3.3.3, The R Foundation for Statistical Computing Platform). The most widely used pre-processing techniques is divided into two categories: scatter-correction methods and spectral derivatives. Since many workers are of the openion that the soil properties can be related to absorbance and reflectance and their first and second derivatives and it has been reported that the absorption features in reflectance spectra were enhanced by derivative spectroscopy (Tsai, 1998), the reflectance data was transformed to absorbance through the expression, absorbance=$\log_{10}$ ($\frac{1}{\text{Reflectance}}$).

First and second derivatives were obtained from reflectance and absorbance data. The spectral derivative method consists of first derivatives (FD) and second derivatives (SD) of the reflectance spectrum using the equation (1) and (2) respectively.

$$\frac{\delta P}{\delta \lambda} = \mathbf{R}\left(\lambda\iota\right) - \frac{\mathbf{R}\left(\lambda \mathbf{i} - \mathbf{1}\right)}{\lambda \mathbf{i} - \lambda \mathbf{i} - \mathbf{1}} \qquad \ldots\ldots\ldots (\mathbf{1})$$

$$\frac{\mathbf{d^2 R}}{\mathbf{d}\lambda^{\mathbf{2}}} = \frac{\mathbf{d}}{\mathbf{d}\lambda}\left(\frac{\mathrm{dR}}{\mathbf{d}\lambda}\right) \qquad \ldots\ldots\ldots (\mathbf{2})$$

Where R, spectral reflection/absorbance; Yi, i^{th}wavelength / band. Further among scatter-correction methods, multiplicative scatter correction (MSC), standard normal variate (SNV) and other smoothing methods include averaging spectra, and median filters, first order derivative, second order derivative and the Savitzky–Golay transformation were used to reduce noise in spectral signals.

2.3 Selection of optimum spectral band width

A correlation analysis was made to establish the relationship between electrical conductivity with reflectance of individual band of each spectral data set using SPSS software (version 23.0). The correlation analysis was used to elucidate the most sensitive spectral regions for electrical conductivity.

2.4 Multivariate Regression Models

Regression analysis was done in R software and a number of regression models such as Partial Least Square Regression (PLSR) , Random Forest Regression (RF) , Support Vector Regression (SVR) , Multivariate Adaptive Regression Splines (MARS) were analysed using different R package 'pls', 'randomForest', 'kernlab' and 'earth, plotmo, plotrix, TeachingDemos' (Clyde *et al.,* 2017), for PLSR, RF, SVR and MARS respectively of R 3.3.3 (The R Development Core Team, 2017).

2.5 Model Evaluation

The coefficient of determination (R^2) in validation dataset, root mean square error of prediction (RMSEP) and ratio of performance deviation (RPD) were used to evaluate models. Ranking was made on the basis of RPD values; higher the RPD better was the model performance. When two models had same RPD values, R^2values were referred to, and models with higher R^2value better explained the fitted data. When two models had same RPD and R^2 values, the RMSE values were referred to, and models with lower RMSE gave better prediction/ validation of data than those with higher RMSE. The R^2, PMSE and RPD were calculated as:

$$R^2 = 1 - \left(\frac{\sum_i n(Y_{\text{Pred}} - Y_{\text{measured}})^2}{\sum_i n\ Y_i - Y_{\text{mean}}{}^2}\right) \ldots\ldots\ldots \quad (3)$$

$$RMSE = \sqrt{\frac{\sum_i n(Y_{\text{Pred}} - Y_{\text{measured}})^2}{n-1}} \ldots\ldots\ldots \quad (4)$$

$$RPD = \frac{\text{SD}_{\text{val.}}}{\text{RMSE}} \ldots\ldots\ldots\ldots \quad (5)$$

Where, Ypred = predicted values; Ymean = mean of measured values; Y meas = measured values; n= number of predicted or measured values with I = 1,2,... n; SDval= standard deviation of measured values in the validation dataset; and RMSEP = root mean square error of prediction in validation dataset. The procedure used for model calibration and validation is presented:

2.6 Statistical Analysis

The Kolmogorov–Smirnov test was used to assess the normality of all variables, and the Quantile–Quantile (Q–Q) plots coupled with the skewness values, were used to evaluate the normality of the data sets (Vasu *et al.,* 2017). The measured EC was subjected to a descriptive analysis and minimum, maximum, mean, standard deviation, coefficient of variation, kurtosis and skewness were determined using SPSS version 23.0.

3. Result and discussion

3.1. Descriptive statistics of soil electrical conductivity

The data for electrical conductivity was not normally distributed as evident from the Q-Q plot (Fig 2a), where, considerable deviation from the straight diagonal line at both ends can be observed and hence the data was log transformed (to the base 10) to make it normally distributed (Fig 2b). The descriptive statistics revealed considerable variability of soil electrical conductivity (Table 1). The minimum and maximum values of electrical conductivity were 0.01 to 1.71 dS m^{-1}, with a mean value of 0.46 dS m^{-1} and the values of skewness and kurtosis were 1.0 and 2.34 respectively. Development of soil salinity is sometimes geogenic, being affected by parent material and arid climate, but is largely anthropogenic in the Indo-Gangetic plains,

being affected by fertilization, crop production or management history. Soil salinity is generally measured as electrical conductivity in the saturation paste or its liquid extracts, with soils having $EC_e > 4$ dS m^{-1} being referred to as saline. It may be reiterated that the procedure for measuring EC in the present study is that which is routinely followed in soil testing laboratories where EC is measured in the soil suspension (1: 2.5) used to measure pH, and would be considerably higher when measured in saturation extract, normally used to differentiate saline and non-saline soils. Moreover, soil salinity is an important parameter of estimating soil quality, is highly variable, changes rapidly over small distances spatially, and hence a large number of samples are required to adequately characterize soil salinity across landscapes (Aldabaa et al., 2015). Further any increase in electrical conductivity, has great bearing on efficient soil and water management and sustainable crop production (Bilgili et al., 2011). To alleviate the cost of extensive sampling, followed by sample preparation and detailed laboratory analysis, hyperspectral reflectance spectroscopy is a lucrative alternative for rapid characterization of salt content in soil, and justifies investigation.

3.2 Spectral preprocessing influence

Spectra obtained from Vis-NIR and MIR spectrometers were subject to preprocessing, such as absorbance, first order derivative, second order derivative, multiple scatter correction and standard normal variate. Spectra pre-treatment is a mathematical manipulation that enhances the spectral information and eliminates the physical effect of light scattering, which can be due to particles of different sizes and shapes of samples (Minasny and McBratney, 2008) and is thus the most important step before any chemometric modeling. Different pre-processing transformations have been applied in numerous studies to transform soil spectral data, remove noise, accentuate features, and prepare them for chemometric modelling. However, the first derivative, second derivative, SNV and MSC manipulation did not greatly enhances some of the spectral features compared to reflectance. Moreover reflectance (unprocessed spectra) presented the best performance as compared to other preprocessing methods, irrespective of the models used (PLSR, RF, SVR or MARS) (Table 2 & 3) and was thus considered to be the most robust spectral preprocessing method based on its predictive performance for EC. Some earlier results (Moros et al. 2009) also suggest that calibration models in which spectra were not preprocessed are more sensitive to changes compared to models for which preprocessing was applied and Nawar et al. (2016) re-confirmed it, and used no preprocessing for prediction. Reflectance has also been successfully used in other studies, to estimate soil properties (Viscarra Rossel et al. 2006, Nawar et al. 2016). Vibhute *et al.,* (2018) reported electrical conductivity to be better calibrated (R^2 = 0.80 and RMSE = 2.07) before pre-treatments than after pretreatment of spectra and Nocita et al. (2014) applied continuous removal reflectance to predict the soil properties by diffuse reflectance spectroscopy from soil samples throughout the European Union. The present study demonstrates that reflectance (unprocessed spectra) (Fig 3 a & b) is better than any preprocessing tool for prediction of EC regardless of the method applied and demonstrates its suitability for prediction of EC, both in the Vis-NIR and MIR spectral regions.

3.3. Correlation between reflectance data and soil properties

Absorption of radiation at molecular vibrational frequencies in the visible-infrared region forms the basis of this spectroscopic technique and a plot of correlation coefficient in the Vis-NIR and MIR region (Fig. 4 a & b) was used to establish the most sensitive spectral region for predicting of electrical conductivity. In the Vis-NIR region, peaks were observed in the visible region at 460-500 nm (with low correlation coefficient) and NIR region at 1890-1906 nm (correlation coefficient= +0.16) followed by a broad shoulder. Gaikwad (2020) reported conspicuous absorption in the region close to wavelength 427, 487 and 1917 nm and weak absorption features near 950, 1414, 2206, 2380 and 2460 nm; whereas Margate *et al.,* (2001) reported most sensitive spectral regions for determination of EC in soils of south Spain being 390-400, 615-625, 685-695, 800-810, 950-960, 1410-1420, 1935-1945, and 2350-2360 nm. The absorption features close to 1400 and 1900 nm represent the stretching of oxygen (O)-hydrogen (H) and bending of H-O-H of the free water and its overtones (Nawar et al., 2014). In the MIR region, five peaks had correlation coefficient > 2, three of them in the positive quadrant (5275-5280, 6660-6670, 7305-7310) and two in the negative quadrant (4200-4310, 8290-8300) and could be important spectral regions for predicting EC using models.

3.4 Calibrations and predictability of models

The calibration was carried out using randomly selected 196 samples from the dataset and validated on 84 samples using four algorithms, namely, partial least square (PLS), random forest (RF), multivariate adaptive regression splines (MARS) and support vector regression (SVR) methodology. In the calibration Vis-NIR data set, the values of R^2 and RMSE for PLSR model was 0.93, 0.12; using the RF model was 0.84, 0.15; while using the SVR model was 0.80, 0.21 and MARS was 0.86, 0.12. In the calibration MIR data set, the values of R^2 and RMSE values for PLSR, RF, MARS and SVR models were 0.94, 0.26; 0.84, 0.25; 0.80, 0.25 and 0.91, 0.25, respectively. R^2 is an important statistical measure which represents the proportion of the difference or variance in statistical terms for a dependent variable which can be explained by an independent variable or variables, and in short, determines how well data fit the regression model; whereas lower RMSE indicates better fit of data. From the calibration datasets it is clear that PLSR model outperformed other models in having higher R^2 and lower RMSE values (Table 2 and 3).

The predictive performance of PLSR, RF, SVR and MARS models for EC in the Vis-NIR range was evaluated and the respective values for PLSR were (R^2 = 0.84, RMSE=0.21 , RPD=2.44); for RF were (R^2 = 0.81, RMSE = 0.20, RPD=1.95); for MARS were (R^2 = 0.73, RMSE = 0.27, RPD=1.81) and for SVR were (R^2 = 0.78, RMSE = 0.22, RPD=2.09). In the MIR dataset, the corresponding values for PLSR were (R^2 = 0.55, RMSE = 0.35, RPD=1.40); for RF were (R^2 = 0.52, RMSE = 0.20, RPD=1.43); for MARS were (R^2 = 0.44, RMSE = 0.37, RPD=1.29); and for SVR were (R^2 = 0.53, RMSE = 0.35, RPD=1.39) respectively. The threshold RPD values used to test model performance were the ones developed by Chang *et al.,*(2001), where excellent models have RPD $>$ 2, fair models have RPD between 1.4 and 2, and non-reliable models with RPD $<$ 1.4. Accordingly, PLSR was considered as an excellent model in the Vis-NIR range (RPD = 2.44) and RF as fairly good in the MIR range (RPD=1.43) (Table 2 and 3).

PLSR model has been successfully used in this study and has been used for estimating soil salinity and other properties of soil elsewhere in the world, e.g., New South Wales, Australia (Janik et al., 2009), the island of Texel in the northwest of The Netherlands (Farifteh et al., 2007a), the Yellow River delta region in China (Weng et al., 2008) and the Hetao Irrigation District of Inner Mongolia in China (Qu *et al.* , 2009). PLSR first decomposes the spectra into a set of eigenvectors and scores and performs a regression with soil attributes in a separate step, thus actually using the soil information during the decomposition process. The advantages of PLSR is its linearity and it takes advantage of the correlation that exists between the spectra and the soil properties; thus, the resulting spectral vectors are directly related to the soil attribute (Geladi and Kowalski, 1986). It is robust in terms of data noise and missing values, and balances the two objectives of explaining response and predictor variation and performs the decomposition and regression in a single step. Sidike *et al.,*(2014) showed that an accurate prediction of soil salinity can be made based on the PLSR method (R^2 = 0.992, RMSE = 0.195) and Farifteh et al., (2007) suggested that PLSR analyses offered accurate to good prediction of EC.

RF is a group of al algorithms that have been developed as an extension of Classification and Regression Tree analysis to enhance the prediction performance and have been mainly used for classification problems (Olson et al. 2017). The RF is a fast, simple data-driven statistical approach that has been used in digital soil mapping and has shown good accuracy and is reported to be resistant to over-fitting and usually performs well in problems with a low sample-to-feature ratio (Wei *et al.,*2012), but could not outperform PLSR in data calibration for both spectral ranges and validation in the Vis-NIR range in the present study. SVR, which is a machine learning algorithm based on the statistical learning theory which seeks to maximize the ability to generalize using the structural risk minimization principle (Filgueiras et al. 2014) and MARS, which splits the data into sub regions (splines) with different interval ending knots, which are the points in the slopes where the regression coefficients change, and fits the data in each sub region using a set of adaptive piece wise linear regressions (Friedman, 1991); both did not perform better than PLSR and RF in this study.

The scatter plots of measured and predicted values for soil electrical conductivity in the calibration Vis-NIR and MIR datasets (Fig 5 and Fig 7) showed good relation between these two variables with high R^2 values in both datasets. The scatter plots of measured and predicted EC in the validation NIR and MIR datasets (Fig 6 and Fig 8) also suggest good model validation with high R^2 values. On comparing the RPD values

of Vis-NIR and MIR validation datasets, higher RPD values were obtained in the Vis-NIR region and hence this region may be better suited for prediction of EC than MIR region. Soriano *et al.,* (2014) reported that Vis-NIR spectroscopy shows better result (R^2= 0.60) in prediction of EC than MIR (R^2= 0.27) as observed in our study. Kodaira *et al.,* (2013) reported that EC was generally poorly predicted by both MIR (R^2= 0.26) and NIR spectroscopy (R^2= 0.57) but, Minasny et al. (2009) predicted EC with good accuracy in the MIR region using large variation of values in the dataset used.

5. Conclusion

The study was conducted in the Indo-Gangetic plain region to evaluate the performance of reflectance spectroscopy in the Vis-Nir and MIR regions for estimation of electrical conductivity which is known to affect crop productivity and soil quality. The procedure adopted for testing EC was one that is commonly used in the soil testing laboratories of India and results suggested wide variation of electrical conductivity in the study region. Spectral regions in 460-500 and 1890-1906 nm in the Vis-NIR region and 4200-4310, 5275-5280, 6660-6670, 7305-7310 and 8290-8300 nm in the MIR region were identified as sensitive for estimation of EC. The PLSR model outperformed other models in calibration and validation in the Vis-NIR range and the PLSR model in calibration and RF model in validation was better in the MIR range (followed by PLSR). Among the Vis-NIR and MIR regions, on the basis of higher RPD values, the data fitted better in the Vis-NIR region and would be the region of choice for predicting EC in the study area. Thus reflectance spectroscopy in the Vis-NIR range with PLSR algorithm is very well suited to replace conventional method of estimating electrical conductivity in the intensively cultivated Indo-Gangetic plain regions of India.

Conflict of Interest Statement

The authors declare that there is no conflict of interest.

Data Availability Statement

Data available on request from the authors

References

Aldabaa, A. A. A., Weindorf, D. C., Chakraborty, S., Sharma, A. & Li, B. (2015). Combination of proximal and remote sensing methods for rapid soil salinity quantification. *Geoderma*, *239–240*, 34–46. http://dx.doi.org/10.1016/j.geoderma.2014.09.011.

Bilgili A.V., van Es. H. M., Akbas F, Durak, A., Hively, W. D. (2010). Visible near infrared reflectance spectroscopy for assessment of soil properties in a semi-arid area of Turkey. *Journal Arid Environment, 74* , 229–238.

https://doi.org/10.1016/j.jaridenv.2009.08.011

Bilgili, A. V., Cullu, M. A., Van Es, H. M., Aydemir, A. & Aydemir, S., (2011). The use of hyperspectral visible and near infrared reflectance spectroscopy for the characterization of salt-affected soils in the Harran plain, Turkey. *Arid Land Research and Management,25,* 19–37. https://doi.org/10.1080/15324982.2010.528153**.**

Chang, C.-W., Laird, D. A., Mausbach, M. J. & Hurburgh C. R. (2001). Near-infrared reflectance spectroscopy - Principal components regression analyses of soil properties. *Soil Science Society of America Journal, 65* , 480-490. **https://doi.org/10.2136/ sssaj2001.652480x**

Clay, D. E., Chang, J., Malo, D. D., Carlson, C. G., Reese, C., Clay, S. A., Ellsbury, M., Berg, B. (2001). Factors influencing spatial variability of soil apparent electrical conductivity. *Communications in Soil Science and Plant Analysis,32,* 2993–3008.

https://doi.org/10.1081/CSS-120001102

Clyde, M. (2017). BAS: Bayesian Adaptive Sampling for Bayesian Model Averaging. R Package Version 1.4.6. Available online: https://CRAN.R–project.org/web/packages/BAS.

Farifteh, J., van der Meer, F., Atzberger, C. & Carranza, E. J.M. (2007). Quantitative analysis of salt-affected soil reflectance spectra: a comparison of two adaptive methods PLSR and ANN. *Remote Sensing of Environment, 110*, 59–78. https://doi.org/10.1016/j.rse.2007.02.005

Felicísimo, Á. M., Cuartero, A., Remondo, J. & Quirós, E. (2012). Mapping landslide susceptibility with logistic regression, multiple adaptive regression splines, classification and regression trees, and maximum entropy methods: a comparative study. *Landslide, 10* , 175–189. https://doi.org/10.1007/s10346-012-0320-1

Filgueiras, P. R., Sad, C. M. S., Loureiro, A. R., Santos, M. F. P., Castro, E. V. R., Dias, J. C. M. & Poppi, R. J. (2014). Determination of API gravity, kinematic viscosity and water content in petroleum by ATR-FTIR spectroscopy and multivariate calibration. *Fuel,* 116, 123–130. http://dx.doi.org/10.1016/j.fuel.2013.07.122

Friedman, J. H.(1991). Multivariate adaptive regressions splines.*Annals of Statistics* , *19* , 1–67.

Gaikwad, B. (2020). Using hyperspectral remote sensing to monitor the properties of salt-affected soils (Doctoral dissertation, National Institute of Abiotic Stress Management, India).

Geladi, P. & Kowalski, B. R. (1986). Partial least-squares regression: a tutorial. *Analytica Chimica Acta* , *185* , 1-17.

https://doi.org/10.1016/0003-2670(86)80028-9

Ghosh A. K., A. P. Singh, S Singh. (2012). A handbook of soil fertility assessment. Kavari publication House, Varanasi, India.

Janik, L. J., Forrester, S. T. & Rawson, A. (2009). The prediction of soil chemical and physical properties from mid-infrared spectroscopy and combined partial least-squares regression and neural networks (PLS-NN) analysis. *Chemometrics and Intelligent Laboratory Systems,97,* 179–188. https://doi.org/10.1016/j.chemolab.2009.04.005 Jin, P., Li, P., Wang, Q. & Pu, Z. (2015). Developing and applying novel spectral feature parameters for classifying soil salt types in arid land. *Ecological Indicators*, *54*, 116-123. https://doi.org/10.1016/j.ecolind.2015.02.028

Karlen, D. L., Tomer, M. D., Neppel, J. & Cambardella, C. A. (2008). A preliminary watershed scale soil quality assessment in north central Iowa, USA. *Soil and Tillage Research, 992* , 291–299.

https://doi:10.1016/j.still.2008.03.002

Kodaira, M. & Shibusawa, S. (2013). Using a mobile real-time soil visible–near infrared sensor for high resolution soil property mapping.*Geoderma*, 199, 64–79. https://doi.org/10.1016/j.geoderma.2012.09.007

Li, H. Y., Shi, Z., Webster, R. & Triantafilis, J. (2013). Mapping the three dimensional variation of soil salinity in a rice-paddy soil.*Geoderma, 195–196* , 31–41.

https://doi.org/10.1016/j.geoderma.2012.11.005

Margate, D. E. & Shrestha, D. P. (2001). The use of hyperspectral data in identifying desert - like soil surface features in Tabernas area, Southeast Spain. In Proceedings of the 22nd Asian conference on remote sensing, 5-9 November 2001, Singapore CRISP, SISV, AARS. pp. 736-741.

Metternicht, G. & Zinck, J. A. (1997). Spatial discrimination of salt-and sodium-affected soil surfaces. *International Journal of Remote Sensing* , *18* , 2571-2586.

https://doi.org/10.1080/014311697217486

Minasny, B. & McBratney. A. B. (2008) Regression rules as a tool for predicting soil properties from infrared reflectance spectroscopy.*Chemometrics and Intelligent Laboratory Systems, 94,* 72–79. https://doi.org/10.1016/j.chemolab.2008.06.003 Minasny, B., Tranter, G., McBratney, A. B., Brough, D. M. & Murphy, B. W. (2009) Regional transferability of mid-infrared diffuse reflectance spectroscopic prediction for soil chemical properties. *Geoderma,153*, 155–162. https://doi.org/10.1016/j.geoderma.2009.07.021 Moros, J., Mart ınez-Sa nchez, M.J., Pe rez-Sirvent, C., Garrigues, S. & de la Guardia, M. (2009). Testing of

the region of Murcia soils by near infrared diffuse reflectance spectroscopy and chemometrics. *Talanta*, *78,* 388–398. https://doi.org/10.1016/j.talanta.2008.11.041

Mouazen, A. M., Kuang, B., De Baerdemaeker, J. & Ramon, H. (2010). Comparison among principal component, partial least squares and back propagation neural network analyses for accuracy of measurement of selected soil properties with visible and near infrared spectroscopy. *Geoderma* , *158* , 23-31. https://doi.org/10.1016/j.geoderma.2010.03.001

Nawar, S., Buddenbaum, H., Hill, J. & Kozak, J. (2014). Modeling and mapping of soil salinity with reflectance spectroscopy and landsat data using two quantitative methods (PLSR and MARS). *Remote Sensing* , *6,* 10813-10834.

https://doi.org/10.3390/rs61110813

Nawar, S., Buddenbaum, H., Hill, J., Kozak, J. & Mouazen, A. M. (2016). Estimating the soil clay content and organic matter by means of different calibration methods of vis-NIR diffuse reflectance spectroscopy. *Soil and Tillage Research* , *155,* 510-522.

https://doi.org/10.1016/j.still.2015.07.021

Nocita, M., Stevens, A., Toth, G., Panagos, P., van Wesemael, B. & Montanarella, L., (2014). Prediction of soil organic carbon content by diffuse reflectance spectroscopy using a local partial least square regression approach. *Soil Biology and Biochemistry, 68* , 337–347.

https://doi.org/10.1016/j.soilbio.2013.10.022

Nocita, M., Stevens, A., van Wesemael, B., Brown, D. J., Shepherd, K. D., Towett, E., Montanarella, L. (2015). Soil spectroscopy: an opportunity to be seized. *Global Change Biology* , *21* , 10-11. https://doi.org/10.1111/gcb.12632

Olson, R. S., La Cava, W., Mustahsan, Z., Varik, A., Moore, J. H., (2017). Data-driven advice for applying machine learning to bioinformatics problems. arXiv:1708.05070 **[q-bio.QM]**

Qu, Y. H., Duan, X., Gao, H. Y., Chen, A. P., An, Y.Q., Song, J. L., Zhou, H. M. & He, T. (2009). Quantitative retrieval of soil salinity using hyperspectral data in the region of Inner Mongolia Hetao irrigation district. *Spectroscopy and Spectral Analysis,29* , 1362–1366. https://doi.org/10.3964/j.issn.1000-0593(2009)05-1362-05

Rhoades, J. D., Chanduvi, F. & Lesch, S. M., (1999). Soil Salinity Assessment: Methods and Interpretation of Electrical Conductivity Measurements. Food and Agricultural Organiza- tion, Rome, Italy.

Shepherd, K. D. & Walsh, M. G (2002). Development of reflectance spectral libraries for characterization of soil properties. *Soil Science Society of America Journal, 66,* 988–998.

https://doi.org/10.2136/sssaj2002.9880

Sidike, A., Zhao, S. & Wen, Y. (2014). Estimating soil salinity in Pingluo County of China using QuickBird data and soil reflectance spectra. *International Journal of Applied Earth Observation and Geoinformation* , *26* , 156-175

https://doi.org/10.1016/j.jag.2013.06.002

Sonmez, S., Buyuktas, D., Okturen, F. & Citak, S. (2008). Assessment of different soil to water ratios 1:1, 1:2.5, 1:5 in soil salinity studies. *Geoderma* , *144* , 361–369.

https://doi.org/10.1016/j.geoderma.2007.12.005

Soriano, P., Moruno, F., Boscaiu, M., Vicente, O., Hurtado, A., Llinares, J. V., Estrelles, E. (2014). Is salinity the main ecologic factor that shapes the distribution of two endemic Mediterranean plant species of the genus Gypsophila? *Plant and Soil, 384* , 363-379.

https://doi.org/10.1007/s11104-014-2218-2

Stockle, C. O. (2013). Environmental impact of irrigation: a review. http://www.swwrc.wsu.edu/newsletter/fall2001/IrrImpact2.pdf

The R Development Core Team. (2017). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna. Available online: https://www.R–project.org/

Tsai, F., Philpot, W. (1998). Derivative analysis of hyperspectral data. *Remote Sensing of Environment* , *66* , 41-51.

https://doi.org/10.1016/S0034-4257(98)00032-7

Vasu, D., Singh, S. K., Sahu, N., Tiwary, P., Chandran, P., Duraisami, V. P. & Kalaiselvi, B. (2017). Assessment of spatial variability of soil properties using geospatial techniques for farm level nutrient management. *Soil and Tillage Research* , *169* , 25-34.

https://doi.org/10.1016/j.still.2017.01.006

Vibhute, A. D., Kale, K. V., Mehrotra, S. C., Dhumal, R. K. & Nagne, A. D. (2018). Determination of soil physicochemical attributes in farming sites through visible, near-infrared diffuse reflectance spectroscopy and PLSR modeling. *Ecological Processes* , *7* , 26.

https://doi.org/10.1186/s13717-018-0138-4

Viscarra Rossel, R. A., Walvoort, D. J. J., Mcbratney, A. B., Janik, L. J. & Skjemstad, J. O. (2006). Visible, near infrared, mid infrared or combined diffuse reflectance spectroscopy for simultaneous assessment of various soil properties. *Geoderma*, *131*, 59–75. https://doi.org/10.1016/j.geoderma.2005.03.007

Wei, G., Zhao, X. (2012). Some dependent aggregation operators with 2-tuple linguistic information and their application to multiple attribute group decision making. *Expert Systems with Applications* , *39* 5, 5881-5886.

https://doi.org/10.1016/j.eswa.2011.11.120

Weng, Y. L., Gong, P. & Zhu, Z. L., (2008). Soil salt content estimation in the Yellow River Delta with satellite hyperspectral data.*Canadian Journal of Remote Sensing, 343* , 259–270.

https://doi.org/10.5589/m08-017

Yong-Ling, W. E. N. G., Peng, G. & Zhi-Liang, Z. (2010). A spectral index for estimating soil salinity in the Yellow River Delta Region of China using EO-1 Hyperion data. *Pedosphere* , *20* 3 , 378-388.

https://doi.org/10.1016/S1002-0160(10)60027-6

Zhang, T. T., Zeng, S. L., Gao, Y., Ouyang, Z. T., Li, B., Fang, C. M. & Zhao, B. (2011). Using hyperspectral vegetation indices as a proxy to monitor soil salinity. *Ecological Indicators* , *116* , 1552-1562. https://doi.org/10.1016/j.ecolind.2011.03.025

List of Figures

Figure 1. Study area showing part of Uttar Pradesh state of India, covering the districts of Varanasi, Chanduali, Sant Ravidas Nagar and Mirzapur, localized between 82°30' and 83deg30' East and 24deg30' and 25deg30' North.

Figure 2. Quantile- Quantile plot for (a) electrical conductivity (untransformed) and (b) $\log_{10}$ transformed electrical conductivity

Figure 3. (a) Reflectance Spectra in the visual- near infrared (Vis-NIR) range (350-2500 nm) of soils.

Figure 3. (b) Reflectance Spectra in the middle infra-red (MIR) range of soils.

Figure 4 (a). Correlation between electrical conductivity and reflectance in the visual- near infrared (Vis-NIR) region

Figure 4 (b). Correlation between EC and reflectance in the middle infra-red (MIR) region

Figure 5 . Calibration model developed for EC in the NIR region using (a) Partial Least Square Regression (PLSR) (b) Random Forest Regression (RF) (c) Support Vector Regression (SVR) and (d) Multivariate Adaptive Regression Splines (MARS)

Figure 6 . Validation model developed for EC in the NIR region using (a) Partial Least Square Regression (PLSR) (b) Random Forest Regression (RF) (c) Support Vector Regression (SVR) and (d) Multivariate Adaptive Regression Splines (MARS)

Figure 7 . Calibration model developed for EC in the MIR region using (a) Partial Least Square Regression (PLSR) (b) Random Forest Regression (RF) (c) Support Vector Regression (SVR) and (d) Multivariate Adaptive Regression Splines (MARS)

Figure 8 . Validation model developed for EC in the MIR region using (a) Partial Least Square Regression (PLSR) (b) Random Forest Regression (RF) (c) Support Vector Regression (SVR) and (d) Multivariate Adaptive Regression Splines (MARS)

Hosted file

`Figures.docx` available at https://authorea.com/users/313230/articles/474625-evaluation-of-multivariate-regression-models-to-predict-electrical-conductivity-using-vis-nir-and-mir-spectra

Hosted file

`Tables EC -NIR (1) final.docx` available at https://authorea.com/users/313230/articles/474625-evaluation-of-multivariate-regression-models-to-predict-electrical-conductivity-using-vis-nir-and-mir-spectra