Rapid Prediction of Soil Quality Indices using Near Infrared Spectroscopy (NIRS)

Project Report
By

Amitakshar Biswas (bmat1904)
Arghyadip Chakraborty (bmat1907)
Kuntal Das (bmat1916)
Soumyajyoti Kundu (bmat1938)

BMath 2nd Year
Indian Statistical Institute, Bangalore Centre

Table of Contents

1. Project Abstract

To determine soil macronutrients and other quality indices, conventional and laborious procedures were employed. However this method is time consuming, involves chemical materials and is laborious. Thus an alternative, fast and environment friendly method is required to determine the macronutrients and other quality indices in agricultural soil. The aim of this study is to apply Near Infrared Spectroscopy (NIRS) to determine soil macronutrients namely nitrogen (N), phosphorus (P) and potassium (K). Diffuse reflectance spectrum of soil samples were acquired and recorded in the wavelength range of 1000 to 2500 nm. Near infrared spectrum were enhanced using de-trending (DT) method. Prediction models, used to predict N, P and K, were established using Principal Component Regression (PCR) algorithm followed by k-fold Cross Validation. The result showed that NIRS method can determine all three quality indices with good accuracy and robustness. Maximum correlation coefficient (r) for N and K prediction models were achieved using DT correction method with $r = 0.85$ for N prediction, $r = 0.76$ for P prediction and $r = 0.90$ for K prediction. Based on the obtained results, we may conclude that NIRS can be applied as an alternative rapid and simultaneous method in predicting soil quality indices.

2. Introduction

In precision farming practices, it is important to monitor soil quality condition and health. As a matter of fact, plants can grow ideally on healthy soil which has the physical and chemical properties that are suitable for plant growth. Soil chemical properties are usually related to macronutrients needed by plants, with the amount needed being different for each growing phase. Thus it is very crucial to provide soil macronutrients since they are commonly required for plants to grow and develop. Macronutrients normally consist of nitrogen, phosphorous and potassium (NPK) and can be added by fertilization practices. Yet, this fertilization must be performed optimally in order to avoid unwanted impact. Excessive use of fertilizers can cause pollution to the environment because it can create artificial nutrient deposits that are not utilized by plants. Therefore, the determination and quantification of soil macronutrients (NPK) is necessary to monitor and take preventative actions. A reliable and environment friendly method is therefore needed to rapidly predict the amount of soil macronutrients in agricultural soils and diagnose suspected areas as well as control the rehabilitation process.

During the last few decades, near infrared technology has been widely used and had become most promising methods of analysis in many fields including in soil and agriculture science due to its advantages; simple sample preparation, rapid process and environment friendly procedure since no chemicals are involved and used. More importantly, it has the potential ability to determine multiple parameters simultaneously. Quantification of soil parameters by near infrared spectroscopy (NIRS) has become an interesting and attractive topic for research in soil science targeting several issues associated with agriculture and the environment. As reported by several authors, NIRS has been proved and employed to determine organic matter in soil, phosphorous species in soil, ammonia concentration and hazardous contamination in agricultural soil.

Numerous studies and publications on the application of near infrared spectroscopy (NIRS) shows that NIRS was feasible to be applied as a rapid and non-destructive tool for quality attributes precision in agricultural sectors. The NIRS can be used to predict several quality parameters on intact mangoes, oranges and apples, meat and dairy products, animal feed, coffee qualities, cocoa and wheat products. Prediction model performance was sufficiently robust and accurate with correlation coefficient (r) range 0.83 – 0.99 and residual

predictive deviation (RPD) index was 1.54 – 5.18 which is categorized as sufficient to excellent prediction model performances.

Based on advantages and excellence of NIRS performance, we performed a study to apply the NIRS in predicting soil quality indices (N, P and K). These macro nutrients are crucial to be monitored in precision farming practices to ensure plant growth optimally. In this study, we attempted to apply spectra enhancement method namely de-trending (DT) and compare reduction result obtained from raw original spectra data. The prediction models were developed based on near infrared spectroscopic data and actual reference data using principal component regression (PCR).

3. Materials and Methods

3.1. Soil Samples

Soil samples (approximately 100g) were collected as top soil samples (0 – 20 cm depth) from 10 different site locations in Aceh Besar district, Aceh Province of Indonesia. In each site, 2 soil samples were taken from 2 rice-paddy field and 2 samples from 2 nearby cropland. Thus, a total of 40 soil samples were collected and stored for a day to equilibrate, then air-dried for one week and sieved through 2 mm nylon sieve in order to remove stones, insects, large debris, pebbles and other unwanted materials.

3.2. Instrument Setup

Near infrared spectra data of soil sample were acquired and measured using a benchtop Fourier transform near infrared (NIR) spectroscopy (Thermo Nicolet Antaris II) with an integrating sphere accessory. The instrument was controlled using integrated software: Thermo Integration and Thermo Operation. The light source of halogen lamp irradiated soil samples from down to up through a quartz window (1 cm of diameter), which was embedded in the top of the NIR instrument. Soil sample was packed in a sample cylindrical quartz cup (10 mm in depth) to ensure full light penetration. The sample cylindrical cup was filled with 20 g of soil samples and levelled using a smooth edge. It was then fixed on the quartz window by a swivel bracket.

3.3. Spectra Data Acquisition

Near infrared (NIR) spectra data were acquired and recorded as absorbance spectra data in the presence of energies in wavenumbers 4000 - 10000 cm^{-1} or in wavelength range from 1000 to 2500 nm. During spectra data acquisition, sample cup was spinning around slowly in order to obtain the averaged spectrum of each soil sample. Background spectra correction was taken once every 10 sample acquisitions. The spectral resolution was 8 cm^{-1} and optical gain was set to 4x. The final spectra data was taken as an average of 64 successive data acquisition. The final dataset consisted of 1557 wavelength variables for each of the 40 soil samples. Each wavelength variable contains the value of the diffuse reflectance spectra of that soil sample in the presence of energy of that specific wavelength. This spectra dataset was used for further analysis in prediction model development.

Table 1: Spectra Dataset (40 rows x 1557 columns)

	999.897	1000.669	1001.442	1002.216	...	2490.618	2495.412	2500.225
1	0	0	-9.577275e-05	-1.967847e-04	...	0.000950	0	0
2	0	0	-1.102686e-04	-1.443744e-04	...	0.000975	0	0
3	0	0	-2.209246e-04	-1.493692e-04	...	0.001180	0	0
4	0	0	-1.443744e-04	-1.459956e-04	...	0.000929	0	0
.	.	.	.	.	...	.	.	.
.	.	.	.	.	...	.	.	.
.	.	.	.	.	...	.	.	.
38	0	0	-4.302859e-05	-7.316470e-05	...	0.000957	0	0
39	0	0	-2.064228e-04	-1.362562e-04	...	0.000839	0	0
40	0	0	-1.058698e-04	2.174377e-05	...	0.000818	0	0

3.4. Soil Properties Data Measurement

After spectra data acquisition is complete, soil properties (N, P and K) of the soil samples were measured and determined using standard chemical laboratory methods. Soil nitrogen (N) content was determined using Kjeldahl method and expressed in percentage of their weight to the total weight of dry soil sample. The soil phosphorous (P) content was determined by means of $HClO_4 - H_2SO_4$ heating extraction and a combination of molybdenum-blue colorimetric method and expressed in ppm. The soil potassium (K), expressed in cmol/kg content, was determined by calcining and extracting with NaOH and then measuring it using an atomic absorption flame photometer. All chemical analyses for soil properties were carried out in duplicate and averaged.

Table 2: Descriptive Statistics of the actual measured soil fertility properties

Descriptive Statistics	N	P	K
Mean	0.15	14.49	0.88
Max	0.52	40.92	2.58
Min	0.02	1.68	0.26
Range	0.50	39.24	2.32
Std. Deviation	0.14	11.46	0.51
Variance	0.02	131.26	0.26
RMS	0.21	18.38	1.02
Skewness	1.25	0.97	1.44
Kurtosis	0.47	-0.16	2.14
Media	0.09	9.77	0.69
Q1	0.04	5.86	0.55
Q3	0.24	21.92	0.16

N: nitrogen, P: phosphorous, K: potassium, Q1: first quartile, Q3: third quartile

3.5. De-trending

De-trending involves removing the effects of trend from a data set to show only the differences in values from the trend. Removing a trend from the data set allows us to focus on the fluctuations and identify the important factors. For our dataset we used de-trending to remove background noise. Spectral data acquired from the near infrared instrument generally contains background information and noises which interfere and affect desired relevant soil quality information such as macronutrient contents (N, P and K). Interfering spectral parameters, such as light scattering, path length variations and random noise resulted from variable physical sample properties or instrumental effects need to be removed or minimized in order to obtain accurate, robust and stable calibration models. These noises were corrected using de-trending enhancement method. The de-trending pre-treatment method not only removes nonlinear trends in spectroscopic data but also reduces amplification due to light scattering and offset due to additive chemical effects. Thus spectra correction and enhancement helps in building more accurate and robust regression models. The data after de-trending looks something like:

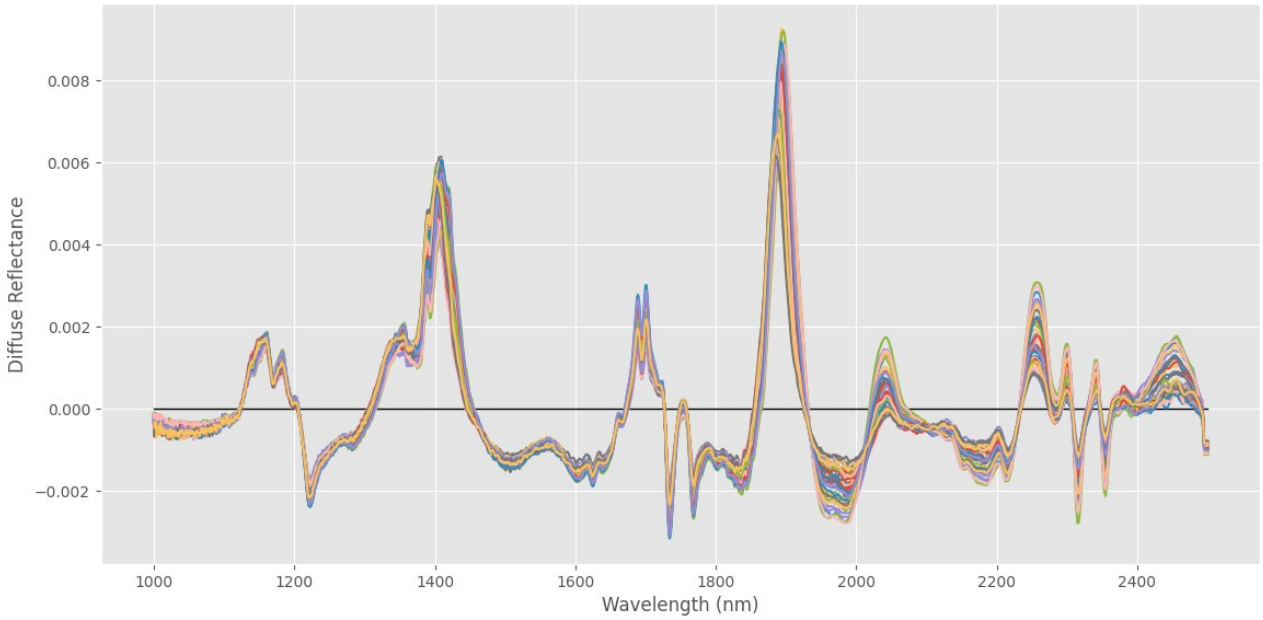

Figure 1: Typical diffuse reflectance spectra data of soil samples after DT enhancement

3.6. Cross-Validation

Model validation is the task of confirming that the outputs of a statistical model are acceptable with respect to the real data-generating process. Cross-validation is a model validation technique for assessing how the results of a statistical analysis will generalize to an independent data set. In cross-validation the data set is divided into two parts: the training or calibration set which is used to regress the prediction model and the testing or validation set against which the model is tested. The goal of cross-validation is to test the model's ability to predict new data that was not used in estimating it, in order to flag problems like overfitting (occurs when the model is closely fit to a limited set of data points) or selection bias (occurs when proper randomization is not achieved in selection of groups) and to give an insight on how the model will generalize to an independent dataset. The advantage of using cross-validation is that it does not wastes any data like other validation techniques which divides the data into 3 sets.

For our purpose we have used a type of cross-validation called k-fold cross validation. In k-fold cross-validation, the original sample is randomly partitioned into k equal sized subsamples. Of the k subsamples, a single subsample is retained as the validation data for testing the model, and the remaining $k - 1$ subsamples are used as training data. The cross-validation process is then repeated k times, with each of the k subsamples used exactly once as the validation data. The k results can then be averaged to produce a single estimation. The advantage of this method is that all observations are used for both training and validation, and each observation is used for validation exactly once. We have used the very commonly used 10-fold cross-validation.

3.7. Principal Component Analysis

Principal Component Analysis (PCA) is the process of computing the principal components and using them to perform a change of basis on the data, sometimes using only the first few principal components and ignoring the rest. It is commonly used for dimensionality reduction by projecting each data point onto only the first few principal components to obtain lower-dimensional data while preserving as much of the data's variation as possible. PCA can be thought of as fitting a p-dimensional ellipsoid to the data, where each axis of the ellipsoid represents a principal component. If some axis of the ellipsoid is small, then the variance along that axis is also small.

To find the axes of the ellipsoid, we must first subtract the mean of each variable from the dataset to center the data around the origin. Then, we compute the covariance matrix of the data and calculate the eigenvalues and corresponding eigenvectors of this covariance matrix. Then we must normalize each of the orthogonal eigenvectors to turn them into unit vectors. Once this is done, each of the mutually orthogonal, unit eigenvectors can be interpreted as an axis of the ellipsoid fitted to the data. This choice of basis will transform our covariance matrix into a diagonalised form with the diagonal elements representing the variance of each axis. The proportion of the variance that each eigenvector represents can be calculated by dividing the eigenvalue corresponding to that eigenvector by the sum of all eigenvalues.

3.8 Principal Component Regression

Principal Component Regression (PCR) is a regression analysis technique that is based on principal component analysis (PCA). PCR is used for estimating the unknown regression coefficients in a standard linear regression model but instead of regressing the dependent variable on the explanatory variables directly, the principal components of the explanatory variables are used as regressors. Typically only a subset of all the principal components is used for regression, making PCR a kind of regularized procedure and also a type of shrinkage estimator. Often the principal components with higher variances (the ones based on eigenvectors corresponding to the higher eigenvalues of the sample variance-covariance matrix of the explanatory variables) are selected as regressors. However, for the purpose of predicting the outcome, the principal components with low variances may also be important, in some cases even more important.

The main use of PCR is to overcome the multicollinearity problem which arises when two or more of the explanatory variables are close to being collinear. PCR can aptly deal with such situations by excluding some of the low-variance principal components in the regression step. In addition, by usually regressing on only a subset of all the principal components, PCR can result in dimension reduction through substantially lowering the effective number of parameters characterizing the underlying model.

For our purpose we built PCR models using the near infrared spectra data as the independent (X) variable and the actual measured soil fertility concerned (N, P or K) as the

dependent variable. Using PCA the wavelength variables are projected to a fewer number of principal components and subsets of the principal components are used to regress the prediction model using PCR. The model with lowest RMSE value is considered as it provides the best accuracy and robustness. Again PCR models were regressed using the DT enhanced data so that the models can be compared for accuracy and robustness.

3.9. Prediction Model Performance

Prediction performance were quantified and judged for their accuracies and robustness using several statistical indicators:

- Coefficient of determination(R^2): It is the proportion of the variance in the dependent variable that is predictable from the independent variable. It normally range from 0 to 1 and a higher value of R^2 indicates better accuracy of the prediction model.

$$R^2 = \frac{\sum(\hat{y}_i - \bar{y})^2}{\sum(y_i - \bar{y})^2}$$

- Coefficient of correlation(r): It is a numerical measure of correlation i.e. a statistical relationship between two variables. It ranges from -1 to $+1$, where ± 1 indicates the strongest possible agreement and 0 indicates the strongest possible disagreement. A higher value of $|r|$ indicates better accuracy of the prediction model.

$$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}}$$

- Root Mean Square Error($RMSE$): It is the quadratic mean of the differences between predicted values and observed values. It is always non-negative, and a value of 0 would indicate a perfect fit to the data. A lower value indicates better accuracy of the prediction model.

$$RMSE = \sqrt{\frac{1}{n}\sum(\hat{y}_i - y_i)^2}$$

- Residual Prediction Deviation(RPD): It is the standard deviation of observed values divided by the Root Mean Square Error($RMSE$). It takes both the prediction error and the variation of observed values into account, providing a metric of model validity that is more objective than the $RMSE$ and more easily comparable across model validation studies. The greater the RPD, the better the model's predictive capacity.

$$RPD = \frac{\sigma_y}{RMSE}$$

3.10. Using R and python to automate

Calculating results, handling the large data and regressing models is practically impossible manually. So we used R and python to do the calculations, regress models and draw the plots. Following R packages were used:

- ✓ readxl – to read data from the .xlsx files
- ✓ prospectr – to use the detrend function to de-trend our data
- ✓ pls – for pcr and statistical indicator functions
- ✓ caret – to use the train function to regress the models. Also it allows to directly use cross-validation for the models

Following python packages were used:

- ✓ pandas – to use the pandas data frame for handling our data
- ✓ matplotlib – to plot the spectra data graph
- ✓ scipy – to use the detrend function to de-trend our data

*The data used in this project, the R and python scripts used and the plots are available at: https://github.com/arghyadipchak/nirs

4. Soil Macronutrients Prediction

As mentioned earlier, macronutrient contents (N, P and K) of soil samples were predicted using PCR by regressing the near infrared spectra data (X-variables) and actual reference N, P, K data obtained using laboratory methods (Y-variables). The prediction results are presented in the following table:

Table 3: Prediction performances for macronutrient contents of soil samples using PCR

Macro nutrients	Spectra Correction	Statistical Indicators			
		R^2	r	$RMSE$	RPD
N	Raw	0.69	0.83	0.12	1.20
	DT	0.72	0.85	0.11	1.26
P	Raw	0.59	0.77	10.35	1.11
	DT	0.58	0.76	11.41	1.00
K	Raw	0.79	0.89	0.36	1.43
	DT	0.81	0.90	0.36	1.44

N: nitrogen content, P: phosphorous content, K: potassium content, DT: de-trending, R^2: coefficient of determination, r: coefficient of correlation, RMSE: root mean square error, RPD: residual predictive deviation index

At first we attempted to predict N, P and K using raw original spectra data. As shown in Table 1, the maximum correlation achieved was 0.89 for K predictions. Moreover, using the raw original spectral data, the maximum RPD index was 1.43 for K predictions which categorized as good prediction performance, while for P and N reduction, the RPD index were 1.11 and 1.20, which were categorized as sufficient and coarse prediction performance respectively.

Prediction model performances were improved for both N and K when the models were developed using de-trending spectra data. The correlation coefficients were significantly increased for N and K while it decreased slightly for P. The maximum correlation coefficient was 0.90 for K predictions, while for N and P prediction, the r coefficient were 0.85 and 0.76 respectively. Moreover, the RPD index was also improved for N and K predictions when the models were established using DT spectrum but it decreased for P predictions. The highest RPD index was achieved for K prediction (1.44) which categorized as excellent prediction performance. Scatter plot derived between actual reference soil macro nutrients and predicted ones are presented as follows:

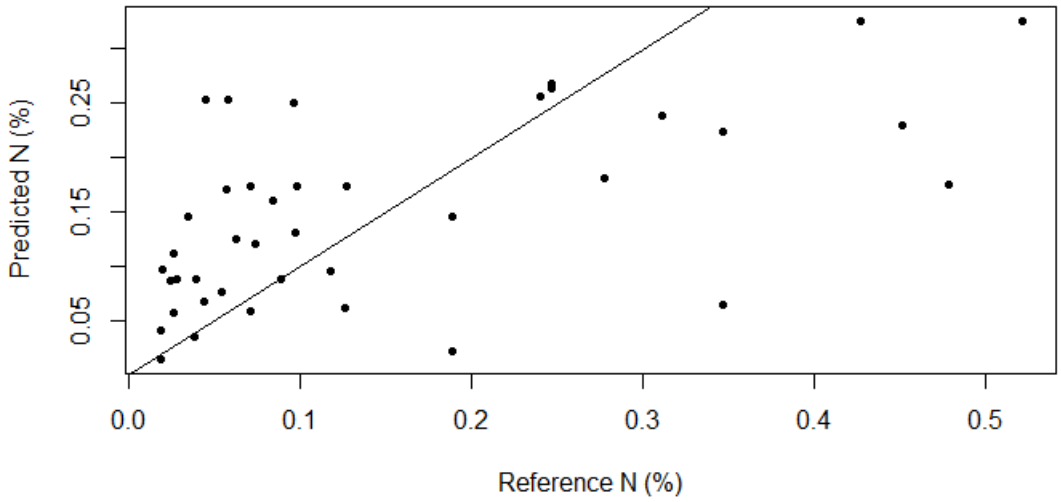

Figure 2: Scatter plot between reference and predicted N content of soil samples using DT spectra

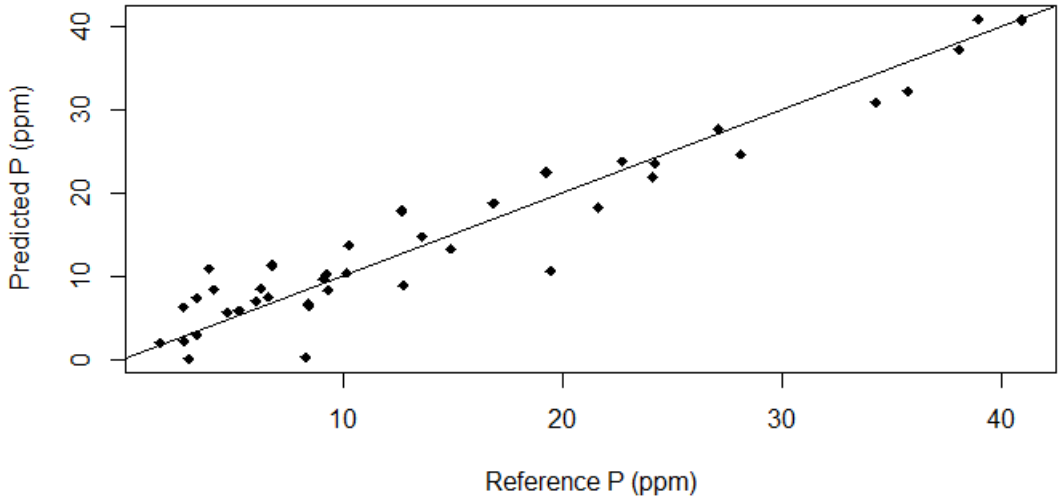

Figure 3: Scatter plot between reference and predicted P content of soil samples using DT spectra

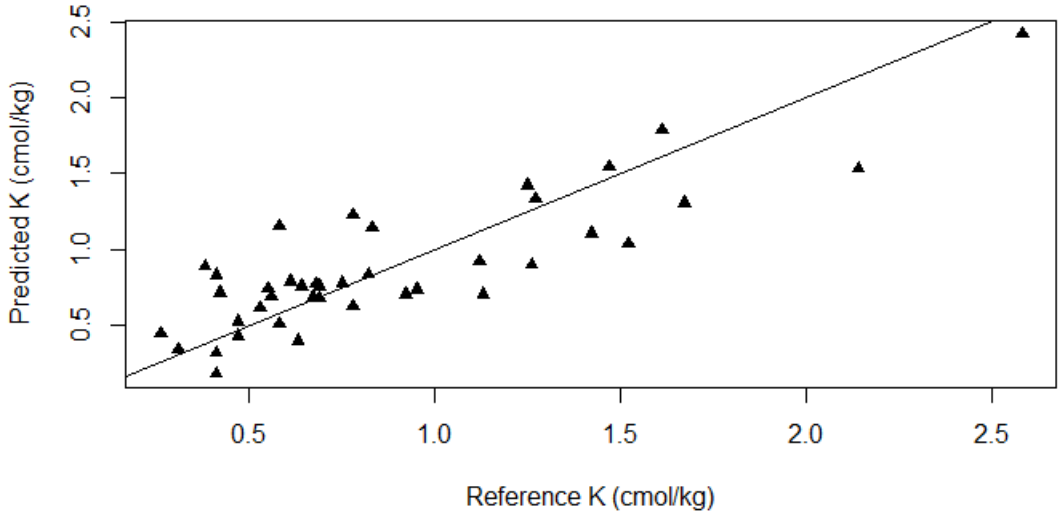

Figure 4: Scatter plot between reference and predicted K content of soil samples using DT spectra

As shown in Figure 2 to 4, it is clear that NIRS can be used to predict N, P and K contents of soil samples. This method can be used to monitor soil condition and take further actions required to maintain soil quality. Spectra enhancement has a significant impact to prediction performances. It obviously improved accuracy and robustness for some macronutrients parameters of soil samples.

5. Conclusion

Based on achieved prediction results, we may conclude that near infrared spectroscopy (NIRS) can be applied in precision farming practices and can be employed as a rapid and environment friendly method to predict soil quality indices. Achieved present study shows that NIRS technology was feasible to be used to monitor and rapidly determine the N, P and K contents of soil samples with good prediction accuracy and robustness.

6. Alternate Models and Macronutrients

It is also possible to use Partial Least Square Regression (PLSR) models to predict macronutrient contents of soil samples. In fact in some cases it might provide better results with more R^2 and RPD index. And not only N, P or K we can also predict other soil macronutrients like calcium (Ca) or magnesium (mg) and even the pH value of the soil using this method. One can simply measure these soil qualities using laboratory methods and regress PCR models with the spectra data and these actual measured soil fertility properties.

7. Acknowledgement

We are sincerely thankful to our Professor Rituparna Sen for giving us the opportunity to do this project and also for helping us out. We are thankful to our friends as well for helping us and our group mates who worked hard to make this project.

8. Resources

- Original Article: https://doi.org/10.1088/1755-1315/365/1/012043
- Supplementary Article: https://doi.org/10.1016/j.dib.2020.105469
- PCR in R:
  - https://rpubs.com/esobolewska/pcr-step-by-step
  - https://27411.compute.dtu.dk/filemanager/27411/uploads/eNotepdfs/eNote4-PCRinR.pdf
- PCR in Python: https://nirpyresearch.com/principal-component-regression-python/
- PCA, PCR, De-trending, etc:
  - https://www.wikipedia.org
  - https://learnche.org/pid/latent-variable-modelling/principal-components-regression
  - https://iq.opengenus.org/principal-component-regression/
- Cross-validation: https://www.analyticsvidhya.com/blog/2018/05/improve-model-performance-cross-validation-in-python-r/
- Documentation of various R packages:
  - https://www.rdocumentation.org/packages/caret/versions/6.0-86
  - https://www.rdocumentation.org/packages/readxl/versions/1.3.1
  - https://www.rdocumentation.org/packages/pls/versions/2.7-3
  - https://antoinestevens.github.io/prospectr