Estimation of Soil Organic Carbon for Sustainable Agriculture using Deep Learning

Singh S..[1], Kasana Singara S.[1]

[1]Thapar Institute of Engineering and Technology, Patiala, Punjab, INDIA

e-mail: simran.jita@gmail.com

Abstract

The organic carbon percentage is concomitant indicating the mineralization of nutrients and the ability of the soil to hold nutrients cations, structural stability, and water holding capacity. It is necessary to know the quantity of carbon for healthy soil and avoid the production related problems which can affect the sustainable agriculture model. In existing approaches, to quantitively calculate soil carbon, sample collection and in-situ laboratory testing are performed. In this work, a novel framework is proposed which is based on Partial Least Square Regression and Long Short-Term Memory networks to quantify soil organic carbon from the LUCAS dataset. Samples of LUCAS dataset are used as input to this framework. The samples are pre-processed by PLS to reduce their dimensions. These pre-processed samples are then passed to the LSTM, a Deep learning framework to build an efficient prediction model. The proposed framework performed more accurately, and its effectiveness is shown by comparing it with existing regression models.

Keywords: Deep Learning, Long Short-Term Networks, Silica, Organic Carbon, Hyperspectral Data

1. Introduction

Carbon is considered to be the most crucial content of the soil as its presence determines, among others, the water holding capacity and microbial activity of the soil. It acts as a pool for nitrogen, phosphorous and other nutrients to enhance plant productivity. It also provides cohesive property to the soil which protects it against erosion. The greenhouse effect is reduced when the soil captures excess carbon from the atmosphere. So, it is imperative to know about the quantity of carbon in the soil so that crucial decision for sustainable agriculture can be taken. Many studies have been conducted to quantify the carbon present in the soil.

Most of the existing works are based on regression techniques like Multiple Linear Regression (MLR), Multivariate Adaptive Regression Splines (MARS), Principal Component Regression (PCR) and Partial Least Square Regression (PLSR). PLSR is considered the most robust in all cases as it can handle missing and noisy data generated during the acquisition process [Geladi et al. 1988]. It can easily find the relations between spectra obtained from the dataset and soil properties. With the advances in the field of machine learning, Daniel [Daniel et al. 2003] used Artificial Neural Network to quantify soil parameters including carbon in Very Near-Infrared Region (VNIR).

In this paper, a framework is proposed using the Land Use/Cover Area frame Survey (LUCAS) [Toth *et al.* 2013] to quantify the soil organic carbon. In the Proposed Framework (PF), PLS is employed to reduce the dimensions of Hyperspectral Data (HSD) obtained from LUCAS as it is considered to be a robust technique. To take advantage of the sequential nature of HSD, a Deep learning framework named Long Short-Term Memory Network is used to quantify the soil organic carbon.

2. Background of the Proposed Work

2.1 Partial Least Squares

PLS is one of the most robust dimensionality reduction algorithms [Geladi *et al.* 1988]. Its main task is to maximize the variance in the resultant output. It is similar to Principal Component Analysis except it requires the responses of the input data to function more accurately.

PLS tries to maximize the co-variance and finds a linear decomposition of the predictors (P) and the responses (R). The linear decomposition of P and R is set such that:

$$P = T.X^T + G$$
$$R = U.Y^T + H$$

Where T=P-scores, U=R-scores, X=X-loadings, Y=Y-loadings, G=P-Residuals, H=R-Residuals.

Afterward, decomposition is performed to maximize the covariance between T and U. In decomposition, the first factor extracted can be defined as:

$$P_1 = P - t.t^T.P$$
$$R_1 = R - t.t^T.R$$

Where t=P.V and V is the eigenvector corresponding to the eigenvalue of $P^T RR^T P$. This process is repeated until the desired factors are obtained.

2.2 Long Short-Term Memory Networks

LSTM [Hochreiter *et al.* 1997] is a new form of recurrent neural network which performs better on sequential problems as shown in Figure 1. LSTMs improves from the traditional recurrent networks by solving the problem of vanishing gradients. The forget gate remembers the vital information which gives

LSTMs the ability to learn the sequences efficiently. The transitions functions are defined as follows:

Input gates: $l_t = \sigma(W_l.[O_{it-1}, x_{it} + b_i])$

Forget gates: $f_t = \sigma(W_f.[O_{it-1}, x_{it}] + b_f)$

New Candidates: $\hat{C}_t = tanh(W_C[O_{it-1}, x_{it}] + b_C)$

Cell States: $C_t = f_t \circ C_{t-1} + l_t \circ \hat{C}_t$

Output gates: $m_t = \sigma(W_{output}[O_{it-1}, x_{it}] + b_{output}])$

Next hidden state: $O_{it} = m_t \times tanh(C_t)$

Where σ is the sigmoid function, $tanh$ is the tangent hyperbolic function, W is the weight, $\circ$ is the Hadamard product, b is the bias, t is the present time state.

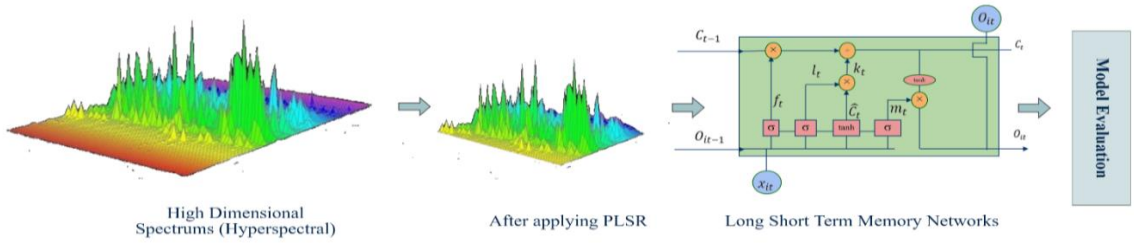

Figure 1. Proposed Framework

3 Experimental Results and Analysis

Experiments are performed on the LUCAS dataset which contains 19036 hyperspectral signatures with the original quantity of soil organic carbon. Initially, the higher dimensions of hyperspectral data are reduced using the robust PLS model. So, 75% hyperspectral signatures along with the quantity associated are given to PLS for supervised learning. Afterward, the learned dimensionality reduction is applied to the whole dataset to transform the spectrums to lower dimensional spectral signatures. Then, 65% of the transformed data is given as input to LSTMs to learn the sequence information. Validation is done on 10% data to avoid any overfitting of the trained model. In the end, a trained LSTM is obtained which is used to predict the soil organic carbon of the remaining 25% data. The predicted output is noted and compared with the original quantity to evaluate the model. It can be seen from Table 1 that the Proposed Framework (PF) is compared with popular prediction models like Partial Least Square Regression (PLSR), Principal Component Regression, Support Vector Regression (SVR) and Multiple Linear Regression (MLR) on the basis of Coefficient of determination (R^2), Root Mean Square Error (RMSE), Lin's Concordance Coefficient ($\boldsymbol{\rho_C}$) and Pearson Correlation Coefficient (r) which can be defined as follows:

$$\boldsymbol{R^2 = 1 - \frac{SS_E}{SS_T}}$$

Where $\boldsymbol{SS_E}$ is the error sum of squares and $\boldsymbol{SS_T}$ is the Total sum of squares.

$$\boldsymbol{RMSE = \sqrt{\frac{1}{n}\sum_{j=1}^{n}(x-y)^2}}$$

Where x is the actual value and y is the predicted value.

$$\boldsymbol{\rho_C = \frac{2\rho\sigma_x\sigma_y}{\sigma_x^2 + \sigma_y^2 + (\mu_x - \mu_y)^2}}$$

Where $\boldsymbol{\mu}$ is mean and $\boldsymbol{\sigma}$ is variance.

$$\boldsymbol{r = \frac{N\sum XY - (\sum X \sum Y)}{\sqrt{[N\sum x^2 - (\sum x)^2][N\sum y^2 - (\sum y)^2]}}}$$

Where N is the number of pair scores, $\boldsymbol{\sum XY}$ is the product of paired scores, $\boldsymbol{\sum X}$ is the sum of X scores, $\boldsymbol{\sum x^2}$ is the sum of the square of x scores. $\boldsymbol{\sum Y}$ is the sum of Y scores, $\boldsymbol{\sum y^2}$ is the sum of square of y scores.

Table 1. Performance comparison of Proposed Framework with other popular models.

	PF	PLSR	PCR	SVR	MLR
R^2	0.93	0.91	0.44	0.38	0.91
RMSE	26.65	30.42	77.87	82.03	30.45
ρ_C	0.97	0.96	0.63	0.48	0.96
r	0.97	0.96	0.67	0.9	0.96

The proposed framework performs better in comparison with other models with the best R^2 of 0.93. The excellent performance of the proposed framework against PCR also shows that PLS components extracted are more efficient than the principal components. The proposed framework also shows the best $\boldsymbol{\rho_C}$and r and the least RMSE.

4 Conclusion

In this work, a novel framework using PLS and LSTM is proposed to quantify the soil organic carbon. PLS is employed to reduce the hyperspectral dimensions efficiently. Then, the resultant dataset from PLS is processed by LSTM to predict the carbon. The predicted outputs are very promising with at most 244%, 202% and 168% improvement in R^2, $\boldsymbol{\rho_C}$ and r respectively and has shown at most 2 times less error in RMSE demonstrating that the use of deep learning can outperform all the existing regression and machine learning models.

References

Geladi, P. (1988). Notes on the history and nature of partial least squares (PLS) modelling. Journal of Chemometrics, 2(4), 231-246.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. Neural computation, 9(8), 1735-1780.

Tóth, G., Jones, A., & Montanarella, L. (2013). The LUCAS topsoil database and derived information on the regional variability of cropland topsoil properties in the European Union. Environmental monitoring and assessment, 185(9), 7409-7425.

Daniel, K. W., Tripathi, N. K., & Honda, K. (2003). Artificial neural network analysis of laboratory and in situ spectra for the estimation of macronutrients in soils of Lop Buri (Thailand). Soil Research, 41(1), 47-59.