Fusing environmental variables into soil spectroscopy modeling using a novel two-step regression method

S H Adi\textsuperscript{a}, S Grunwald\textsuperscript{b}, C Tafakresnanto\textsuperscript{a}

\textsuperscript{a} Indonesian Agency for Agricultural Research and Development
\textsuperscript{b} Soil and Water Sciences Department, University of Florida, USA

setyonohari@pertanian.go.id

Abstract. Soil spectroscopy modelling has been extensively studied as the cost-effective proximal sensing method for soil total and organic carbon predictions. Soil carbon properties were highly predictable due to the existence of active carbon molecular bonds within the Visible/Near-Infrared (VNIR) wavelength region. However, prediction results are highly variable for soil properties without active molecular bonds within the VNIR region, such as soil pH, sum of bases (SB), and cation exchange capacity (CEC). This research is intended to enhance soil organic carbon (SOC), nitrogen (N), pH, SB and CEC prediction accuracies by fusing categorical environmental variables (soil sample depth class, soil order, landform, and parent material) with continuous soil spectral data. We introduce a novel two-step regression method (2Step-R) to properly integrate the mixed type variables utilizing Partial Least Squares Regression (PLSR) and Ridge Regression in the modelling. Results from our analysis showed that the novel 2Step-R method was capable to improve the standard PLSR prediction model performances from fair (with ratio of performance to deviation or RPD between 1 and 1.4) to acceptable (RPD between 1.4 to 2), particularly for N, pH, and SB predictions. Slight model performance improvements were achieved for SOC and CEC predictions, although RPD values were within the acceptable range. In conclusion, the 2Step-R method is promising to enhance soil prediction performances and offers flexibility to include different types of ancillary model covariates suited to mix categorical soil-environmental and continuous spectral data.

Keywords: Soil spectroscopy; Regression; Proximal sensing; Soil property predictions

1. Introduction
Soil visible/near-infrared (VNIR) spectroscopy has been widely known as the low-cost alternative to wet-chemical soil laboratory analysis to estimate soil properties value [1]. This method utilizes multivariate statistical analysis to predict soil property content based on the corresponding soil spectra. Soil spectroscopy offers a rapid, non-destructive, and chemical waste-free soil analysis with reliable prediction accuracy [2]. Among soil properties, carbon contents (i.e., total and/or organic) are well estimated using spectroscopy method due to the existence of active carbon molecular bonds (e.g., C-O and C-H) within the VNIR wavelength region at 350 – 2500 nm [3]. Recent studies reported model validation coefficient of determination values (R2) for soil carbon predictions between 0.5 to 0.9, depending on the combination of the selected data preprocessing and statistical analysis [4,5]. However, prediction results for other soil properties without VNIR active spectral signatures (i.e., soil pH, sum of bases-SB, and cation exchange capacity-CEC) were highly variable [6].
Attempts to reduce spectroscopy modeling result variabilities include but not limited to implementation of enhance data preprocessing [7], data spiking [8], statistical modeling [9,10], and data fusion [11]. This article particularly dealing with statistical model enhancement in relation with fusing environmental data in soil spectroscopy. Partial Least Squares Regression (PLSR) is extensively utilized in spectroscopy modeling (chemometrics) because the interpretability of the analysis result [12]. By default, PLSR only accept numeric variables. Fusing categorical/discrete soil-environmental variables into PLSR will need additional step to convert the discreet categories into dummy values (0 and 1 values). However, this technique potentially violates model normality assumptions that could potentially reduce prediction accuracy. The purpose of this article, therefore, is to introduce a new modeling technique to properly fuse different type of variables in soil spectroscopy to enhance prediction result accuracy.

This study focuses on soil VNIR spectroscopy modeling of SOC, N, pH, SB, and CEC in East Java agricultural fields, Indonesia. Chemometric models were developed utilizing two statistical regression methods, including the PLSR [13] and ridge regression-RR [14,15] that were implemented using R [16]. This study implemented a two-step regression (2Step-R) method for proper inclusion of important ancillary categorical variables in soil VNIR spectroscopy modeling, including soil order, topographical landform, parent material, and soil sample depth. Therefore, this research objectives were to (1) develop and evaluate VNIR chemometric models using 2Step-R method for SOC, N, pH, SB, and CEC predictions in East Java agricultural fields, Indonesia, and (2) assess the model performances derived from the modified statistical methods, in comparison with the standard PLSR method.

2. Materials and Methods
This research considers East Java province, Indonesia, as the study area. The area is located between 110° 54’ and 114° 37’ east longitude and between 8° 48’ and 6° 44’ south latitude, with a total area of about 42,000 km2. The dominant geological formation in this study site is volcanic rocks (63%), with Vertisols (44%) and Inceptisols (32%) the major soil order [17,18]. The map of the study area with the distribution of soil samples is presented in Figure 1.

Figure 1. Map of study area and soil survey locations in East Java, Indonesia [19].
2.1. Materials

This research utilized the East Java soil survey data that were collected from agricultural fields in 2016 [19]. There was a total of 316 soil samples collected from 170 unique observation sites, with up to 6 sampling depths down to 1.2 meters (each sampling depth is about 20 cm). Sampling locations were determined to represent each unique polygon area derived from the intersection of soil forming factor spatial data, including climate classification, land use (agricultural field), topographic factor (slope and landform), and parent material. Five laboratory-measured soil properties were considered in this study, including soil organic carbon (SOC, %), nitrogen (N, %), pH, sum of bases (SB, cmolc kg⁻¹; i.e., the sum of Ca, Mg, K, and Na concentrations), and cation exchange capacity (CEC, cmolc kg⁻¹). SOC was measured using the Walkley and Black (1934) method, while N was extracted based using Kjeldahl (1883) method. Furthermore, pH was measured based on the diluted soil samples in deionized water at 1:5 ratio, while SB and CEC were determined by reacting each sample with 1 Normal ammonium acetate solution at pH 7. Soil spectra for the chemometric modeling were measured on air-dried (16 hours) and sieved (2 mm) soil samples. Measurements were performed using HR-1024i spectrometer (Spectra Vista Corporation, Poughkeepsie, NY). Ten automatic internal measurements taken for each soil samples at two different random probe positions were averaged to reduce spectral noises.

This research considered four environmental variables as ancillary predictors for soil spectroscopy modeling. These variables include soil taxonomy order [17], Digital Elevation Model (DEM)-generated landform based topographic position index [22], and parent material [23]. For SOC prediction, soil sample depth information was also considered as an ancillary variable, because soil organic carbon was known to vary with soil depth [24]. These variables were combined prior to the analysis to form a new classification based on the unique combination between soil order, landform, parent material, and sample depth class (only for SOC prediction).

2.2. Methods

Soil property data were log-transformed and normalized to the mean of 0 and the standard deviation of 1 prior to the analysis. Furthermore, soil spectra were noise-filtered using the Savitzky-Golay algorithm by utilizing the “signal” package in R [25,26]. The filtered spectra were averaged for each 10 nm spectral band and centered to the mean of 0. Moreover, all spatial data were projected into the Asia Lambert Conformal Conic projection prior to the data extractions. The landform classification was generated from the 90 meter resolution Shuttle Radar Topography Mission (SRTM) DEM utilizing “raster” and “RSAGA” packages in R [22,27,28]. Other spatial data consists of soil order, parent material, and soil depth class) were used as is without any modification. The ancillary spatial data, including soil order, landform, and parent material, were extracted into each soil observation geolocation utilizing the “raster” package in R [28].

This study introduces a newly developed two-step (sequential) regression process (2Step-R) that integrates PLSR [13] and RR [14,15]. The general representation of the 2Step-R technique is as follows:

\[
\text{Step 1: } \hat{y}_{PLSR} \sim f(X_{cont}) \\
\text{Step 2: } \hat{y}_{RR} \sim f(X_{cat}, \hat{y}_{PLSR})
\]

This technique was implemented using the combination of “pls” and “glmnet” package for the PLSR and RR implementation in R, respectively [29,30]. Systematic data splitting with 2 to 1 ratio of calibration and validation dataset was performed for each soil variable prior to the chemometric modeling. The splitting procedure was implemented using a systematic iterative procedure on the ordered (i.e., lowest to highest) soil property values. Furthermore, this study utilized ratio of performance to deviation (RPD) for validation model performance evaluations [31].
3. Results
The model validation performance metrics of the standard and modified PLSR methods for soil property predictions are presented in Table 1. In general, the median values of the performance metrics show better prediction performances of the modified PLSR compared with the standard PLSR models. The modified PLSR models were capable of improving the standard PLSR model prediction performances for all soil properties, most noticeably for SOC, N, pH, and SB predictions. Furthermore, validation plots comparing the predicted (modified PLSR) and laboratory-measured soil property values are presented in Figure 2. Each trend line in both figures shows the tendency of the modified PLSR models to overestimate low soil property values but underestimate the high values. Note that soil property values are presented in its original measurement units (non-transformed).

Table 1. Ratio of performance to validation values for soil property predictions in East Java

| Methods    | SOC (%) | N (%) | pH | SB (cmolc kg⁻¹) | CEC (cmolc kg⁻¹) | Mean | Min. | Median | Max. |
|------------|---------|-------|-----|-----------------|-----------------|------|------|--------|------|
| PLSR.Std   | 1.67    | 1.32  | 1.33| 1.24            | 1.80            | 1.47 | 1.24 | 1.33   | 1.80 |
| PLSR.Mod   | 1.85    | 1.55  | 1.45| 1.62            | 1.89            | 1.67 | 1.45 | 1.62   | 1.89 |

Abbreviations: SOC, soil organic carbon; N, Kjeldahl nitrogen; SB, sum of bases; CEC, cation exchange capacity; PLSR.Std, standard Partial Least Squares Regression; PLSR.Mod, modified PLSR.

Figure 2. Validation plots comparing the predicted (modified PLSR, y-axis) and laboratory-measured soil property values (x-axis).
4. Discussion

Recent VNIR spectroscopy studies to predict topsoil SOC using the PLSR model in Lombok (island), Indonesia, located further east of the study area, showed decent prediction accuracies with cross-validation RPD of about 2 [32,33]. Furthermore, a VNIR study to predict soil properties using PLSR in Brazil with the same tropical climate showed RPD for SOC, pH, SB, and CEC of about 1.84, 1.18, 1.00, and 1.17, respectively [24]. Therefore, these studies suggest that the chemometric models in this research produced comparable soil property prediction performances well within the results of the previous VNIR spectroscopy studies. A similar tendency of relatively better SOC and CEC prediction accuracies compared to N, pH, and SB predictions was also observed.

This research has demonstrated a novel method to fuse auxiliary environmental variables in soil spectral data through the two-step regression technique. The auxiliary predictor variables provided additional information that were not captured by the soil VNIR spectral data (e.g., soil sample depth, soil order, landform, and parent materials) to predict soil properties. The combined environmental and VNIR spectral data have shown to markedly improve standard chemometric model performances that solely use the VNIR spectra predictor dataset (i.e., standard PLSR), specifically for soil N, pH, and SB predictions. Therefore, this novel method is promising to enhance soil prediction performances. It offers flexibility to include different types of ancillary model covariates suited to mix categorical soil-environmental and continuous spectral data types.

Results from this study have also presented evidence that “acceptable” spectroscopy models were attainable for agricultural soil property predictions (i.e., SOC, N, pH, SB, and CEC) in East Java, Indonesia. This research has further demonstrated a promising application of soil spectroscopy as a cost-effective alternative method to conventional wet-chemistry laboratory analysis. Limited soil data availability is the main obstacle of agricultural land management in Indonesia. Therefore, soil spectroscopy is poised to enhance soil monitoring programs in Indonesia.

Acknowledgement

We would like to thank Dr. Hikmatullah for providing Soil Survey Report from his project in 2016 as the main data source for this study. We would also like to extend our gratitude to Indonesia Center for Agricultural Land Resources Research and Development for providing spectrometer instrument for this research.

References

[1] Gredilla A, de Vallejuelo S F-O, Elejoste N, de Diego A and Madariaga J M 2016 Non-destructive Spectroscopy combined with chemometrics as a tool for Green Chemical Analysis of environmental samples: A review TrAC Trends Anal. Chem. 76 30–39
[2] Bellon-Maurel V and McBratney A 2011 Near-infrared (NIR) and mid-infrared (MIR) spectroscopic techniques for assessing the amount of carbon stock in soils–Critical review and research perspectives Soil Biol. Biochem. 43 1398–1410
[3] Stenberg B, Viscarra-Rossel R A, Mouazen A M and Wetterlind J 2010 Visible and near infrared spectroscopy in soil science Adv. Agron. 107 163–215
[4] Dotto A C, Dalmolin R S D, Grunwald S, ten Caten A and Pereira Filho W 2017 Two preprocessing techniques to reduce model covariables in soil property predictions by Vis-NIR spectroscopy Soil Tillage Res. 172 59–68
[5] Knox N M, Grunwald S, McDowell M L, Bruland G L, Myers D B and Harris W G 2015 Modelling soil carbon fractions with visible near-infrared (VNIR) and mid-infrared (MIR) spectroscopy Geoderma 239 229–239
[6] Nocita M, Stevens A, van Wesemael B, Brown D J, Shepherd K D, Towett E, Vargas R and Montanarella L 2015 Soil spectroscopy: an opportunity to be seized Glob. Change Biol. 21 10–11
[7] Dotto A C, Dalmolin R S D, ten Caten A and Grunwald S 2018 A systematic study on the application of scatter-corrective and spectral-derivative preprocessing for multivariate prediction of soil organic carbon by Vis-NIR spectra Geoderma 314 262–274

[8] Jiang Q, Li Q, Wang X, Wu Y, Yang X and Liu F 2017 Estimation of soil organic carbon and total nitrogen in different soil layers using VNIR spectroscopy: Effects of spiking on model applicability Geoderma 293 54–63

[9] Morellos A, Pantazi X-E, Moshou D, Alexandridis T, Whetton R, Tziotzios G, Wiebensohn J, Bill R and Mouazen A M 2016 Machine learning based prediction of soil total nitrogen, organic carbon and moisture content by using VIS-NIR spectroscopy Biosyst. Eng. 152 104–116

[10] Sorenson P, Small C, Tappert M, Quideau S, Drozdowski B, Underwood A and Janz A 2017 Monitoring organic carbon, total nitrogen, and pH for reclaimed soils using field reflectance spectroscopy Can. J. Soil Sci. 97 241–248

[11] Knadel M, Thomsen A, Schelde K and Greve M H 2015 Soil organic carbon and particle sizes mapping using vis–NIR, EC and temperature mobile sensor platform Comput. Electron. Agric. 114 134–144

[12] Soriano-Disla J M, Janik L J, Viscarra-Rossel R A, Macdonald L M and McLaughlin M J 2014 The performance of visible, near-, and mid-infrared reflectance spectroscopy for prediction of soil physical, chemical, and biological properties Appl. Spectrosc. Rev. 49 139–186

[13] De Jong S 1993 SIMPLS: an alternative approach to partial least squares regression Chemom. Intell. Lab. Syst. 18 251–263

[14] Hoerl A 1962 Application of ridge analysis to regression problems Chem. Eng. Prog. 58 54–59

[15] Hoerl A E and Kennard R W 1970 Ridge regression: Biased estimation for nonorthogonal problems Technometrics 12 55–67

[16] R Core Team 2017 R: A Language and Environment for Statistical Computing (Vienna, Austria: R Foundation for Statistical Computing)

[17] Indonesia Center for Agricultural Land Resources Research and Development 2000 Map of Indonesia Land System

[18] Sigit S 1965 Geologic map of Indonesia - Peta geologi Indonesia

[19] Hikmatullah, Tafakresnanto, Chendy and Sulaeman, Yyi 2016 Framing the Land Resources Geospatial Information to Support Agricultural Area Development - Penyusunan Informasi Geospatial Mendukung Pengembangan Kawasan Pertanian (Bogor, Indonesia: Indonesian Center for Agricultural Land Resources Research and Development (ICALRRD))

[20] Walkley A and Black I A 1934 An examination of the Degtjareff method for determining soil organic matter, and a proposed modification of the chromic acid titration method Soil Sci. 37 29–38

[21] Kjeldahl J 1883 Neue methode zur bestimmung des stickstoffs in organismen körpren Z. Für Anal. Chem. 22 366–382

[22] Jarvis A, Reuter H I, Nelson A and Guevara E 2008 Hole-filled SRTM for the globe Version 4, available from the CGIAR-CSI SRTM 90m Database

[23] Hartmann J and Moosdorf N 2012 The new global lithological map database GLiM: A representation of rock properties at the Earth surface Geochem. Geophys. Geosystems 13 1–37

[24] Pinheiro É F, Ceddia M B, Clingensmith C M, Grunwald S and Vasques G M 2017 Prediction of Soil Physical and Chemical Properties by Visible and Near-Infrared Diffuse Reflectance Spectroscopy in the Central Amazon Remote Sens. 9 1–22

[25] Savitzky A and Golay M J 1964 Smoothing and differentiation of data by simplified least squares procedures. Anal. Chem. 36 1627–1639

[26] Signal Developers 2014 signal: Signal processing
[27] Brenning A 2008 Statistical geocomputing combining R and SAGA: The example of landslide susceptibility analysis with generalized additive models J. Böhner, T. Blaschke, L. Montanarella (Eds.), SAGA — Seconds Out. Hamburger Beiträge zur Physischen Geographie und Landschaftsökologie vol 19 pp 23–32

[28] Hijmans R J 2016 raster: Geographic Data Analysis and Modeling

[29] Friedman J, Hastie T and Tibshirani R 2010 Regularization paths for generalized linear models via coordinate descent J. Stat. Softw. 33 1–24

[30] Wehrens R and Mevik B 2007 The pls package: principal component and partial least squares regression in R J. Stat. Softw. 18 1–23

[31] Chang C-W, Laird D A, Mausbach M J and Hurburgh C R 2001 Near-infrared reflectance spectroscopy–principal components regression analyses of soil properties Soil Sci. Soc. Am. J. 65 480–490

[32] Kusumo B, Sukartono, S and Bustan, B 2018 The rapid measurement of soil carbon stock using near-infrared technology IOP Conference Series: Earth and Environmental Science vol 129 (IOP Publishing) pp 1–7

[33] Kusumo B, Sukartono S and Bustan B 2018 Rapid Measurement of Soil Carbon in Rice Paddy Field of Lombok Island Indonesia Using Near Infrared Technology IOP Conference Series: Materials Science and Engineering vol 306 (IOP Publishing) pp 1–7