ISSN 0367-6315
Korean J. Soil Sci. Fert. 45(4), 497-502 (2012)
http://dx.doi.org/10.7745/KJSSF.2012.45.4.497

Article

Predicting Organic Matter content in Korean Soils Using Regression rules on Visible-Near Infrared Diffuse Reflectance Spectra

Hyen Chung Chun, Suk Young Hong*, Kwan Cheol Song, Yihyun Kim, Byung Keun Hyun, and Budiman Minasny[1]

National Academy of Agricultural Science, Suwon, 441-707, Korea
[1]Department of Agriculture, Food and Natural Resources, The University of Sydney, Sydney, NSW 2006, Australia

This study investigates the prediction of soil OM on Korean soils using the Visible-Near Infrared (Vis-NIR) spectroscopy. The ASD Field Spec Pro was used to acquire the reflectance of soil samples to visible to near-infrared radiation (350 to 2500 nm). A total of 503 soil samples from 61 Korean soil series were scanned using the instrument and OM was measured using the Walkley and Black method. For data analysis, the spectra were resampled from 500-2450 nm with 4 nm spacing and converted to the 1st derivative of absorbance (log (1/R)). Partial least squares regression (PLSR) and regression rules model (Cubist) were applied to predict soil OM. Regression rules model estimates the target value by building conditional rules, and each rule contains a linear expression predicting OM from selected absorbance values. The regression rules model was shown to give a better prediction compared to PLSR. Although the prediction for Andisols had a larger error, soil order was not found to be useful in stratifying the prediction model. The stratification used by Cubist was mainly based on absorbance at wavelengths of 850 and 2320 nm, which corresponds to the organic absorption bands. These results showed that there could be more information on soil properties useful to classify or group OM data from Korean soils. In conclusion, this study shows it is possible to develop good prediction model of OM from Korean soils and provide data to reexamine the existing prediction models for more accurate prediction.

Key words: Soil carbon, Diffuse reflectance spectroscopy, Infrared spectroscopy

Introduction

Soil organic matter (OM) content is an important indicator for soil quality, fertility, and soil health. For the past decade, soil OM content has received a lot of interest because it is closely related to the potential of soil carbon sequestration (Bellon-Maurel and McBratney, 2011). In addition, soil OM promotes good soil structure and soil health. However, soil OM is highly variable across scales, and it is difficult to measure or predict continuously over a region (Pozdnyakova et al., 2005). Traditional measurement of soil carbon based on chemical oxidation or dry combustion is time consuming and expensive (Bellon-Maurel and McBratney, 2011). To cover the large variation of soil OM in a region, we need to analyze a large number of soil samples, thus the traditional methods are expensive to carry out for regional soil assessment. Reeves (2010) suggested that there is a need for a new rapid method which can give good quality data needed to cover soil's variation in a region.

Currently there is a large variety of methods for measuring soil carbon. Among these relatively new techniques, infrared spectroscopic techniques are promising, because they are low cost and are easy to use, which can be feasible for acquiring data for a large region. The use of infrared spectroscopy in agriculture started in 80's for measuring fruits and vegetations qualities. Near-infrared (NIR) spectroscopy has become well established in agricultural field (Wetzel, 1983). Recently, NIR spectroscopy techniques have been developed as a useful quantitative tool for the prediction of various soil properties; including soil moisture, soil organic carbon, nitrogen content, and soil texture (Dalal and Henry, 1986; Morra et al., 1991; Reeves et al., 2002).

While spectroscopic techniques are easy to use, they produce a huge amount of data which can be difficult to

Received : 2012. 6. 18 Accepted : 2012. 8. 7
*Corresponding author : Phone: +82312900340
E-mail: syhong @korea.kr

handle. Researchers resolve this problem by applying data reduction methods, such as principal component analysis, partial least squares or rule-based regression (Bellon-Maurel and McBratney, 2011). Because of the complex and overlapping absorption of soil constituents in the infrared spectra, it is unfeasible to predict OM from the reflectance at selected wavelengths. The use of these chemometric methods in the past twenty years made it possible to predict various soil properties from the whole spectra. Partial least squares regression (PLSR) is a linear model, which uses both the spectral and known property data during the calculation of the principal components (Wold et al., 1984). The known property data, spectral information and organic matter content, are projected onto a latent variable, and a second orthogonal variable is derived from the residuals. This process is repeated until the model is complete. The method has advantages in reducing spectral dimension, noise, and avoiding the need for wavelength selection. PLSR has been used routinely in NIR spectroscopy for predicting soil properties. However, because of its linear nature, the PLS model have limitations in its prediction power.

The rule-based regression is a data mining technique that builds a model that contains one or more rules that relates the independent variables (spectra) to a dependent variable (soil OM). If a case satisfies all conditions of a rule, then the linear equation is determined for the dependent variable. Cubist is the software implementin the rule-based piece-wise regression model (Quinlan, 1992). This type of model has just recently been introduced in handling soil spectral data by Minasny and McBratney (2008). It is attractive as it produces descriptive models that can help better understand the complicated structure and relationships in data. It was found to give high prediction accuracy, the model is easy to interpret, has automatic variable selection that makes it parsimonious, and respects the upper and lower boundary values of the predictant. The objectives of this study are to predict soil OM content for Korean soils using visible-near infrared spectra, to develop prediction models using PLSR and Cubist, and to validate accuracy between two models for finding the optimal prediction model.

Materials and Methods

Soils Soil samples were taken from all over South Korea region based on the soil series information. A total of 580 samples from 61 (out of 123) Korean soil series were taken during 2009-2011. For each soil order (soil profile), about 4 kg of soil was taken from each horizon using a small shovel. Each soil profile has an average of 4 horizons. The soil samples were collected from Inceptisols (61% among the whole samples), Alfisols (16%), Ultisols (12%), Andisols (10%), and Mollisols (1%). All samples were transferred to soil testing laboratory in National Academy of Agricultural Sciences, Suwon (Korea). The samples were air dired for two weeks to make sure all samples were fully dired. After laboratory measurement, the samples were dried again at 60°C for 5 hours for the spectrum scanning. All soil samples were ground and sieved (< 2 mm) to reduce aggregated particles for the reflectance spectra scanning and laboratory OM measurements. An amount of 500 g of soils were measured for soil OM content in the laboratory with the Walkley-Black method. The rest of the samples were used for the spectra scanning.

After laboratory measurement, only 503 samples were selected for spectroscopy scanning. This is because some samples showed unreasonable OM contents based on the laboratory analysis, and therefore these samples were excluded. The soil samples were placed into a 3.2 cm wide and 1 cm height sample holder without compression and leveled for the spectra reading. The visible-near infrared spectra from 350 nm to 2500 nm were measured using the ASD FieldSpec Pro (Applied Spectral Devices, Boulder, CO). During the reflectance measurement, a halogen lamp was installed to equalize the radiation energy into the soil samples.

Spectral preprocessing and OM transformation Averaging multiple measurements of a target is a good practice to compensate for variations, and so that scans with spectral artifacts can be removed. The samples were scanned 50 times at each spectrum and average values of each sample were used in this study. The original spectrum with 1 nm interval bands was resampled at every 4 nm from 500-2,450 nm. Reflectance values < 500 nm and > 2,450 nm were remove because of the low signal to noise ratio. The spectra were smoothed and the 1st derivative of the absorbance spectra (log[1/reflectance]) was calculated with the Savitsky-Golay algorithm. In addition, the raw soil OM data were normalized using a square root transformation for the statistical analyses. Transformed OM data were used for all of the model developments.

Prediction & Validation In this study, two prediction methods were applied: partial least-squares regression

(PLSR) and a regression rule model Cubist. PSLR was employed to quantify OM from VIS/IR spectra. The PLSR was performed with the JMP statistical program (SAS, N.C., USA). The Cubist model consists of a collection of rules of the form of:

```
If      A[w_c1] > c1 and A[w_c2] > c2
Then    y = b0 + b1 * A[w_1] + b2 * A[w_2] + ···
```

where A[w] refers to the 1st derivative absorbance value at wavelength w, b are parameters of a linear model, c are the value of the conditions, and y is the target variable (square root of OM). A rule indicates that, whenever a case satisfies all the conditions, the linear model is appropriate for predicting the value of the target attribute.

The collected profile data were split randomly into two parts: 363 samples (from 44 profiles) were used for developing the prediction model, while the rest 140 samples (from 16 profiles) were used for validation of the prediction accuracy.

Results and Discussion

The results of the measured soil OM data are shown in Fig. 1 & Fig. 2. The mean value of measured SOM from all soil samples was 17.28 ± 25.11g/kg. Andisols have the highest OM concentration (70.07 ± 34.43 g/kg), followed by Mollisols (23.76 ± 0.00 g/kg), Inceptisols (13.84 ± 10.60 g/kg), Entisols (10.64 ± 14.74 g/kg), Ultisols (9.56 ± 3.86 g/kg) and Alfisols (7.21 ± 3.76 g/kg). However the distribution was skewed, so the data were normalized using a square root transformation (Fig. 2). In Fig. 3, the average values of OM scanning from Spectra shows that Alfisols had the highest reflectance values while the Andisols had the smallestwhich correspond to soil color

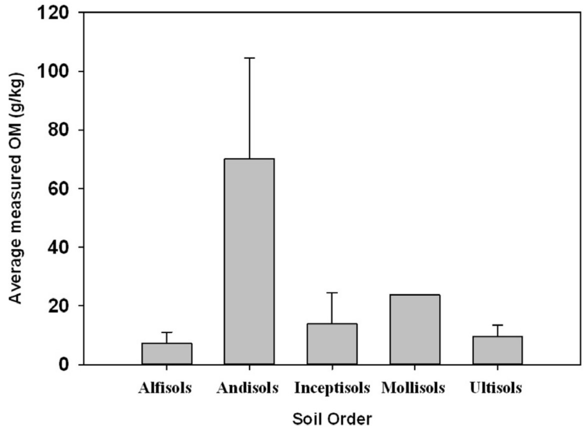
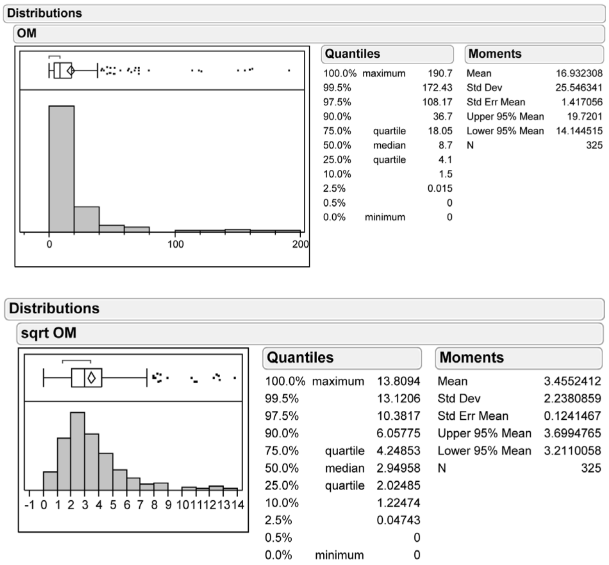


Fig. 1. The average and standard deviation of organic matter (OM) contents by laboratory measuring from each soil order.



```
Quantiles                          Moments
100.0%  maximum    190.7           Mean             16.932308
99.5%              172.43          Std Dev          25.546341
97.5%              108.17          Std Err Mean      1.417056
90.0%              36.7            Upper 95% Mean   19.7201
75.0%   quartile   18.05           Lower 95% Mean   14.144515
50.0%   median     8.7             N                      325
25.0%   quartile   4.1
10.0%              1.5
2.5%               0.015
0.5%               0
0.0%    minimum    0
```

```
Quantiles                          Moments
100.0%  maximum    13.8094         Mean             3.4552412
99.5%              13.1206         Std Dev          2.2380859
97.5%              10.3817         Std Err Mean     0.1241467
90.0%              6.05775         Upper 95% Mean   3.6994765
75.0%   quartile   4.24853         Lower 95% Mean   3.2110058
50.0%   median     2.94958         N                      325
25.0%   quartile   2.02485
10.0%              1.22474
2.5%               0.04743
0.5%               0
0.0%    minimum    0
```

Fig. 2. Histogram of OM content from the Korean soil database (top) and the data renormalization by square root transformation (bottom).

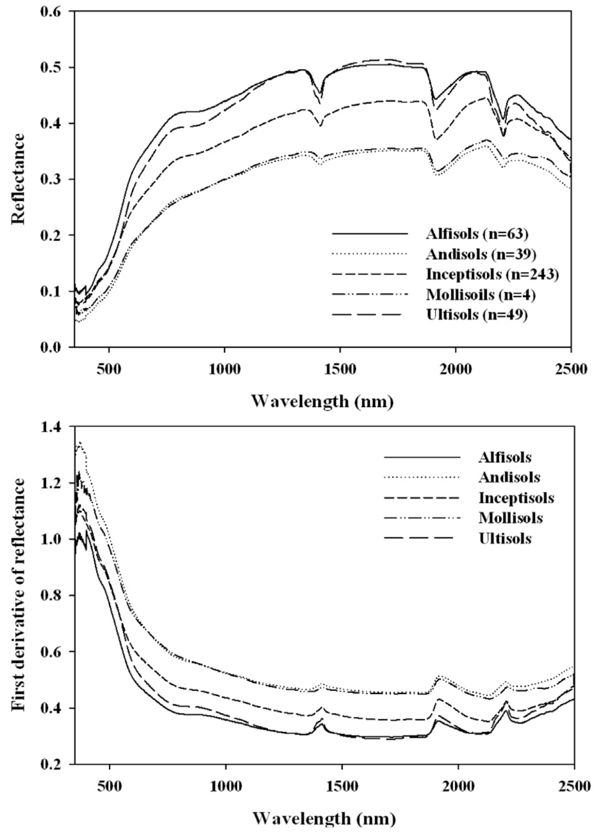


Fig. 3. The average values of 50 times scanning by spectrascopy from all soil samples (Top) and the first derivative of the reflectance after transformed to (log[1/reflectance]) (Bottom).

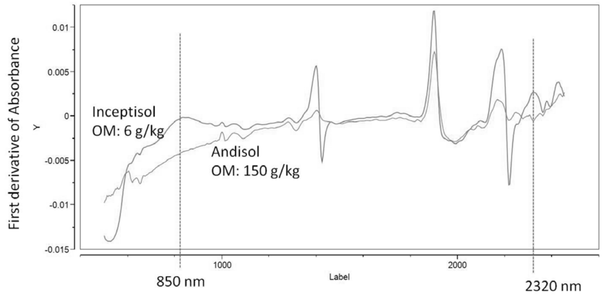


Fig. 4. Separation of absorbance at 850 nm and 2320 nm from organic matter readings of two soil orders (Inceptisol and Andisol).

and organic matter. The result of the PSLR was;

$$PLSR\ Sqrt(OM) = 0.70 + 0.79 * Sqrt(OM)\ R^2=0.870 \qquad (1)$$

where Sqrt represents square root function. The detailed discussion of the validation for PSLR would be discussed later in this paper.

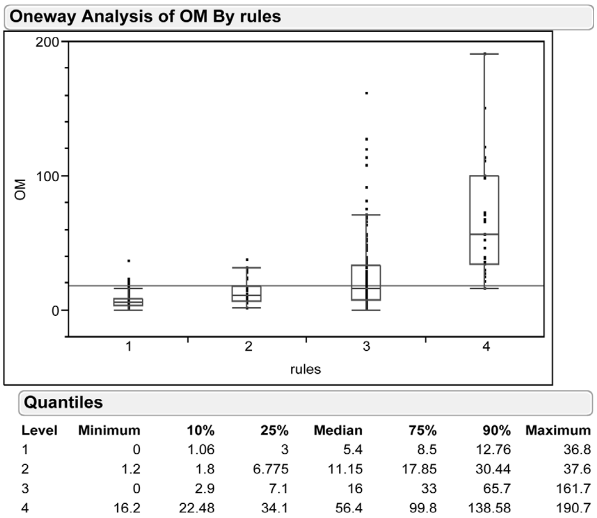
Cubist is a rule-based model and one of the advantages is providing data grouping in the process. In Fig. 4, Inceptisols and Andisols are shown as examples of separation in the spectra reading. As explained above in Materials and Methods section, Cubist provides linear regression models as many as data groupings (rules). In this study, the stratification occurred at 850 and 2320 nm and the 4 rules were determined based on these separations. These Cubist analyses provided four equations to predict OM for each rule;

```
Rule 1: [170 cases, mean 2.3143947, range 0 to 6.0663,
        est err 0.6424281]
    if
        852 > -0.0006437
        2320 > 0.00056882
    then
        Sqrt_OM = 2.9199813 - 3267 724 + 2414 844 +
                  2512 728 + 1523 1920 - 1332 1924
                  - 1679 816 - 1777 852 - 501 2148 +
                  602 2140 + 356 1424 + 1417 720 -
                  1348 708 + 1252 2020 - 413 2232 +
                  440 824 + 529 716 + 350 2296 - 64
                  2216 - 459 1940 + 106 2208 + 134
                  1388 - 223 624 - 304 2316 + 197
                  828 + 175 680 + 176 700 + 177
                  2248 - 33 556 + 108 1360 + 35 544
                  - 119 2320 + 43 2344 + 34 2340
Rule 2: [31 cases, mean 3.4962814, range 1.095445 to
        6.131884, est err 0.9603595]
    if
        852 > -0.0006437
        2320 <= 0.00056882
    then
        Sqrt_OM = 2.4285159 + 9750 716 - 6821 724 -
                  3419 700 + 313 2140 - 239 2148 +
                  154 1424 + 454 2020 - 50 2216 +
                  83 2208 - 357 1940 + 201 1920 +
                  321 2248 - 192 2232 - 133 624 +
                  148 2296 - 91 1924 + 84 1360 + 28
                  544 - 93 2320
```

```
Rule 3: [139 cases, mean 4.4711714, range 0 to 12.71613,
        est err 0.9374755]
    if
        852 <= -0.0006437
        2316 > 4.83718e-005
    then
        Sqrt_OM = 1.0737783 + 10255 844 - 12518
                  724 - 8057 852 + 9626 728 - 6432
                  816 + 5432 720 - 5167 708 + 1604
                  1920 + 769 2140 + 1687 824 +
                  2028 716 - 1091 1924 + 376 1424
                  - 415 2148 + 515 1388 - 948 2344
                  - 1164 2316 + 920 2340 + 754 828
                  + 671 680 + 673 700 - 127 556 -
                  401 2232 - 218 1916 - 60 2196
```

```
Rule 4: [23 cases, mean 7.8844032, range 4.024922 to
        13.80942, est err 1.0692692]
    if
        852 <= -0.0006437
        2316 <= 4.83718e-005
    then
        Sqrt_OM = 3.8367835 - 2856 852 + 1471 844 - 1243 724
                  + 684 1920 + 956 728 - 639 816 -
                  391 1924 + 539 720 - 513 708 - 228
                  1916 - 80 2196 + 76 2140 + 168
                  824 + 201 716 + 37 1424 - 41 2148
                  + 51 1388 - 94 2344 - 116 2316 +
                  91 2340 + 75 828 + 67 680 + 67
                  700 - 13 556 - 40 2232            (2)
```

Based on Cubist separations, the 4 rules correspond to OM concentration at mean values of 6 ± 5, 14±10, 25±28, and 69 ± 45 g/kg (Fig. 5). There was an expectation to find separation or characterization of soil orders based on 4



Level	Minimum	10%	25%	Median	75%	90%	Maximum
1	0	1.06	3	5.4	8.5	12.76	36.8
2	1.2	1.8	6.775	11.15	17.85	30.44	37.6
3	0	2.9	7.1	16	33	65.7	161.7
4	16.2	22.48	34.1	56.4	99.8	138.58	190.7

Fig. 5. Organic matter data from spectroscopy grouping by Cubist; Oneway analysis results of the organic matter grouping.

rules from Cubist. Unfortunately, we found no apparent relation between the grouping results and soil orders.

In order to validate OM, PLSR analysis using cross-validation showed that the 15 components accounted for 90% of variation in the prediction. This PLSR was applied to predict OM using 363 measured samples. In the training data, PLSR shows a good prediction for OM content with R^2 = 0.870 (Table 1). The PLSR and Cubist models were validated by predicting it to the validation set. Although PLSR showed a better prediction compared to Cubist on the training data, it has a lower accuracy on the validation dataset. R^2 value for PLSR is 0.503, while for Cubist is 0.701 (Table 1). This indicates that PLSR overfitted the data.

Cubist appears to provide a better prediction for OM. The results of Cubist model were plotted to compare the actual measured values and the predicted ones. Cubist model displays a better prediction following the 1:1 line when compared to the PLSR (Fig. 6). In addition, the MSE for Cubist is almost half of the PLSR prediction, indicating a higher accuracy.

Minasny and McBratney (2008) stated that one of the advantages in Cubist models provide data grouping. It is possible for Cubist to separate data into more detailed groups to improve the accuracy of the prediction. There was an assumption that soil orders may affect the stratification of the model. However, we do not find any pattern that soil order can be related directly to the rules grouping estimated by Cubist. Although Ultisols which has a low OM content only occurs in Rule 1, Inceptisols can occur in all 4 rules, and Andisols which has a special characteristic of allophanic minerals also can occur in all rules. In this case, the stratification for OM prediction has no relation with soil orders. It mainly reflects the organic components of the soils. In order to find detail relations between spectral data and soil orders, other properties such as soil color or texture are needed to consider in developing prediction models.

Table 1. Goodness of fit for the prediction of the square root of OM spectroscpty using PLSR and Cubist.

	Mean Error (ME) $(g/kg)^{0.5}$	Mean Squared Error (MSE) (g/kg)	R^2
Training set (n=363)			
PLSR	0.000	0.715	0.870
Cubist	0.078	0.847	0.849
Validation set (n=140)			
PLSR	-0.041	2.128	0.503
Cubist	0.043	1.194	0.701

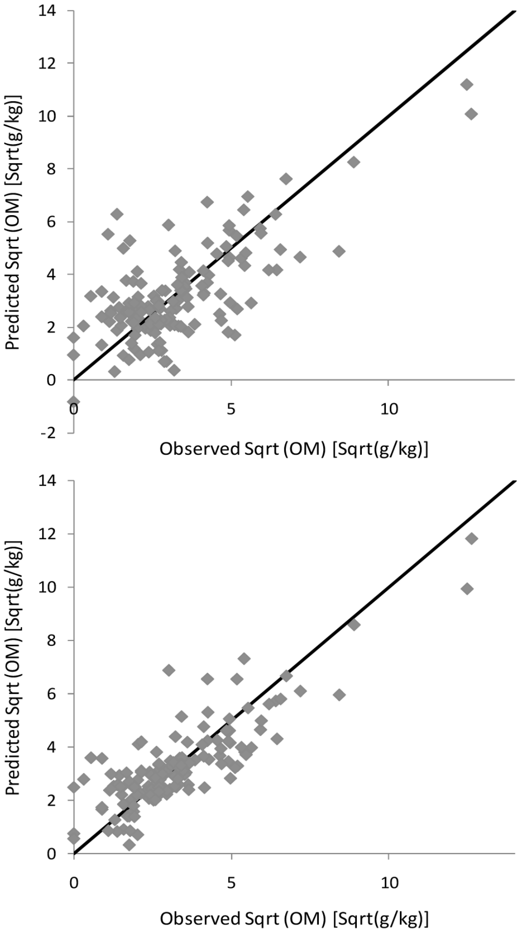


Fig. 6. Scatter plot of PLS (top) and Cubist (bottom) model to predict OM.

Conclusion

This study concluded that the rule-based regression model performed better that PLSR to predict OM for Korean soils. Since Korean soils display extreme variety in a relatively small scale all over the country, it would be better using a rule-based regression model, which is useful to classify large dataset and provide clear linear relationships among data. This study will be useful information to create digital map of OM and its change in Korea.

Acknowledgments

This study was funded by 'Estimation and mapping soil properties using remote sensing techniques', Rural Development Administaion, Republic of Korea (PJ006909) and 'Estimation of carbon storage in cropland using remote sensing', National Research Foundation of Korea.

References

Bellon-Maurel, V. and A. McBratney. 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.

Dalal, R.C. and R.J. Henry. 1986. Simultaneous determination of moisture, organic carbon, and total nitrogen by near infrared reflectance spectrophotometry. Soil Sci. Soc. Am. J. 50:120-123.

Minasny, B. and A. B. McBratney. 2008. Regression rules as a tool for predicting soil properties from infrared reflectance spectroscopy. Chemometr. Intell. Lab. 94:72-79.

Morra, M.J, M.H. Hall, and L.L. Freeborn. 1991. Carbon and nitrogen analysis of soil fractions using near infrared reflectance spectroscopy. Soil Sci. Soc. Am. J. 55:288-291.

Pozdnyakova, L, D. Giménez, and P. Oudemans. 2005. Spatial analysis of cranberry yield at three scales. Agron. J. 97:49-57.

Quinlan, J.R. 1992. Learning with continuous classes, Proceedings of the 5th Australian Joint Conference on Artificial Intelligence, World Scientific, Singapore, pp. 343-348.

Reeves III, J. B. 2010. Near-versus mid-infrared diffuse reflectance spectroscopy for soil analysis emphasizing carbon and laboratory versus on-site analysis: Where are we and what needs to be done? Geoderma 158:3-14.

Reeves, III J.B., G.W. McCarty, and T. Mimmo. 2002. The potential of diffuse reflectance spectroscopy for the determination of carbon inventories in soil. Environ. Pollut. 116:264-277.

Wetzel, D.L. 1983. Near-infrared reflectance analysis: Sleeper among spectroscopic techniques. Anal. Chem.55:1165-1176.

Wold, S., A. Ruke, H. Wold and W.J. Dunn. 1984. The collinearity problem in linear regression, the partial least squares (PLS) approach to generalized inverses. SIAM J. Sci. Stat. Comp. 5:735 -743.