Development of field-scale soil organic matter content estimation models in Eastern Canada using airborne hyperspectral imagery

Y. Uno[1], S.O. Prasher[1]*, R.M. Patel[1], I.B. Strachan[2], E. Pattey[3] and Y. Karimi[1]

*[1]Department of Bioresource Engineering and [2]Department of Natural Resource Sciences, Macdonald Campus of McGill University, 21111 Lakeshore Road, Ste-Anne-de-Bellevue, Quebec H9X 3V9, Canada; and [3]Eastern Cereal and Oilseed Research Centre, Research Branch, Agriculture and Agri-Food Canada, K.W. Neatby Building, 960 Carling Avenue, Ottawa, Ontario K1A 0C6, Canada. Email: *shiv.prasher@mcgill.ca*

Uno, Y., Prasher, S.O., Patel, R.M., Strachan, I.B. and Pattey, E. 2005. **Development of field-scale soil organic matter content estimation models in Eastern Canada using airborne hyperspectral imagery.** Canadian Biosystems Engineering/Le génie des biosystèmes au Canada **47**: 1.9 - 1.14. Accurate estimation of within-field soil organic matter (SOM) is currently an important priority for precision agriculture, given its importance in defining precise fertilizer and pesticide management practices. In this study, the potential of airborne hyperspectral imagery in estimating within-field SOM was assessed for a 30 ha clay-loam soil field in Ottawa, Ontario, Canada. Various multivariate statistical techniques, including Principal Component Analysis and Stepwise Regression, as well as Artificial Neural Networks, were employed to generate a predictive model for SOM. The high prediction accuracy obtained (NRMSE = 9.98% for PCA-SMLR; 12.08% for PCA-ANN models) suggests that hyperspecral remote sensing can be an effective tool in describing the variability of SOM on a field scale. However, further studies are needed before this methodology can be applied for other soil types and locations.
Keywords: soil organic matter, airborne hyperspectral remote sensing, precision agriculture, multivariate statistical analysis, artificial neural networks.

En agriculture de precision, l'estimation de la matière organique du sol (MOS) à même le champ est présentement, étant donné son importance dans le développement d'un bon aménagement des application de fertilisants et de pesticides, une priorité de pointe. Cette étude visa à établir le potentiel de l'utilisation d'un appareil de détection hyperspectrale aéroporté pour estimer la teneur en MOS d'une parcelle de 30 ha de loam argileux près d'Ottawa, Ontario, Canada. Plusieurs modèles statistiques furent employés, soit l'analyse en composantes principales suivi de régression par degrés (PCA-SMLR), ou suivi d'analyse par réseau de neurones formel (PCA-ANN), afin de développer un modèle servant à estimer la MOS.L'exactitude élevée de prédiction (NRMSE = 9.98% pour PCA-SMLR; 12.08% pour PCA-ANN) indique qu'une détection hyperspectrale aéroporté pourrait être très efficace pour prédire la quantite et la variabilité de la MOS dans un champ. Cependant, des études supplémentaires sont nécessaires avant que cette méthode ne soit mise en œuvre avec d'autres type de sols ou en d'autres régions.

INTRODUCTION

Soil organic matter (SOM), a useful indicator of soil fertility and a crucial factor in the soil dynamics of various agrochemicals, is one of the most influential of agricultural soil parameters (Krishnan et al. 1981; Page 1974; Pitts et al. 1986). Accurate estimation of SOM levels can allow an accurate field-specific estimation of appropriate quantities and types of agrochemical inputs, thus optimizing field productivity and minimizing groundwater contamination risks (Fernandez et al. 1988; Ingleby and Crowe 2001; Sudduth and Hummel 1996). In particular, the development of SOM field maps is currently an important aspect of precision agriculture, as such maps can yield information useful in the implementation of variable rate technologies (VRTs).

Although many new techniques are currently being developed, grid sampling, followed by geostatistical methods such as kriging, is most commonly used in mapping SOM levels on a field-scale (Chen et al. 2000; Fox and Sabbagh 2002). However, some problems arise with this approach: (i) it normally requires a large number of soil samples, and (ii) SOM determinations are time consuming (Chen et al. 2000; Fox and Sabbagh 2002).

However, remote sensing can directly correlate spectral information with SOM content at a pixel level. Use of airborne remote sensing is potentially an effective alternative to describe the within-field variability of SOM. Unlike conventional aerial photography, the use of an airborne digital multi- or hyper-spectral imaging system, or a Charged Coupled Device (CCD) camera, have great potential in accurately estimating SOM levels, since they offer a large amount of spectral information (Chen et al. 2000; Plant et al. 2000). Although in past years, high cost of remote sensing data acquisition operation has precluded such approaches from being applied for agricultural purposes, recent advances in technology and reductions in costs have increased the viability of these methodologies for a number of precision agricultural applications.

It appears that although the potential of airborne remote sensing for SOM estimation is quite high, significant technical breakthroughs are still needed in order to render it a fully operative technology. In particular, the development of accurate prediction models is one of the most important issues. Past studies have shown that various modeling methods, such as Stepwise Regression (Dalal and Henry 1986; Hummel et al. 2001; Ingleby and Crowe 2000; Sudduth and Hummel 1996, 1991), Partial Least Square (PLS) analysis (McCarty et al. 2002; Sudduth and Hummel 1991), and ANNs (Ingleby and Crowe 2001), can be used effectively. However, most of these studies were conducted under laboratory conditions, with minimal sources of noise, such as differences in soil moisture and where organic residues were removed prior to the analysis. Few studies were conducted outdoors where aerial images were used (Bajwa et al. 2001; Chen et al. 2000; Vinogradov 1981). Another problem is that prediction models, whether derived from laboratory analyses or aerial images, cannot be generalized, since the spectral signature of soil is a complicated function of various chemical substances and parent materials with different absorption peaks (Ben-Dor et al. 1997). Indeed, past studies demonstrate that the performance of models is largely influenced by the experimental site and conditions (Dalal and Henry 1986; Sudduth and Hummel 1996). Therefore, it is currently an essential task to test these methodologies on different soil types.

The goal of this study was to develop a field-scale SOM estimation model using airborne hyperspectral remote sensing. Two different modeling methods, Artificial Neural Networks (ANNs) and Stepwise Multiple Linear Regression (SMLR) were employed to extract the full potential of hyperspectral remote sensing, while Principal Component Analysis (PCA) was used to reduce the redundant information of hyperspectral imagery.

METHODOLOGY

Experimental site and data acquisition

The experiment was conducted at the Greenbelt Farm near Ottawa, Ontario, Canada. This 30-ha corn (*Zea mays* L. Pioneer 3893) field consisted mainly of a clay-loam soil with a relatively high SOM level (range 3.35-6.63%, mean 4.83%, 0.77 standard deviation). Airborne hyperspectral imagery was obtained through a Compact Airborne Spectrographic Imager (CASI) sensor (ITRES) on May 2000 to avoid the influence of crop emergence on the spectral signature. Seventy-two spectral bands between 408.73 and 947.07 nm were obtained each with an interval of about 7.5 nm. However, band 72 (947.07nm) was later eliminated due to excessive noise. The measured radiances were converted into reflectance values through a series of pre-processing (geometric and atmospheric corrections, etc.) steps. Details on these corrections are given by Goel et al. (2003). The data extraction processes were conducted with the help of *ENVI* (Research Systems Inc., Boulder, CO), a software for remote sensing data analysis. Based on a preliminary analysis, a low-pass filter with a kernel size of 7 x 7 pixels (i.e. spatial resolution of 14 x 14 m) was used to remove noise.

A total of 50 samples was collected from the field. The SOM concentrations were determined by measuring ignition loss after heating the soils for 4 hours at 525°C. The spectral reflectance values for model development were extracted from the CASI data for these 50 locations. Three samples were later removed, since a different soil type (sandy soil) was apparent in some areas of the experimental site. Thus a total of 47 soil samples was used in this study.

Multivariate statistical analysis

The extracted spectral reflectance values were first analyzed by multivariate statistical methods. First, linear correlation coefficients were calculated between SOM levels and individual spectral bands. Afterward, principal component analysis (PCA) was conducted to reduce these 71 input variables into three important principal components (PCs). Correlation coefficients between these three principal components and SOM concentration were also calculated. The PCA was conducted with the default settings of the *Clementine Data Mining System* (SPSS 2001).

Model development

Two different modeling strategies were employed: (i) a Stepwise Multiple Linear Regression (SMLR) model was coupled with PCA (PCA-SMLR model), and (ii) an ANN model (feed-forward network with a back-propagation learning algorithm) was also coupled with the same PCA (PCA-ANN model). Past laboratory analysis had demonstrated these two modeling methods to be effective in the development of SOM estimation models from spectral information (Ingleby and Crowe 2000, 2001; Sudduth and Hummel 1991). Furthermore, ANNs were found to be a useful tool in development of non-linear models (Atkinson and Tatnall 1997).

For the SMLR model, the stepwise entry and exclusion significance levels were set to 0.050 and 0.10, respectively. Factor scores calculated from the top three principal components (PCs) (PC-1, 2, and 3) were used as inputs for this SMLR model, since the results of the principal component analysis inferred that the remaining 68 PCs did not have significant information for the SOM estimation. This modeling process was all conducted with the *Clementine Data Mining System* (SPSS 2001).

Except as stated below, the default "Quick" option was used for the development of ANN models with this system. Under the default settings, the optimum number of processing elements (PEs) in the hidden layer are automatically identified through an algorithm proprietary to the software, and based on the relationships between prediction accuracies and network structure (Integral Solutions 1998; SPSS 2001). A single hidden layer was selected to keep the model simple. To avoid the risk of overfit, the non-default "prevent-overtraining" option was used, where 75% of calibration samples (35 samples) are randomly selected for training, and the remaining 25% (12 samples) are allocated to testing during model building. The optimum learning cycles were also automatically determined by the software. The details of the algorithms are given in Integral Solutions (1998) and SPSS (2001). As in the case of SMLR models, three PCs were used as the inputs for the ANN model development.

Performance analysis

Since the number of samples was small, a ten-fold cross validation was conducted. In this procedure, the 47 samples were initially divided into 10 subsets (Groups A to J, with 4 or 5 samples per group). Next, 9 of the 10 subsets (e.g. Groups A to I) were allocated to calibration, and the remaining subset (Group J) was kept for validation. After finishing the validation, a different subset of 9 groups (e.g. Groups A to H and J) were selected for calibration, and the remaining group (Group I) was used for validation. This calibration and validation process was repeated 10 times with all possible combinations. It should be noted that the validation data set was not seen by the model during model development.

Various statistical evaluation methods, such as regression parameters, root mean squared error (RMSE), normalized root mean squared error (NRMSE) – RMSE represented as a percent of the mean observed values, and Nash-Sutcliffe coefficient

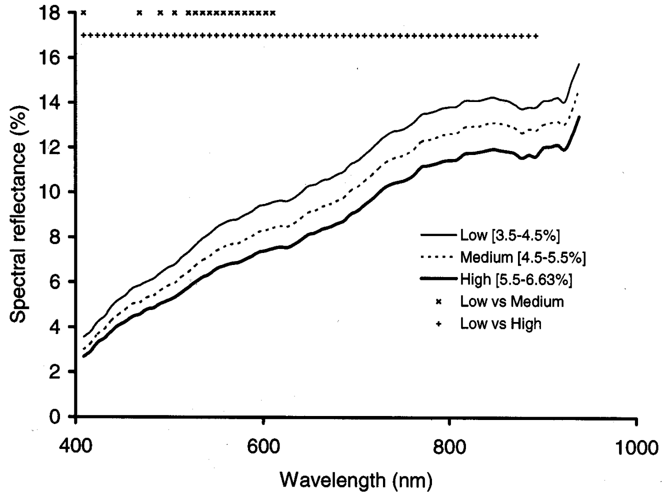

Fig. 1. Spectral signature of soils with three different soil organic matter levels. [x] indicates that the reflectance values of soil with low organic matter content is significantly higher (P<0.05) than that with medium organic matter content. [+] indicates that the reflectance values of soil with low organic matter content is significantly higher (P<0.05) than that with high organic matter content.

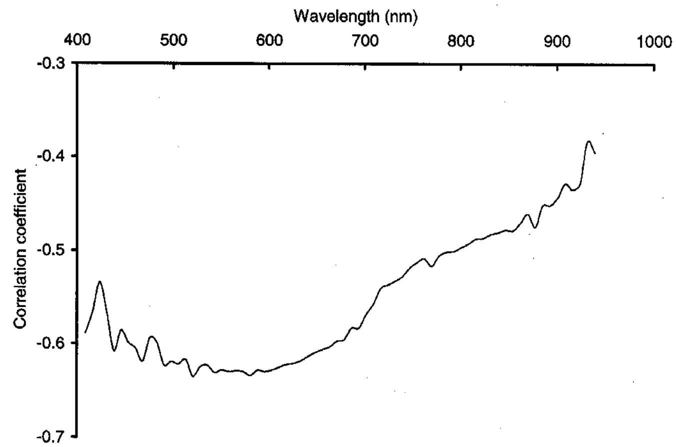

Fig. 2. Correlation coefficients (r) between the spectral reflectance values and percent soil organic matter content as a function of wavelength.

(NSC), were used to evaluate the performance of models. Regression parameters, i.e. intercept, slope, and correlation coefficient (r), represent the quality of relationship between the observed and predicted values. The values of intercept, slope, and r closer to 0.0, 1.0, and 1.0, respectively, represent very good model performance. Similarly, lower RMSE and NRMSE also represent good model performance. NSC value of 1.0 represents a perfect model, while a value of 0.0 would indicate a prediction no better than simply taking the mean of measured values. Negative NSC values show that the prediction is worse than using the observed mean (Nash and Sutcliffe 1970).

Finally, the topology of the developed ANN model was analyzed by using the "Net Node Browser" option of the software. The number of neurons in input, hidden, and output layers, and relative importance of each input variable were observed for all 10 models developed with different calibration datasets. For the calculation of relative importance, the "sensitivity analysis" option was used. In this option, the contribution of each input variable is calculated by measuring the variability of output values produced by the change of input values in testing samples. Further information is given in the *Clementine 8.0 Algorithms Guide* (SPSS 2003).

RESULTS and DISCUSSIONS

Spectral signature and soil organic matter content

The spectral signatures of soil with different SOM levels are presented in Fig. 1. At many wavebands, the spectral reflectance of soil with low SOM was statistically different (P<0.05) from the ones with medium and high SOM levels. Linear correlation coefficients between reflectance and SOM were moderately high ("-" indicates inverse relationship, $-0.4 < r < -0.6$) throughout the visible and NIR regions (Fig. 2). This indicates that SOM does contribute to the reflectance of almost all wavebands. The correlation coefficients are in the same range as those reported by Bajwa et al. (2001) for Illinois soils.

The results of the PCA (Table 1) show that approximately 97% of the variance was represented by the first principal component. This value is quite high, as compared with some other research studies (78% by Kaleita and Tian 2002). However, it should be noted that in our study: (i) the principal components were extracted from the spectral reflectance values of bare soil, and that all other objects in the image (i.e. buildings, vegetation) were excluded while extracting reflectance values from the image, and (ii) a low-pass filter was used before data extraction, which means that various noise in the image was removed before the PCA was performed (Kaleita and Tian (2002) had used a ground-based spectroradiometer and made point measurements). Without the low-pass filter, the first three PCs accounted for 92.3, 4.6, and 1.5, respectively, in this study.

Model performance and structure

The prediction accuracies obtained with the two modeling strategies are summarized in Figs. 3 and 4 and in Table 2. For both the models, PCA-SMLR and PCA-ANN, the regression parameters, intercept and slope were not different from their ideal values of 0 and 1, respectively(Figs. 3 and 4). These statistics indicated that both the models did a good job of

Table 1. Eigenvalues of top five principal components (PCs) and correlation coefficients between the principal components and percent organic matter.

Factors	Eigenvalues (%)	Eigenvalues (%, cumulative)	Correlation coefficient (r) with OM (%)
PC-1	96.98	96.98	-0.538
PC-2	1.618	98.60	-0.333
PC-3	1.040	99.64	0.152
PC-4	0.123	99.76	-0.020
PC-5	0.046	99.81	-0.082

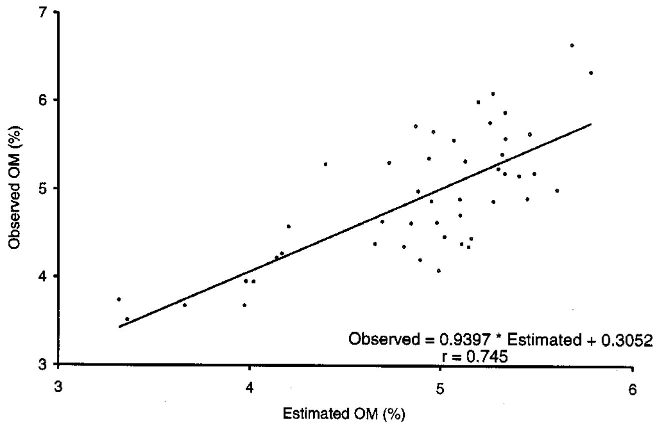

Fig. 3. Observed and estimated soil organic matter with SMLR model for ten-fold cross validation. The line represents the least-squares linear fit to the data. Regression parameters, intercept and slope, are not significantly different from 0 and 1, respectively.

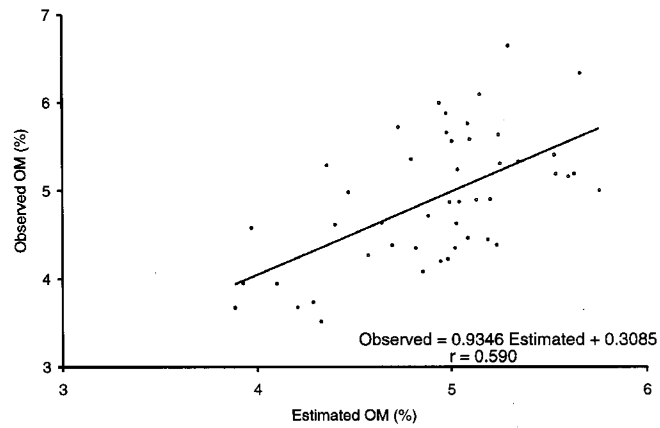

Fig. 4. Observed and estimated soil organic matter with ANN model for ten-fold cross validation. The line represents the least-squares linear fit to the data. Regression parameters, intercept and slope, are not significantly different from 0 and 1, respectively.

Table 2. Performance of the developed SMLR and ANN models. The NRMSE values shown were calculated based on the linear equations given in Figs. 3 and 4.

Methods	RMSE	NRMSE (%)	NSC
SMLR	0.490	9.98	0.553
ANN	0.592	12.08	0.346

simulating SOM. For all other evaluation parameters (RMSE, NRMSE, r, and NSC), the PCA-SMLR model outperformed the PCA-ANN model (Table 2). However, significance of these differences was not clear, since cross-validation methods generally produce high deviations when applied to a small number of samples, although it can generally be regarded as an unbiased estimator (Weiss and Kulikowski 1991). Further analysis is recommended by increasing the number of samples to define the significance of this difference.

From a practical point of view, the prediction accuracies obtained with these two methods (NRMSE = 9.98% for SMLR and 12.08% for ANN) generally seem to be acceptable for a number of agricultural applications. In particular, a graphical analysis shows that the PCA-SMLR model clearly described the variability of SOM levels in this experimental site (Fig. 3). The ANN models also described the variability in SOM (Fig. 4). However, it should also be noted that the relatively low NRMSEs are partially due to the high SOM levels with relatively low variability at this site (4.83% mean and 0.77 standard deviation). The low NSC value obtained with the PCA-ANN model 0.346 suggests that the effectiveness of the prediction is not very good.

The structures of the ANN models developed in this study are summarized in Table 3. According to the *Clementine Data Mining System* (SPSS 2001), the optimum number of PEs in the hidden layer was consistently determined to be four. The sensitivity analysis showed that the relative importance of the second principal component was the highest on average (relative importance = 0.453), followed by the first principal component (0.408), and the third principal component (0.114). However, the difference between the first principal component and second principal component was not significant based on the Wilcoxon matched pair test ($P > 0.05$). Further research is needed to analyze the stability of the ANN structure by using a larger number of samples. The equations developed with the SMLR models with a 10-fold cross validation procedure are summarized in Table 4. The results obtained with the 10-fold validation were consistent, as indicated by the low standard deviation values. Also, the third principal component was consistently rejected for all calibration data sets because of its insignificant contribution to the model.

Table 3. Structures of the developed ANN models and the relative importance of each input variable obtained in the 10-fold cross validation.

Model	Number of neurons			Sensitivity Analysis (Relative importance)		
	Input layer	Hidden layer	Output layer	PC-1	PC-2	PC-3
Model A	3	4	1	0.468	0.613	0.153
Model B	3	4	1	0.285	0.265	0.053
Model C	3	4	1	0.335	0.300	0.109
Model D	3	4	1	0.459	0.450	0.077
Model E	3	4	1	0.498	0.362	0.062
Model F	3	4	1	0.225	0.469	0.281
Model G	3	4	1	0.376	0.447	0.030
Model H	3	4	1	0.470	0.492	0.149
Model I	3	4	1	0.467	0.602	0.103
Model J	3	4	1	0.494	0.525	0.117
Average				0.408	0.453	0.114

Table 4. Coefficients and intercepts of the developed SMLR models obtained with the 10-fold cross validation. The values are base on the equation: SOM[%] = α + β*[PC-1] + γ*[PC-2] + δ*[PC-3]

Models	α	β	γ	δ
Model A	4.9417	-0.3601	-0.4484	N.A. (Rejected)
Model B	4.9556	-0.4410	-0.3711	N.A. (Rejected)
Model C	4.9530	-0.3926	-0.4427	N.A. (Rejected)
Model D	4.8941	-0.4159	-0.4251	N.A. (Rejected)
Model E	4.9572	-0.3889	-0.4903	N.A. (Rejected)
Model F	4.9140	-0.3619	-0.3995	N.A. (Rejected)
Model G	4.9108	-0.3459	-0.4352	N.A. (Rejected)
Model H	4.9285	-0.3910	-0.4227	N.A. (Rejected)
Model I	4.9660	-0.3932	-0.4515	N.A. (Rejected)
Model J	4.9697	-0.4105	-0.4121	N.A. (Rejected)
Average	4.9391	-0.3901	-0.4299	N.A
S.D.	0.0259	0.0286	0.0324	N.A.

N.A.: Not applicable

SUMMARY

In this study, the potential of airborne hyperspectral remote sensing for the estimation of within-field variability of SOM was explored for a clay-loam soil located near Ottawa, Ontario, Canada. Multivariate statistical techniques and ANNs were employed for model development to explore the potential of airborne hyperspectral imagery. The correlation coefficients between the observed and estimated values of SOM were 0.745 and 0.590 for SMLR and ANN models, respectively. The corresponding NRMSE values were 9.98 and 12.08. The results show that, with some ground truthing, aerial hyperspectral imagery can provide useful information for SOM estimation. It was also shown that PCA, SMLR analysis, and ANN models are effective methodologies for extracting this information from aerial hyperspectral imagery data. Further research is, however, recommended for different soil types, locations, and soil conditions, such as different moisture/nutrient levels, residue cover, and surface roughness. As shown by some previously conducted laboratory analyses (McCarty et al. 2002), use of broadband sensors, which include the whole range of near-infrared (700 - 2500 nm) and mid-infrared region (2500 - 25000 nm), could prove to be helpful in improving prediction accuracies and increase the generalization ability of the models.

REFERENCES

Atkinson, P.M. and AR.L. Tatnall. 1997. Neural networks in remote sensing. *International Journal of Remote Sensing* 18(4): 699-709.

Bajwa, S.G., L. Tian, D. Bullock, K. Sudduth, N. Kitche and H. Palm. 2001. Soil characterization in agricultural fields using hyperspectral image data. ASAE Paper No. 01-1106. St. Joseph: MI: ASAE.

Ben-Dor, E., Y. Inbar and Y. Chen. 1997. The reflectance spectra of organic matter in the visible near-infrared and short wave infrared region (400-2500 nm) during a controlled decomposition process. *Remote Sensing of Environment* 61: 1-15.

Blackburn, G.A. and E.J. Milton. 1997. An ecological survey of deciduous woodlands using airborne remote sensing and geographical information system (GIS*). International Journal of Remote Sensing* 18(9): 1919-1935.

Chen, F., D.E. Kissel, L.T. West and W. Adkins. 2000. Field-scale mapping of surface soil organic carbon using remotely sensed imagery. *Soil Science Society of America Journal* 64: 746-753.

Dalal, R.C. and R.J. Henry. 1986. Simultaneous determination of moisture, organic carbon, and total nitrogen by near infrared reflectance spectrophotometry. *Soil Science Society of America Journal* 50: 120-123.

Fernandez, R.N., D.G. Schulze, D.L. Coffin and G.E. Van Scoyoc. 1988. Color, organic matter, and pesticide absorption relationships in a soil landscape. *Soil Science Society of America Journal* 52: 1023-1026.

Fox, G.A.and G.J. Sabbagh. 2002. Estimation of soil organic matter from red and near-infrared remotely sensed data using soil line Euclidean techniques. *Soil Science Society of America Journal* 66: 1922-1929.

Goel, P.K., S.O. Prasher, J.-A. Landry, R.M. Patel, A.A. Viau and J.R. Miller. 2003. Estimation of crop biophysical parameters through airborne and field hyperspectral remote sensing. *Transaction of the ASAE* 46(4): 1235-1246.

Hummel, J.W., K.A. Sudduth and S.E. Hollinger. 2001. Soil moisture and organic matter prediction of surface and subsurface soils using an NIR soil sensor. *Computers and Electronics in Agriculture* 32: 149-165.

Ingleby, H.R. and T.G. Crowe. 2000. Reflectance models for predicting organic carbon in Saskatchewan soils. *Canadian Agricultural Engineering* 42(2): 57-63.

Ingleby, H.R. and T.G. Crowe. 2001. Neural network models for predicting organic matter content in Saskatchewan soils. *Canadian Biosystems Engineering* 43: 7.1-7.5.

Integral Solutions. 1998. *Clementine Reference Manual*. Southampton, Hampshire, UK: Integral Solutions Limited..

Kaleita, A.L. and L. Tian. 2002. Remote sensing of site-specific soil characteristics for precision farming. ASAE Paper No. 021078. St. Joseph, MI: ASAE.

Krishnan, P., B.J. Butler and J. Hummel. 1981. Close-range sensing of soil organic matter. *Transaction of the ASAE* 24(2): 306-311.

McCarty, G.W., J.B. Reeves IV, V.B. Reeves, R.F. Follett and J. Kimble. 2002. Mid-Infrared and Near-Infrared diffuse reflectance spectroscopy for soil carbon measurement. *Soil Science Society of America Journal* 66: 640-646.

Nash, J.E. and J.V. Sutcliffe. 1970. River flow forecasting through conceptual models: Part I – A discussion of principles. *Journal of Hydrology* 10: 282-290.

Page, N.R. 1974. Estimation of organic matter in Atlantic coastal plain soils with a color-difference meter. *Agronomy Journal* 66: 652-653.

Pitts, M.J., J.W. Hummel and B.J. Butler. 1986. Sensors utilizing light reflection to measure soil organic matter. *Transaction of the ASAE* 29(2): 422-428.

Plant, R.E., D.S. Munk, B.R. Roberts, R.L. Vargas, D.W. Rains, R.L. Travis and R.B. Hutmacher. 2000. Relationship between remotely sensed reflectance data and cotton growth and yield. *Transaction of the ASAE* 43(3): 535-546.

SPSS. 2001. *Clementine Version 6.0 User's Guide.* Chicago, IL: SPSS Inc.

SPSS. 2003. *Clementine 8.0 Algorithms Guide.* Chicago, IL: SPSS Inc.

Sudduth, K.A. and J.W. Hummel. 1991. Evaluation of reflectance methods for soil organic matter sensing. *Transaction of the ASAE* 34(4): 1900-1909.

Sudduth, K.A. and J.W. Hummel. 1996. Geographic operating range evaluation of a NIR soil sensor. *Transaction of the ASAE* 39(5): 1599-1604.

Vinogradov, B.V. 1981. Remote sensing of the hummus content of soils. *Soviet Soil Science* 13(6): 103-113.

Weiss, S.M. and C.A. Kulikowski. 1991. *Computer Systems That Learn: Classification and Prediction Methods from Statistics, Neural Nets, Machine Learning, and Expert Systems.* San Mateo, CA: M. Kaufmann Publishers Inc.