Assessment of reclaimed soils by unsupervised clustering of proximal sensor data

<table>
<thead>
<tr>
<th>Journal:</th>
<th>Canadian Journal of Soil Science</th>
</tr>
</thead>
<tbody>
<tr>
<td>Manuscript ID</td>
<td>CJSS-2018-0057.R1</td>
</tr>
<tr>
<td>Manuscript Type:</td>
<td>Article</td>
</tr>
<tr>
<td>Date Submitted by the Author:</td>
<td>30-Aug-2018</td>
</tr>
<tr>
<td>Complete List of Authors:</td>
<td>Sorenson, Preston; University of Alberta, Renewable Resources&lt;br&gt;Small, Christina; InnoTech Alberta, Ecosystems and Plant Sciences - Reclamation&lt;br&gt;Quideau, Sylvie; University of Alberta, Renewable Resources&lt;br&gt;Underwood, Andrew; InnoTech Alberta, Ecosystems and Plant Sciences - Reclamation&lt;br&gt;Janz, Arnold; Alberta Environment,</td>
</tr>
<tr>
<td>Keywords:</td>
<td>Proximal soil sensing, Soil organic matter, Unsupervised classification</td>
</tr>
<tr>
<td>Is the invited manuscript for consideration in a Special Issue?:</td>
<td>Not applicable (regular submission)</td>
</tr>
</tbody>
</table>
Assessment of reclaimed soils by unsupervised clustering of proximal sensor data

P.T. Sorenson¹, C. Small², S.A. Quideau¹, A. Underwood², A. Janz³

¹ Department of Renewable Resources, 442 Earth Science Building, University of Alberta, Edmonton, AB Canada T6G 2E3
² InnoTech Alberta, 250 Karl Clark Road, Edmonton, AB T6N 1E4
³ Alberta Environment and Parks, 10th Floor 9888 Jasper Avenue, Edmonton Alberta, T5J 5C6
⁴ Corresponding author: Preston Sorenson (preston.sorenson@ualberta.ca)
Abstract

The application of soil proximal sensors on reclaimed sites presents a novel method for assessing the quality of reclaimed landscapes; improving assessment reliability, information management, and environmental assurance. One proximal sensing system that could be used to provide high spatial resolution measurements of soil parameters is an on-the-go optical sensor that collects data at two wavelengths: 660 nm and 940 nm. Proximal soil sensing data were collected at 27 sites, where organic matter, cation exchange capacity and soil water content were collected from 221 soil samples from 0 to 15 cm. The proximal soil sensor data were then automatically clustered using a combination of self-organizing maps and random uniform forests. Overall, the proximal sensor data combined with this data analysis approach created maps with either three or four soil zones. On average, soil zones had statistically significant differences in organic matter, cation exchange capacity and water content. This system could be used to map out zones with significant soil variation as part of reclamation monitoring, and then used to guide laboratory analytical sampling. Future work should focus on development of on-the-go reflectance spectroscopy systems to provide quantitative soil data with high spatial resolution.

Key Words: Proximal soil sensing; soil organic matter; unsupervised classification
Introduction

Evaluation and assessment of disturbed soils is required for ensuring soil quality for both environmental management and reclamation. The energy industry in Alberta is responsible for soil disturbance at over 470,000 wellsites, 39,000 other oil and gas facilities and 24 open pit coal mines (Alberta Energy Regulator 2016). These primary facilities have additional soil disturbances because of associated development activities such as borrow pits, transmission lines and access roads. As per the regulatory requirements in Alberta, all sites and associated facilities must be eventually reclaimed and restored to a state of equivalent land capability (Government of Alberta 2010). Heightened expectations from stakeholder groups and the public, regarding the need for more timely and comprehensive information on the status of provincial reclamation activities, is pressuring the government and industry to pursue new technologies and applications for acquiring information on the status of reclaimed sites.

A potential solution to the need for more timely and spatially comprehensive information is the use of soil proximal sensors. The potential of reflectance spectroscopy to measure key soil parameters such as organic carbon and total nitrogen has been extensively investigated within the broader context of soil analysis (i.e. Ben-Dor and Banin 1995; Chang et al. 2001; Chang and Laird 2002; Rossel and Behrens 2010; Sorenson et al. 2018) and more specifically for reclamation monitoring (Sorenson et al. 2017). Additionally, simpler lower cost two-band reflectance sensors have been explored for potential use in precision agriculture for
mapping changes in soil organic matter contents at high spatial resolutions (Kweon et al. 2013; Piikki et al. 2016).

Determining soil organic carbon is part of wellsite reclamation assessment procedures in Alberta (Alberta Environment 2010). Previous work in Alberta has identified that wellsite construction can be associated with a decrease in soil organic carbon and overall changes in soil properties (Hammermeister et al. 2003). Soil bulk density has been found to be higher on a number of reclaimed wellsites, which in turn affects infiltration and soil water content (Hammermeister et al. 2003). Soil reclamation strategies on oil and gas wellsites often focus on increasing soil porosity and water holding capacity (McConkey et al. 2012). Cation exchange capacity (CEC) is closely related to organic matter content in soil, and both parameters are closely tied to soil fertility (Parfitt et al. 1995). Therefore, all three of these parameters can be affected by construction activities, and the ability to detect changes in these parameters is important for monitoring if a site is moving toward reclamation success.

Soil proximal sensing generates large volumes of data, which necessitates the development of techniques to process this information into a format that facilitates decision making. While building site-specific calibration models with two-band optical reflectance sensors has met some degree of success, universal calibration equations have not been successfully developed so far (Kweon et al. 2013). Rather than quantitative analysis, two-band reflectance data could potentially be used for qualitative measurements using unsupervised machine learning tools. One such technique for unsupervised data analysis and dimensionality
reduction is self-organizing maps (Wehrens and Buydens 2007). Self-organizing maps are similar to multi-dimensional scaling, but rather than trying to preserve relative distance between points, self-organizing maps focus on preserving the topology of the data structure, and concentrate more on mapping similarity rather than dissimilarity (Wehrens and Buydens 2007). Self-organizing maps have been used successfully to solve other challenges associated with complex soil data analysis, such as the assessment of soil biological quality (Mele and Crowley 2008).

The main objective of this research was to determine if proximal sensing data collected with a two-band reflectance sensor with bands in the red and near-infrared could be used to map relative differences in key soil attributes. While the spectral features for organic matter and water are strongest in the short-wave infrared region, some features are present in near infrared region and increases in both properties decrease overall reflectance (Rossel and Behrens 2010). Specifically, this study focused on combining two unsupervised machine learning methods, self-organizing maps and random uniform forests, to classify soil proximal sensing data. The overall goal of this research is to validate automated spatial zoning of soil data to support reclamation assessments.

**Materials and Methods**

**Studied Soils**
Sampling locations included 27 sites (Table 1) at locations ranging from southeast of Calgary, Alberta to south of Lethbridge, Alberta (Figure 1). All sites contained disturbed and undisturbed soils as they consisted of reclaimed oil and gas wellsites, with the center of each assessed area consisting of the wellsite. An area surrounding the wellsite, which included undisturbed agriculture soils, was also included in the assessment area. At each site, 5 to 10 samples were collected for laboratory analyses. In total, 221 samples were collected from the 27 sites. Each site has samples collected from disturbed and undisturbed soils.

**Proximal sensing data**

Proximal soil sensing data were acquired using the Veris Technologies OpticMapper®, which consists of a two-band reflectance sensor along with a GPS unit to collect location and elevation data. Reflectance data were collected by the sensor in two bands, within the red portion of the electromagnetic spectrum from 650 nm to 670 nm and within the near infrared region from 930 nm to 950 nm. A fluted coulter on the optic mapper cuts through crop residues and opens a slot in the soil where measurements take place. All measurements take place on the surface of the exposed soil in the slot. Data were collected in northeast to southwest passes spaced 10 m across the site. Calibration checks for the instrument were performed as per manufacturer specifications. Light and dark reference panels were used to check sensor performance prior to data collection and sensor response was in accordance to manufacturer specifications for each site prior to data collection.

**Laboratory Analyses**
Each of the 221 soil samples collected was analyzed for organic matter content and for CEC; 179 of the samples were additionally analyzed for gravimetric water content. Gravimetric water content was analyzed rather than volumetric water content, as previous reflectance spectroscopy research has demonstrated that gravimetric water content can be more accurately measured than volumetric water content with reflectance-based measurement methods (Ji et al. 2016). Samples were collected the same day the proximal sensor measurements were taken. All soil samples were collected from a depth of 0 to 15 cm to correspond to the range of the optical measurements that were collected, based on the depth of the furrow created by the fluted coulter. Organic matter content was determined by the loss on ignition method described in Nelson and Sommers (1996). CEC was quantified using the methods described in Hendershot et al. (2008), and gravimetric water content was determined using the method described in Topp et al. (2008).

**Statistical Analysis**

**Data Processing**

All data processing was performed using R (R Core Team 2018). The raw reflectance data and elevation point data were converted to rasters using the inverse distance weighting function in the gstat package of R (Pebesma 2004). The reflectance and elevation data were then smoothed using a 3x3 focal median window to reduce noise using the raster package in R (Hijmans 2016). Following smoothing, the ratio of near infrared to red reflectance was
calculated. Slope and topographic position index values were calculated from the elevation data using the terrain function in the raster package of R (Hijmans 2016).

**Model Development**

For each site, a raster stack was created with a raster for each of the following parameters: red reflectance, near infrared reflectance, ratio of red to near infrared reflectance, elevation, slope, and topographic position index. The raster stack was first processed using a self-organizing map with the kohonen package in R (Wehrens and Buydens 2007). A 10 by 10 hexagonal grid topology was specified for the self organizing map. Following initial clustering of the data into the self organizing map grid, final clustering into a smaller number of clusters was performed using an unsupervised random uniform forest in R (Ciss 2015). The advantage of this analysis is that it can be used for unsupervised clustering of data. Additionally, random uniform forests will identify the optimal number of clusters for the data without user specification of cluster numbers.

Random uniform forests uses a multiple step approach to perform the unsupervised clustering. The random uniform forest first grows a forest of decision trees using random subsampling and random cut points according to a continuous uniform distribution, followed by multidimensional scaling and clustering with either k-means or hierarchical clustering (Ciss 2015). In this study the hierarchical clustering step was used, and the optimal number of clusters was automatically selected based on where the maximum lagged difference in cluster heights occurred. The unsupervised clustering using random forests was performed on the self-
organizing map grid cell codes, leading to a final cluster number for each grid cell and each
associated raster cell in each grid cell. An example of the clustering results from a site are
illustrated in Figure 2.

Model Evaluation

Model performance was evaluated on the unsupervised classification results using a
linear mixed effects model with the nlme package in R (Pinheiro et al. 2016). For each site, the
cluster that each laboratory analytical sample was collected from was determined. The clusters
were then relabelled from lowest to highest to correspond to the highest to lowest organic
matter concentrations, CEC or gravimetric water content as the specific number sequence is
randomized during the classification process. This step was necessary to allow for the
comparison of relative effect differences among clusters across all sites. Organic matter, CEC
and gravimetric water content data were then scaled by subtracting the mean and dividing by
standard deviation. This transformation permitted comparison between sites regardless of the
differences in magnitude across sites. Mean centered data were then compiled and analyzed
using a linear mixed effects model with site as random factor. The data were normally
distributed and showed homogeneity of variance. Linear mixed effects models were run for
organic matter, CEC, and gravimetric water content to test if significant differences were
present among clusters.
Results and Discussion

The clustering exercise yielded three clusters for most sites (Table 2), with 26 of 27 sites including three clusters and only one site clustering into four clusters. The average cluster size was 2.40 ha compared to an average site size of 14.40 ha (Table 1). Soil organic matter values ranged from 2.28 to 8.27 percent with an average value of 4.54 percent. Overall, there was a significant difference among clusters in terms of their organic matter content (Overall F-value = 22.33, p value of <0.01). Additionally, each cluster had significant differences in organic matter content relative to all other clusters (Table 3). The greatest differences in organic matter content was between clusters one and two, with cluster three only having slightly more organic matter compared to cluster two (Figure 3a). On average, organic matter contents in cluster one were 0.52 standard deviations below the mean, while cluster two included organic matter contents 0.15 standard deviations above the mean and cluster three included organic matter contents 0.46 standard deviations also above the mean. For the one site with four clusters (cluster four) organic matter contents were 1.13 standard deviations above the mean.

The CEC ranged from 1.24 to 71.76 meq/100g with an average value of 27.90 meq/100g across all sites (Table 1). The average gravimetric water content was 18 percent with values ranging from 8 to 29 percent. There was a significant difference among clusters in terms of CEC (F-value = 7.29 p-value<0.01) and gravimetric water content (F-value = 28.60, p-value<0.01). There were significant differences amongst all clusters for both parameters, with the exception of cluster four for CEC, which was not significantly different from cluster three (Table 3).
average cluster one had CEC values 0.31 standard deviations below the mean, CEC values for cluster two were 0.16 standard deviations above the mean and for cluster three 0.57 standard deviations above the mean (Figure 3b). For gravimetric water content, cluster one had average values 0.04 standard deviations below the mean, cluster two had average values at the mean, and cluster three had average values 0.25 standard deviations above the mean (Figure 3c).

Previous work investigating the same sensors as the ones used in this study had been successful in building site-specific calibration models (Kweon et al. 2013; Piikki et al. 2016). However, calibration models that generalized to new areas could not be built in either of these two studies. In this present study, site-specific calibration models could not be successfully built for any of the analyzed soil parameters. Specifically, Piikki et al. (2016) used the optic mapper sensor with portable x-ray fluorescence and electromagnetic induction to build site-specific calibrations for total carbon. However, predictions for new sites were poor for total carbon, and soil texture could not be modelled with any accuracy. Kweon et al. (2013) concluded that field specific calibrations are possible and sufficiently accurate based on leave-one-out cross validation results. However, they could not develop a universal calibration model. While a quantitative model could not be developed in our study either, our results indicated that an unsupervised classification approach can be used to successfully classify a reclaimed area into zones with different organic matter, CEC and moisture contents based on soil reflectance data from two distinct bands along with elevation data. A possible explanation for why a quantitative model could not be built for this study is that 30-40 samples were collected from
each site in Piikki et al. (2016), and with more data collected from each site a site-specific calibration model may have been possible with this study. Additionally, compared to the Kweon et al. (2013) study, the average standard deviation in organic matter was higher in this study, likely due to the presence of disturbed soils. More calibration samples are needed to build an accurate calibration model for these sites.

It is important to note that while results from the current study demonstrate that proximal sensor data can be used to automatically zone soil with different parameters, the methodological approach used does not provide any information on the direction or magnitude of difference among classes. Alternative sensor arrangements have been used by other researchers to quantitively measure soil organic carbon and other soil parameters under field conditions. Viscarra Rossel et al. (2017) successfully used reflectance spectroscopy and gamma ray attenuation to measure soil organic stocks from soil cores collected in the field. Reflectance spectroscopy has been combined with x-ray fluorescence to successfully take multi-parameter measurements of soil (Duda et al. 2017). Additionally, field reflectance spectroscopy has been used under similar site conditions to successfully measure soil organic carbon and nitrogen, but not soil pH (Sorenson et al. 2017).

Construction activities have been documented to lower soil organic matter content and increase bulk density (Hammermeister et al. 2003). These changes in turn influence both the porosity of the soil along with the CEC, which is closely linked to organic matter content (Parfitt et al. 1995). Overall, these results show that two band reflectance measurements, along with
elevation data, have a role to play for monitoring soil reclamation when combined with the appropriate data analysis techniques. While quantitative models could not be successfully built using the sensors in this study, automated clustering of the data allowed the identification of zones with different organic matter, CEC and water contents (Table 3). In practice, follow-up investigations would be required for quantitative determination of soil organic matter, CEC and water content for each soil zone. This analysis could consist of either collecting samples for conventional laboratory analysis, or by collecting further point data using a higher spectral resolution reflectance spectroscopy system (Sorenson et al. 2017).

Ultimately, while there is value in qualitatively delineating reclaimed soils into zones to identify variance in key soil characteristics, quantitative measurements would be an improvement. Rather than using two reflectance bands, high resolution visible light near infrared reflectance spectroscopy could be utilized to provide quantitative measurements. Soil organic carbon, CEC and water content have all been shown to be successfully measurable using reflectance spectroscopy (Soriano-Disla et al. 2014). Point spectroscopy has been used on some of these sites to measure soil carbon quantitatively in-field as well (Sorenson et al. 2017).

**Conclusion**

The results of this study indicate that broad-band reflectance data combined with unsupervised classification approaches can be used to qualitatively and automatically map a field into soil zones with differences in organic matter, CEC and water contents. However, more site-specific data than were collected in this study or alternative types sensors are needed to
quantitatively map soil organic matter content. Future work should focus on integrating reflectance spectroscopy sensors, rather than on two-band reflectance sensors, into a hardware configuration similar to the Optic Mapper® to provide high spatial resolution quantitative measurements of soil parameters for reclamation monitoring.

Acknowledgements

We would like to thank Alberta Environment and Parks (formally the Alberta Environmental Monitoring and Evaluation Agency (AEMERA)) and InnoTech Alberta for providing financial assistance, soil samples and soil proximal soil data for the project. We would also like to acknowledge MITACS for providing financial support for this project with a MITACS Accelerate to S.A. Quideau (no IT08009).
References


Tables

Table 1. Soil and Site Characteristics. The first number in each column indicates the mean for a given parameter. The second numbers are the range of values observed across all sites.

Table 2. Cluster numbers at each site and average cluster sizes

Table 3. Linear Mixed Effects Model Results for examining if significant differences exist between Organic Matter Content, Cation Exchange Capacity and Gravimetric Water Content across clusters.
Table 1. Mean values of soil and site characteristics, with the range of values observed across all sites indicated in parentheses.

<table>
<thead>
<tr>
<th>Sites</th>
<th>Organic matter content (%)</th>
<th>Cation exchange capacity (meq/100g)</th>
<th>Gravimetric water content (%)</th>
<th>Area (ha)</th>
<th>Slope (%)</th>
<th>Topographic Position Index</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>4.54 (2.28 – 8.27)</td>
<td>27.9 (1.24 – 71.76)</td>
<td>18 (8 – 29)</td>
<td>14.4 (7.8 – 27)</td>
<td>2 (0 – 8)</td>
<td>0 (0.67 – 0.68)</td>
</tr>
<tr>
<td>27</td>
<td>8.27)</td>
<td>71.76)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 2. Cluster numbers at each site and average cluster sizes

<table>
<thead>
<tr>
<th>Number of clusters</th>
<th>Number of sites</th>
<th>Average cluster size (ha)</th>
<th>Minimum cluster size (ha)</th>
<th>Maximum cluster size (ha)</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>26</td>
<td>2.4</td>
<td>0.02</td>
<td>9.4</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>1.5</td>
<td>0.44</td>
<td>2.7</td>
</tr>
</tbody>
</table>


Table 3. Results from the linear mixed effects model examining if significant differences exist between organic matter content, cation exchange capacity and gravimetric water content across clusters.

<table>
<thead>
<tr>
<th>Cluster</th>
<th>Parameter</th>
<th>t-Value</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>estimate</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Organic matter</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intercept</td>
<td>-0.52</td>
<td>-5.76</td>
<td>&lt;0.01</td>
</tr>
<tr>
<td>2</td>
<td>0.68</td>
<td>5.67</td>
<td>&lt;0.01</td>
</tr>
<tr>
<td>3</td>
<td>1.07</td>
<td>6.53</td>
<td>&lt;0.01</td>
</tr>
<tr>
<td>4</td>
<td>1.95</td>
<td>4.46</td>
<td>&lt;0.01</td>
</tr>
<tr>
<td>Cation exchange capacity</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intercept</td>
<td>-0.31</td>
<td>-2.87</td>
<td>&lt;0.01</td>
</tr>
<tr>
<td>2</td>
<td>0.47</td>
<td>2.99</td>
<td>&lt;0.01</td>
</tr>
<tr>
<td>3</td>
<td>0.88</td>
<td>4.22</td>
<td>&lt;0.01</td>
</tr>
<tr>
<td>4</td>
<td>1.01</td>
<td>1.86</td>
<td>0.08</td>
</tr>
<tr>
<td>Gravimetric water content</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intercept</td>
<td>-0.44</td>
<td>-5.76</td>
<td>&lt;0.01</td>
</tr>
<tr>
<td>2</td>
<td>0.62</td>
<td>6.28</td>
<td>&lt;0.01</td>
</tr>
<tr>
<td>3</td>
<td>1.08</td>
<td>8.00</td>
<td>&lt;0.01</td>
</tr>
<tr>
<td>4</td>
<td>1.49</td>
<td>4.12</td>
<td>&lt;0.01</td>
</tr>
</tbody>
</table>
Figures

Figure 1. Location of study sites sampled using the Veris® Optic Mapper in Alberta, Canada. Research site locations are indicated with the red triangles. Source: © 2018 Google Imagery.

Figure 2. Results from the self organizing map and random uniform forest. The map on the left illustrates the results from the self organizing map. The image on the right illustrates the results of the random uniform forest clustering of the self organizing map results. The black polygon within each image illustrates the area disturbed by construction activities.

Figure 3a Average organic matter content for each unsupervised cluster across all study sites. Organic matter contents have been scaled and centered around the mean for each site to allow comparisons across sites with different magnitudes of organic matter contents. No distribution can be provided for class 4 as only one site had four classes after the unsupervised cluster analysis.

Figure 3b Average cation exchange capacity (CEC) for each unsupervised cluster across all study sites. CEC values have been scaled and centered around the mean for each site to allow comparisons across sites with different magnitudes of CEC values. All clusters have significantly different CEC values except for clusters three and four. No distribution can be provided for class 4 as only one site had four classes after the unsupervised cluster analysis.

Figure 3c Average moisture content for each unsupervised cluster across all study sites. Moisture contents have been scaled and centered around the mean for each site to allow
comparisons across sites with different magnitudes of moisture contents. No distribution can be provided for class 4 as only one site had four classes after the unsupervised cluster analysis.
Figure 1. Location of study sites sampled using the Veris® Optic Mapper in Alberta, Canada. Research site locations are indicated with the red triangles. Source: © 2018 Google Imagery.
Results from the self organizing map and random uniform forest. The map on the left illustrates the results from the self organizing map. The image on the right illustrates the results of the random uniform forest clustering of the self organizing map results. The black polygon within each image illustrates the approximated area disturbed by construction activities.
Figure 3a  Average organic matter content for each unsupervised cluster across all study sites. Organic matter contents have been scaled and centered around the mean for each site to allow comparisons across sites with different magnitudes of organic matter contents. No distribution can be provided for class 4 as only one site had four classes after the unsupervised cluster analysis. The circles indicate values that are 1.5 times the interquartile range above the third quartile or below the first quartile.
Figure 3b  Average cation exchange capacity (CEC) for each unsupervised cluster across all study sites. CEC values have been scaled and centered around the mean for each site to allow comparisons across sites with different magnitudes of CEC values. All clusters have significantly different CEC values except for clusters three and four. No distribution can be provided for class 4 as only one site had four classes after the unsupervised cluster analysis.
Figure 3c  Average moisture content for each unsupervised cluster across all study sites.

Moisture contents have been scaled and centered around the mean for each site to allow comparisons across sites with different magnitudes of moisture contents. No distribution can be provided for class 4 as only one site had four classes after the unsupervised cluster analysis.