THE INTERNATIONAL SOCIETY OF
PRECISION AGRICULTURE PRESENTS THE
13th INTERNATIONAL CONFERENCE ON
PRECISION AGRICULTURE

July 31-August 4, 2016 • St. Louis, Missouri USA

Integrated Analysis of Multilayer Proximal Soil Sensing Data

N. Dhawale[1], V. Adamchuk[1], H. Huang[1], W. Ji[1], S. Lauzon[1], A. Biswas[2], P. Dutilleul[3]

[1]Bioresource Engineering Dept., McGill University (Ste-Anne-de-Bellevue, QC, H9X 3V9, Canada)

[2]Natural Resource Science Dept., McGill University (Ste-Anne-de-Bellevue, QC, H9X 3V9, Canada)

[3]Plant Science Dept. McGill University (Ste-Anne-de-Bellevue, QC, H9X 3V9, Canada)

A paper from the Proceedings of the
13th International Conference on Precision Agriculture
July 31 – August 4, 2016
St. Louis, Missouri, USA

Abstract. *Data revealing spatial soil heterogeneity can be obtained in an economically feasible manner using on-the-go proximal soil sensing (PSS) platforms. Gathered georeferenced measurements demonstrate changes related to physical and chemical soil attributes across an agricultural field. However, since many PSS measurements are affected by multiple soil properties to different degrees, it is important to assess soil heterogeneity using a multilayer approach. Thus, analysis of multiple layers of geospatial data leads to: 1) delineation of relatively heterogeneous field areas characterized by a particular combination of individual sensor measurements, and 2) identification of field locations representing these different combinations to be used for traditional soil sampling and analysis required for site-specific sensor calibration. The objective of this research was to develop an algorithm that would accomplish both functions. It was expected that delineated field areas would be spatially contiguous with relatively low variance for each sensor measurement. The algorithm was based on the adapted stepwise grouping method using a neighborhood search analysis (NSA). In addition, a circular area search method was implemented to define field locations that best represent each delineated field partition. The algorithm was evaluated using PSS data of varying quality from over 20 agricultural fields from Eastern Canada. To demonstrate its performance for this conference paper, field elevation, apparent soil electrical conductivity and soil pH maps from two experimental sites were used. D-optimality criterion was applied to individual sensor values corresponding to the set of selected representative field locations to evaluate the quality of these selections.*

Keywords. *proximal soil sensing, data clustering, sensor fusion, geospatial data management*

Introduction

Conventional systematic soil sampling and analysis is a laborious and time consuming practice. Therefore, proximal soil sensing (PSS) has been used to increase the quality of thematic soil maps while minimizing the number of required traditional soil samples (Viscarra Rossel et al., 2011). To pursue site-specific crop management, layers of geospatial data are frequently aggregated into groups (clusters or zones) to represent significantly different growing conditions (Fraisse et al., 2001; Ping and Dobermann, 2003). Treating these field partitions according to local needs can significantly improve profitability and reduce the environmental footprint of crop production. Therefore, spatial data clustering applied to agricultural landscapes is an important process (Li and Wang, 2010). Similar techniques are widely used in remote sensing (Deng et. al. 2003), neuroanatomy analysis (Prodanov et. al, 2007), and other areas of application. Several different spatial clustering algorithms have been applied to PSS measurements as well. For example, Management Zone Analyst (MZA) described by Fridgen et al. (2004) is a publicly available tool closely related to the popular k-means clustering algorithm. Since MZA clusters depend on the selection of initial centroids, results are not repeatable. Thus, this method requires cross-validation to select the best data grouping among several runs (Abdul-Nazeer and Sebastian, 2009). Although the method allows multidimensional data analysis, it often produces spatially discontinuous zones, and also the number of zones is determined subjectively (Kerby et al., 2007; Shatar and McBratney, 2001).

To achieve spatial continuity of formed clusters, grouping together non-adjacent data can be restricted. The so-called neighborhood search analysis (NSA) has been developed to seek the emergence of data groupings with relatively similar measurements within a group, and the greatest possible difference between the group average values of these measurements. Dhawale et al. (2014) successfully tested this approach using a single PSS data layer. However, it is a challenge to bring multiple data layers together because of the need for the weighting of different data layers. Therefore, the objective of this research was to develop a robust algorithm that could handle multilayer PSS data and produce an unspecified number of spatially contiguous field partitions while relying solely on the information embedded within the specified PSS dataset.

Materials and Methods

Neighborhood Search Analysis Algorithm

PSS data is usually collected following a series of parallel passes at a fixed width between them, which results in different distances between measurements in the direction of travel and perpendicular to the direction of travel. In many instances, site-specific management decisions are made based on the spatial scale comparable with the width of PSS passes (typically between 10 and 20 m). As a result, the dimension of each individual data element containing values representing each data layer may be similar to the width of sensor passes. In this algorithm, all valid (filtered) PSS measurements were first projected into linear units from geographic coordinates. Then, a series of square grids covering the entire field area was produced. PSS measurements within each individual grid were averaged and assigned to the location of its center. A median 5 filter was used to determine values for occasional grid cells within which no PSS measurements were present. This technique was also used to remove potential outliers. After this step, every data element (called grid cell throughout this paper) had a maximum of eight immediate neighbors with only one value representing each involved data layer.

To assure spatial continuity of data groupings, a stepwise, or hierarchical, clustering method was implemented in this research. Such data grouping concept can be initiated by either assuming that each data element is its own group and then merge neighboring groups according to their similarities (growing the groups), or start by assuming that all of the data initially belong to one large group and then partition out groups of data elements that are different from their surrounding groups (splitting the groups). The algorithm presented in this paper was based on the second strategy. It seeks the delineation of field areas representing conditions different from what could be considered "the rest of the field" and does not imply that every data element has to be assigned to a newly formed group. Certain field locations might be similar to the average field conditions, which does not signify the need for further explorations and/or site-specific treatment.

Another important assumption was that the number of field areas that could be delineated was unknown prior to the analysis. Input data layers were selected by users, but no priority was given to any of them. Nevertheless, a data layer with a strong spatial structure (relatively low data variance at a short separation distance) naturally provides a more reliable characterization of potential data groupings than a data layer that varies from one measurement to its neighbor.

To construct an objective function to be optimized through the data grouping process, mean squared error (MSE) was calculated for each individual data layer *k* according to:

$$MSE_k = \frac{\sum_{j=1}^{m}\sum_{i=1}^{n_j}\left(X_{ij} - \overline{X}_j\right)^2}{N - m} \tag{1}$$

where X_{ij} is a sensor value for the i_{th} grid cells within the j_{th} group; $\overline{X}_j$ is the mean of j_{th} group; N is the total number of grid cells; m is the number of groups; n_j is the number of grid cells within the j_{th} group. It should be noted that:

$$N - m = \sum_{j=1}^{m}\left(n_j - 1\right) \tag{2}$$

Since the algorithm initially assumes that all data elements belong to the same group number 1, named "the rest of the field". $MSE_k(m=1)$ represents the variance of k_{th} data layer across the entire field. Considering that the area of the field is substantially greater than the area of a grid cell, $MSE_k(m=1)$ can be called Farthest Distance Variance (FDV_k). In such situation, the portion of data variance accounted for by distributing N grid cells among m groups can be calculated according to:

$$R_k^2 = 1 - \frac{MSE_k}{FDV_k} \tag{3}$$

The maximum value of R^2_k can be obtained when MSE_k is as small as possible and it is approaching 1 when the number of groups increases. Since the result can be considered less favorable if at least one data layer *k* is not adequately accounted for, it is reasonable to employ the integration operator OR instead of more common AND. This excludes the need to assign weight factor to each individual data layer when adding corresponding MSE_k estimates. In mathematical term, this would mean that the product of all R^2_k should be maximized. Therefore, the objective function (*OF*) was defined as:

$$OF = \prod_{k=1}^{K} R_k^2 \tag{4}$$

where *K* is the number of PSS data layers.

In this research, the smallest number of data elements that could be grouped was assigned to be nine (3 x 3) grid cells square window. Therefore, the maximum accountable variance is the variance of PSS measurements between immediate neighbors. This, so-called Shortest Distance Variances (SDV_k) can be found using:

$$SDV_k = \frac{1}{w}\sum_{j=1}^{w}\sum_{i=1}^{9}\frac{\left(X_{ij} - \overline{X}_j\right)^2}{8} \tag{5}$$

where *w* is the total number of 3x3 square windows of grid cells.

Since SDV_k represents the smallest MSE_k value, the maximum value of R^2_k is calculated as:

$$R^2_{k\,\max} = 1 - \frac{SDV_k}{FDV_k} \tag{6}$$

This $R^2_{k\,max}$ parameter can range between 0 and 1. It is equal to 0 when data layer k is either uniform, or highly variable so that $SDV_k = FDV_k$. In such a case, the data layer should not be able to affect changes in the OF. Alternatively, when $R^2_{k\,max}$ is close to 1, the data layer has a strong spatial structure ($SDV_k << FDV_k$) and OF must be sensitive to the change of MSE_k corresponding to that particular data layer. In mathematical terms, this goal can be achieved by multiplying all R^2_k values raised to $R^2_{k\,max}$ power:

$$OF = \prod_{k=1}^{K} {R^2_k}^{R^2_{k\,\max}} = \prod_{k=1}^{K} \left(1 - \frac{MSE_k}{FDV_k}\right)^{\left(1-\frac{SDV_k}{FDV_k}\right)} \tag{7}$$

The resultant OF indicates the overall quality of grid cell groupings. It varies from 0 to 1 and approaches high values when every spatially structured layer of PSS measurements is separated among spatially continuous groups of grid cells with the minimum internal group variance. Such groups represent different combinations of average PSS measurements obtained with different sensors that diverge from average field conditions.

To facilitate the formation of grid cell groups that would maximize the OF, the NSA algorithm was implemented using MATLAB R2015b (The MathWorks, Inc., Natick, Massachusetts, USA) following the flowchart illustrated in Figure 1. Figure 2 illustrates the initial version of the graphical user interface (GUI).

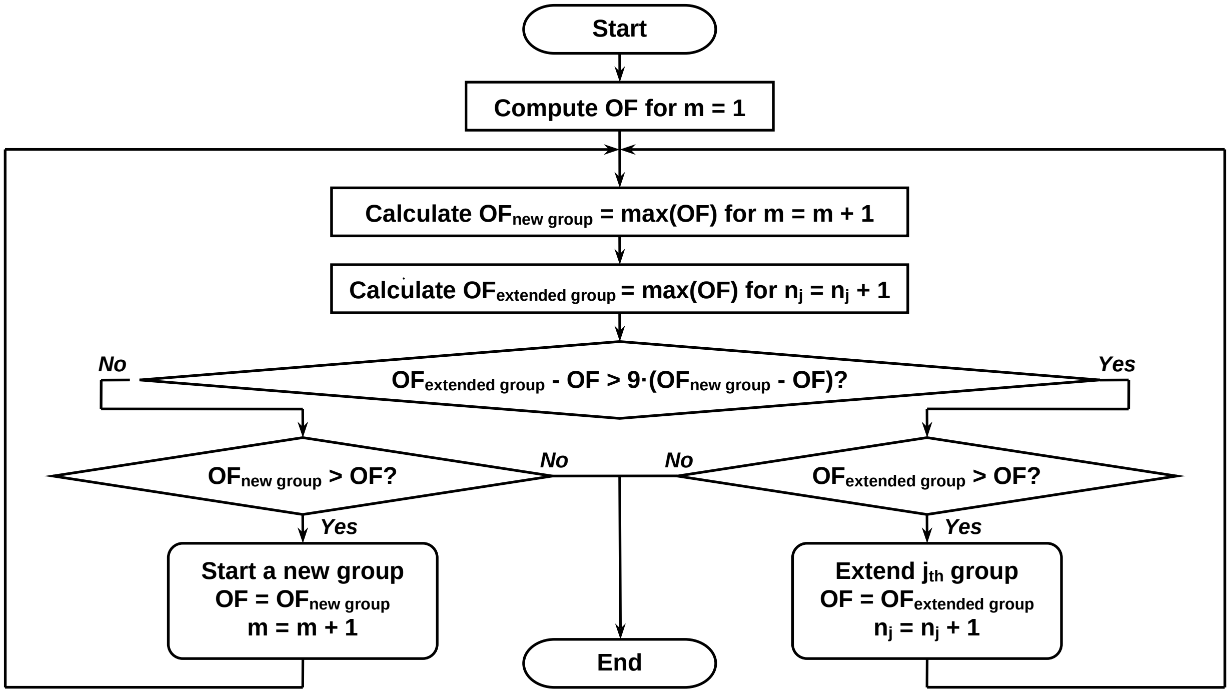
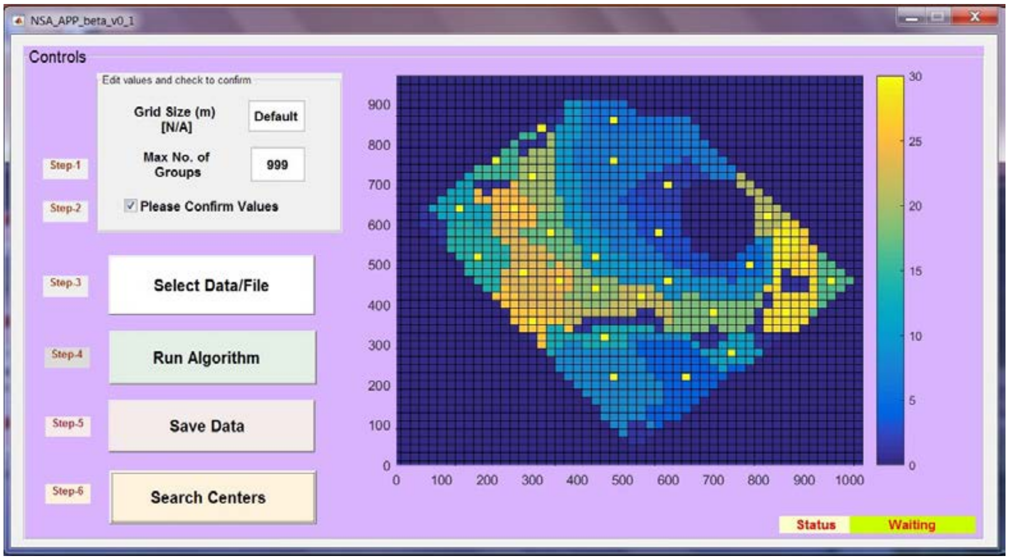

Start
Compute OF for m = 1
Calculate OF$_{new\ group}$ = max(OF) for m = m + 1
Calculate OF$_{extended\ group}$ = max(OF) for $n_j = n_j + 1$
No
Yes
OF$_{extended\ group}$ - OF > 9·(OF$_{new\ group}$ - OF)?
OF$_{new\ group}$ > OF?
No
No
OF$_{extended\ group}$ > OF?
Yes
Yes
Start a new group
OF = OF$_{new\ group}$
m = m + 1
End
Extend j_{th} group
OF = OF$_{extended\ group}$
$n_j = n_j + 1$

Figure 1. NSA algorithm flowchart.



NSA_APP_beta_v0_1

Controls

Edit values and check to confirm

Grid Size (m) [N/A]: Default

Max No. of Groups: 999

☑ Please Confirm Values

Status: Waiting

Figure 2. NSA application GUI.

The interactive process of grid cell grouping starts with the assumption that all grid cells belong to the group labelled “1” designated as “the rest of the field”. Grid cells can be grouped together only when they are adjacent. This assumption is typically referred to as “rook’s rule”. Only nine neighboring grid cells in a 3x3 configuration can form a new group. The beginning of a new group as well as the merger of a new grid cell to an existing group is accepted when the resultant value of OF is greater than its current value. Since the formation of a new group starts with nine grid cells, the OF increase must be nine times greater than the highest OF increase when an existing group is extended. The process continues until any further increase in OF is not feasible, or the total number of groups reaches the maximum that optionally can be defined by the user.

Circular Area Search Method

Once the grid cell grouping is finalized, it is desirable to define one representative location per group. These locations can be used for soil sampling, profiling, temporal monitoring (through wireless sensor networks) and other practices necessary for the model-based representation of spatio-temporal landscape behavior. The NSA produces groups of grid cells covering significant field areas (at least the size of a 3 x 3 grid cells square) that may emerge in any part of the field and cover every combination of PSS measurements that may characterize these specific areas.

According to Adamchuk et al. (2011), an additional requirement for representative locations is relative soil homogeneity around these locations. This way, it could be safe to assume that the soil sample or wireless sensor measurement actually represents the same soil conditions as the PSS measurement and the effect of short-distance variability is minimized. Naturally, such a location could be expected in the middle of an established group of grid cells. However, the shape of the groups is not well defined and frequently, its centroid may be found outside of the group. To avoid this difficulty, a circular area search method was implemented.

Figure 3 illustrates the principle of defining representative locations for a group of square grid cells as the location around which the largest circular area can be established that is entirely inside the group's boundary. In other words, this is the location that has the largest number of nearby grid cells belonging to the same group.

Algorithm Performance Illustration

Although the algorithm has been evaluated using more than 20 different fields across Eastern Canada with different types of PSS measurements, two common agricultural fields from Eastern Ontario were selected to illustrate performance of the NSA algorithm in this paper. Both fields were mapped using the Veris® Mobile Sensor Platform (MSP, Veris Technologies, Inc., Salina, Kansas, USA) equipped with a real-time kinematic (RTK) global navigation satellite system (GNSS) receiver.

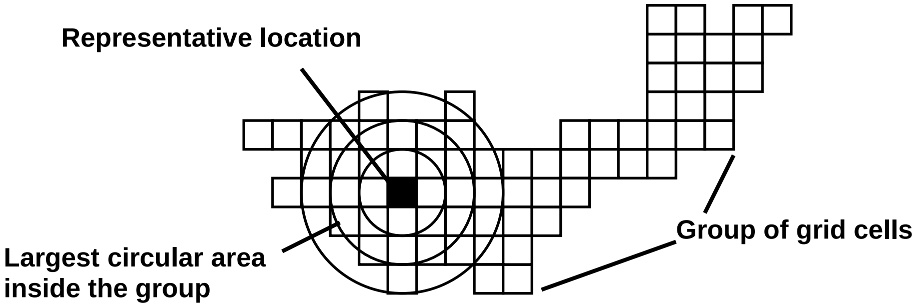


Figure 3. Illustration of the circular area search method.

Figure 4 illustrates both the 39-ha ST and 48-ha NX sites along with the dotted path of the instrument. For this paper, only field elevation, shallow apparent soil electrical conductivity (EC_a) and soil pH measurement were used. Both the elevation and EC_a data were collected at 1 Hz logging interval resulting in a relatively high density of measurements along the travel path. Soil pH measurements were recorded at different logging frequency (typically between 10 and 20 s per measurement) depending on the stability of Antimony ion-selective electrodes. In both cases, approximately 15 m intervals between consecutive travel passes was maintained during field mapping and, therefore, 20-m grid cell averaging for median 5 filtering has been applied as the pre-processing step within the NSA application.

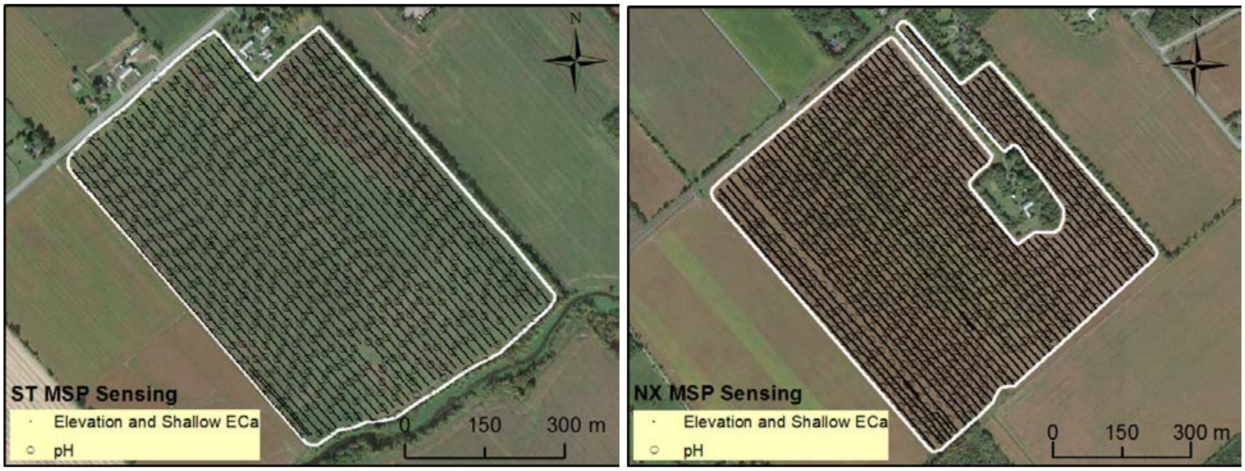


Figure 4. Two experimental sites used to demonstrate the performance of the NSA algorithm.

Although there is no direct evaluation of the quality of grid cell grouping as the NSA was developed in response to the number of algorithm design criteria that emerged from the practical use of the PSS instruments, it is expected that representative locations would represent the entire range of every PSS measurement data layer that has a strong spatial structure. Traditionally, the quality of such a distribution is evaluated using D-optimality criteria that seek to minimize the determinant of the inverse of information matrix X'X. In the case of a linear model that would mean an evenly spread distribution of measurements along the entire range. This paper compares D-optimality estimated for the set of representative locations selected using the NSA algorithm against an equally numbered random selection of grid cells from each experimental site.

Results and Discussion

Figures 5 and 6 illustrate the results produced by the NSA algorithm and circular area search method resulting from the three layers of data representing PSS measurements at both sites.

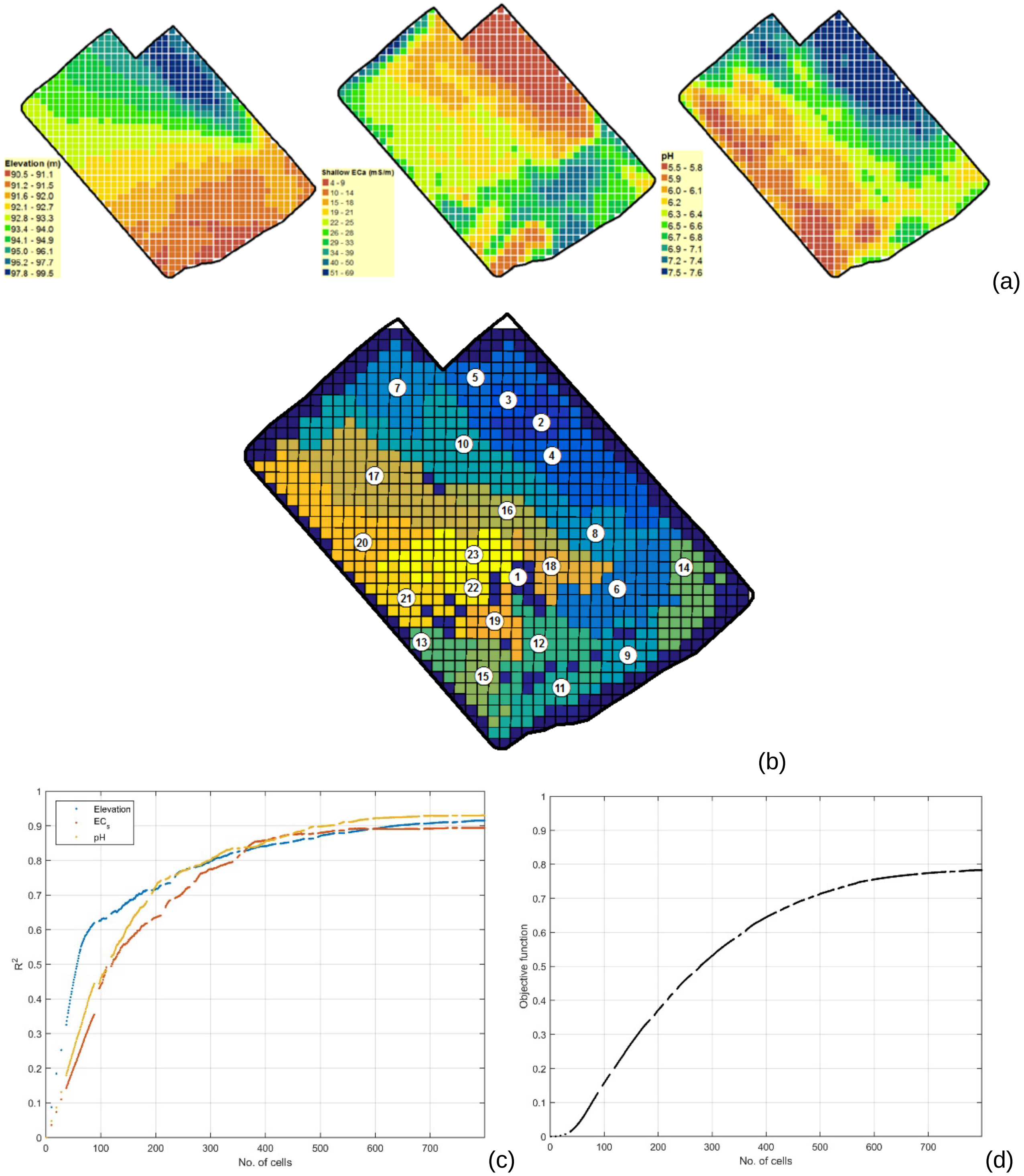


Figure 5. ST field analysis results, including: (a) the three input data layers (field elevation, shallow EC$_a$ and soil pH), (b) final groupings with labeled representative locations, (c) R^2_k for each data layer *k*, and (d) the overall OF versus the number of grid cells classified at each algorithm cycle.

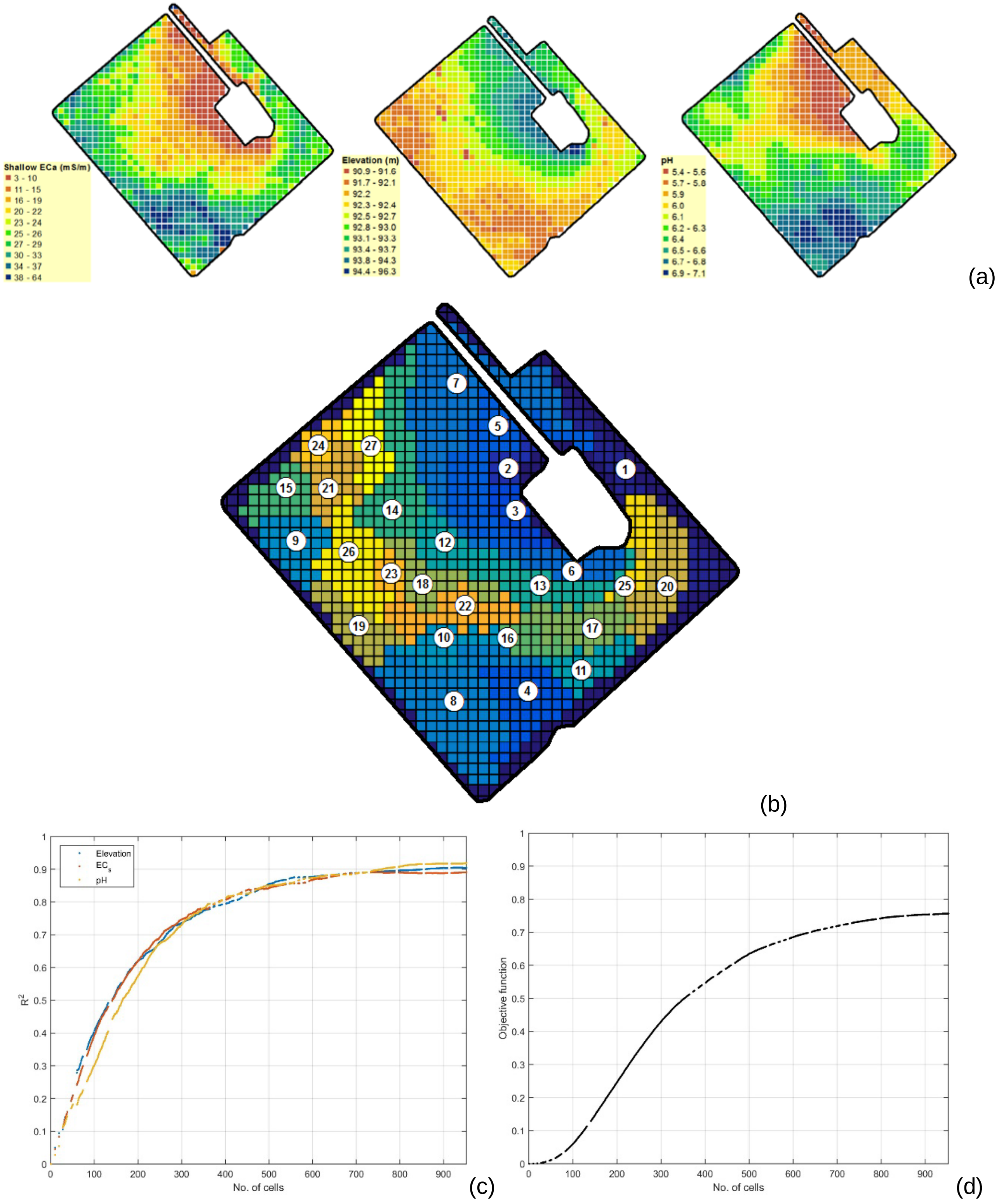


Figure 6. NX field analysis results, including: (a) the three input data layers (field elevation, shallow EC$_a$ and soil pH), (b) final groupings with labeled representative locations, (c) R^2_k for each data layer *k*, and (d) the overall OF versus the number of grid cells classified at each algorithm cycle.

As can be seen from both maps of grid cell groupings, representative locations starting with number "2" in consecutive order, indicate the greatest anomaly field conditions identified with the three poorly correlated PSS data layers. Establishment of the first ten or so groups is responsible for a more than 50% increase in individual R^2_k as well as the overall OF. In fact, after assigning only about 200 grid cells (among about 1000 per field) R^2_k for each data layer exceeds 0.7, which means that over 70% of field variance is accounted for by this grouping. Establishment of groups with higher ordinary numbering may be viewed unnecessary as the parts of the field fragmented at that stage can be considered relatively uniform. However, if OF continues to grow, some potential applications, such as stratified space-based field sampling, may benefit from these additional groupings.

Table 1 illustrates the summary of D-optimality estimates for 23 (ST site) and 27 (NX site) locations that were selected randomly (5000 times), or according to the NSA algorithm. It has been shown that in both cases, estimated D-optimality performed on locations selected using NSA and the circular area search method were ranked among the top third of possible random combinations.

Table 1. Comparison of D-optimality criteria for the two experimental sites.

Data layers	NX field (27 locations)			ST field (23 locations)		
	Elevation	Shallow EC_a	pH	Elevation	Shallow EC_a	pH
	D-optimality = $\lvert(X'X)^{-1}\rvert$					
NSA (max OF)	0.003008	$1.631 \cdot 10^{-5}$	0.005292	0.004048	$2.211 \cdot 10^{-5}$	0.011186
Random (max)	0.004108	$1.246 \cdot 10^{-4}$	0.040810	0.024190	$8.890 \cdot 10^{-5}$	0.325400
Random (median)	0.000480	$2.224 \cdot 10^{-5}$	0.006441	0.005007	$2.432 \cdot 10^{-5}$	0.015570
Random (min)	0.000188	$6.458 \cdot 10^{-5}$	0.003617	0.001372	$8.944 \cdot 10^{-6}$	0.006450
	Ranking (from 0 to 100) with respect to 5000 random selections					
NSA ranking	7	20	19	25	35	28

Conclusions

The spatial clustering algorithm developed in this study is based on a neighborhood search analysis method and seeks to minimize variance inside each group of interpolated grid pixels corresponding to an unlimited number of sensor-based data layers. The circular area search method was implemented to define representative locations within each delineated group of grid cells. Preliminary testing of the algorithm was illustrated using field elevation, shallow EC_a and soil pH PSS data layers obtained from two agricultural fields. The algorithm produced robust results revealing diverse growing conditions. D-optimality criteria applied to the representative locations defined within each group were highly ranked as compared to the random selection of the same number of grid cells. Further testing of this algorithm will involve different types of data, preprocessing operations and post-processing interpretation as well as a comparison with more traditional spatial clustering algorithms previously applied to PSS data.

Acknowledgements

This research was supported in part by funds provided through the National Science and Engineering Research of Canada (NSERC) Discovery Grant and through Ontario Ministry of Agriculture, Food and Rural Affairs (OMAFRA) New Directions Research Program (NDRP). Special thanks to Paul Hermans of DuPont Pioneer (Richmond, Ontario, Canada) for providing Veris MSP data and to the cooperating farmers Jeremy Nixon of J & H Nixon Farms and Michael Schouten of Schouten Dairy Farms.

References

Adamchuk, V.I., Viscarra Rossel, R.A., Marx, D.B., and Samal, A.K. 2011. Using targeted sampling to process multivariate soil sensing data. Geoderma, **163**(1-2), 63-73.

Abdul-Nazeer, K.A. and Sebastian, M.P. 2009. Improving the accuracy and efficiency of the k-means clustering algorithm. In: Proceedings of the World Congress on Engineering, **23**(1), 1-3.

Deng, X., Wang, Y., and Peng, H. 2003. The clustering of high resolution remote sensing imagery. Journal of Electronics and Information Technology, **25**(8), 1073-1080.

Dhawale, N.M., Adamchuk, V.I., Prasher, S.O., Dutilleul, P.R.L. and Ferguson, R.B. 2014. Spatially constrained geospatial data clustering for multilayer sensor-based measurements. In: Proceedings of Joint International Conference on Geospatial Theory, Processing, Modeling and Applications, pp. 187-190.

Fraisse, C.W., Sudduth, K.A., and Kitchen, N.R. 2001. Delineation of site-specific management zones by unsupervised classification. Transactions of the ASAE, **44**(1), 155-166.

Fridgen, J.J., Kitchen, N.R., Sudduth, K.A., Drummond, S.T., Wiebold, W.J., and Fraisse, C.W. 2004. Management Zone Analyst (MZA): software for subfield management zone delineation. Agronomy Journal, **96**, 100-108.

Kerby, A., Marx, D., Samal, A., and Adamchuk, V. 2007. Spatial clustering using the likelihood function. In: ICDM Workshops, Seventh IEEE International Conference, pp. 637-642.

Li, Z. and Wang, X. 2010. Spatial clustering algorithm Basedon Hierarchical-Partition Tree. International Journal of Digital Content Technology and its Applications, **4**(6), 26-35.

Ping, J.L. and Dobermann, A. 2003. Creating spatially contiguous yield classes for site-specific management. Agronomy Journal, **95**, 1121-1131.

Prodanov, D.P., Nagelkerke, N., and Marani, E. 2007. Spatial clustering analysis in neuroanatomy: Applications of different approaches to motor nerve fiber distribution. Journal of Neuroscience Methods, **160**(1), 93-108.

Shatar, T.M. and McBratney, A. 2001. Subdividing a field into contiguous management zones using a k-zones algorithm. In: Proceedings of the 3rd European Conference on Precision Agriculture, pp. 115-120.

Viscarra Rossel, R.A., Adamchuk, V.I., Sudduth, K.A., McKenzie, N.J., and Lobsey, C. 2011. Proximal soil sensing: an effective approach for soil measurements in space and time, Chapter 5. Advances in Agronomy, **113**, 237-283.