SPATIALLY CONSTRAINED GEOSPATIAL DATA CLUSTERING FOR MULTILAYER SENSOR-BASED MEASUREMENTS

N. M. Dhawale [a, *], V. I. Adamchuk [a], S. O. Prasher [a], P. R. L. Dutilleul[b], R. B. Ferguson[c]

[a] Department of Bioresource Engineering, McGill University, Macdonald Campus, Ste-Anne-de-Bellevue,
Quebec, H9X 3V9, Canada - nandkishor.dhawale@mail.mcgill.ca
[b] Department of Plant Science, McGill University, Macdonald Campus, Ste-Anne-de-Bellevue,
Quebec, H9X 3V9, Canada- pierre.dutilleul@mcgill.ca
[c] Department of Agronomy and Horticulture, University of Nebraska-Lincoln, Lincoln,
Nebraska, 68508, USA- rferguson1@unl.edu

KEYWORDS: precision agriculture; proximal soil sensing; geospatial data clustering; management zones

ABSTRACT:

One of the most popular approaches to process high-density proximal soil sensing data is to aggregate similar measurements representing unique field conditions. An innovative constraint-based spatial clustering algorithm has been developed. The algorithm seeks to minimize the mean squared error during the interactive grouping of spatially adjacent measurements similar to each other and different from the other parts of the field. After successful implementation of a one soil property scenario, this research was to accommodate multiple layers of soil properties representing the same area under investigation. Six agricultural fields across Nebraska, USA, were chosen to illustrate the algorithm performance. The three layers considered were field elevation and apparent soil electrical conductivity representing both deep and shallow layers of the soil profile. The algorithm was implemented in MATLAB, R2013b. Prior to the process of interactive grouping, geographic coordinates were projected and erroneous data were filtered out. Additional data pre-processing included bringing each data layer to a 20x20 m raster to facilitate multi-layer computations. An interactive grouping starts with a new “nest” search to initiate the first group of measurements that are most different from the rest of the field. This group is grown using a neighbourhood search approach and once growing the group fails to reduce the overall mean squared error, the algorithm seeks to locate a new “nest”, which will grow into another group. This process continues until there is no benefit from separating out an additional part of the field. Results of the six-field trial showed that each case generated a reasonable number of groups which corresponded to agronomic knowledge of the fields. The unique feature of this approach is spatial continuity of each group and capability to process multiple data layers. Further development will involve comparison with a more traditional k-means clustering approach and agronomic model calibration using a targeted soil sampling.

1. INTRODUCTION

1.1 General Instructions

While conventional soil sampling techniques are laborious and time consuming, proximal soil sensing (PSS) allows rapid and inexpensive collection of high-density data (Viscarra Rossel and McBratney, 1998; Viscarra Rossel *et al.*, 2010). To pursue various site-specific management practices, spatial data is frequently split into groups (clusters or zones) to represent significantly different growing conditions (Fraise *et al.*, 2001; Ping and Dobermann, 2003). Geo-spatial data clustering is an important process (Li and Wang, 2010), which is widely used in remote sensing (Deng, *et. al.* 2003), neuroanatomy analysis (Prodanov, *et. al,* 2007), and other areas. Several different spatial clustering algorithms have been developed to group geospatially dense PSS-based measurements of soil attributes into management zones. For example Management Zone Analyst (Fridgen *et al.*, 2004) represents a publicly available tool accepted by a number of practitioners. The algorithm is based on computing a distance matrix and performing clustering over this new distance matrix. It is closely related to the popular k-means clustering algorithm, where quality of the resulting clusters heavily depends on the selection of initial centroids and the results are not repeatable. However, this method requires cross-validation to select the best among several runs. (Abdul-Nazeer and Sebastian, 2009). Although the method allows multidimensional data analysis, complexity and frequently occurring discontinuities of management zones make this technology non-robust for potential users (Kerby *et al.*, 2007; Shatar and McBratney, 2001).

Spatial continuity of formed clusters can be achieved by restricting grouping measurements that are not adjacent to each other (Dhawale *et al.*, 2012) through so called Neighbourhood Search Analysis (NSA). This is a form of clustering built on the principle of growing new groups of data points or grid cells with a fixed size through minimization of the mean squared error (MSE). Since previous trials with one measured soil attribute revealed positive outcomes, the **objective** of this study was to advance an algorithm to allow multiple data layers to be used for delineating spatially constrained groups of high-density soil sensor-based measurements. Field elevation and apparent soil electrical conductivity (ECa) at two depths obtained from six agricultural fields with different levels of spatial structure were used to illustrate the performance of the algorithm developed.

[*] Corresponding author

This contribution has been peer-reviewed.
doi:10.5194/isprsarchives-XL-2-187-2014

2. MATERIALS AND METHODS

2.1 Data collection

Six production fields from Nebraska were mapped using Veris 3100 (Veris Technologies, Salina, Kansas, USA) galvanic contact soil EC_a mapping unit equipped with an RTK-level, AgGPS 442 (Trimble Navigation Ltd., Sunnyvale, California, USA) and a global navigation satellite system (GNSS) receiver.. The three data layers were: 1) field elevation 2) deep soil EC_a (~0-90 cm) obtained with a wide pair of Wanner array electrodes and 3) shallow soil EC_a (~0-30 cm) obtained using the narrow pair of electrodes. Table 1 summarises data from the six fields. It has been noted that the two layers of EC_a represent similar but not identical spatial patterns, while field elevation does not always correspond to the overall pattern of changing EC_a. Therefore, the ideal map of field partitioning would delineate areas with different combinations of the three values significantly different from the average field conditions.

Field ID	Area, ha	Mean	Range	SD
		Field elevation, m		
1	25	1.50	3.20	0.53
2	46	4.95	17.82	3.99
3	50	7.10	11.54	2.09
4	55	8.07	27.44	5.68
5	67	4.22	8.09	1.60
6	44	6.15	10.59	2.15
		Shallow ECa, mSm^{-1}		
1	25	0.73	1.58	0.28
2	46	3.99	13.14	1.67
3	50	6.21	11.64	1.84
4	55	2.44	9.04	1.72
5	67	7.25	9.32	1.88
6	44	2.29	7.42	0.82
		Deep ECa, mSm^{-1}		
1	25	7.62	27.66	3.76
2	46	30.24	86.90	14.39
3	50	4.10	8.68	1.71
4	55	16.31	61.97	12.06
5	67	51.01	80.77	14.07
6	44	25.72	81.74	14.36

Table 1. Summary of data from agricultural fields

2.2 Data pre-processing

All data processing was accomplished using MATLAB R2013b (The MathWorks, Inc. Natick, Massachusetts, USA). To obtain three 2D matrices representing each field, sensor-based data pre-processing involved four steps: 1) removing erroneous data using predefined threshold values of physically feasible measurements, 2) 1D data smoothing using a 5-point moving average technique, 3) projection of local coordinates according to Adamchuk (2001), and 4) 20x20 m averaging of all measurements inside each grid cell. Field elevation data were relative to the lowest grid cell found in every field. The resulting rectangular matrix representing each field covered the entire spatial domain. Grid cells outside field boundaries were assigned zero values. Therefore, no grid cells inside the fields were without corresponding sensor measurements. Smaller grid cell size would also be possible, but require more computation power. The selected resolution using the total of 600-1500 grid cells per field was considered reasonable to reveal field macro-variability.

2.3 Data clustering algorithm

The data clustering algorithm was constructed using an assumption that treating a group of adjacent grid cells separately form the rest of the field would reduce the MSE between individual cell values and the average for corresponding groups:

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

where X_{ij} = sensor-value for i^{th} grid cell within j^{th} group
$\dot{X}_j$ = sensor-value average for j^{th} group
k = the number of grid cell groups
n_j = the number of grid cells within j^{th} group
N = the total number of non-zero grid cells

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 have at least one common side. This assumption is typically referred to as “rook’s rule”. Only nine neighbouring grid cells in a 3x3 configuration can form a new group. The beginning of a new group as well as a merger of a new grid cell to an existing group is accepted when the result produces the lowest MSE. Group enlargement as well as the search for a new group stops when neither action could result in further MSE decrease. Figure 1 illustrates the flowchart of the algorithm developed.

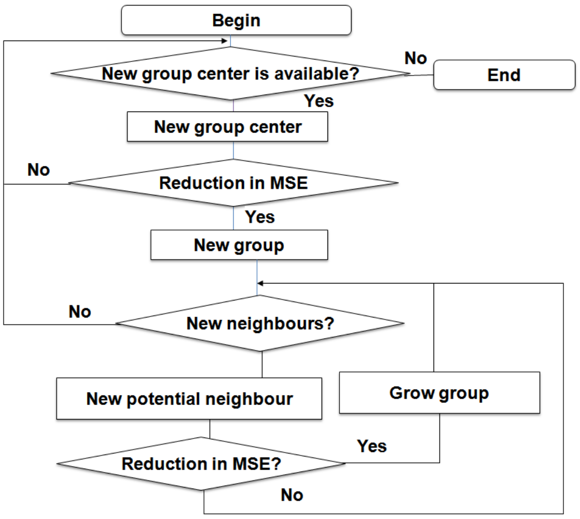

Figure 1. Algorithm flow chart

One minus the ratio of the MSE calculated using equation (1) and the initial MSE (considering that k = 1) indicates the fraction of variability accounted for by the grouping and is equivalent to the coefficient of determination (R^2) typically used to quantify the quality of a linear regression model:

$$R^2 = 1 - \frac{MSE}{MSE_{k=1}} \tag{2}$$

As shown by Dhawale *et al.* (2012), this algorithm can be successfully used for a single data layer. To achieve multilayer analysis, MSE for each data layer should be minimized and R^2 maximized. This can be realized by multiplying R^2 values.

Thus, perfect recognition of spatial variability would mean $R^2 = 1$ (measured values within each group are exactly the same), and $R^2 < 1$ once a fraction of the variability is not accounted for. Therefore, the product of three R^2 (elevation and two depths of EC) will be small if at least one of the three multipliers is relatively low. Since $MSE_{k=1}$ is a constant value, the same grid cell grouping result will occur when minimizing the product of three MSE estimates as when maximizing the product of three R^2 values. Quality partitioning of an agricultural field would occur when R^2 for all the data layers would be relatively high with the smallest possible number of identified groups of relatively homogeneous grid cells different from their surroundings. Since the two EC_a measurements frequently correlate, the influence of field elevation in this study was made similar to the influence of EC_a by raising the elevation R^2 estimate to the second power:

$$R^2_{Product} = R^2_{ShallowEC_a} \cdot R^2_{DeepEC_a} \cdot \left(R^2_{Elevation}\right)^2 \qquad (3)$$

Therefore, the algorithm shown in Figure 1 was implemented to maximize product R^2 instead of the MSE for a single data layer. No formal statistical analysis and comparison with more traditional spatial clustering techniques were performed at this preliminary stage.

3. RESULTS AND DISCUSSION

Figure 2 illustrates one of the six fields from the initial trial of the developed algorithm. Areas of the field representing low elevation in the east and high elevation in the west were delineated first with two additional groups emerging later. Figure 3 illustrates that R^2 values increased as new groups were formed. Apparently, delineation of groups 2 and 4 were primarily caused by the spatial variability of soil EC_a, while groups 3 and 5 emerged predominantly due to differences in field elevation. The algorithm did not locate any new groups of 3x3 grid cells that could further increase the R^2 product.

Figure 4 illustrates grid cell grouping for all the fields resulting in a total of 2-8 groups per field. Figure 5 summarises resulting R^2 values. The products of these values are shown in Figure 6. Fields 2 and 4 revealed only one group of grid cells that could be separated from the rest of the field while Fields 3 and 4 had 6 and 7 groups, respectively. At the same time, the algorithm produced groups with relatively strong three data layer partitioning for Fields 1, 3, 4, and 6. However, sub-division of Fields 2 and 5 was mainly dominated by field elevation, which resulted in relatively low R^2 products. In both cases, soil EC_a measurements differed significantly among neighbouring cells, indicating relatively poor spatial structure.

Although the strength of this algorithm is spatial continuity of each group of grid cells, group edges may need smoothing for improved field manageability. Since grid cells poorly associated with their neighbours occur mostly due to field anomalies or erroneous measurements, edge smoothing will always reduce R^2 product objective function. The next step in this research will include a comparison of resulting field partitioning with equivalent processing that can be conducted using more traditional k-means-type clustering algorithms (Fraise *et al.*, 2001; Ping and Dobermann, 2003) with suitable pre- and post-processing techniques.

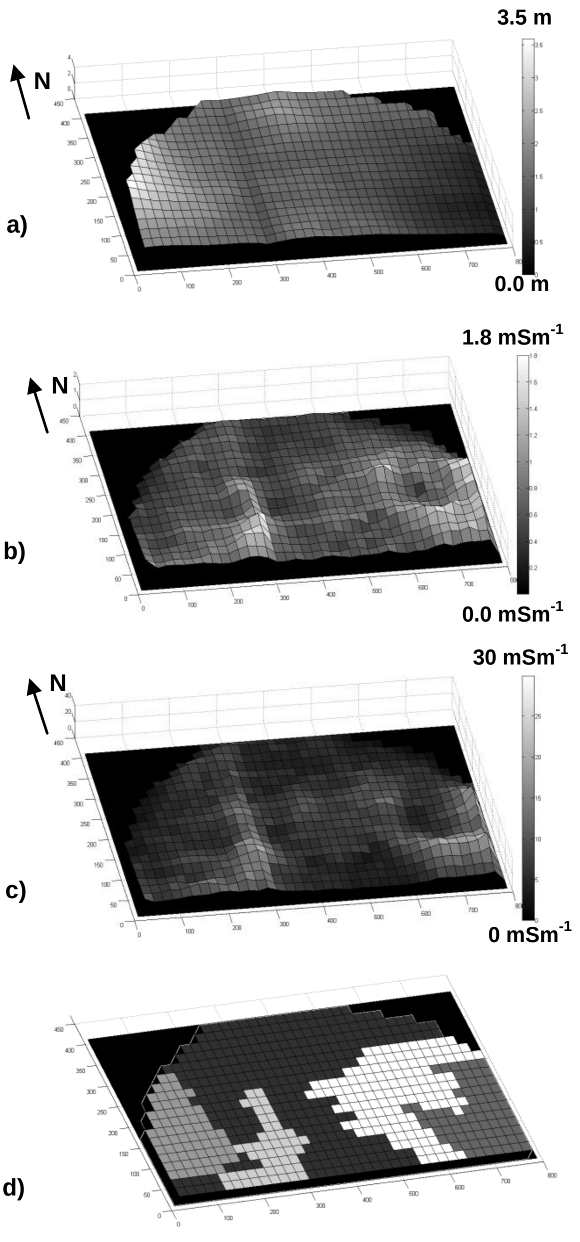

Figure 2. Maps of field elevation (a), shallow EC_a (b), deep EC_a (c), and delineated groups of grid cells (d) for Field 1.

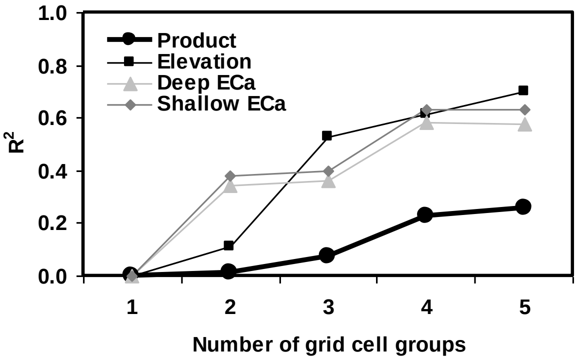

Figure 3. Change in R^2 product with number of delineated grid cell groupings.

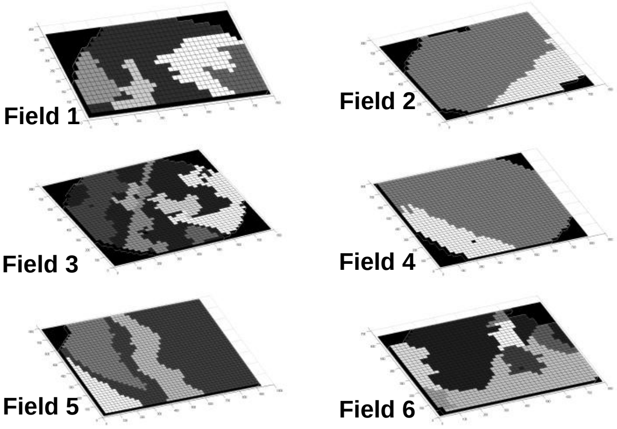

Figure 4. Maps of partitioned fields.

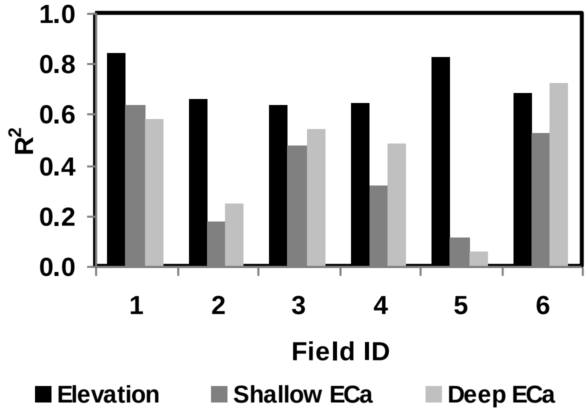

Figure 5. R^2 values for three data layers used to partition the six experimental fields.

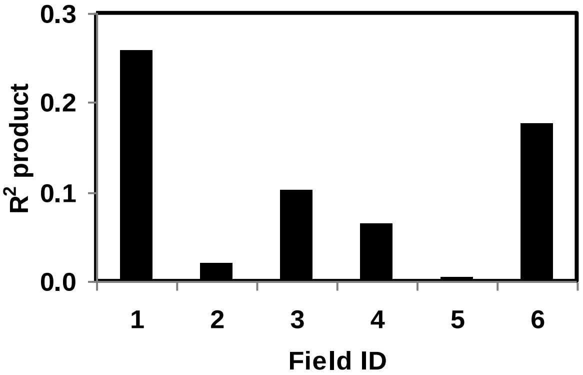

Figure 6. R^2 products for the six fields.

4. SUMMARY

The spatial clustering algorithm developed in this study is based on a neighbourhood search method and seeks to minimize variance inside each group of interpolated grid pixels corresponding to an unlimited number of sensor-based data layers. Preliminary tests of the algorithm using six production fields illustrated algorithm robustness when delineating field areas with different field elevations and soil EC_a measurements. Each spatially constrained group of grid cells with the exception of the first group designated as "the rest of the field" emerged in response to every unique combination of data values relatively constant within each group.

4.1 References

Adamchuk, V.I. 2001. Untangling the GPS String. University of Nebraska Co-operative Extension,EC-01-157, University of Nebraska, Lincoln, Nebraska, USA.

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), pp. 1-3.

Deng, X., Wang, Y., and Peng, H. 2003. The Clustering of High Resolution Remote Sensing Imagery. *Journal of Electronics & Information Technology*. 25(8), pp. 1073-1080.

Dhawale, N., V. I. Adamchuk, S.O. Prasher, and P.R.L. Dutilleul. 2012. Spatial data clustering using neighbourhood analysis. *Paper No. 121337939*. St. Joseph, Michigan: ASABE.

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

Fridgen, J.J., N.R. Kitchen, K.A. Sudduth, S.T. Drummond, W.J. Wiebold, and C.W. Fraisse. 2004. Management Zone Analyst (MZA): software for subfield management zone delineation. *Agron. J.*, 96, pp. 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 Based on Hierarchical-Partition Tree. *International Journal of Digital Content Technology and its Applications*. 4(6), pp. 26–35.

Ping, J. L., and Dobermann, A. 2003. Creating spatially contiguous yield classes for site-specific management. *Agronomy Journal*. 95, pp. 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), pp. 93-108.

Shatar, T. M., and McBratney, A. 2001. Subdividing a field into contiguous management zones using a k-zones algorithm. *In: (Eds.) G. Grenier and S. Blackmore, Proceedings of the 3rd European Conference on Precision Agriculture., Agro-Montpellier ENSAM.*, France, pp. 115-120.

Viscarra Rossel, R. A., and McBratney, A. B. 1998. Laboratory evaluation of a proximal sensing technique for simultaneous measurement of clay and water content. *Geoderma*. 85, pp.19-39.

Viscarra Rossel, R. A., McBratney, A. B., and Minasny, B., (Eds.). 2010. *Proximal Soil Sensing, Progress in Soil Science Series*, Springer-Verlag, New York. New York. USA.

4.2 Acknowledgements

This research was supported in part by funds provided through the National Science and Engineering Research of Canada (NSERC) Discovery Grant; Agricultural Research Division of the University of Nebraska-Lincoln; and Deere and Company.