Unsupervised classification method for placing sampling points

By Michael Nørremark, Aarhus University, Department of Engineering

E-mail: Michael.Norremark@eng.au.dk

Revised February 10th, 2016

Introduction

The objective of this research was to apply unsupervised classification statistical methods based on soil surface topographic attributes (TA) and soil electrical conductivity (EC) to identify homogenous sampling points for large plot experiments. The procedure is inspired by the work of Fridgen et al. (2004) applied to similar research objective in Terra et al. (2006). The procedure consists of: i) TA and EC data outlier detection and removal, II) georeference of TA to EC measurement positions by nearest neighbour, iii) normalisation of data into 0-100 scalar, iv) *k*-means clustering, v) sample size estimation, vi) stratified sampling point mapping.

Materials and methods

The electrical conductivity (EC) of the fields were surveyed during March 2013 using an EM38 (Geonics Limited, Mississauga, Ontario, Canada) just prior the establishment of the experimental plots. The EM38 uses the principle of electromagnetic induction to quantify soil EC in milliSiemens per meter (mS/m). The instrument was operated in the dipole mode, providing an effective measurement depth of approximately 1.4 m in the vertical mode and 0.7 m in horizontal mode Measurements in the field were performed using a mobile system that included an all–terrain vehicle, a sled for carrying the EM38DD, a sub meter accurate DGPS receiver, and a computer for data acquisition. Geo-referenced EC data (mSiemens/m) were recorded at 0.5 s intervals at 0 to 70 cm (horizontal) and 0 to 140 cm (vertical) depths with a vehicle travelling at speed of 6 km/h in transects spaced 5-6 m apart. The fields were at fallow during the survey and soil moisture conditions were near field capacity.

The topographic attributes were requested from the Danish Agency for Data Supply and Efficiency (http://download.kortforsyningen.dk) and downloaded as vector data provided in the LAS file format. The LAS file was converted to ASCII file format using the LASTool (Rapidlasso GmbH, Gilching, Germany). The requested topographic attributes were modelled into point resolution of 1.6 m.

Outliers in data were identified and removed according to Chauvenet's criterion (Taylor, 1997) prior to statistical analysis. The EC data was normally distributed and no transformation was done before statistical analysis. Data outlier detection and removal was not done to the TA data as the data was representing soil surface slopes which are not normally distributed. For n being the sample size, the probability α:

$$\alpha_j = 1 - \frac{1}{2n} \qquad \text{(eq.1)}$$

for retention of data set, j, distributed about the mean is related to a maximum deviation, δ, away from the mean by using the Gaussian probability table. For a given α, the non dimensional maximum deviation τ, is determined per data set, j, from the table where:

$$\tau_j = \frac{(z_{i,j} - \bar{z}_j)}{\sigma_j} \qquad \text{(eq.2)}$$

and σ is the standard deviation. Therefore, all measurements that deviated from the mean by more than τσ was rejected per data set. Figure 1 and 2 shows the distribution of EC measurements before and application of the Chauvenet's criterion, respectively.

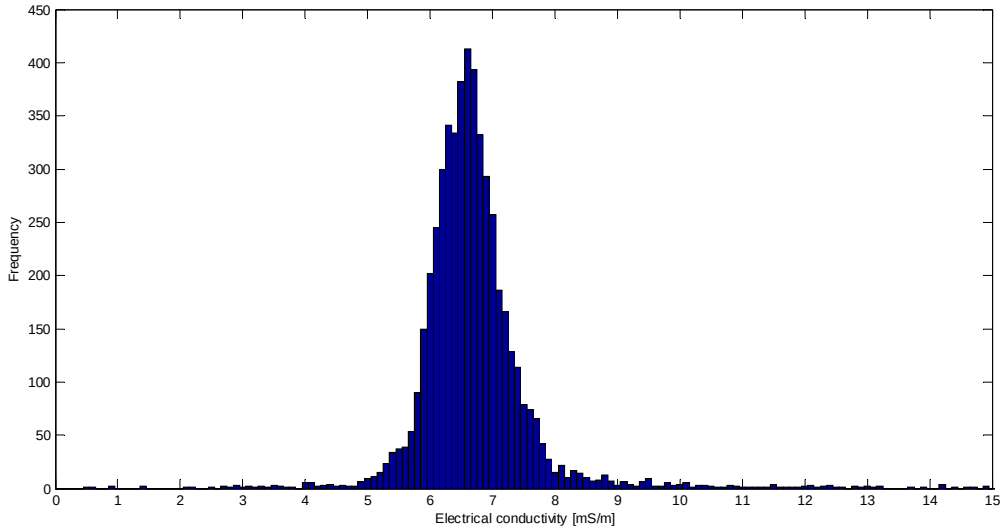


Figure 1. Histogram of horizontal EC 0-0.7 m.

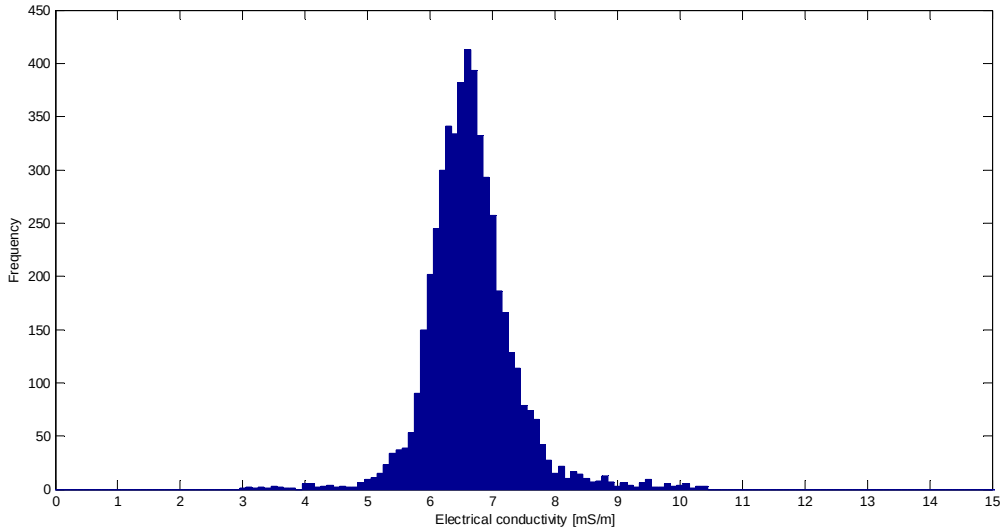


Figure 2. Histogram of horizontal EC 0-0.7 m after outlier removal by Chauvenets' criterion .

A fuzzy *c*-mean classification methodology was studied by Fraisse et al. (2001), adapted to software in Fridgen et al. (2004), that was further adapted in Terra et al. (2006). However, in the present study the classification was based on *k*-mean clustering using MatLab. A comparative analysis of *k*-means and fuzzy *c*-means algorithms concluded that *k*-mean produces close clustering results to *c*-means (Ghosh & Dubey, 2013). The outlier processed data, i = 1,….n, for each factor, j=1,2, (i.e. horizontal and vertical) needs to be normalised before *k*-means clustering, done by the following equation, where n is the sample size:

$$\hat{z}_{i,j} = \frac{(z_{i,j} - \min z_{i,j})100}{\max z_{i,j} - \min z_{i,j}} \qquad \text{(eq.3)}$$

In order to combine the global positions of TA data (X_i,Y_i) with the global positions of EC data (X_l,Y_l) a nearest neighbour approach was used. A nearest neighbour (NN) query was performed to estimate altitude (A) by retrieving the Northing and Easting global position values belonging to the TA data set that was nearest to the Northing and Easting global position values belonging to the EC data (example shown in Fig. 3 & 4). The matching was done according to the minimum distance between the pairs of X_i, Y_i and X_l, Y_l:

$$A_i = \arg\min_i \left| \sqrt{(X_i - X_l)^2 - (Y_i - Y_l)^2} \right| \qquad \text{(eq.4)}$$

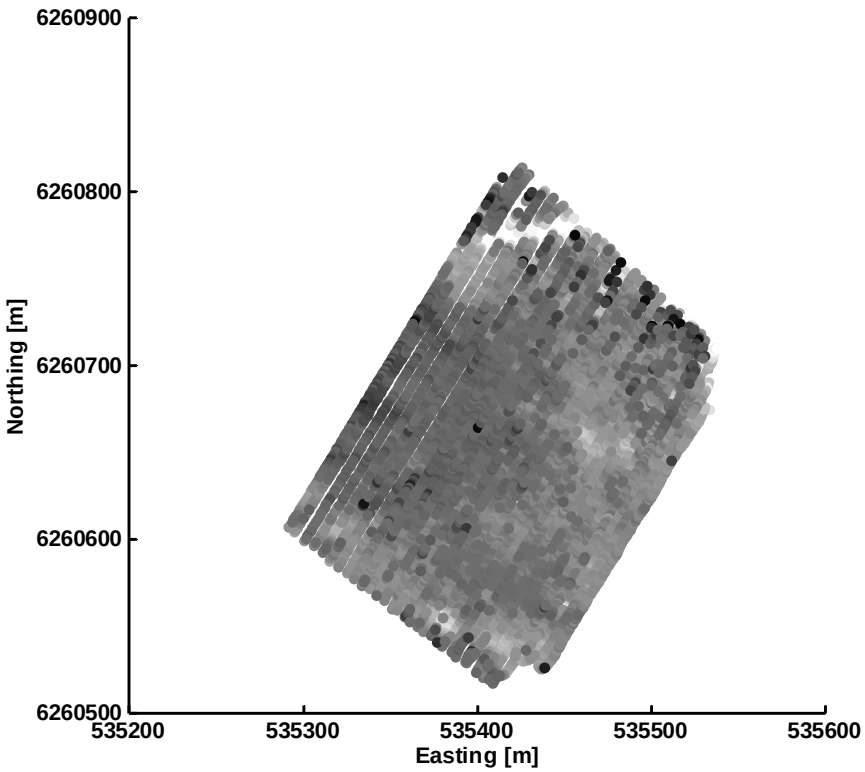


Figure 3. Horizontal EC data; Black: low EC level (3.8 mS/m), light gray: high EC level (9.6 mS/m)

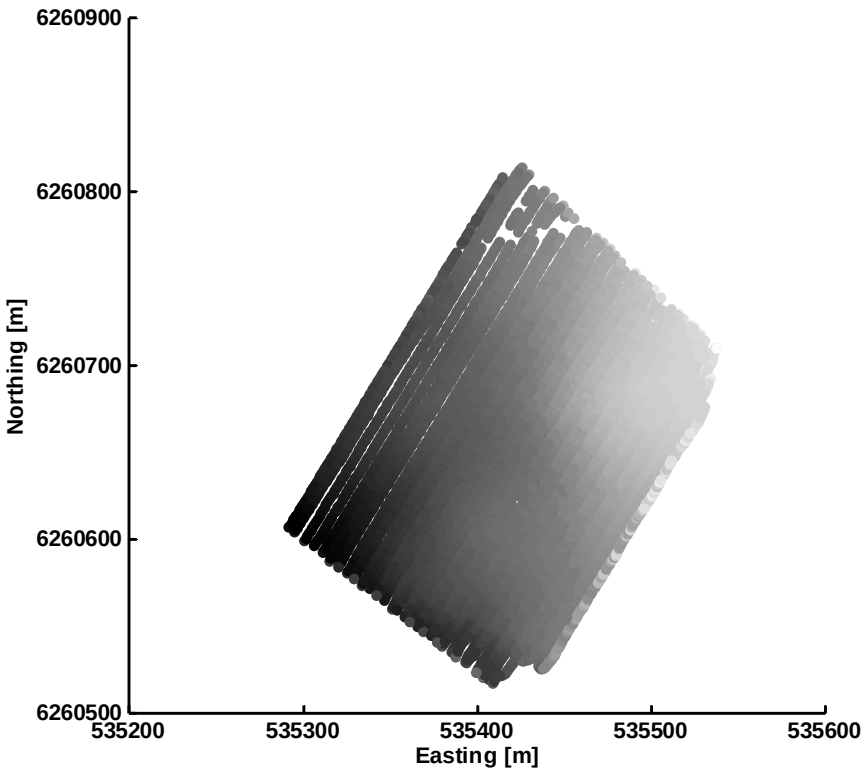


Figure 4. Topographic height above sea level at positions where EC data was acquired. Light gray 57.5 m, black 45.1m above se level

The EC data contained multiple classes, and it gives meaning to include the classes into higher level classes that are related to some soil characteristics (e.g. Domsch & Giebel, 2004, Neudecker et al., 2001). However, these studies concluded that the level of the mean clay content cannot be identified from the relative values of the soil electrical conductivity. Despite this, such a map is suitable for approximate decision making in precision agriculture, and thus for dividing the field plot into strata for sampling. The strata used for stratified sampling were clustered using *k*-mean algorithm, which is a straightforward and effective algorithm for finding clusters in data. Clusters were created with a fuzzy *k*-means unsupervised classification of multivariate data using MatLab (Mathworks).

The field plot area was subdivided using the *k*-mean clustering procedure. It classifies X_i, Y_i based on $Z_{i,j}$ and A_i into *k* centroids of group, one for each cluster. The cluster analysis was performed with data that explained multivariate data variability on the fields, i.e. TA and EC data. The *k*-means place centroids as much as possible far away from each other initially, followed by taking each point belonging to a given data set which associate to the nearest centroid. The algorithm developed in MatLab proceeds as follows following steps explain k-mean clustering algorithm in brief:

1. Determining *k* (number of clusters)
2. Initializing *k* centroid

3. Calculating distances to centroid and assigning to a cluster according to distance of point to cluster centroid.
4. Updating centroid attributes value
5. Repeating steps 3 and 4 until no data point is reassigned to clusters

Dataset consists of points $Z_{i,j}$, where j represents the attributes, i.e. horizontal and vertical EC data and altitude value of the i^{th} point. At the beginning of *k*-means algorithm a chosen point, o, in the dataset are taken as centroids by setting their attributes to random values. *K*-means used the distance metric; Euclidean distance. Suppose $w = \{w_1, w_2, \ldots, w_o\}$ to be calculated centroids. Assign centroid value, w, to the position and compute distance between $Z_{i,j}$ and $w_{k,j}$ for all w centroids. The Euclidean distance is calculated as below:

$$d_{i,j} = \sum_{i=1}^{j} \left| \left(Z_{i,j} - w_{j,k} \right)^2 \right| \qquad \text{(eq.5)}$$

Euclidean distance metric is used to define a nearest centroid to a point and the point is assigned to a cluster with the nearest centroid. Assign $Z_{i,j}$ to the cluster with minimum distance and for each $w_{k,j}$ centroid value move the position of $w_{k,j}$ to the mean of points in corresponding k^{th} cluster which contains m points:

$$w_{k,j} = \frac{\sum_{i=1}^{j} z_{i,k}}{m} \qquad \text{(eq.6)}$$

Where $Z_{i,k}$ is the points in cluster, this is what is done in step 4. Iterating steps 3 and 4 finishes when no data point changes cluster, i.e. total sum of distances d between i^{th} data points and their k^{th} cluster centroids does not change. A silhouette index, s_i, were used to determine the optimal number of clusters (= strata) for the field plot area. For each datum i, let a(i) be the average dissimilarity of i with all other data within the same cluster. Suppose a_i as how well data point i is assigned to its cluster k (the smaller the value, the better the assignment). The a_i define the average dissimilarity of data i to a cluster k as the average of the distance from i to all points in k. Let b_i be the minimum average dissimilarity of i to any other cluster, of which i is not a member. The cluster with this lowest average dissimilarity is said to be the "neighbouring cluster" of data point i because it is the next best fit cluster for point i. The silhouette index is defined:

$$s_i = 1 - \frac{a_i}{b_i} \qquad \text{(eq.7)}$$

The optimum number of clusters was determined by inspection of the silhouette index for each relevant number of clusters.

K-means summary: Firstly, the number of classes which the data set should be partitioned into is inputted, and *k* records are randomly assigned to be the initial cluster center. Then, for each record, it finds the nearest cluster center. For each of the *k* clusters, it finds the cluster centroid, and update the location of each cluster center to the new value of the centroid. Repeat steps until convergence or termination. $Z_{i,j}$ and A_i data were clustered using *k*-mean algorithm and final 3 classes were outputted. Then, classification result using adjustable threshold were merged according the *k*-mean algorithm result. Final classification result is shown in the figure below. Comparing the classification result and field data, they are in accord with each other.

Stratified sampling by placing sample points in strata that is defined by EC and TA data. Sample size, N, was estimated based on historical N-min data from 25 x 25 m grid soil sampling on two field locations in Denmark (Field 41: 568300 E, 6202400 N (13.3 ha) and Field 49: 569800 E 6201100 N (15.8 ha)) (Philipp Trénel, pers. comm.). The average N-min was 42.2 ± 13.4 kg N/ha and 45.1 ± 18.4 kg N/ha for field 41 and 49 respectively. The required risk, α, that the computed mean value (or mean difference) is outside the interval was set to 5 % (P=0.90). The desired (half-) width of the confidence interval, w, was set to 3 kg N/ha. The w was adjusted to balance the costs of measurements of the treatment effects (i.e. soil sample analysis, soil resistance measurements, yield analysis, etc.). The sample standard deviation, σ_m, was derived from choosing the field 49 having largest variation in N-min and m number of samples.

$$N = \left(t_{\alpha/2,m-1} \cdot \sigma_m / w\right)^2 \qquad \text{(eq.7)}$$

The eq. 6 provided the result of 6 samples per ha, but was adjusted to balance the costs per experimental site.

The chosen sample size was then divided in equal numbers per individual found strata derived from the *k*-means clustering. The placement was also done in accordance with planned tracks for the auto steering of machine operations, and furthermore also in alignment with the machine tire width, wheel base (tractors, combines and implements) such that the sample points were not trafficked at any time during the experiments.

Plot treatments were considered as fixed effects and sample points within each cluster as repeated observations.

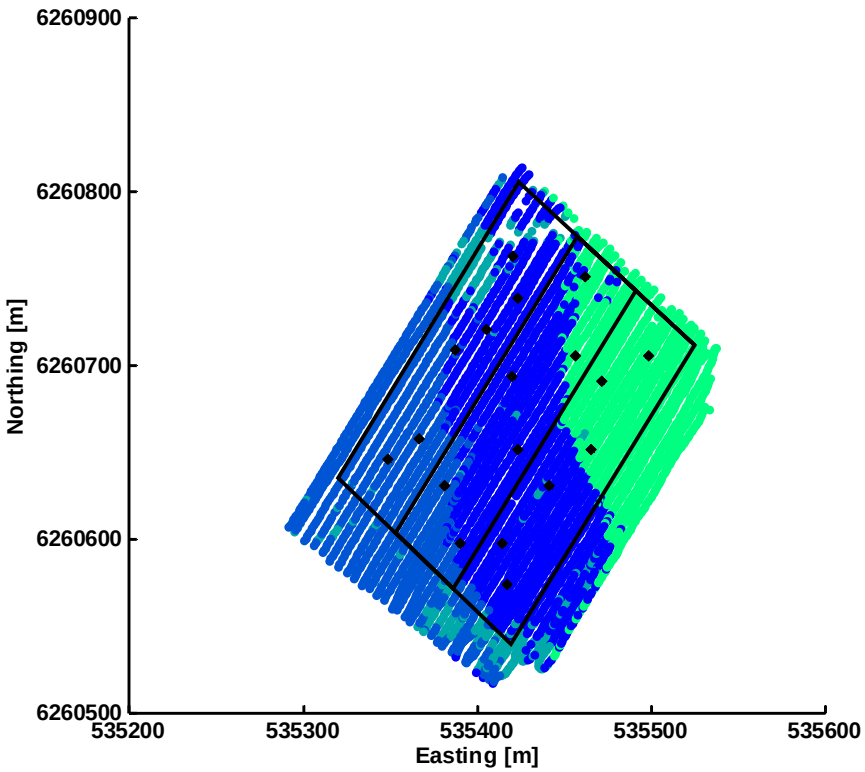


Figure 5. The four strata and positioning of measurement points in each of three treatment plots according to geometrics of controlled traffic farming experiments. All features, j, were used for the *k*-means clustering.

Due to the dominance of the downhill slope towards South West to the k-means clustering, a clustering was done only on EC data, with following results where the silhouette analysis showed optimum of 6 clusters:

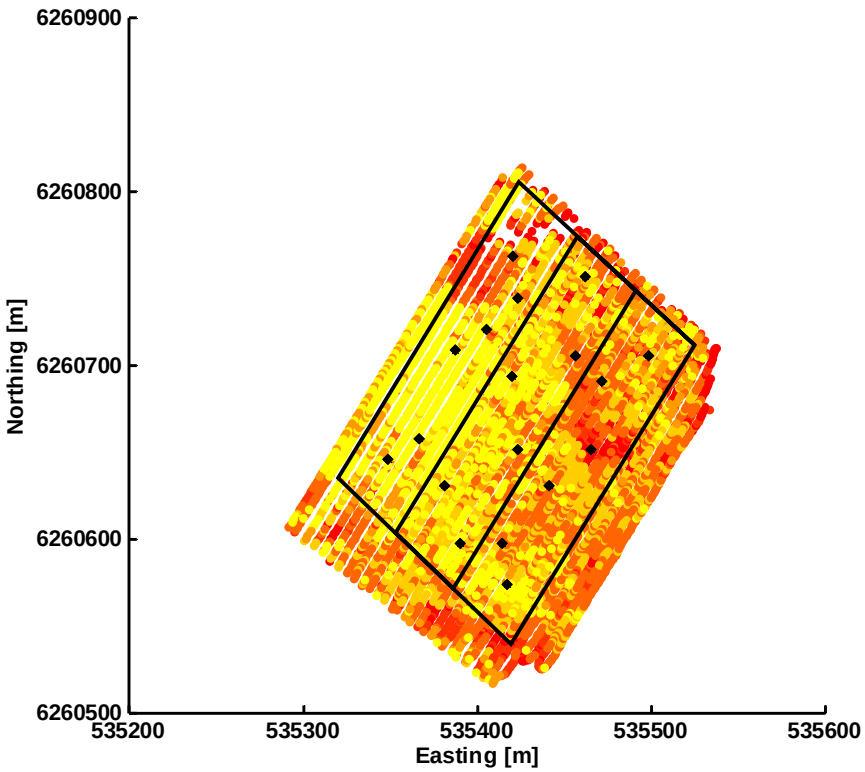
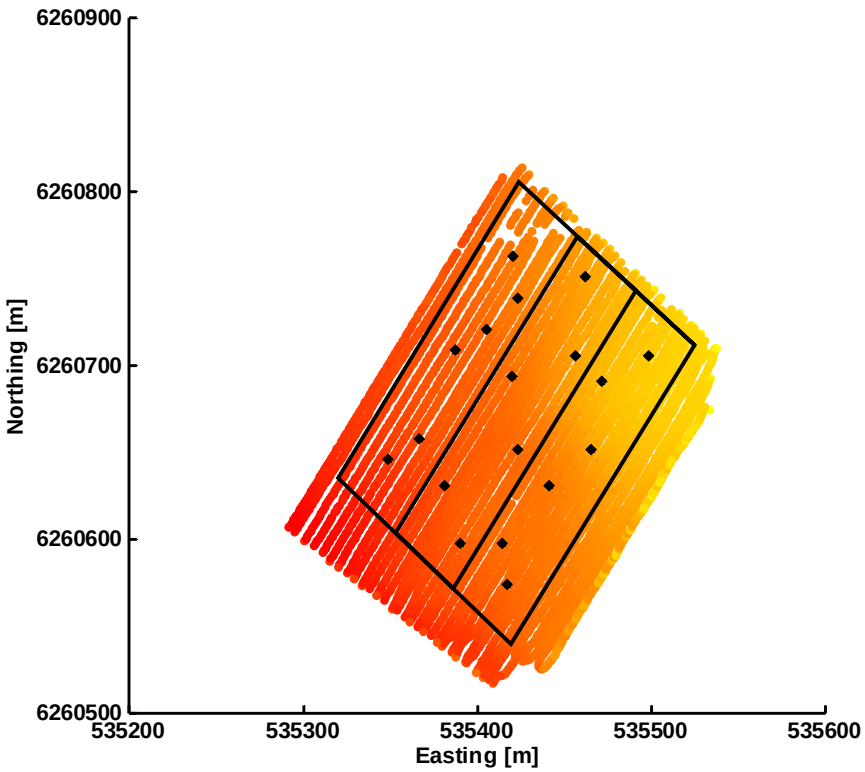


Figure 5. The six strata and positioning of measurement points in each of three treatment plots according to geometrics of controlled traffic farming experiments. Only features from EC data were used for the *k*-means clustering.


6260900
6260800
6260700
6260600
6260500
Northing [m]
535200
535300
535400
535500
535600
Easting [m]

References

Domsch, H., Giebel, A. (2004) Estimation of Soil Textural Features from Soil Electrical Conductivity Recorded Using the EM38. Precision Agriculture5, 389–409.

Fraisse, C.W. , Sudduth, K.A. , Kitchen, N.R. (2001) Delineation of site-specific management zones by unsupervised classification of topographic attributes and soil electrical conductivity. Transactions of the ASAE 44(1), 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. Agronomy Journal 96, 100-108.

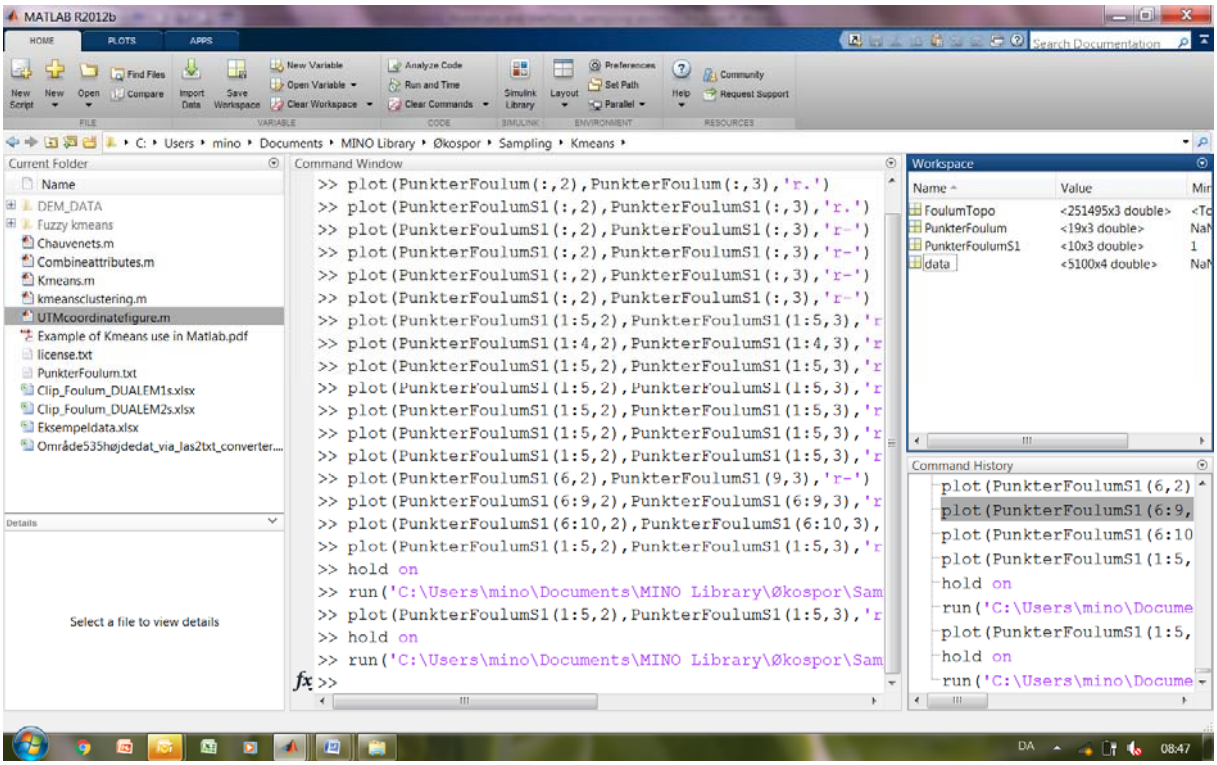
Ghosh, S., Dubey, S.K. (2013) Comparative Analysis of K-Means and Fuzzy C-Means Algorithms. International Journal of Advanced Computer Science and Applications 4 (4), 35-39.

Neudecker, E., Schmidhalter, U., Sperl, C. and Selige, T. 2001. Site-specific soil mapping by electromagnetic induction. *In:* Proceedings of Third European Conference on Precision Agriculture, edited by G. Grenier and S. Blackmore (agro Montpellier, France), pp. 271–276.

Taylor, J.R. (1997) An Introduction to Error Analysis, California: University Science Books, 2st, edition, 1997.

Terra, J.A. , Shaw, J.N. , Reeves, D.W. , Raper, R.L. , Van Santen, E. , Schwab, E.B. , Mask, P.L. (2006) Soil Management and Landscape Variability Affects Field-Scale Cotton Productivity. Soil Science Society of America journal 70(1), 98-107.

Appendices


MATLAB R2012b
HOME
PLOTS
APPS
Search Documentation
C: ▸ Users ▸ mino ▸ Documents ▸ MINO Library ▸ Økospor ▸ Sampling ▸ Kmeans ▸
Current Folder
Name
DEM_DATA
Fuzzy kmeans
Chauvenets.m
Combineattributes.m
Kmeans.m
kmeansclustering.m
UTMcoordinatefigure.m
Example of Kmeans use in Matlab.pdf
license.txt
PunkterFoulum.txt
Clip_Foulum_DUALEM1s.xlsx
Clip_Foulum_DUALEM2s.xlsx
Eksempeldata.xlsx
Område535højdedat_via_las2txt_converter....
Details
Select a file to view details
Command Window
>> plot(PunkterFoulum(:,2),PunkterFoulum(:,3),'r.')
>> plot(PunkterFoulumS1(:,2),PunkterFoulumS1(:,3),'r.')
>> plot(PunkterFoulumS1(:,2),PunkterFoulumS1(:,3),'r-')
>> plot(PunkterFoulumS1(:,2),PunkterFoulumS1(:,3),'r-')
>> plot(PunkterFoulumS1(:,2),PunkterFoulumS1(:,3),'r-')
>> plot(PunkterFoulumS1(:,2),PunkterFoulumS1(:,3),'r-')
>> plot(PunkterFoulumS1(1:5,2),PunkterFoulumS1(1:5,3),'r
>> plot(PunkterFoulumS1(1:4,2),PunkterFoulumS1(1:4,3),'r
>> plot(PunkterFoulumS1(1:5,2),PunkterFoulumS1(1:5,3),'r
>> plot(PunkterFoulumS1(1:5,2),PunkterFoulumS1(1:5,3),'r
>> plot(PunkterFoulumS1(1:5,2),PunkterFoulumS1(1:5,3),'r
>> plot(PunkterFoulumS1(1:5,2),PunkterFoulumS1(1:5,3),'r
>> plot(PunkterFoulumS1(1:5,2),PunkterFoulumS1(1:5,3),'r
>> plot(PunkterFoulumS1(6,2),PunkterFoulumS1(9,3),'r-')
>> plot(PunkterFoulumS1(6:9,2),PunkterFoulumS1(6:9,3),'r
>> plot(PunkterFoulumS1(6:10,2),PunkterFoulumS1(6:10,3),
>> plot(PunkterFoulumS1(1:5,2),PunkterFoulumS1(1:5,3),'r
>> hold on
>> run('C:\Users\mino\Documents\MINO Library\Økospor\Sam
>> plot(PunkterFoulumS1(1:5,2),PunkterFoulumS1(1:5,3),'r
>> hold on
>> run('C:\Users\mino\Documents\MINO Library\Økospor\Sam
>>
Workspace
Name
Value
FoulumTopo
<251495x3 double>
PunkterFoulum
<19x3 double>
PunkterFoulumS1
<10x3 double>
data
<5100x4 double>
Command History
plot(PunkterFoulumS1(6,2)
plot(PunkterFoulumS1(6:9,
plot(PunkterFoulumS1(6:10
plot(PunkterFoulumS1(1:5,
hold on
run('C:\Users\mino\Docume
plot(PunkterFoulumS1(1:5,
hold on
run('C:\Users\mino\Docume
DA
08:47