Determining the management zones with hierarchic and non-hierarchic clustering methods

J. Galambošová[1], V. Rataj[1], R. Prokeinová[2], J. Prešinská[1]

[1]*Department of Machines and Production Systems, Faculty of Engineering, Slovak University of Agriculture in Nitra, Nitra, Slovak Republic*
[2]*Department of Statistics and Operation Research, Faculty of Economics and Management, Slovak University of Agriculture in Nitra, Nitra, Slovak Republic*

Abstract

Galambošová J., Rataj V., Prokeinová R., Prešinská J., 2014. **Determining the management zones with hierarchic and non-hierarchic clustering methods**. Res. Agr. Eng., *60* (Special Issue): S44–S51.

Delineation of the management zones of a field is commonly used in precision agriculture technology. There are many techniques used to identify management zones. The most used technique is *k*-means clustering, where the number of clusters is managed by the user. The paper deals with clustering the yield data and electromagnetic data of a 17 ha field using the Ward's method followed by the *k*-means clustering method. The cubic clustering criterion was used to determine the number of clusters. Based on results, it can be concluded that it is beneficial to combine the *k*-means clustering method with the hierarchic method (Ward's method).

Keywords: Ward's method; k-means clustering; CCC; yield data; EMI

Precision farming technology is based on a site-specific approach to field treatments. Before implementation of this technology, the variability of the field has to be assessed in the first step. Operations can be conducted variably, or the site-specific approach can be used. For the latter one, management zones should be determined, and the field operation can be conducted based on them. Mostly, fertiliser application or soil tillage operation is carried out. However, other application such as e.g. variable irrigation (Chiericati et al. 2007) is possible.

Based on Khosla et al. (2010), there are numerous techniques for delineating management zones. Some of them are based on single soil or crop property or a combination of several that are known to affect crop productivity and yield. As Ortega and Santibanez (2007) reviewed, there are several approximations for the development of site-specific management zones. The authors reported that the first approach is based on soil and/or relief information, including topographic maps, direct soil sampling, non-invasive soil sampling by electrical conductivity equipment, and soil organic matter or organic estimated by remote sensing. The second approach is based on yield maps, combining data from several seasons, while the third is the integration of the two previous approaches and considers soil and/or relief information plus the use of yield maps. Zhang et al. (2009) developed a ZoneMAP web application, which designs the management zones based on satellite imagery, which the authors suggested as a preliminary basis when a yield map is not available. Delin and Berglund (2005) suggested to create management zones based on risk levels for drought and waterlogging, to be used in site-specific N application (based on information on soil electrical conductivity and elevation). Fleming and Westfall (2000) concluded that

Supported by the European Union within the frame of the Project No. ITMS 26220220014.

farmer-developed management zones appear to be effective in identifying different management zones; however, ground verification is needed to develop accurate Variable rate technology (VRT) maps from the zones. The need for confirmation of specific soil characteristics was reported also by King et al. (2005) when they evaluated the analysis of yield map sequences and electromagnetic induction (EMI) soil sensing as potentially cost-effective methods for identifying and mapping "management zones" (MZ) within fields. Rataj and Galambošová (2006) proposed to use the cluster analyses for identifying high- and low-profitability zones.

Guastaferro (2010) compared the techniques for identification of MZs: (1) the ISODATA method, (2) the fuzzy *c*-means algorithm and (3) a non-parametric density algorithm. They concluded that all the methods have advantages and disadvantages. However, they suggest to manage the variation within one year, and to combine the use of MZs with crop-based in-season remote sensing when using them in that particular conditions. The hierarchic as well as non-hierarchic clustering procedures can be used for management zones determination (Rus, Kruse 2011; Tiwari, Misra 2011). The most common is the use of a cluster procedure using the k-means or fuzzy *k*-mean method (Minasny, McBratney 2002; Ortega, Santibáñes 2007).

However, the estimation of a number of clusters is needed in advance. Statistical determination of the number of clusters or practical field-management considerations can be used as proposed by Taylor et al. (2003). Li et al. (2008) used fuzzy performance index (FPI) and normalized classification entropy (NCE) to determine the optimal cluster numbers. Fridgen (2000) reported that the measures of cluster performance indicated no advantage of dividing fields into more than four or five management zones. Moreover, year-to-year differences in an appropriate number of management zones were attributed to weather and crop type.

Based on the literature review, the most used technique in precision agriculture is *k*-means clustering (non-hierarchical method), which requires the estimation of the number of clusters based on previous knowledge of the farmer (expert knowledge) or management considerations are included.

The aim of this paper is to point out to the possibility of the hierarchic method as complementary to the non-hierarchic clustering method in order to (a) estimate a statistically significant number of management zones of a given field which can be used as an input for the non-hierarchical method and (b) to interpret the results of clustering process with the support of statistics.

MATERIAL AND METHODS

At the first stage, the cubic clustering criterion (CCC) will be used to estimate the statistically significant number of clusters. The CCC criterion can be used to estimate the number of clusters using Ward's method and the k-means method (SAS Institute Inc. 1983). The values of CCC greater than 2 or 3 indicate good clusters; values between 0 and 2 indicate potential clusters, but they should be considered with caution. Very negative values of the CCC, such as –30, may be due to outliers (SAS Institute Inc. 1983). The clustering analyses will be conducted using two procedures. In the first step, the Ward's method (hierarchic method) will be used followed by the non-hierarchic method (*k*-means clustering).

Ward's method. This method involves an agglomerative clustering algorithm. It starts out with *n* clusters of size 1 and continues until all the observations are included into one cluster. This method attempts to minimize the Sum of Squares (SS) of any two (hypothetical) clusters that can be formed at each step. The ESS is considered as a measure of homogeneity of the cluster. This method is regarded as very efficient; however, it tends to create clusters of small size. At the first step, when each object represents its own cluster, the distances between those objects are defined by the chosen distance measure (The Pennsylvania State University 2004).

Error Sum of Squares:

$$\mathrm{ESS} = \sum_i \sum_j \sum_k |X_{ijk} - \bar{x}^2_{i.k}| \quad (1)$$

where:

X_{ijk} – value for variable *k* in observation *j* belonging to cluster *i*

R-Square – proportion of variation explained by a particular clustering of the observations

$$r^2 = \frac{\mathrm{TSS} - \mathrm{ESS}}{\mathrm{TSS}} \quad (2)$$

where: TSS is Total Sum of Squares

$$\mathrm{TSS} = \sum_i \sum_j \sum_k |X_{ijk} - \bar{x}^2_{..k}| \quad (3)$$

As a result, a dendrogram is plotted. Here, each stage of clustering processed is displayed and the R-Square is plotted at *y* axis.

Table 1. Basic statistics of the input data

Parameter/unit	No. of samples	Average	Minimum	Maximum	Standard deviation
EMI (mS/m)	31	49.51	39.42	56.64	4.74
Yield in 2009 /(t/ha)	31	6.93	2.52	9.8	0.82
Yield in 2010 (t/ha)	31	1.99	0.11	3.08	0.37
Yield in 2011 (t/ha)	31	7.73	5.67	10.46	0.78

EMI – electromagnetic induction

***k*-Means method**. Based on McQueen (1967), the procedure follows a simple and easy way to classify a given data set through a certain number of clusters (assume *k* clusters) fixed a priori. The main idea is to define *k* centroids, one for each cluster. These centroids should be placed in a cunning way because different location causes different result. So, the better choice is to place them as much far away from each other as possible. The next step is to take each point belonging to a given data set and associate it to the nearest centroid. When no point is pending, the first step is completed and an early groupage is done. At this point we need to re-calculate k new centroids as barycentres of the clusters resulting from the previous step. After we have these k new centroids, a new binding has to be done between the same data set points and the nearest new centroid. A loop has been generated. As a result of this loop we may notice that the *k* centroids change their location step by step until no more changes are done. In other words centroids do not move any more. Finally, this algorithm aims at minimizing an objective function, in this case a squared error function. The objective function is

$$J = \sum_{j=1}^{k} \sum_{i=1}^{n} ||x_i^{(j)} - c_j||^2 \quad (4)$$

where: $||x_i^{(j)} - c_j||^2$ – chosen distance measure between a data point $x_i^{(j)}$ and the cluster centre c_j is an indicator of the distance of the *n* data points from their respective cluster centres (A Tutorial on Clustering Algorithms).

For clustering procedure, Ward's method conducted in SAS (version 4.3, SAS Institute Inc., Cary, USA) as well as the k-means clustering method conducted in Statistica (Statistica CZ 10; StatSoft, Tulsa, USA) was used.

Data used. The above-described methods were applied to data obtained from yield monitoring during three seasons (2009 – spring barley; 2010 – oilseed rape; 2011 – winter wheat) at a 17 ha experimental field. The information on field variability was extended by data on electromagnetic induction (EMI) measured by Geonics EM38 (Geonics Limited, Mississauga, Ontario, Canada). In order to estimate the management zones within the given field, 31 monitoring points were designed across the field and data from all datasets were allocated

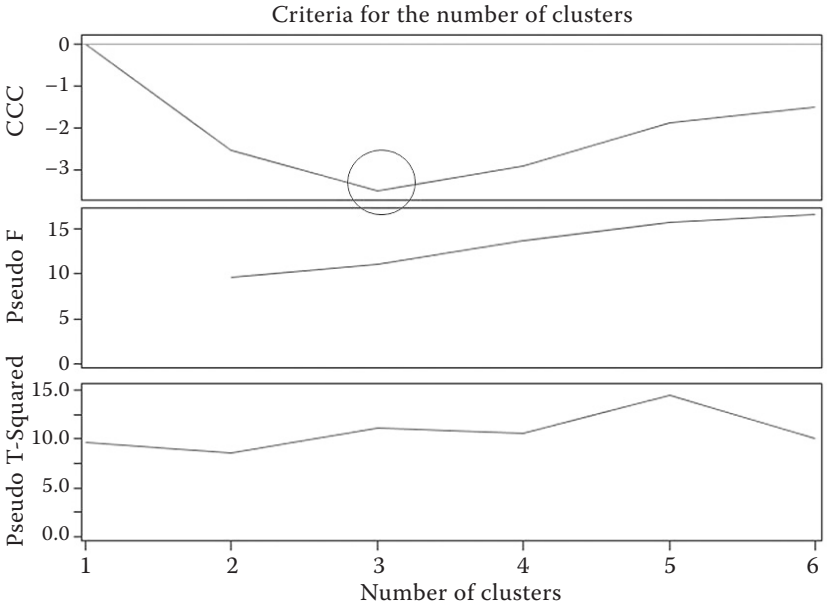


Fig. 1. Estimation of the number of clusters
CCC – cubic clustering criterion; pseudo T-Squared; pseudo F – output statistics

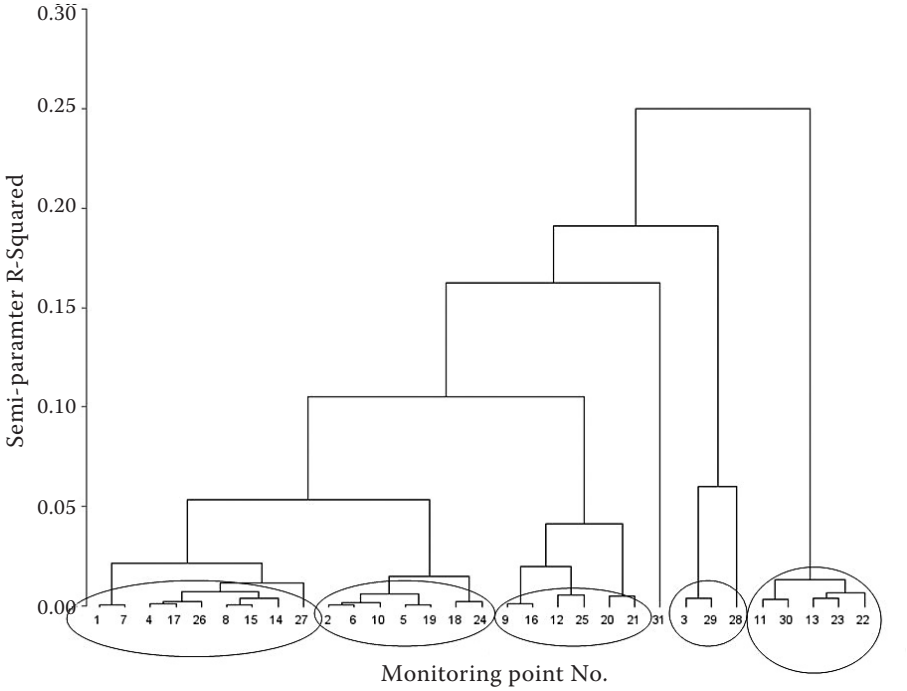


Fig. 2. Output of Ward's method – dendrogram

to these points. This means the values of yield and EMI for these 31 monitoring places were an input dataset for the analyses.

In order to be able to process the yield data from different seasons and of different crops, the preprocessing of these data was conducted and standardized normal values were calculated based on the following formula:

$$SNV = \frac{(x - \bar{x})}{SD} \times 100 \ \% \quad (5)$$

where:

SNV – standardized normal values

x – actual value of parameter

$\bar{x}$ – arithmetic mean

SD – standard deviation

The results of cluster analyses have to be interpreted from the spatial aspect. Only the members of a cluster (in our case the monitoring point) lying in a near distance and creating one area can be used for creating management zones. Therefore, the data were displayed with support of the geographical information system (GIS) ArcGIS 10.1. (ESRI, Redlands, USA) and the results were interpreted.

RESULTS AND DISCUSSION

The experimental field is represented by 31 monitoring points, which were designed across the field. All the data from yield maps and EMI maps were subtracted for these points. The data were preprocessed, and the SNV value (standardized normal value) was calculated (Table 1). First of all, the CCC was calculated, and the appropriate number of clusters was selected. The results are given in Fig. 1. According to CCC criterion, the number of statistically significant clusters was estimated for 3. As it was proposed in the methodology, the hierarchical method (Ward's method) was applied at first

Table 2. Results of cluster analyses – cluster members and Euclidean distances

Cluster 1	Member	11	12	13	22	23	30				
	Euclidean distance	49.20	40.94	27.03	50.95	36.12	44.62				
Cluster 2	Member	2	4	5	6	8	9	10	14	15	16
	Euclidean distance	47.13	27.43	39.20	36.63	8.11	63.16	31.05	45.55	20.12	83.07
Cluster 2	Member	17	18	19	21	24	25	26	27	31	
	Euclidean distance	45.61	76.75	30.98	118.97	64.16	50.67	44.74	73.96	192.48	
Cluster 3	Member	1	3	7	28	29					
	Euclidean distance	40.67	53.49	50.13	106.14	56.35					

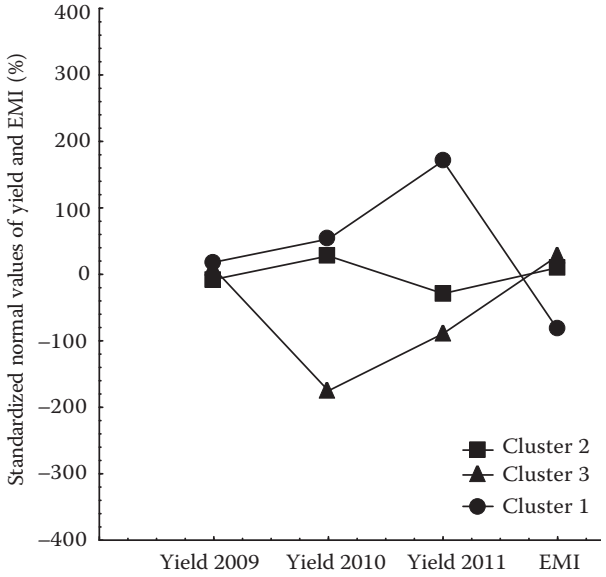


Fig. 3. The means of each cluster
EMI – electromagnetic induction

in order to explore the data. The principle of this method is to cluster the data based on homogeneity. Based on the dendrogram (Fig. 2), it is obvious that in the first stage of the clustering process, the monitoring point number 31 was not assigned to any cluster, which means that the yield performance during the years was not similar to any of the other places. Ward's method allowed identifying the most homogenous places (represented by monitoring points) and also outliers.

In the second step, the *k*-means clustering method was applied to the data. The principle of this method is to create clusters which are as heterogeneous as possible. As it is the non-hierarchical

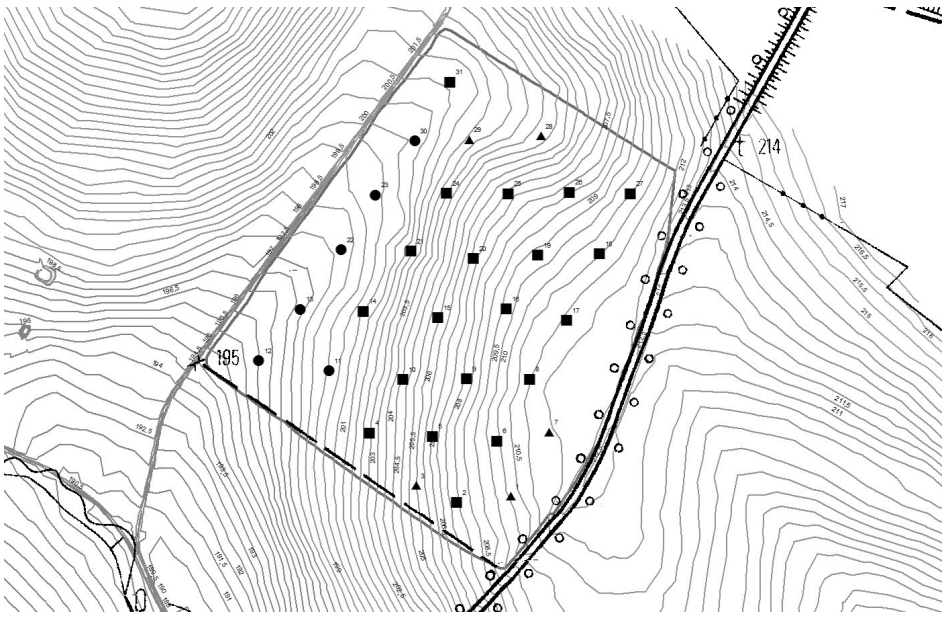
Table 3. Standardized normal values (SNV) for yield and electromagnetic induction for the areas of cluster No. 3

Monitoring point	SNV (%)			EMI
	2009	2010	2011	
1	31.52	–102.75	–98.03	–0.36
3	–83.78	–161.36	–52.91	64.80
7	17.29	–84.33	–88.43	–14.03
28	167.78	–315.43	–83.63	34.60
29	–89.02	–211.60	–122.98	53.51

EMI – electromagnetic induction

method, the number of required clusters has to be set up at the beginning. As explained above, the number of clusters was selected based on CCC criterion. The results are shown in Fig. 3, where the means of each cluster are defined, and in Table 2, where the members of each cluster are defined. The data were plotted in GIS, and their spatial localisation was considered. Based on Fig. 4, it can be concluded that two clusters create compact areas of the field and one of them not. The clusters can be interpreted as follows:

(a) The cluster No. 1 comprises areas where yield was above the average of the field in all the three seasons. This area could be considered as a high-yielding zone of the field. The results were compared with the digital terrain model, and it was shown that this zone lays in terrain depreciation with better water availability, which is confirmed also by the EMI map (Fig. 5). The values of EMI displayed in light colour reached values from 23.1 to 51.12 mS/m, the dark colour goes for values from 51.12 to 79.11 mS/m.



● cluster 1
■ cluster 2
▲ cluster 3

Fig. 4. The spatial display of cluster and its comparison with digital terrain model

Silty loam soil is typical for the entire field; therefore the change of EMI values is expected to be due to the change of the moisture content.

Monitoring point 31 lays at the same area; however, due to waterlogging problems, the yield in 2009 and 2011 was extremely low and so this location was assigned to a different management zone.

(b) Average-yielding zone – cluster No. 2 – almost all the rest of the field.

(c) Low-yielding areas – cluster No. 3 – the performance of these monitoring points is given in Table 3. The cluster No. 3 comprises only 5 monitoring points, which are not located next to each other.

When further analysing the results and looking at the dendrogram (Fig. 2), the monitoring points of cluster 2 can be described as follows:

Monitoring points 1 and 7 are characterized by good performance in 2009, but low yield in 2010 and 2011 as well as similar values of conductivity. Monitoring points 3 and 29 are characterised by an extremely low yield in all the three years; the cause is soil compaction at the headlands of the field. The monitoring point number 28 was not clustered in the first stage of the analyses with any point and is characterised by an extremely high yield in 2009 and an extremely low yield in 2010 and 2011. The

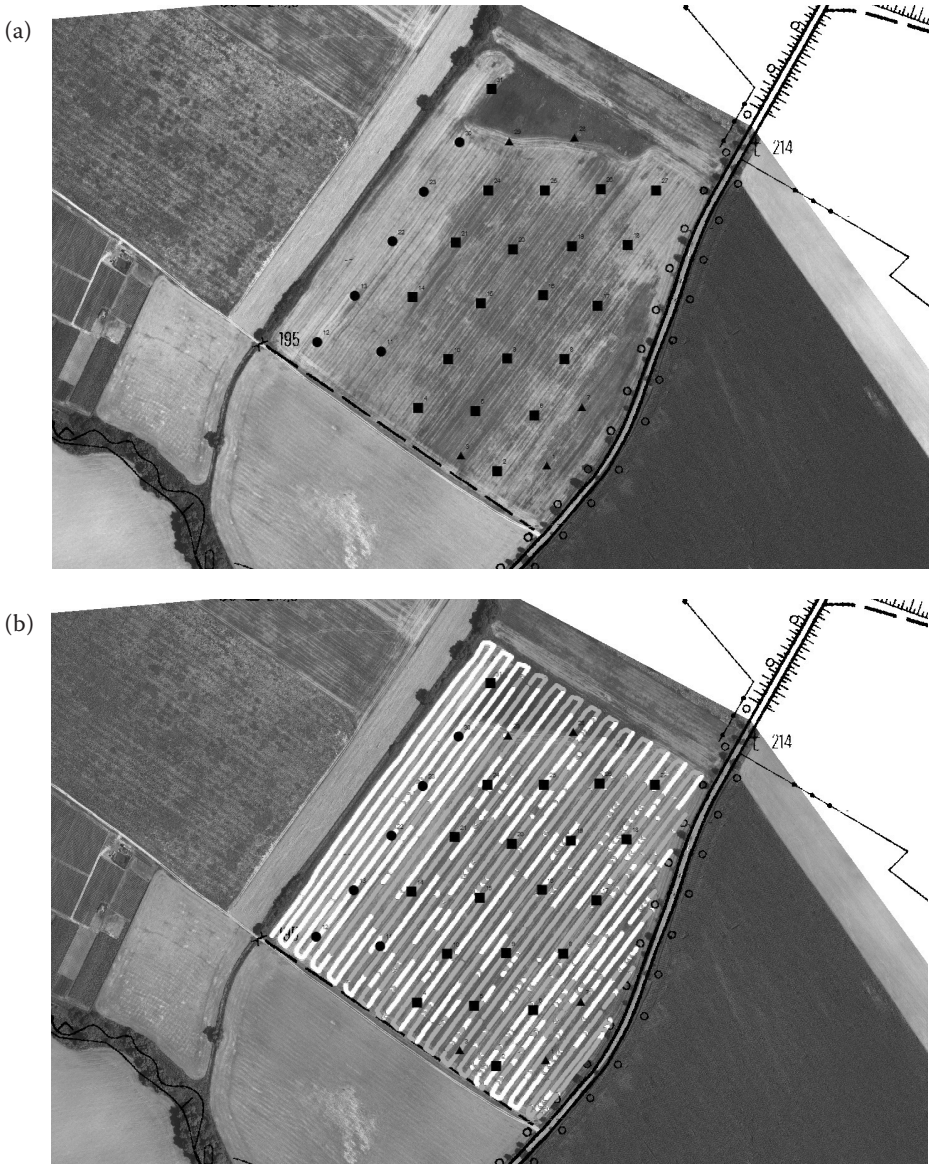


Fig. 5. Location of waterlogged area which caused the yield problems in (a) 2010 and 2011, and (b) electromagnetic induction data of the field

● cluster 1; ■cluster 2; ▲ cluster 3

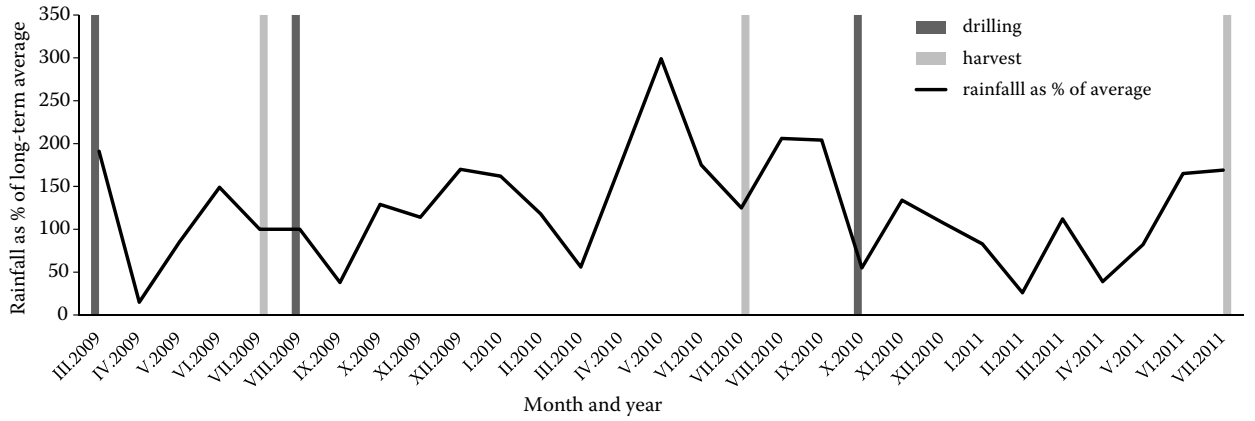


Fig. 6. Rainfall and timing of drilling and harvest operation

extremely low yields at these areas were caused by waterlogging problems in 2010 (Fig. 5) caused by extreme rainfall in 2010 (in May 2010 there was 300% rainfall compared to average), followed by problems with crop establishment in 2011 as the areas were waterlogged. The average rainfall as well as the dates of drilling and harvesting operations are given in Fig. 6.

Therefore, it can be concluded that the different performance of these areas (represented by the monitoring points) was caused by waterlogging and compaction. When an appropriate management operation would be conducted (drainage and sub-soiling), the areas would be included in the management zones No. 1 and No. 2.

Furthermore, monitoring point No. 31 was assigned to the cluster No. 2, which is characterised by an average-yielding performance along years. However, this is not characteristic for this monitoring point. The yield in 2009 and 2011 reached extremely low values at this area. Only in 2010 the yield reached values above the average. Looking at the results of analyses (Table 2), the Euclidean distance of this point within the cluster reaches the value of 192.48 so this point is the most distant point from the cluster centre. Therefore, it can be stated that there was a problem at this area. Again, waterlogging problems caused the differences, and this monitoring point should be included in management zone (cluster) 1 after an appropriate management operation is conducted at this area.

CONCLUSION

The results of these analyses showed that it is beneficial to use both the hierarchical and non-hierarchical clustering methods when determining the management zones from yield maps. The hierarchical method allows determining the statistically significant number of clusters as well as to help to interpret the data.

Also, Ward's method can be used as input information before conducting the k-means clustering method.

Further testing over a broader scope of fields and crop production systems is needed to confirm these results.

References

A tutorial on clustering algorithms. Available at http://home.dei.polimi.it/matteucc/Clustering/tutorial_html/kmeans.html

Chiericati M., Morari F., Sartori L., Ortiz B., Perry C., Vellidis G., 2007. Delineating management zones to apply site-specific irrigation in the Venice lagoon watershed. In: Stafford J.V. (ed.), Proceedings from 6th European Conference on Precision Agriculture 07 (6ECPA). Skiathos, Greece: 599–605.

Delin S., Berglund K., 2005. Management zones classified with respect to drought and waterlogging. Precision Agriculture, *6*: 321–340.

Flemming K.L., Westfall D.G., Wiens D.W., Brodahl M.C., 2000. Evaluating farmer defined management zone maps for variable rate fertilizer application. Precision Agriculture, *2*: 201–215.

Fridgen J.J., 2000. Delineation and analysis of site-specific management zones. In: Conference on Geospatial Information in Agriculture and Forestry, Lake Buena Vista, Florida, 10–12 January, 2000.

Guastaferro F., Castrignano A., De Benedetto D., Sollitto D., Troccoli A., Cafarelli B., 2010. A comparison of different algorithms for the delineation of management zones. Precision Agriculture, *11*: 600–620.

Khosla R., Westfall D.G., Reich R.M., Mahal J.S., Gangloff W.J., 2010. Spatial variation and site-specific management zones. In: Oliver M., 2010. Geostatistical Applications for Precision Agriculture. London, Springer.

King J.A., Damplney P.M.R., Lark R.M., Wheeler H.C., Bradley R.I., Mayr T.R., 2005. Mapping potential crop management zones within fields: Use of yield-map series and patterns of soil physical properties identified by electromagnetic induction sensing. Precision Agriculture, *6*: 167–181.

McQueen B.,1967. Some methods for classification and analysis of multivariate observations. In: Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability. Berkeley, University of California Press, *1*: 281–297

Minasny B., McBratney A.B., 2002. FuzME Version 3.0. Sydney, Australian Centre for Precision Agriculture. The University of Sydney.

Li Y., Shi Z., Wu Ch., Li H., Li F., 2008. Determination of potential management zones from soil electrical conductivity, yield and crop data. Journal of Zhejiang University Science B, 9: 68–76

Ortega R.A., Santibáñez O.A., 2007. Determination of management zones in corn (*Zea mays* L.) based on soil fertility. Computers and Electronics in Agriculture, *58*: 49–59.

Rataj V., Galambošová J., 2006. Farm production planning based on 4-years site-specific information. In: Agricultural Engineering for a Better World: XVI CIGR World Congress, September 3–7, 2006. Bonn, Düsseldorf, VDI Verlag GmbH: 375–376.

Russ G., Kruse R., 2011. Exploratory hierarchical clustering for management zone delineation in precision agriculture. Advances in Data Mining. Applications and Theoretical Aspects Lecture Notes in Computer Science, *6870*: 161–173.

SAS Institute Inc., 1983. SAS Technical Report A-108, Cubic Clustering Criterion. SAS Institute, Carry.

Taylor J.C., Wood G.A., Earl R., Godwin R.J., 2003. Soil factors and their influence on within-field crop variability II: Spatial analysis and determination of management zones. Biosystems Engineering, *84*: 441–453.

The Pennsylvania State University, 2004. Ward's method. Available at http://sites.stat.psu.edu/~ajw13/stat505/fa06/19_cluster/09_cluster_wards.html

Tiwari M., Misra B., 2011. Application of cluster analysis in agriculture – A review article. International Journal of Computer Applications, *36*: 43–47.

Zhang X., Shi L., Jia X., Seielstad G., Helgason C., 2010. Zone mapping application for precision-farming: A decision support tool for variable rate application. Precision Agriculture, *11*: 103–114.

Received for publication April 15, 2013
Accepted after corrections July 29, 2014

Corresponding author:

Ing. Jana Galambošová, PhD., Slovak University of Agriculture in Nitra, Faculty of Engineering, Department of Machines and Production Systems, Tr. Andreja Hlinku 2, 949 76 Nitra, Slovak Republic
phone: +421 37 6414 344, e-mail: jana.galambosova@uniag.sk