Variable selection and data clustering methods for agricultural management zones delineation

[1]Gavioli, A.; [2]Souza, E. G.; [1]Bazzi, C. L.; [1]Betzek, N. M.; [3]Schenatto, K.

[1]Federal University of Technology – Paraná (UTFPR), Medianeira, Brazil. [2]Western Paraná State University (UNIOESTE), Cascavel, Brazil. [3]Federal University of Technology – Paraná (UTFPR), Santa Helena, Brazil.

A paper from the Proceedings of the
14th International Conference on Precision Agriculture
June 24 – June 27, 2018
Montreal, Quebec, Canada

Abstract.

Delineation of agricultural management zones (MZs) is the delimitation, within a field, of a number of sub-areas with high internal similarity in the topographic, soil and/or crop characteristics. This approach can contribute significantly to enable precision agriculture (PA) benefits for a larger number of producers, mainly due to the possibility of reducing costs related to the field management. Two fundamental tasks for the delineation of MZs are the variable selection and the cluster analysis. There are several methods proposed to execute them, but due to their complexity, they need to be run by computer systems. In this context, the objective of this paper is to present two computational modules developed to enable an efficient execution of these tasks. The variable selection module provides 5 algorithms, based on spatial correlation analysis of crop and field variables, principal component analysis (PCA), and multivariate spatial analysis based on Moran's index and PCA (MULTISPATI-PCA). The data clustering module provides 17 clustering algorithms: average linkage, bagged clustering, centroid linkage, clustering large applications, complete linkage, fuzzy analysis clustering, fuzzy c-means, hard competitive learning, hybrid hierarchical clustering, k-means, McQuitty's method, median linkage, neural gas, partitioning around medoids, spherical k-means, unsupervised fuzzy competitive learning, and Ward's method. The algorithms were programmed in R statistical software routines, with the main objective of guaranteeing flexibility and speed in execution. This software was tested for the delineation of management zones for several agricultural fields. However, to exemplify its use in this paper, we consider data obtained from 2012 to 2015 in an agricultural area in the municipality of Céu Azul, Brazil, where soybean crop was cultivated. The computational modules developed proved to be adequate and efficient to define MZs. In addition, they are more comprehensive than other free-to-use software in terms of the diversity of variable selection and data clustering methods.

Keywords. *precision agriculture, principal component analysis, software for agriculture.*

Introduction

Several data clustering methods can be applied in the process of delineating management zones (MZs), and they can use a large number of variables in this process. However, to obtain better quality MZs, it is important to preselect the really necessary variables. Because variable selection and cluster analysis are complex execution activities, they depend on the use of specific software. In this context, the objective of this work is to present two computational modules developed to enable variable selection and class definition tasks, for the MZs delineation. As an advantage in relation to other software of similar purpose, these modules provide five methods for variable selection based on principal component analysis (PCA), MULTISPATI-PCA (Dray et al. 2008), and analysis of spatial correlation between variables (Reich et al. 1994), as well as 17 data clustering methods.

Material and Methods

The five algorithms for variable selection and the 17 clustering algorithms were programmed in routines of the R statistical software. The algorithms implemented in the variable selection module were suggested by Gavioli et al. (2016), and were called PCA-All, MPCA-All, PCA-SC, MPCA-SC, and Spatial-Matrix. In the data clustering module were implemented 17 methods: average linkage, bagged clustering, centroid linkage, clustering large applications, complete linkage, fuzzy analysis clustering, fuzzy c-means, hard competitive learning, hybrid hierarchical clustering, k-means, McQuitty's method, median linkage, neural gas, partitioning around medoids, spherical k-means, unsupervised fuzzy competitive learning, and Ward's method. In order to evaluate the performance of the clustering methods in relation to the quality of the classes generated, the ANOVA (Tukey's test), the variance reduction index (VR), and the average silhouette coefficient (ASC) were implemented in the clustering module. The R software packages denoted geoR, gstat, ade4 and spdep were employed to apply the PCA and the MULTISPATI-PCA-based methods. The cclust, cluster, e1071, fastcluster, fclust, hybridHclust, optpart, and skmeans packages were employed to enable the execution of the clustering algorithms. To exemplify the operation of the two modules, data collected between 2012 and 2015 from a 15.5 ha agricultural area (municipality of Céu Azul, Brazil) were used, where soybean was cultivated. For this area, a sample grid with 40 points was generated and the variables altitude, soil texture, slope, density, and soil penetration resistance were determined.

Results and Discussion

When executing MPCA-All, MPCA-SC, PCA-All or PCA-SC in the variable selection module, the principal components (PCs) generated are presented with the respective values of the percentage representation of the original data variance and the cumulative percentage of that variance. When the user chooses MPCA-SC or PCA-SC, original variables which have significant spatial correlation with the yield variable (used to validate the generated MZs) are automatically highlighted by the software, and then only those variables are used for the PCs generation. This module also allows users to simultaneously select up to four methods of PCs generation to run with the same input data, thus enabling a comparison that determines which one provides the best results. The comparison criterion is the largest percentage representation of the original data variance in the first PCs. For the example considered, MPCA-All, MPCA-SC, PCA-All, and PCA-SC were performed, and MPCA-SC achieved the best performance (Table 1).

Table 1. Percentage representation of the variance of the original variables corresponding to the first and second principal components obtained by the MPCA-SC method.

Principal component	Percentage of the original variance	Cumulative percentage of the original variance
First PC	71	71
Second PC	29	100

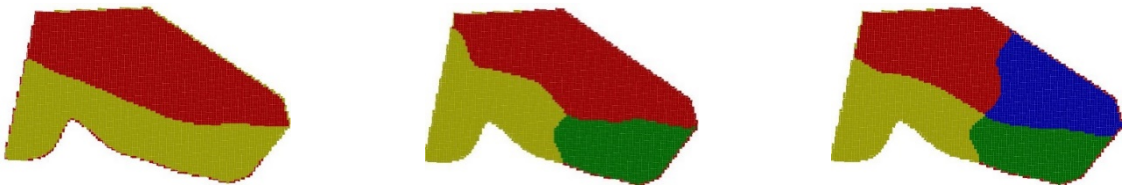
In the clustering module, one of the 17 clustering algorithms and the variables of interest (original

area variables or PCs) can be selected to be used to define two or more MZs. This module also allows users to define specific parameters of the selected clustering method, but to facilitate the use commonly used values for such parameters are automatically suggested. In the example considered, the K-means algorithm was selected to generate two, three, and four classes, from the values interpolated by ordinary kriging of the first two PCs (obtained using MPCA-SC method). The maximum number of 500 iterations was used to execute the K-means algorithm. After execution, this module displays the performance of the algorithm for each number of classes. The Tukey test (5% significance level) showed that it is possible to divide the area into only two MZs with statistically distinct productive potentials (Table 2). This division promoted a satisfactory reduction of the yield variance, because VR = 33.8%, with subareas that presented high internal homogeneity (ASC = 0.59).

Table 2. Performance of the K-means clustering algorithm to generate two, three, and four classes.

Number of classes	ANOVA (Tukey's test)	VR(%)	ASC
2	a b	33.8	0.59
3	a b a	23.8	0.46
4	a a b b	35.8	0.39

After completing the K-means execution for this case study, the SDUM (Software for Defining Management Zones) software (Bazzi et al. 2013) was used to generate thematic maps corresponding to two, three, and four MZs (Fig. 1).



2 MZs 3 MZs 4 MZs

Fig 1. Thematic maps representing two, three, and four management zones for the 15.5 ha agricultural area.

Conclusion

The developed computational modules provide flexibility through the provision of five variable selection and 17 clustering algorithms. These functionalities are not found in the FuzME (Minasny and McBratney 2002) and MZA (Fridgen et al. 2004) software, which are widely used in the MZs definition process.

References

Bazzi, C. L., Souza, E. G., Uribe-Opazo, M. A., Nóbrega, L. H. P., & Rocha, D. M. (2013). Management zones definition using soil chemical and physical attributes in a soybean area. *Agricultural Engineering*, 33(5), 952-964.

Dray, S., Saïd, S., & Débias, F. (2008). Spatial ordination of vegetation data using a generalization of Wartenberg's multivariate spatial correlation. *Journal of Vegetation Science,* 19(1), 45-56.

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

Gavioli, A., Souza, E. G., Bazzi, C. L., Guedes, L. P. C., & Schenatto, K. (2016). Optimization of management zone delineation by using spatial principal components *Computers and Electronics in Agriculture,* 127, 302-310.

Minasny, B., & Mcbratney, A. B. (2002). *FuzME 3.0*. Sydney, AU: The University of Sydney.

Reich, R. M., Czaplewski, R. L., & Bechtold, W. A. (1994). Spatial cross-correlation of undisturbed, natural shortleaf pine stands in northern Georgia. *Environmental and Ecological Statistics,* 1, 201-217.