Computers and Electronics in Agriculture 127 (2016) 302–310

Contents lists available at ScienceDirect

Computers and Electronics in Agriculture

journal homepage: www.elsevier.com/locate/compag

Original papers

Optimization of management zone delineation by using spatial principal components

Alan Gavioli [a,b], Eduardo Godoy Souza [b,*], Claudio Leones Bazzi [a], Luciana Pagliosa Carvalho Guedes [b], Kelyn Schenatto [c]

[a] Department of Computer Science, Federal University of Technology of Paraná (UTFPR), Medianeira, Paraná, Brazil
[b] PGEAGRI, Technological and Exact Sciences Center, State University of West Paraná (UNIOESTE), Cascavel, Paraná, Brazil
[c] Department of Computer Science, Federal University of Technology of Paraná (UTFPR), Santa Helena, Paraná, Brazil

ARTICLE INFO

Article history:
Received 18 February 2016
Received in revised form 6 May 2016
Accepted 23 June 2016

Keywords:
Fuzzy C-means
Moran's index
MULTISPATI-PCA
PCA
Precision agriculture

ABSTRACT

Definition of management zones is the delimitation of sub-areas with similar topographic, soil and crop characteristics within a field. Among the many variables that can be used for this definition, those that are stable and spatially correlated with yield are more often recommended for use. Clustering algorithms such as Fuzzy C-means are also frequently applied to define management zones. Three variable selection techniques that can be applied with Fuzzy C-means are spatial correlation analysis, principal component analysis (PCA), and multivariate spatial analysis based on Moran's index PCA (MULTISPATI-PCA). In this study, the efficiency of each of these three techniques used in conjunction with the clustering method was assessed. Furthermore, a new variable selection approach, named MPCA-SC, based on the combined use of Moran's bivariate spatial autocorrelation statistic and MULTISPATI-PCA, was proposed and tested. The evaluation was performed by using data collected from 2010 to 2014 from three agricultural areas in Paraná State, Brazil, with corn and soybean crops, generating two, three, and four classes. The delineated management zones were different according to the method used, and MPCA-SC provided the best performance for the Fuzzy C-means algorithm and the best variance reduction values of the data after the delimitation of the sub-areas. Furthermore, MPCA-SC provided management zones with greater internal homogeneity, making them more viable for implementation from the viewpoint of field operations.

1. Introduction

Management zones (MZs) are defined as the delimitation of sub-areas within a field. Such definition allows for these sub-areas to be uniformly managed. A MZ shows similar characteristics of soil and topography, and therefore, require similar amounts of agricultural supplies (Moral et al., 2010; Schepers et al., 2004). This delineation can contribute significantly to enable precision agriculture for a larger number of producers, because the homogeneous rate in each sub-area enables the use of conventional agricultural machines.

The MZs can also represent indicators for soil and planted crops sampling, reducing the number of samples to be analyzed without compromising on the reliability of the results. Yield data, chemical and physical data of the soil, topographic data and data on the apparent electrical conductivity of the soil, vegetation indexes, and combinations of these data, may be used to define MZs (Fraisse et al., 2001). However, it is recommended that stable variables (attributes) correlated with yield be used for delimiting the sub-areas (Doerge, 2000). This is so because the variables used for the definition are intended to be used for several years; hence, chemical attributes are eliminated. For this process of delimitation, is also customary to employ clustering algorithms such as Fuzzy C-means, also known as Fuzzy K-means (Fridgen et al., 2004; Fu et al., 2010; Hornung et al., 2006; Li et al., 2013; Zhang et al., 2013).

Weighting and selection of variables are difficult tasks in cluster analysis. The capacity of cluster software to process a large number of variables tends to encourage users to use many in this process. However, one should be aware that the choice of variables and that of the weights assigned to them often influence the determination of clusters (Gnanadesikan et al., 1995).

Three variable selection techniques that can be applied in combination with the Fuzzy C-means algorithm are as follows: spatial correlation analysis (Reich, 2008; Schepers et al., 2004), applied as

* Corresponding author at: Rua Universitária, 1619, Cascavel, Paraná CEP: 85819-110, Brazil.
E-mail addresses: alan@utfpr.edu.br (A. Gavioli), eduardo.souza@unioeste.br (E.G. Souza), bazzi@utfpr.edu.br (C.L. Bazzi), luciana_pagliosa@hotmail.com (L.P.C. Guedes), kelynschenatto@gmail.com (K. Schenatto).

http://dx.doi.org/10.1016/j.compag.2016.06.029
0168-1699/

described by Bazzi et al. (2013) and Schenatto et al. (2016); principal component analysis (PCA) (Hotelling, 1933), used by Fraisse et al. (2001), Li et al. (2007), Moral et al. (2010), and Cohen et al. (2013); and multivariate spatial analysis based on Moran's index PCA (MULTISPATI-PCA) (Dray et al., 2008), applied by Córdoba et al. (2013, 2016), and Peralta et al. (2015).

For spatial correlation analysis, Moran's bivariate spatial autocorrelation statistic (Ord, 1975) is used to evaluate whether the variables have correlation and spatial autocorrelation. Thereafter, the variables without spatial dependence, those with no correlation with yield, and redundant variables are eliminated.

PCA is a multivariate analysis technique that allows identifying the variables that account for most of the total variance in data sets. When using PCA, a new set of synthetic variables named principal components (PCs), which are uncorrelated among themselves and commonly denoted as linear combinations of the original variables, are obtained from the original variables through some transformations (Johnson and Wichern, 2007).

MULTISPATI-PCA aims to add a spatial restriction on the traditional PCA, enabling it to be executed considering the existence of spatial dependence in sets of georeferenced data. This technique relies on introducing a spatial weighting matrix, which is constructed using Moran's bivariate spatial autocorrelation statistic, to the PCA. Its advantage over the PCA is that the scores obtained with MULTISPATI-PCA maximize the spatial autocorrelation between points, while those obtained with PCA maximize the total variance (Córdoba et al., 2013; Dray et al., 2008).

Therefore, the scores generated with MULTISPATI-PCA show strong spatial structures in the first PCs, while the PCA scores may show spatial structures in any component, even in the last, which in practice are generally disregarded (Arrouays et al., 2011).

The aim of this study was to evaluate the efficiency of spatial correlation analysis, PCA, and MULTISPATI-PCA techniques, when used jointly with the Fuzzy C-means algorithm to define MZs. In addition, a new approach of generating synthetic variables for defining MZs, based on the joint use of spatial correlation analysis and MULTISPATI-PCA, was proposed and assessed.

2. Materials and methods

2.1. Data sets

Data collected between 2010 and 2014 from three commercial agricultural areas with corn and soybean crops (Fig. 1), located in Paraná State, Brazil, were used. The soils were classified as typical dystroferric Red Latosol (Embrapa, 2006) and grown in a no-till system. Field A extends for 15 ha, and is located in the municipality of Céu Azul (central geographical location 25°06′32″S and 53°49′55″W, and an average elevation of 460 m). Field B extends for 9.9 ha, and is located in the municipality of Serranópolis do Iguaçu (central geographical location 25°24′28″S and 54°00′17″W, and an average elevation of 355 m). Field C extends for 19.8 ha, and is located in the municipality of Cascavel (central geographical location 24°57′08″S and 53°33′59″W, and an average elevation of 650 m).

Table 1
Variables collected by year, for each experimental area.

Variable (attribute)	Field A			Field B			Field C	
	2012	2013	2014	2012	2013	2014	2010	2011
SPR 0–0.1 m (MPa)	X	X	X	X	X	X	X	
SPR 0.1–0.2 m (MPa)	X	X	X	X	X	X	X	
SPR 0.2–0.3 m (MPa)	X	X	X	X	X	X	X	
pH	X			X			X	
Elevation (m)	X			X			X	
Slope (°)	X						X	
Density (g cm^{-3})	X						X	
Sand (%)	X			X			X	
Silt (%)	X			X			X	
Clay (%)	X			X			X	
OM (%)	X			X				
Soybean yield (t ha^{-1})	X	X	X	X	X	X	X	X
Corn yield (t ha^{-1})					X	X		

SPR: soil penetration resistance; OM: organic matter.

Only those variables considered stable (Table 1) were used for defining the classes, to meet the recommendation of Doerge (2000). Irregular sampling grids were used to assign 40 (2.67 points ha^{-1}), 42 (4.24 points ha^{-1}), and 68 (3.43 points ha^{-1}) sample points to areas A, B, and C, respectively, with the sampling points located in the central imaginary line between the contours present in each area.

Soil samples were collected at depths of 0–0.2 m. The soil penetration resistance (SPR) was determined for the depths 0–0.1 m, 0.1–0.2 m, and 0.2–0.3 m, using an electronic meter of soil compaction Falker PenetroLOG PLG1020. The data of elevation of the three areas were obtained using an electronic total station of high precision Topcon GPT-7505, and subsequently, the slopes were calculated depending on the elevation of the sampling points.

Soybean yield data for area A was determined by means of a harvesting monitor attached to a CASE IV harvester. As for areas B and C, yield was determined by hand harvesting of a 1 m^2 sample area in each of the sample points. In all cases, yield values were corrected to 13% water content.

To meet the requirement of stability of the yield data, which is normally heavily influenced by climate and rainfall, the data of soybean yield for the three areas, and data of corn yield for area B, were standardized through the standard score technique (Eq. (1); Larscheid and Blackmore, 1996). Then, the arithmetic average of the standardized values of available years was calculated, generating a single variable corresponding to the average of standard yield.

$$P_{iN} = \frac{(P_i - \overline{P})}{S} \tag{1}$$

where P_{iN} is the standardized value for the sample point i; P_i is the original value of the sample point i; $\overline{P}$ corresponds to the arithmetic

Fig. 1. The three experimental areas: field A: Céu Azul, Paraná, Brazil; field B: Serranópolis do Iguaçu, Paraná, Brazil; field C: Cascavel, Paraná, Brazil.

average of all the original values of the points to be standardized; and S corresponds to the standard deviation of the original values.

2.2. Variable selection

Six approaches for selecting variables for defining MZs were compared:

(1) All-Attrib: no disposal of stable variables.
(2) Spatial-Matrix: after calculating Moran's bivariate spatial autocorrelation statistic (Eq. (2); Czaplewski and Reich, 1993) among all the variables by using the software for management zones definition SDUM (Bazzi et al., 2013), variables were selected by the procedure proposed by Bazzi et al. (2013): (a) elimination of variables with no significant spatial autocorrelation at 95% significance; (b) removal of the variables that were not correlated with yield; (c) decreasing ordination of the remaining variables, considering the degree of correlation with yield; and (d) elimination of variables which are correlated with each other, with preference to the withdrawal of those variables with lower correlation with yield.

$$I_{XY} = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} W_{ij} * X_i * Y_j}{W\sqrt{m_X^2 * m_Y^2}} \quad (2)$$

where W_{ij} is the spatial association matrix, calculated by $W_{ij} = (1/(1 + D_{ij}))$; D_{ij} is the distance between points i and j; X_i is the value of variable X transformed, at point i; Y_j is the value of the variable Y transformed, at point j; W corresponds to the sum of the degrees of spatial association, obtained from the W_{ij} matrix, for $i \neq j$; m_X^2 corresponds to the sample variance of X; and m_Y^2 corresponds to the sample variance of Y. Note that the transformation of a variable Z should be interpreted as the procedure performed on their values so that it is on an average equal to zero, applying the equation $Z_k = (z_k - \overline{Z})$, wherein $\overline{Z}$ is the sample mean of Z;
(3) PCA-All (traditional PCA): calculation of PCs from all stable variables, such that the amount of PCs selected was based on the criterion of representation of at least 70% of the total variability of the data associated with the original variables (Johnson and Wichern, 2007).
(4) MPCA-All (traditional MULTISPATI-PCA): calculation of spatial principal components (SPCs) from all stable variables, such that the amount of SPCs selected was also based on the criterion of representation of at least 70% of the total variability of the original data.
(5) PCA-SC: PCA with same parameters as in approach (3), however, applied only on the stable variables that showed significant spatial correlation with the yield of each area.
(6) MPCA-SC (the new approach developed): MULTISPATI-PCA with same parameters as in approach (4), but also applied only on the stable variables that were significantly correlated with the yield of each area.

The PCA-All, MPCA-All, PCA-SC, and MPCA-SC approaches were applied to the data of each area by developing a routine in the statistical software R (R Core Team, 2014), including the packages *geoR*, *gstat*, *ade4* (Chessel et al., 2004), and *spdep* (Bivand, 2012). The package *spdep* provided the function *dnearneigh* in order to identify the neighbors of each sample point (required by MPCA-All and MPCA-SC). This function uses euclidean distance to compute the distance of a point from another and returns a list of neighbors for each point, based on the value set to neighborhood radius. This distance was determined experimentally for each location: area A, 240 m radius; area B, 120 m radius; and area C, 200 m radius.

The object-relational database system PostgreSQL 9.0.5, maintained by the PostgreSQL Global Development Group, was used for data storage. The software PostGIS 1.5.5, a spatial database extender for PostgreSQL maintained by the PostGIS Project Steering Committee, was also applied. Furthermore, the software pgAdmin III, maintained by the pgAdmin Development Team, was used for managing the databases that were created.

2.3. Interpolation and definition of MZs

Data interpolation in advance is important for generating MZs with smoother contours and for greater reduction in data variance (Schenatto et al., 2016). The authors found that the kriging interpolator had the best performance, but the advantage of using this interpolator over the inverse square distance method was small. Further, the software SDUM has the limitation that it cannot interpolated by kriging, but it is the only one free software that can both interpolate and define MZs. Because of this, data of the selected variables were interpolated using the inverse square distance method with pixels in an area of 5 × 5 m and 10 neighbors. After interpolation, resulting data were used as the input for the Fuzzy C-means algorithm, considering error parameter equals to 0.0001 and weight index equals to 1.3, thus generating two, three, and four classes. For interpolating data, defining classes, and delineating MZ maps, the software SDUM was used. For All-Attrib and Spatial-Matrix approaches, variables were standardized before interpolation (Eq. (3); Mielke Jr. and Berry, 2007), with the objective of maintaining the same data range, regardless of the used variable.

$$P_{in} = \frac{P_i - Median}{Range} \quad (3)$$

where P_i is the value of the pixel i to be standardized, and P_{in} is the standardization result.

2.4. Evaluation of MZs

The performance of the variable selection approaches was assessed using six indexes:

(1) Variance Reduction (VR) (Li et al., 2007; Ping and Dobermann, 2003): is calculated for the standardized average yield, with the expectation that the sum of the variances of the data from MZs generated is smaller than the total variance (Eq. (4)).

$$VR = \left(1 - \frac{\sum_{i=1}^{c} W_i * V_{mz_i}}{V_{field}}\right) * 100 \quad (4)$$

where c is the number of MZs; W_i is the proportion of the area of i-th MZ to the total area; V_{mz_i} is the data variance of the i-th MZ; and V_{field} is the data variance corresponding to the area as a whole.
(2) Fuzziness Performance Index (FPI) (Fridgen et al., 2004): it allows determining the degree of separation between the fuzzy c groups generated from a data set. FPI varies between 0 and 1, such that the closer this value to 0, the lower is the degree of sharing of elements among the generated groups (Eq. (5)).

$$FPI = 1 - \frac{c}{(c-1)}\left[1 - \sum_{j=1}^{n}\sum_{i=1}^{c}(m_{ij})^2/n\right] \quad (5)$$

where c is the number of groups; n is the number of elements in the data set; and m_{ij} is the element of the fuzzy pertinence matrix $\boldsymbol{M}$.

(3) Modified Partition Entropy (MPE) (Boydell and McBratney, 2002): it is an estimate of the level of difficulty of organization of c groups, such that the closer the value to 0, the lower is the difficulty of organizing groups (Eq. (6)).

$$MPE = \frac{-\sum_{j=1}^{n}\sum_{i=1}^{c} m_{ij}\log(m_{ij})/n}{\log c} \tag{6}$$

where c is the number of groups; n is the number of elements in the data set; and m_{ij} is the element of the fuzzy pertinence matrix $\boldsymbol{M}$.

(4) Smoothness Index (SI): it gives the pixel-by-pixel frequency of change of classes in a thematic map in the horizontal and vertical directions and along the diagonal (Eq. (7)). It also characterizes the smoothness of the boundary curves of the MZs. If a map has a completely homogeneous area, the result is SI equals to 100% because of lack of changes in class. On the other hand, if the map is completely generated with random values, the SI will have a value close to 0.

$$SI = 100 - \left(\left(\frac{\sum_{i=1}^{k} NM_{Hi}}{4P_H} + \frac{\sum_{j=1}^{k} NM_{Vj}}{4P_V} + \frac{\sum_{l=1}^{k} NM_{Ddl}}{4P_{Dd}} + \frac{\sum_{m=1}^{k} NM_{Dem}}{4P_{De}}\right) * 100\right) \tag{7}$$

where NM_{Hi} is the number of changes in row i (horizontal); NM_{Vj} is the number of changes in column j (vertical); NM_{Ddl} is the number of changes in diagonal l (right diagonal Dd); NM_{Dem} is the number of changes in diagonal m (left diagonal De); k is the maximum number of pixels in a row, column, or diagonal; P_H is the possibility of changes in horizontal pixels; P_V is the possibility of changes in vertical pixels; P_{Dd} is the possibility of changes in the right diagonal Dd; and P_{De} is the possibility of changes in the left diagonal De .

(5) Analysis of Variance (ANOVA): the yield values were compared between classes by using the normalized average yield, and performing the Tukey's range test to identify whether the generated classes showed significant differences in normalized average yield (first, we confirmed that there was no spatial dependence within each class).

(6) Improved Cluster Validation Index (ICVI): based on the CVI index (Schenatto et al., 2016), the ICVI index is proposed in this work (Eq. (8)) to solve a possible problem when the estimates for FPI, MPE, and VR did not indicate similar methods to the definition of MZs. ICVI lies between 0 and 1, such that the greater the value of VR and lower the values of the FPI and the MPE, the closer will the ICVI be to 0. In a comparison between n clustering methods, the best method is the one with the lowest $ICVI_i$.

$$ICVI_i = \frac{1}{3} * \left(\frac{FPI_i}{Max\{FPI\}} + \frac{MPE_i}{Max\{MPE\}} + \left(1 - \frac{VR_i}{Max\{VR\}}\right)\right) \tag{8}$$

where FPI_i is the FPI value of the i-th variable selection method; MPE_i is the MPE value of the i-th variable selection method; VR_i is the VR value of the i-th variable selection method; and $Max\{Index\_X\}$ represents the maximum value of the $Index\_X$ index among the n variable selection methods.

3. Results and discussion

3.1. Variables selected

The variables selected for defining the classes and the values of Moran's bivariate spatial autocorrelation statistic, between each

Table 2
Variables selected by each of the six approaches, and Moran's index with the normalized average yield.

Field	Variables	MI with NAY	SA	CY	NR	Variable selection approaches					
						All-Attrib	Spatial-Matrix	PCA-All	MPCA-All	PCA-SC	MPCA-SC
A	SPR 0–0.1 m (MPa)	−0.053*	Y	Y	Y	Y	Y	Y	Y	Y	Y
	SPR 0.1–0.2 m (MPa)	−0.017	N	N	N	Y	N	Y	Y	N	N
	SPR 0.2–0.3 m (MPa)	−0.022	N	N	N	Y	N	Y	Y	N	N
	pH	−0.034*	N	Y	N	Y	N	Y	Y	Y	Y
	Elevation (m)	0.100*	Y	Y	Y	Y	Y	Y	Y	Y	Y
	Slope (°)	−0.016	N	N	N	Y	N	Y	Y	N	N
	Density (g cm^{-3})	0.023	N	N	N	Y	N	Y	Y	N	N
	Sand (%)	−0.075*	Y	Y	N	Y	N	Y	Y	Y	Y
	Silt (%)	0.028	N	N	N	Y	N	Y	Y	N	N
	Clay (%)	−0.040*	Y	Y	N	Y	N	Y	Y	Y	Y
B	SPR 0–0.1 m (MPa)	0.039*	Y	Y	Y	Y	Y	Y	Y	Y	Y
	SPR 0.1–0.2 m (MPa)	0.044*	N	Y	N	Y	N	Y	Y	Y	Y
	SPR 0.2–0.3 m (MPa)	−0.014	N	N	N	Y	N	Y	Y	N	N
	pH	−0.029*	N	Y	N	Y	N	Y	Y	Y	Y
	Elevation (m)	0.051*	Y	Y	Y	Y	Y	Y	Y	Y	Y
	Sand (%)	0.007	N	N	N	Y	N	Y	Y	N	N
	Silt (%)	−0.013	Y	N	N	Y	N	Y	Y	N	N
	Clay (%)	0.012	Y	N	N	Y	N	Y	Y	N	N
	OM (%)	−0.037*	N	Y	N	Y	N	Y	Y	Y	Y
C	SPR 0–0.1 m (MPa)	−0.002	N	N	N	Y	N	Y	Y	N	N
	SPR 0.1–0.2 m (MPa)	0.114*	Y	Y	N	Y	N	Y	Y	Y	Y
	SPR 0.2–0.3 m (MPa)	0.102*	Y	Y	N	Y	N	Y	Y	Y	Y
	pH	0.024	N	N	N	Y	N	Y	Y	N	N
	Elevation (m)	0.137*	Y	Y	Y	Y	Y	Y	Y	Y	Y
	Slope (°)	0.011	Y	N	N	Y	N	Y	Y	N	N
	Density (g cm^{-3})	−0.029	N	N	N	Y	N	Y	Y	N	N
	Sand (%)	0.078*	Y	Y	N	Y	N	Y	Y	Y	Y
	Silt (%)	0.021	N	N	N	Y	N	Y	Y	N	N
	Clay (%)	−0.082*	Y	Y	N	Y	N	Y	Y	Y	Y

* Significative value; SPR: soil penetration resistance; OM: organic matter; MI: Moran's bivariate index; NAY: normalized average yield; SA: spatial autocorrelation; CY: correlation with yield; NR: not redundant; Y: yes; N: no.

variable and the normalized average yield value, are listed in Table 2. Because the values are not standardized, even small values of Moran's index can be statistically significant. In this case, the values are important if the statistic is significant at 0.05 level.

It was found that elevation was the variable with a strong spatial correlation with normalized average yield in all three fields. These findings agree with those of Jaynes et al. (2005) and Peralta et al. (2013), which suggests that there is a spatial association between this variable and yield of soybeans and corn.

According to the criterion of spatial correlation matrix used in the Spatial-Matrix approach, the variables selected for areas A and B were elevation and SPR 0–0.1 m, while for the area C, only elevation was selected.

3.2. Creation of the principal components

As expected, when considering all stable variables for obtaining the PCs, the necessary number of components was higher than when only those variables with significant spatial correlation with normalized average yield were selected (variables with value equals to "Y" in Table 2, for PCA-SC and MPCA-SC). This suggests that variables spatially uncorrelated to yield can disrupt the construction of the components, both in the case of PCA-All and MPCA-All. Comparison of the four approaches based on PCA or MULTISPATI-PCA showed that MPCA-SC had the best performance in reducing the dimensionality of data without significant loss of information, and therefore, MPCA-SC ensured the highest cumulative percentage representation of the original variance with smaller number of PCs in the three areas (Table 3). This is because with MPCA-SC, only two SPCs are required for each field, while the other techniques required up to five components.

When comparing PCA-All and MPCA-All, or PCA-SC and MPCA-SC, from the viewpoint of variance and spatial autocorrelation (Table 3), the first spatial component (SPC1) had lower variance and higher spatial autocorrelation than the first component (PC1), in the three fields. This indicates that the spatial autocorrelation indexes increased with the use of MULTISPATI-PCA. Therefore, this technique facilitated the selection of principal components needed for definition of MZs in the fields. Similar results were obtained by Córdoba et al. (2012,2013); in their studies, although they have not reported an approach similar to MPCA-SC, they applied PCA-All and MPCA-All to the variables elevation, SPR, apparent electrical conductivity of the soil, and soybean and wheat yield, in agricultural areas in Argentina.

In the analysis of the coefficients of PCs and SPCs, which act as weights for the original variables in that components (Tables 4–6), the first component (PC1 or SPC1) had higher weighting coefficients, in absolute values, for the variables as follows: elevation and clay to field A; elevation and SPR 0–0.1 m to field B; elevation, clay, and SPR 0.1–0.2 m to field C.

Table 3
Statistics of the principal components for PCA-All, MPCA-All, PCA-SC, and MPCA-SC.

Field	Component	Variance	Percentage	Sum of percentages	Moran's index
A	PCA-All				
	PC1	2.98	27	27	0.23
	PC2	2.57	23	50	0.15
	PC3	1.50	14	64	−0.05
	PC4	1.15	10	74	−0.05
	MPCA-All				
	SPC1	2.81	53	53	0.29
	SPC2	2.45	47	100	0.15
	PCA-SC				
	PC1	2.94	49	49	0.22
	PC2	1.27	21	70	0.09
	MPCA-SC				
	SPC1	2.77	71	71	0.25
	SPC2	1.11	29	100	0.13
B	PCA-All				
	PC1	3.20	32	32	0.01
	PC2	1.93	19	51	0.01
	PC3	1.33	13	64	0.07
	PC4	1.18	12	76	0.03
	MPCA-All				
	SPC1	1.66	35	35	0.19
	SPC2	1.50	32	67	0.11
	SPC3	0.68	15	82	0.08
	PCA-SC				
	PC1	2.56	43	43	0.03
	PC2	1.34	22	65	0.11
	PC3	0.92	15	80	−0.05
	MPCA-SC				
	SPC1	1.67	61	61	0.19
	SPC2	0.64	23	84	0.05
C	PCA-All				
	PC1	3.44	31	31	0.34
	PC2	1.40	13	44	0.03
	PC3	1.27	12	56	0.22
	PC4	1.10	10	66	−0.02
	PC5	0.99	9	75	0.03
	MPCA-All				
	SPC1	3.07	48	48	0.44
	SPC2	1.31	21	69	0.24
	SPC3	1.14	18	87	0.06
	PCA-SC				
	PC1	2.87	48	48	0.62
	PC2	1.12	19	67	0.36
	PC3	0.98	16	83	0.10
	MPCA-SC				
	SPC1	2.63	68	68	0.65
	SPC2	1.21	32	100	0.46

Table 4
Weights for the variables in the PCs and SPCs, for field A.

Variables	Elevation	SPR 0–0.1	pH	Clay	Sand	Silt	Slope	Density	SPR 0.1–0.2	SPR 0.2–0.3
PCA-All										
PC1	0.49	−0.26	−0.39	0.45	−0.49	−0.07	−0.12	0.07	−0.05	0.01
PC2	0.07	−0.41	0.12	−0.30	−0.04	0.46	0.04	−0.10	−0.49	−0.48
PC3	0.25	−0.03	0.23	−0.20	−0.07	0.35	0.49	0.43	0.38	0.25
PC4	−0.16	−0.38	−0.18	0.07	−0.02	−0.08	0.62	−0.58	0.09	0.19
MPCA-All										
SPC1	0.53	0.02	−0.26	0.44	−0.49	−0.18	−0.22	0.20	0.15	0.22
SPC2	0.45	−0.48	0.01	−0.14	−0.17	0.38	0.14	0.05	−0.29	−0.46
PCA-SC										
PC1	0.50	−0.29	−0.38	0.43	−0.50					
PC2	0.08	−0.55	0.24	−0.47	0.16					
MPCA-SC										
SPC1	0.56	−0.07	−0.26	0.52	−0.54					
SPC2	−0.39	0.72	−0.13	0.52	−0.01					

Table 5
Weights for the variables in the PCs and SPCs, for field B.

Variables	Elevation	SPR 0–0.1	SPR 0.1–0.2	OM	pH	SPR 0.2–0.3	Sand	Clay	Silt
PCA-All									
PC1	−0.36	−0.29	−0.43	0.43	0.22	−0.29	−0.26	−0.31	0.34
PC2	0.10	0.19	0.19	−0.20	−0.42	0.22	−0.30	−0.51	0.53
PC3	0.37	−0.21	−0.03	−0.14	0.32	0.21	0.07	−0.22	0.18
PC4	0.37	0.61	0.04	0.16	0.30	−0.51	0.24	−0.18	0.15
MPCA-All									
SPC1	−0.76	0.07	−0.31	0.27	−0.02	0.07	−0.12	−0.01	0.07
SPC2	−0.09	0.04	0.44	−0.10	−0.18	0.85	−0.13	−0.04	0.12
SPC3	0.18	−0.14	−0.16	0.53	−0.40	0.22	0.59	0.11	−0.27
PCA-SC									
PC1	−0.43	−0.43	−0.51	0.49	0.35				
PC2	0.42	−0.10	0.02	−0.11	0.46				
PC3	0.27	−0.65	−0.08	−0.43	−0.54				
MPCA-SC									
SPC1	−0.76	0.05	−0.35	0.30	−0.02				
SPC2	−0.43	−0.09	−0.05	−0.57	0.60				

Table 6
Weights for the variables in the PCs and SPCs, for field C.

Variables	Elevation	SPR 0.1–0.2	SPR 0.2–0.3	Clay	Sand	SPR 0–0.1	Silt	Slope	Density	pH
PCA-All										
PC1	−0.40	−0.45	−0.42	0.27	−0.30	0.01	−0.24	0.04	0.39	−0.06
PC2	0.25	−0.19	−0.22	−0.01	−0.13	−0.63	0.07	0.42	0.05	−0.08
PC3	0.11	0.20	0.19	0.48	−0.51	−0.09	−0.03	−0.35	0.10	0.48
PC4	−0.06	0.29	0.34	−0.16	−0.14	−0.17	−0.49	−0.18	0.36	−0.55
PC5	−0.11	0.01	0.10	0.39	0.12	−0.41	0.54	−0.31	−0.05	−0.42
MPCA-All										
SPC1	−0.41	−0.32	−0.30	0.53	−0.36	0.07	−0.22	−0.15	0.27	−0.01
SPC2	−0.38	−0.37	−0.31	−0.66	0.27	0.01	0.03	0.15	0.11	−0.08
SPC3	0.03	−0.17	−0.12	0.09	−0.25	0.11	0.07	0.85	0.09	−0.04
PCA-SC										
PC1	−0.43	−0.52	−0.49	0.32	−0.30					
PC2	−0.16	−0.20	−0.20	−0.58	0.66					
PC3	0.41	−0.39	−0.49	−0.24	−0.05					
MPCA-SC										
SPC1	−0.44	−0.35	−0.32	0.58	−0.39					
SPC2	−0.39	−0.37	−0.31	−0.66	0.28					

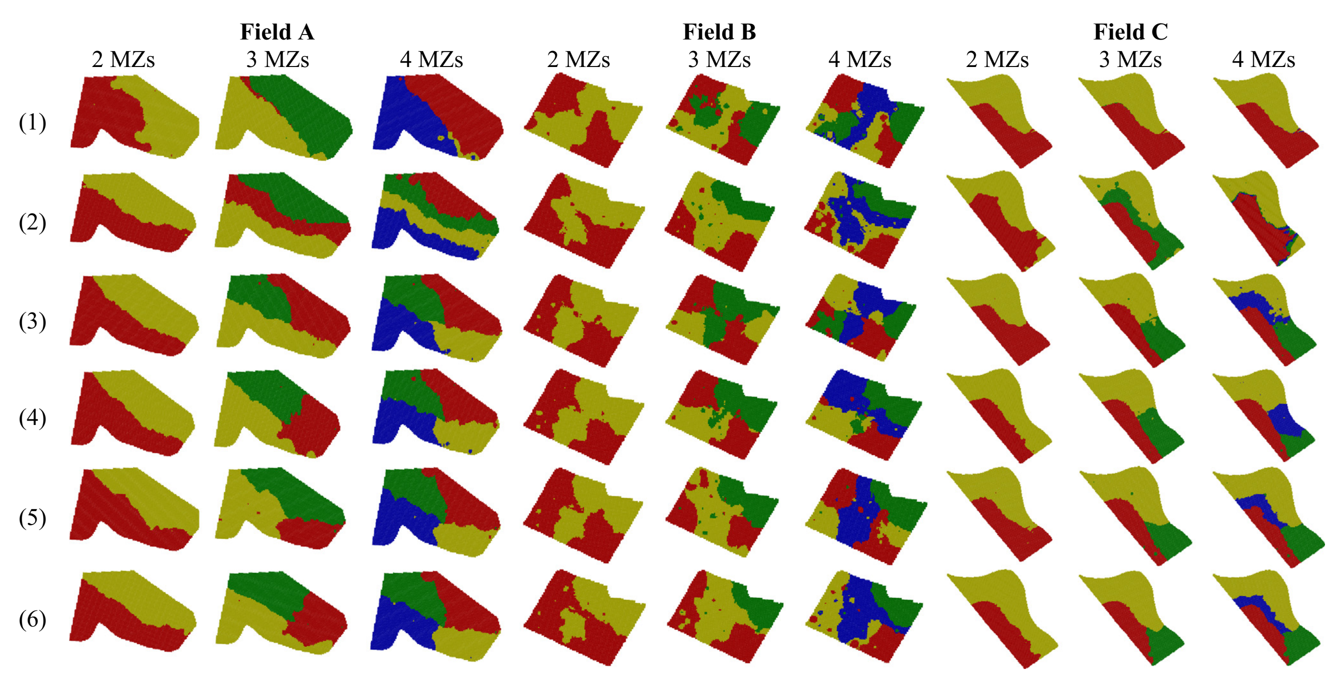

Fig. 2. Thematic maps generated by the six approaches: (1) All-Attrib; (2) Spatial-Matrix; (3) PCA-All; (4) MPCA-All; (5) PCA-SC; (6) MPCA-SC.

Table 7
Results for ANOVA (Tukey's range test), VR, FPI, MPE, SI, and ICVI, for the three fields.

Field	Classes	Approach	ANOVA (Tukey's test)				VR (%)	FPI	MPE	SI (%)	ICVI
			C1	C2	C3	C4					
A	2	All-Attrib	a	a			0.0	0.500	0.079	98.4	1
		Spatial-Matrix	a	b			42.7	0.091	0.018	98.3	0.137
		PCA-All	a	b			42.5	0.185	0.035	98.5	0.273
		MPCA-All	a	b			25.5	0.161	0.030	98.6	0.368
		PCA-SC	a	b			24.4	0.177	0.032	98.4	0.396
		MPCA-SC	a	b			28.8	0.153	0.029	98.6	0.333
	3	All-Attrib	a	a	a		0.0	0.667	0.125	97.7	1
		Spatial-Matrix	a	b	b		22.6	0.156	0.032	96.8	0.307
		PCA-All	a	a	b		39.8	0.287	0.058	97.6	0.298
		MPCA-All	a	a	b		16.7	0.212	0.043	97.5	0.414
		PCA-SC	a	b	a		28.4	0.200	0.042	97.7	0.307
		MPCA-SC	a	b	b		33.6	0.210	0.043	97.7	0.272
	4	All-Attrib	a	a	a	a	0.0	0.750	0.158	97.1	1
		Spatial-Matrix	a	b	b	a	39.1	0.213	0.044	95.0	0.254
		PCA-All	a	b	b	a	28.1	0.314	0.069	96.9	0.427
		MPCA-All	a	ab	b	a	20.8	0.215	0.048	96.5	0.388
		PCA-SC	a	b	a	b	48.9	0.178	0.038	97.0	0.159
		MPCA-SC	a	a	b	b	33.7	0.182	0.041	97.2	0.271
B	2	All-Attrib	a	a			4.1	0.285	0.054	95.7	0.908
		Spatial-Matrix	a	a			5.2	0.146	0.029	95.5	0.573
		PCA-All	a	a			1.7	0.292	0.054	95.7	0.965
		MPCA-All	a	b			15.1	0.255	0.048	95.8	0.612
		PCA-SC	a	a			0.0	0.234	0.045	95.8	0.878
		MPCA-SC	a	b			16.3	0.161	0.032	95.8	0.381
	3	All-Attrib	a	a	a		8.5	0.667	0.132	91.7	0.919
		Spatial-Matrix	a	a	a		11.6	0.153	0.034	94.7	0.385
		PCA-All	a	a	a		2.2	0.357	0.076	94.3	0.683
		MPCA-All	a	ab	b		21.7	0.333	0.071	94.0	0.472
		PCA-SC	a	a	a		17.9	0.327	0.069	93.6	0.500
		MPCA-SC	a	b	a		34.9	0.176	0.038	94.7	0.184
	4	All-Attrib	a	b	ab	ab	22.8	0.536	0.119	91.2	0.774
		Spatial-Matrix	a	ab	b	ab	15.3	0.239	0.052	89.8	0.476
		PCA-All	a	a	a	a	7.7	0.415	0.095	93.1	0.781
		MPCA-All	a	ab	b	ab	-0.2	0.290	0.068	92.9	0.704
		PCA-SC	a	a	b	a	21.2	0.316	0.073	93.6	0.525
		MPCA-SC	a	a	b	a	33.7	0.205	0.046	93.8	0.256
C	2	All-Attrib	a	b			19.4	0.500	0.077	98.8	0.797
		Spatial-Matrix	a	b			31.9	0.495	0.076	97.9	0.659
		PCA-All	a	b			26.9	0.206	0.037	99.0	0.350
		MPCA-All	a	b			23.2	0.162	0.030	98.6	0.329
		PCA-SC	a	b			28.6	0.150	0.027	98.7	0.251
		MPCA-SC	a	b			23.7	0.117	0.021	98.6	0.255
	3	All-Attrib	a	a	a		0.0	0.667	0.122	98.3	1
		Spatial-Matrix	a	a	b		28.0	0.108	0.023	96.4	0.157
		PCA-All	a	b	a		25.4	0.189	0.040	97.8	0.271
		MPCA-All	a	b	b		20.1	0.147	0.031	98.2	0.281
		PCA-SC	a	b	a		28.9	0.127	0.027	97.9	0.168
		MPCA-SC	a	a	b		31.8	0.085	0.017	98.4	0.089
	4	All-Attrib	a	a	a	a	0.0	0.750	0.154	98.1	1
		Spatial-Matrix	a	b	bc	ac	35.9	0.535	0.111	94.7	0.512
		PCA-All	a	ab	b	c	26.6	0.286	0.061	96.3	0.371
		MPCA-All	a	b	ac	bc	26.4	0.166	0.037	97.6	0.267
		PCA-SC	a	b	a	a	38.5	0.175	0.039	97.1	0.175
		MPCA-SC	a	b	c	a	40.0	0.146	0.032	97.3	0.134

C_i: class i.

Variable elevation differed from the other parameters in that it influenced PC1 and SPC1 in the three areas. The result for PC1 is similar to the results obtained by Fraisse et al. (2001), who used PCA for defining MZs in two agricultural areas with corn and soybean crops in the United States. Saleh and Belal (2014) also applied PCA for an area in Egypt and obtained similar results with regard to the influence of elevation on PC1. The influence of clay on PC1 was also observed by Moral et al. (2010), who used PCA for setting MZs in an area in Spain. The considerable influence of the variables elevation and SPR on SPC1 when defining MZs in various fields with wheat crop in Argentina was also detected by Peralta et al. (2015) and Córdoba et al. (2016).

3.3. Thematic maps

For each field, the delineated MZs differed according to the variable selection approach used along with Fuzzy C-means (Fig. 2).

When using the All-Attrib approach for defining three or four classes in field A, field operations are difficult to perform in at least one of the classes owing to its small size and format. The same problem exists in the case of Spatial-Matrix for four classes for field C. Another situation that arose when using All-Attrib was that the approach could not be used to define three or four classes for field C. However, similar problems did not arise when using PCA-All, PCA-SC, MPCA-All, and MPCA-SC.

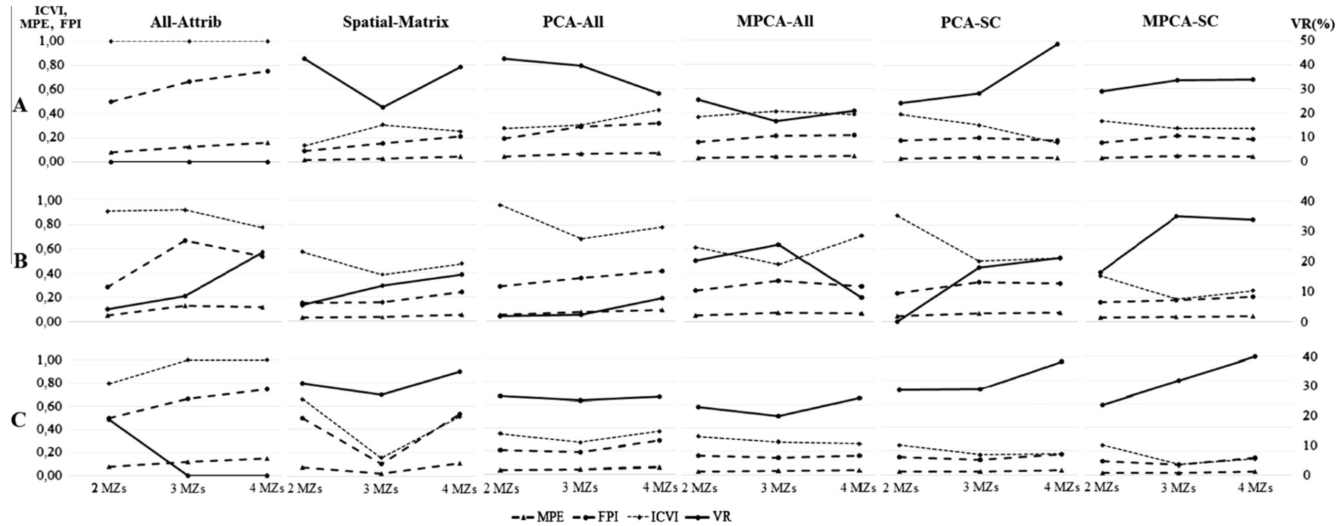

Fig. 3. Graphs for MPE, FPI, ICVI, and VR, for the six approaches assessed, considering two, three and four classes.

The results of the evaluations of the generated classes, according to ANOVA (Tukey's test), VR, FPI, MPE, SI, and ICVI indexes (Table 7), make it possible to state that the division of each area is possible in two classes with statistically different potential yields. For field B, this result was obtained only with MPCA-All and MPCA-SC.

Furthermore, MPCA-SC yielded usually the best results in terms of the variance reduction index; in other words, this approach identified classes with larger differences between the respective normalized average yields and lower internal residual values. Differences in the normalized average yield between classes indicate that soil conditions influence the crop response. As previously mentioned, for all areas, elevation was the variable that had the greatest influence among all variables on SPC1, and therefore, this variable was crucial to the results obtained with MPCA-SC, as found by Córdoba et al. (2013) and Peralta et al. (2015) who used MPCA-All.

The smoothness of the boundary curves of the MZs was assessed by the smoothness index (SI, Table 7). It was confirmed that MPCA-SC usually yielded the best results for all areas, regardless of the number of defined sub-areas. In other words, MPCA-SC yielded sub-areas that were more viable in terms of field operations.

Fig. 3 shows graphically the values of the MPE, FPI, ICVI, and VR, provided by each approach assessed for the three fields. Analysis of the values of the FPI and MPE indexes showed that the MPCA-SC approach was the one that provided the best performance in combination with Fuzzy C-means algorithm when defining the MZs. This is because MPCA-SC showed the lowest values of FPI and MPE. Consequently, this approach is also the one that stood out from the viewpoint of the values of ICVI index.

The combined analysis of FPI, MPE, and ANOVA results confirms the recommendation of the division of each area into two classes, using MPCA-SC to define the variables. If this recommendation is adopted, larger MZs with smoother boundaries are obtained. Córdoba et al. (2016) and Peralta et al. (2015) also used lower FPI and MPE values, as well as easier field operations, as the criteria for choosing two classes.

The use of MPCA-SC allowed identification of the variables that account for global spatial variation. By using this approach, the part of the multidimensional variance that is spatially structured was analyzed. In addition to the works mentioned above, similar discussion about the treatment of multidimensional spatially structured variance by using MULTISPATI-PCA was addressed in the context of ecological data by Dray et al. (2008).

4. Conclusions

A case study of three fields showed that the MPCA-SC approach, which combines spatial correlation analysis with the MULTISPATI-PCA technique, can greatly improve the quality of management zones (MZs). The defined MZs were larger and had smoother boundaries, and consequently, were more viable in terms of field operations.

MPCA-SC conducted, in most situations, distinguished the classes with larger differences between the respective normalized average yield values and lower internal residual values. This approach provided the best dimensionality reduction of the original data without significant loss of information for the three fields.

Acknowledgements

The authors would like to thank the State University of Western Paraná (UNIOESTE), the Federal University of Technology of Paraná (UTFPR), the Araucária Foundation, and the National Council for Scientific and Technological Development (CNPq) for funding.

References

Arrouays, D., Saby, N.P.A., Thioulouse, J., Jolivet, C., Boulonne, L., Ratié, C., 2011. Large trends in French topsoil characteristics are revealed by spatially constrained multivariate analysis. Geoderma 161, 107–114.

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. Engenharia Agrícola 33, 952–964.

Bivand, R., 2012. spdep: Spatial Dependence: Weighting Schemes, Statistics and Models. R Foundation for Statistical Computing.

Boydell, B., McBratney, A.B., 2002. Identifying potential within-field management zones from cotton-yield estimates. Precision Agric. 3, 9–23.

Chessel, D., Dufour, A.B., Thioulouse, J., 2004. The ade4 package-I-one-table methods. R News, 5–10.

Cohen, S., Cohen, Y., Alchanatis, V., Levi, O., 2013. Combining spectral and spatial information from aerial hyperspectral images for delineating homogenous management zones. Biosyst. Eng. 114, 435–443.

Córdoba, M., Balzarini, M., Bruno, C., Costa, J.L., 2012. Principal component analysis with georeferenced data: an application in precision agriculture. FCA UNCUYO 44, 27–39.

Córdoba, M., Bruno, C., Costa, J.L., Balzarini, M., 2013. Subfield management class delineation using cluster analysis from spatial principal components of soil variables. Comput. Electron. Agric. 97, 6–14.

Córdoba, M., Bruno, C., Costa, J.L., Peralta, N.R., Balzarini, M., 2016. Protocol for multivariate homogeneous zone delineation in precision agriculture. Biosyst. Eng. 143, 95–107.
Czaplewski, R.L., Reich, R.M., 1993. Expected Value and Variance of Moran's Bivariate Spatial Autocorrelation Statistic Under Permutation. Department of Agriculture, Fort Collins, USA, p. 13.
Doerge, T.A., 2000. Site-Specific Management Guidelines. Potash & Phosphate Institute, Norcross.
Dray, S., Saïd, S., Débias, F., 2008. Spatial ordination of vegetation data using a generalization of Wartenberg's multivariate spatial correlation. J. Veg. Sci. 19, 45–56.
Embrapa, Brazilian Agricultural Research Corporation, 2006. Brazilian System of Soil Classification. CNPSO, Rio de Janeiro.
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. Int. J. ASABE 44, 155–166.
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 J. 96, 100–108.
Fu, Q., Wang, Z., Jiang, Q., 2010. Delineating soil nutrient management zones based on fuzzy clustering optimized by PSO. Math. Comput. Model. 51, 1299–1305.
Gnanadesikan, R., Kettenring, J., Tsao, S., 1995. Weighting and selection of variables for cluster analysis. J. Classif. 12, 113–136.
Hornung, A., Khosla, R., Reich, R.M., Inman, D., Westfall, D.G., 2006. Comparison of site-specific management zones: soil-color-based and yield-based. Agronomy J. 98, 407–415.
Hotelling, H., 1933. Analysis of a complex of statistical variables into principal components. J. Educ. Psychol. 24, 417–441.
Jaynes, D.B., Colvin, T.S., Kaspar, T.C., 2005. Identifying potencial soybean management zones from multi-year yield data. Comput. Electron. Agric. 46, 309–327.
Johnson, R.A., Wichern, D.W., 2007. Applied Multivariate Statistical Analysis, sixth ed. Pearson, New Jersey.
Larscheid, G., Blackmore, B.S., 1996. Interactions between farm managers and information systems with respect to yield mapping. In: International Conference on Precision Agriculture. American Society of Agronomy, Minneapolis, pp. 1153–1163.
Li, Y., Shi, Z., Li, F., Li, H.Y., 2007. Delineation of site-specific management zones using fuzzy clustering analysis in a coastal saline land. Comput. Electron. Agric. 56, 174–186.
Li, Y., Shi, Z., Wu, H., Li, F., Li, H., 2013. Definition of management zones for enhancing cultivated land conservation using combined spatial data. Environ. Manage. 52, 792–806.
Mielke Jr, P.W., Berry, K.J., 2007. Permutation Methods: A Distance Function Approach. Springer, New York.
Moral, F.J., Terrón, J.M., Silva, J.R.M., 2010. Delineation of management zones using mobile measurements of soil apparent electrical conductivity and multivariate geostatistical techniques. Soil Tillage Res. 106, 335–343.
Ord, J.K., 1975. Estimation methods for models of spatial interaction. J. Am. Stat. Assoc. 70, 120–126.
Peralta, N., Costa, J.L., Franco, M.C., Balzarini, M., 2013. Delimitación de zonas de manejo con modelos de elevación digital y profundidad de suelo. Interciencia 38, 418–424.
Peralta, N.R., Costa, J.L., Balzarini, M., Franco, M.C., Córdoba, M., Bullock, D., 2015. Delineation of management zones to improve nitrogen management of wheat. Comput. Electron. Agric. 110, 103–113.
Ping, J.L., Dobermann, A., 2003. Creating spatially contiguous yield classes for site-specific management. Agronomy J. 95, 1121–1131.
R Core Team, 2014. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.
Reich, R.M., 2008. Spatial Statistical Modeling of Natural Resources. Colorado State University, Fort Collins.
Saleh, A., Belal, A.A., 2014. Delineation of site-specific management zones by fuzzy clustering of soil and topographic attributes: A case study of East Nile Delta, Egypt. IOP Conf. Ser.: Earth Environ. Sci. 18, 1–6.
Schenatto, K., Souza, E.G., Bazzi, C.L., Bier, V.A., Betzek, N.M., Gavioli, A., 2016. Data interpolation in the definition of management zones. Acta Scientiarum 38, 31–40.
Schepers, A.R., Shanahan, F.J., Liebig, M.A., Schepers, J.S., Johnson, S.H., Luchiari, J.A., 2004. Appropriateness of management zones for characterizing spatial variability of soil properties and irrigated corn yields across years. Agronomy J. 96, 195–203.
Zhang, Z., Lu, X., Lv, N., Chen, J., Feng, B., Li, X.W., Ma, L., 2013. Defining agricultural management zones using GIS techniques: case study of drip-irrigated cotton fields. Inf. Technol. J. 12, 6241–6246.