Geoderma 310 (2018) 170–177

Contents lists available at ScienceDirect

Geoderma

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

Enhancing edaphoclimatic zoning by adding multivariate spatial statistics to regional data

Franca Giannini Kurina[a,b,*], Susana Hang[b], Mariano A. Cordoba[a,b], Gustavo J. Negro[b], Mónica G. Balzarini[a,b]

[a] CONICET, Consejo Nacional de Investigaciones Científicas y Técnicas, Argentina
[b] Facultad de Ciencias Agropecuarias, Universidad Nacional de Córdoba, Argentina

ARTICLE INFO

Handling Editor: A.B. McBratney

Keywords:
Multivariate zoning
Spatial principal components
Fuzzy k-means

ABSTRACT

Joint spatial variability of soil and climate variables offers the opportunity to delimit contiguous edaphoclimatic zones. These zones can be useful to improve natural resource management. The aim of this work was to develop a statistical protocol for multivariate zoning at regional scales. A zoning of Córdoba, Argentina, was generated using data from a sample of 355 sites involving edaphic and climatic data (pH, TN, TOC, Na, K, CEC, Cu, Clay, Sand, WHC, elevation, annual precipitation and mean temperature). We proposed a two-step algorithm that considers the spatial correlation of these variables in a clustering of sites. The protocol was run after modeling the spatial pattern of each soil variable to adapt information from different sources and formats to a fine grid. In the first step of the protocol, MULTISPATI-PCA, an extension of the principal component analysis that considers the spatial co-variability between variables, was used to obtain linear combinations of original data. In the second step, such synthetic variables (spatial principal components) were used as input of the fuzzy k-mean clustering method to delineate homogeneous zones. The number of clusters was established by internal validation indices. The use of MULtISPATI-PCA was compared with the more conventional and non-spatial PCA. Results suggest that previous geostatistical interpolation and spatially constrained multivariate analysis create meaningful and spatially coherent zones. Four zones were identified in Córdoba region, Argentina.

1. Introduction

The most widely used tools to differentiate geographical areas with different soil types in a region are soil maps (Buol et al., 1990; Imbellone and Teruggi, 1993; Jarsún et al., 2006). Soil classification has usually been performed using threshold models resulting from a sequence of binary partitions (Burrough et al., 1991) for several variables that are addressed independently. While these tools are very useful, they do not capitalize on the spatial continuum often present in geostatistical data because they do not consider the joint correlation and variation among variables. Moreover, when these models are applied to relatively homogeneous landscapes at a regional scale (i.e. little diverse land uses), the variability due to soil-climate interaction may be masked.

Different univariate geostatistical techniques have been used to model spatial variability of a variable and identify gradients in its values (Cressie and Chan, 1989; Lark, 2000). Nevertheless, when more than one variable is recorded at a site, the spatial co-variability between variables requires less common analyses, such as multivariate geostatistical analyses (Schabenberger and Gotway, 2004). Joint variability of two georeferenced variables has been identified and used to characterize and classify edaphic processes (Cosby et al., 1984). However, it has been demonstrated that the spatial variability analysis of a variable can be improved by incorporating the covariance structure of that variable with respect to an auxiliary variable (Hengl et al., 2004; Wu et al., 2003). Thus, the study of spatial variation pattern not only of that variable but also of the spatial correlations among variables might contribute to the understanding of joint variability and generate zoning at a multidimensional level, i.e. considering a series or set of site variables simultaneously. Currently available data analysis tools consider not only the multivariate nature of data but also their spatiality when data are georeferenced (Wackernagel, 2013). MULTISPATI-PCA (Dray et al., 2008; Arrouays et al., 2011), an extension of principal component analysis (PCA), incorporates spatial co-variability among variables. It is based on linear combinations of the site variables that maximize both spatial autocorrelation and variability rather than only

Abbreviations: sPC, spatial principal components; TN, total nitrogen; WHC, water holding capacity; pp, annual precipitation; Tm, annual mean temperature

* Corresponding author at: CONICET, Consejo Nacional de Investigaciones Científicas y Técnicas, Argentina.
E-mail address: fgkurina@agro.unc.edu.ar (F. Giannini Kurina).

http://dx.doi.org/10.1016/j.geoderma.2017.09.011
Received 10 February 2017; Received in revised form 9 August 2017; Accepted 10 September 2017

variability, as classical non-spatial PCA does.

Another important aspect in multivariate zoning is the scale of analysis. Studies on spatial patterns at the regional scale have shown a level of noise and cover different from that found at the fine scale (Miller and Schaetzl, 2016), which is characterized by a higher density of data per surface unit. Different spatial interpolation techniques (Oliver and Webster, 2014) can be used to perform a re-scaling of row data and associate available data of different variables on a common grid to facilitate the study of correlations (Long, 1998). The potential of MULTISPATI-PCA as applied with spatial interpolated data has been poorly explored at a regional scale.

Different clustering methods can be used with the aim of conglomerating grid nodes (Anderberg, 1973). For continuous phenomena, fuzzy clustering methods (Bezdek et al., 1981) have resulted particularly useful, since they classify the objects as belonging to one group or another. The application of the fuzzy set theory (Burrough et al., 2000) to clustering takes into account the probability of belonging to a cluster rather than only a statistical measure (distance). Irvin et al. (1997) demonstrated the value of the use of fuzzy k-means for classifying spatial data. However, most of the algorithms, even fuzzy clustering ones, have not been developed to handle spatial co-variability between input variables. By coupling MULTISPATI-PCA and fuzzy clustering, the spatial co-variability in the row data can be taken into account for spatial zoning. Site clustering may be implemented either on the set of original variables or on their linear combinations (PCA or MULTISPATI-PCA) (Burrough and Swindell, 1997).

Clustering methods perform unsupervised classification, i.e., they are applied without prior knowledge of the underlying clustering and, therefore, the number of groups that the clustering structure defines is unknown. Although there are numerous indices (Hennig, 2007) to identify a recommended number of groups, there is no clear consensus about which one to use for a particular dataset. In general, cluster validity indices combine information about the data variability within and between clusters and dissimilarity measurements taken from the matrix containing the distances between the objects that need to be grouped. Recently, with the increase in computational power, stability indices have also been used as validation measures obtained from information intrinsic to the data (internal validation); these indices result from comparing the clusters obtained from random successive removal of part of the data. Since there is no consensus on which algorithm to use to determine the number of clusters, it is important to explore the numerous methodological criteria and compare the results with information that was not used during the clustering process (Theodoridis and Koutroumbas, 2008).

The aim of this work was to develop a zoning protocol based on fuzzy clustering analysis of sites characterized by several layers of variables, considering their spatial co-variability, to delimit homogeneous edapho-climatic zones at a regional scale.

2. Material and methods

2.1. Study area

The study area corresponds to Córdoba province, Argentina, between 29° and 35°S and 61° and 65°W (Fig.1). The landscape is composed mostly of plains (~60%), with the remaining territory being north-south mountain ranges to the west of the province. Elevation varies between 79 and 2884 a.s.l. (m). The area is crossed by the 700 mm and 500 mm isohyets, determining an *E*-W humidity gradient from humid, through subhumid, semiarid and arid climates. Mean annual precipitation ranges between 900 and 400 mm, and mean annual temperature between 10 °C and 24 °C. According to the hydrological balance, annual hydric deficit ranges between 80 mm and − 480 mm. According to Soil Taxonomy (Soil Survey Staff, 1975), soils are classified as Mollisols (61%), Entisols (13%), Alfisols (7%) and Aridisols (5%) (Jarsún et al., 2006).

2.2. Database

Soil data were obtained from a previous work conducted in an area of 14.2 million hectares (Hang et al., 2015). Soils of Córdoba province were sampled from the upper 15 cm using a regular 20 × 20 km grid (355 points). Sampling sites corresponded mostly to Mollisols (72%), followed by Entisols (13%), with Aridisols and Alfisols being represented by 5% each. Sampling was mainly conducted in soil used for agriculture (72%), with the remaining 28% corresponding to natural vegetation (grasslands and woodlands) and implanted pastures.

Of the total of soil data available for this work, we selected 10 edaphic variables for zoning: pH, total nitrogen (TN), total organic carbon (TOC), sodium (Na), potassium (K), cation exchange capacity (CEC), cooper (Cu), Sand, Clay, and water holding capacity (WHC), maintaining the chemical, physical and physico-chemical properties most frequently used in soil characterization and discarding highly correlated variables.

In addition to soil data, we used elevation, slope and other data products obtained from the Digital Elevation Model provided by the STRM (Shuttle Radar Topography Mission) (Farr et al., 2007) for each sampling site. We also included site climatic information (mean annual precipitations and temperatures), which was taken from the global database of climatic analysis BIOCLIM (Busby, 1991) for the 1970–2000 period. Elevation and climate data were directly extracted from the databases in raster format for the sites (points) to be classified. We generated the database using the freely available software QGis (QGIS Development Team, 2014).

2.3. Data preprocessing

The zoning protocol was run after modeling the spatial pattern of each variable in order to obtain a fine grid (2.5 × 2.5 km) and adapt information layers for soil and climate covariates. Each soil variable was processed using a spatial structure analysis by regression kriging with the slope, extracted from the Digital Elevation Model (DEM), as a covariate (Hengl et al., 2004). The slope yielded a better fit than other DEM covariates. Experimental semivariograms were calculated using Cressie's robust estimator (Cressie, 1993) and the semivariograms were fitted using the WLS (Weighted Least Squares) estimation method (Cressie, 1985; Oliver and Webster, 2014). Spatial variability was estimated using the “gstat” library (Pebesma, 2004) in R (R Development Core Team, 2016).

2.4. Protocol sequence

2.4.1. Step 1: deriving spatial synthetic variables

A principal component analysis with spatial restriction, MULTISPATI-PCA (Dray et al., 2008), was performed considering as inputs soil, elevation and climate variables associated with the sites of the grid to be classified. MULTISPATI-PCA introduces a spatial weighting matrix to calculate spatial correlations among original data. Spatial autocorrelations are obtained using Moran's index, taking into account the network of neighboring observations of each raw data. Neighbors can be defined using different connection networks (Wartenberg, 1985). We established a maximum distance of 50,000 m to create neighborhoods consistent with the phenomena under study. This value was determined considering the cell size of the grid to be classified and the fitted ranges of the semivariograms of each variable. Spatial principal components (sPCs) necessary for accounting the cumulative variability percentage of at least 80% of the total variability were selected for further clustering. We did not include all the components in order to remove residual variability, i.e. variability that is little explained by repeatable spatial patterns. Such noise is expected to be associated with the last principal components. MULTISPATI-PCA procedure was implemented using the library “ade4” (Chessel et al., 2004) and “spdep” (Bivand et al., 2014) in the software R (R

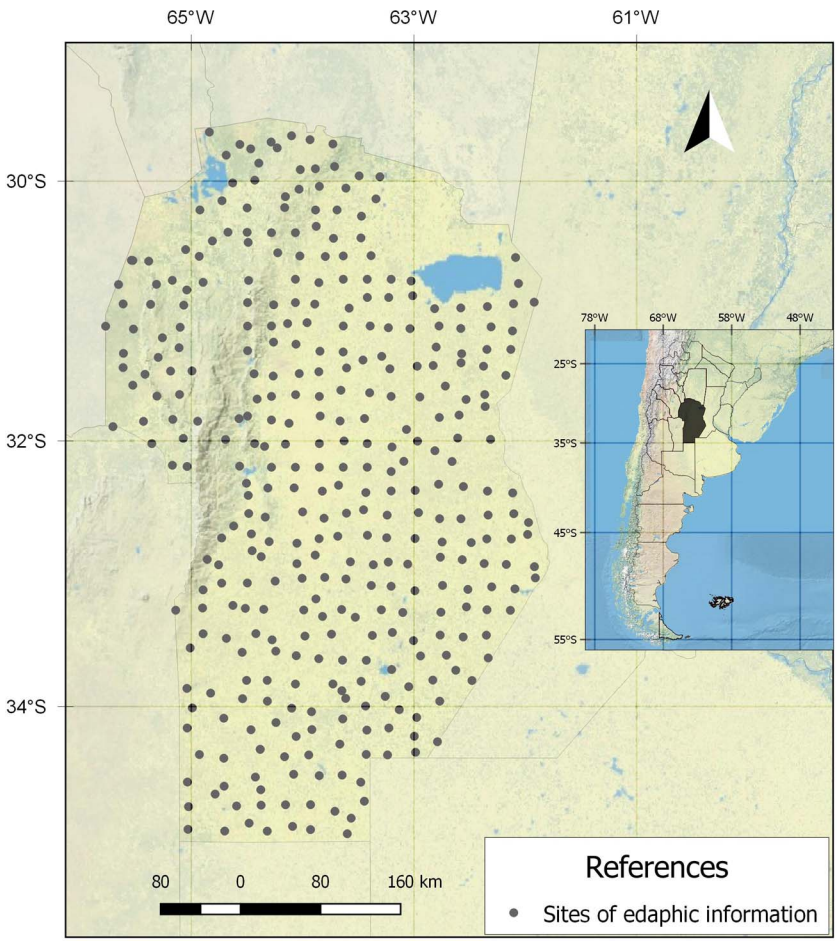

Fig. 1. Study area, Córdoba province, Argentina. Sites of edaphic information (Hang et al., 2015).

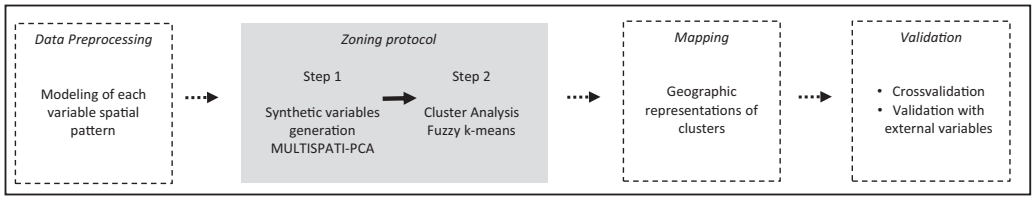

Fig. 2. Zoning protocol sequence scheme.

Development Core Team, 2016).

2.4.2. Step 2: cluster analysis

The sPCs were used as inputs of a non-hierarchical fuzzy k-means cluster analysis (Bezdek et al., 1981). The classification algorithm k-means clusters objects in k groups, maximizing variation among clusters and minimizing variation within the cluster. This method starts with an initial clustering or with a group of seed points (centroids) that will form the centers of the groups. Then, each object is assigned to the group that contains the nearest centroid (mean). Fuzzy classification determines the degree of resemblance of an object to a cluster by its membership to the cluster. We used the squared Euclidean distance and a fuzzyness coefficient k = 2 with the package "e1071" (Meyer et al., 2014) in R (R Development Core Team, 2016). Other coefficient values (1.5 and 2.5) were also evaluated, but they did not improve the clustering reached with a fuzziness coefficient k = 2.

The number of clusters used for zoning was determined considering 17 clustering validity indices included in the "Nbclust" library of R software. Table 4 includes names of indices and their corresponding references. These indices combine information about intracluster compactness and intercluster isolation, i.e. multivariate variability within and between groups. Some indices also involve other statistical properties of the data, such as the number of data objects to be clustered and dissimilarity measurements between objects. A detailed formulation of each of these 17 indices can be found in Charrad et al. (2014). As Charrad et al. (2014) proposed, after calculating several indices we applied the majority rule, i.e., we selected the number of clusters recommended by most of the indices that were calculated for the same data.

2.5. Protocol assessment

To evaluate the proposed methodological workflow (Fig. 2), the fuzzy k-means cluster analysis was computed using as input variables the synthetic variables obtained by MulTISPATI-PCA (sPC) and those derived from the conventional non-spatial PCA (PC). For both types of input variables (sPC and PC), the protocol was evaluated using k-fold cross validation, for k = 10. The membership of 10% of sites, which

have not been used for zoning, to one of the clusters was established considering the Euclidean distance between the site to be classified and the centroid of each cluster. The sites were assigned to the cluster with centroid that was nearest its variable profile, and the correct classification percentage was calculated by comparing the assignment to a given cluster by the protocol with the true location of the point as indicated by its coordinates. Zoning was also validated externally by comparing means of zones for a set of edaphic variables (Zn, Fe, Mn, P and EC) that were not used in the implementation of the zoning protocol. These variables were selected to determine differences among clusters, since all of them result from the same edaphogenetic factors that the original variables. The comparison of centroids between groups was made via a multivariate analysis of variance (Johnson and Wichern, 1998). The objective of this analysis was to determine possible mean differences in the validation variables between the proposed clusters.

3. Results

Soil variables (pH, TN, TOC, Na, K, Cu, CEC, WHC, Sand and Clay), as well as topographic (elevation) and climatic (pp and Tm) variables and the parameters of the selected spatial structure models for each soil variable are presented in Table 1. Coefficients of variation (CV) ranging between 10 (pH) and 273% (Na) exhibited high edaphic variability in the study area. The variables that were closely associated with parental material and weathering (Sand, Clay, CEC and WHC) exhibited a coefficient of variation between 33 and 54%. Variables associated with soil organic matter (SOM), such as TOC and TN, had a similar CV (45%). Noticeably, Elevation exhibited high variability (CV = 76%) and pp. showed a wide range of variation (547 mm), whereas Tm only varied in 7 °C.

The parameters of the semivariagram functions fitted to the soil variables showed that the extension of the spatial correlation (range) varied between 50 and 180 km. The lowest range value was recorded for WHC, whereas high range values were recorded for variables associated with SOM, such as TOC, CEC, and TN. A high percentage of total variability was found to be spatially structured (between 40 and 68%), mainly for K, Cu and Sand.

The application of PCA and MULTISPATI-PCA to the preprocessed data indicated that three PCs would be necessary to represent the edaphoclimatic information, accounting for > 80% of the total variability. Variances and spatial autocorrelation coefficients of each synthetic variable derived from MULTISPATI-PCA and PCA are presented in Table 2. The spatial autocorrelation index was lower in the first PC than in the first sPC. Eigenvectors associated with the synthetic variables derived from MULTISPATI-PCA and PCA were compared (Table 3). The elements in these vectors are the weights of each original variable in the linear combination representing the new variables (synthetic variables). The higher the weight (in absolute value), the higher contribution of the variable to explain variability. For the first sPC, variability was mostly explained by Sand, Clay, WHC and CEC. The most important variables in the second sPC were pH, K, Tm, and pp. The variability of the third sPC was correlated with variations in Elevation and Na (Table 3). The main differences between sPCs and PCs were observed in the assigned role to Tm and Elevation. MULTISPATI-PCA weighted in the first sPC those variables that account for global spatial variation and with greatest spatial autocorrelation.

As result of the fuzzy k-means clustering analysis we identified four clusters of sites, independently of which type of synthetic variable was used. The values of all indices for 2,3,4,5, and 6 clusters obtained from the sPCs are indicated in Table 4. The spatial representation of the four delineated zones is presented in Fig. 3 as obtained from MULTISPATI-PCA (Fig. 3a) as well as from conventional non-spatial PCA (Fig. 3b).

The validation of the complete protocol used for zoning, via cross-validation, yielded a correct classification of 80% of sites when using MULTISPATI-PCA; that value dropped to 69% with the use of the conventional PCA. The comparison of means of zones for external variables that were not used in the classification protocol (Zn, Fe, Mn, P, and EC) indicated statistical differences ($p < 0.05$) between the delimited zones (Table 5).

The characteristics of each of the four delimited edaphoclimatic zones are shown in Table 6. Zone I, located to the west of the province, exhibited the lowest mean annual precipitation (531 mm) and the highest mean pH (7.5). Zone I and zone IV presented the lowest values of TOC ($I = 11.3\ g\ kg^{-1}$ and IV = $9.9\ g\ kg^{-1}$), TN ($I = 0.11\%$ and IV = 0.10%), Clay ($I = 11.7\%$ and IV = 13.4%) and CEC ($I = 12.2\ cmol_c\ kg^{-1}$ and IV = $10.9\ cmol_c\ kg^{-1}$). In turn, Zone IV located to the south of the province, differed from zone I in its low pH (6.5) and high pp. (744 mm); moreover, Zone IV presented the highest Sand content (63.7%) and the lowest Cu content and ($0.9\ mg\ kg^{-1}$). Zone II corresponds to the central mountain range and the piedmont with the highest mean elevation, 761 a.s.l. (m), and lowest mean temperature, 16 °C. In this zone, TOC, TN and Cu had the highest means ($22.3\ g\ kg^{-1}$, 0.21% and $2.2\ mg\ kg^{-1}$, respectively). Zone III covers a

Table 1
Summary statistics for soil and climatic variables, and spatial variability of soil variables in Cordoba ($n = 355$).

Variable	Units	Mean	Min	Max	CV (%)	Fitted semivariogram for soil variables					Prediction Error[d] (%)
						Best Model[c]	Nugget	Partial Sill	Range (km)	RSV	
pH	1:2.5 (s:w)	6.8	5.31	10	10	Sph	0.2	0.2	102	48	8.5
TN	%	0.13	0.04	0.52	45	Sph	0.0	0.0	149	45	30.8
TOC	$g\ kg^{-1}$	13.8	2.8	59.3	45	Exp	13.7	7.3	180	35	33.3
Na	$Cmol\ kg^{-1}$	1.44	0.02	53.7	273	Gau	0.3	0.5	113	64	37.6
K	$Cmol\ kg^{-1}$	1.86	0.4	4.8	31	Sph	0.2	0.1	58	40	24.7
CEC	$Cmol\ kg^{-1}$	17.7	5.3	35.8	33	Sph	7.8	11.4	101	59	4.4
Cu	$mg\ kg^{-1}$	1.63	0	5.9	61	Exp	0.3	0.7	104	68	23.9
Sand	$g\ kg^{-1}$	42	0.6	94.8	57	Gau	95.8	171.5	91	64	30.8
Clay	$g\ kg^{-1}$	18.6	0.1	44.6	43	Gau	20.1	16.4	86	45	32.3
WHC[a]	% p:p	17.7	5	34	33	Exp	8.3	9.6	50	54	22.8
Elevation	m	312	80	1421	76						
pp	mm	738	461	908	14						
Tm[b]	°C	17.2	13.8	20.8	7						

[a] WHC: Water Holding Capacity.
[b] Tm: average daily mean temperature (°C).
[c] Sph: Spherical, Exp: Exponential, Gau: Gaussian.
[d] Root Mean Square Prediction Error expressed as percentage of the mean.

Table 2
Eigenvalues table for MULTISPATI-PCA and classical PCA analysis.

Axis	MULTISPATI-PCA				PCA			
	Eigenvalue	Percentage	Cumulative percentage	Moran's Index	Eigenvalue	Percentage	Cumulative percentage	Moran's Index
1	4.46	0.38	0.38	0.90	5.12	0.39	0.39	0.82
2	2.92	0.3	0.68	0.75	3.91	0.3	0.69	0.78
3	1.46	0.16	0.85	0.69	2.01	0.15	0.85	0.77

Table 3
Contribution of variables to the first components after spatial data interpolation.

Variable	MULTISPATI-PCA			PCA		
	sPC 1	sPC 2	sPC 3	PC 1	PC 2	PC 3
pH	0.04	− 0.43	0.3	0.1	− 0.32	0.4
TN	− 0.13	− 0.31	− 0.19	0.32	− 0.3	− 0.09
TOC	− 0.17	− 0.33	− 0.18	0.29	− 0.32	− 0.11
Na	− 0.04	− 0.05	0.56	4.00E − 03	0.06	0.57
K	− 0.26	0.38	0.01	0.1	0.45	− 0.06
CEC	− 0.39	− 0.14	0.05	0.42	0.01	0.07
Cu	− 0.25	− 0.32	0.09	0.32	− 0.15	0.2
Sand	0.45	− 0.01	− 0.06	− 0.38	− 0.21	− 0.08
Clay	− 0.43	0.03	− 0.05	0.39	0.19	− 0.05
WHC[a]	− 0.44	− 0.04	0.06	0.41	0.15	0.08
Elevation	− 0.02	0.1	0.61	0.09	− 0.44	− 0.15
pp	− 0.29	0.36	− 0.27	0.19	0.35	− 0.33
Tm[b]	0.07	− 0.44	− 0.25	− 0.08	0.22	0.55

[a] WHC: Water Holding Capacity.
[b] Tm: average daily mean temperature (°C).

large area in the east of the province; mean annual precipitations are the highest, 807 mm, and elevation is the lowest, 200 a.s.l. (m), showing the highest contents of Clay, 24.53%, as well as the highest values of K, CEC and WHC (2.3 $cmol_c\ kg^{-1}$, 19.8 $cmol_c\ kg^{-1}$ and 22.7%, respectively).

4. Discussion

The study area covers a wide latitudinal and longitudinal variation range (6 and 4°, respectively), as well as important differences in topography due to the presence of a central mountain range and its piedmont area. These characteristics were reflected in the high coefficients of variation in most of the variables. We also detected variability in the ranges of the semivariance functions fitted for each variable; thus, some variables, such as K and WHC, showed variations at the local level, whereas others showed regional variations, such as pH and TOC. An important proportion (over 40%) of edaphic, topographic and climatic variability was found to be spatially structured, showing that this territory has sufficient spatial heterogeneity to support and justify edaphoclimatic zoning.

4.1. Methodological proposal

The protocol, based on fuzzy-k-means clustering of sites applied to sPCs (Fig. 2), as a tool to capture spatial co-variability, allowed us to summarize not only individual behavior of the variables used but also their spatial co-variation. In the context of this study – the regional scale –, the method produced clusters of sites with spatial consistency, i.e., neighboring sites showed to belong to the same cluster.

Preprocessing raw data via the study of the spatial pattern of each variable allowed us to re-grid and adapt information from different sources and formats. Thus, we went from relatively few observations to regularly spaced and highly dense predictions of the same target phenomena.

The comparison of the results of classical non-spatial PCA and MULTISPATI-PCA showed some advantages of the latter over the former for the spatial analysis. The spatially restricted method maximized the spatial autocorrelation in the first synthetic variable; these results are consistent with the nature of the method (Arrouays et al., 2011). Clustering from sPCs showed a lower cross classification error (20%) than from PCs (32%). By coupling fuzzy K-means with MULTISPATI-PCA we generated more contiguous variation (deleting

Table 4
Optimum number of clusters. Best partition according to 17 internal validation indices.

Index	Number of clusters					Best partition	Author
	2	3	4	5	6		
KL[a]	2.9	0.6	4.9	3.0	1.0	4	(Krzanowski and Lai, 1988)
CH[a]	165	139	149	133	119	2	(Caliński and Harabasz, 1974)
CCC[a]	− 0.3	0.0	4.4	4.8	4.6	5	(Sarle, 1983)
Silhouette[a]	0.3	0.4	0.4	0.3	0.2	4	(Rousseeuw, 1987)
Ratkowsky[a]	0.2	0.2	0.3	0.2	0.2	4	(Ratkowsky and Lance, 1978)
Ptbiserial[a]	0.5	0.5	0.6	0.5	0.5	4	(Milligan, 1980, 1981)
McClain[a]	0.7	0.9	1.1	1.7	2.3	2	(McClain and Rao, 1975)
Dunn[a]	0.05	0.05	0.06	0.04	0.04	4	(Dunn, 1974)
Cindex[b]	0.2	0.3	0.2	0.2	0.2	4	(Hubert and Levin, 1976)
DB[b]	1.4	1.3	1.2	1.2	1.4	4	(Davies and Bouldin, 1979)
SDindex[b]	1.0	0.8	0.8	1.0	1.2	4	(Halkidi et al., 2000)
SDbw[b]	1.1	0.6	0.6	1.0	0.7	4	(Halkidi and Vazirgiannis, 2001)
Hartigan[c]	74	89	36	23	20	4	(Hartigan, 1975)
Scott[c]	426	736	1167	1310	1422	4	(Scott and Symons, 1971)
TrCovW[c]	2.2E + 05	1.2E + 05	7.9E + 04	6.0E + 04	5.3E + 04	3	(Milligan and Cooper, 1985)
Friedman[c]	2.9	4.3	8.2	9.9	12.2	4	(Friedman and Rubin, 1967)
Ball[c]	1180	634	369	264	205	3	(Ball and Hall, 1965)

[a] Value of the index should be maximized.
[b] Value of the index should be minimized.
[c] Difference between sequential levels of the index should be maximized.

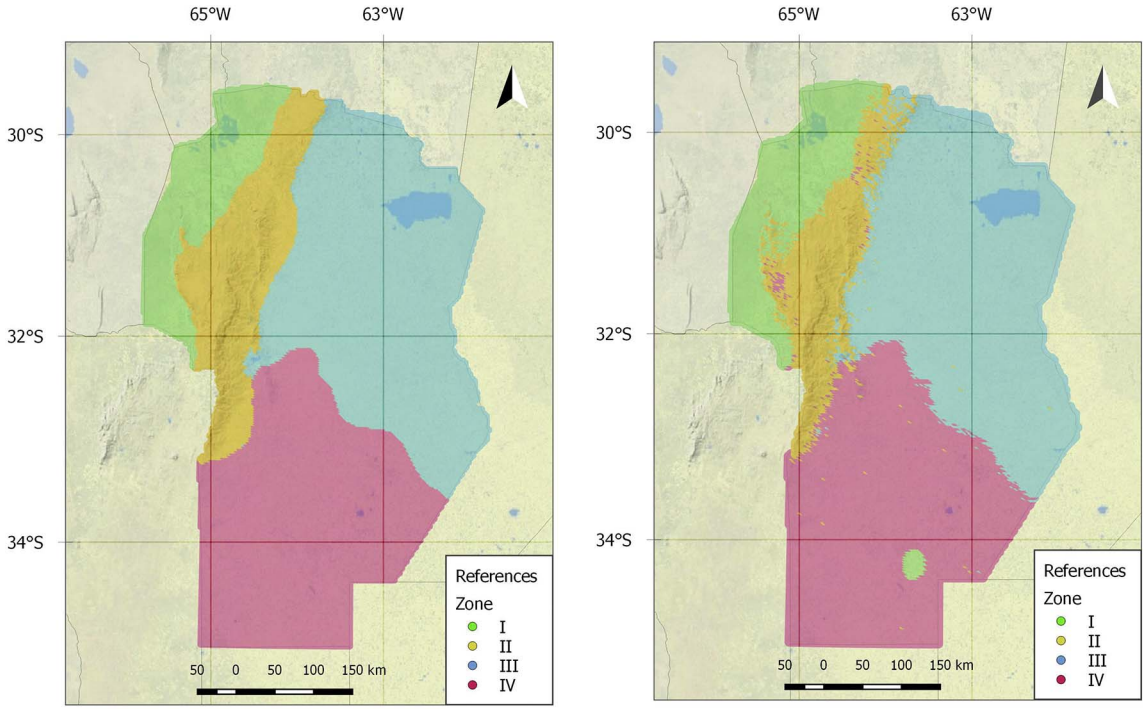

Fig. 3. Zoning of Córdoba, Argentina. a) Zoning map obtained from MULTISPATI-PCA. b) Zoning map obtained from classical PCA.

Table 5
Zones differences for external validation variables (Zn, Fe, Mn, P, and EC).

Zone identification	Zn	Fe	Mn	P	EC 1:2.5 (s:w)	Hotelling test[1]
	(mg kg^{-1})			(ppm)	(dS m^{-1})	
I	1.2	154.6	34.7	41.3	2.4	a
II	1.8	142.4	25.2	82.8	0.2	b
III	1.7	188.3	55.6	124.9	0.5	c
IV	0.7	94.9	36.9	135.9	0.6	d

[1] Different letters are significant to $p < 0.05$.

small spots within a zone) than with non-spatial PCA.

To identify the zones, it was necessary to define the appropriate number of clusters in the partitioning. Different approaches have been described to investigate cluster validity, some of them based on comparing the results of the cluster analysis with other classifications provided externally (external criteria) and others using information obtained from the clustering process itself (internal criteria) (Theodoridis and Koutroubas, 2008). The protocol included the use of several indices to determine the optimum number of clusters (Charrad et al., 2014). The zoning validation with external variables allowed us to confirm that the classification performed also describes the spatial variability of the external validation variables (Zn, Fe, Mn, P and EC).

A generic framework, called scorpan-SSPFe method, has been used to predict digital soil maps. This method is based on seven predictive factors (soil, climate, climatic, organisms, topography, parent material, time factor, spatial or geographic position), a generalization of Jenny's formative factors (McBratney et al., 2003). The use of a large amount of complementary information, which is now easily available and of lower cost than soil sampling, such as that extracted from digital elevation models and satellite images, has stimulated the implementation of automated soil mapping techniques to be used with big data, such as automatic learning techniques. Random forest algorithms (Breiman, 2001), support vector machines and self-organizing maps (Kohonen, 1982) have been used to obtain automatic classifications based on multidimensional information. However, these methods do not provide information about the magnitude of spatial covariances among the original variables, which hinders the understanding of the processes explaining zoning. By contrast, the algorithm we implemented provides measures of the relative contribution of the different variables to zoning. These methods are undoubtedly of different nature, each having advantages and disadvantages, and, importantly, they are complementary. For example, the results of the classifications made using machine learning methods might be used as another form of

Table 6
Soil and climatic variables for each homogenous edaphoclimatical zone in Cordoba, Argentina.

Zone	pH	TN	TOC	Na	K	CEC	Cu	Sand	Clay	WHC	Elevation	pp	Tm
	1:2.5 (s:w)	(%)	(g kg^{-1})	(cmol$_c$ kg^{-1})			(mg kg^{-1})	(g kg^{-1})		(%)	(m)	(mm)	(°C)
I	7.5[a]	0.11	11.3	4.9	1.3	12.2	1.9	57.8	11.7	14.3	332	531	19.5
	(8)	(41)	(51)	(237)	(31)	(26)	(28)	(19)	(63)	(25)	(43)	(12)	(5)
II	7.2	0.21	22.3	1.2	1.4	18.1	2.2	41.3	20.0	19.0	761	673	16.0
	(7)	(43)	(39)	(38)	(38)	(24)	(47)	(41)	(39)	(26)	(26)	(8)	(7)
III	6.8	0.15	15.3	1.1	2.3	19.8	2.0	18.9	24.5	22.7	200	807	17.3
	(8)	(24)	(26)	(118)	(20)	(25)	(48)	(54)	(23)	(16)	(60)	(7)	(4)
IV	6.5	0.10	9.9	0.9	1.7	10.9	0.9	63.7	13.4	12.6	258	744	16.8
	(9)	(30)	(33)	(106)	(24)	(22)	(65)	(21)	(36)	(27)	(66)	(9)	(2)

[a] Values are averages and coefficient of variations (between parentheses).

statistically-based zoning validation.

4.2. Zoning of Córdoba

The general characteristics of the entire territory were captured by the zoning performed. The territory has more or less abrupt gradients of three of the five soil forming factors: climate, parent material and relief (Jenny, 1941). The Pampas loess has a W-E decreasing grain size gradient associated with the direction of the transporting winds (Rocca et al., 2006). Moreover, the $CaCO_3$ levels of the Pampas loess range between 5 and 10% (Gorgas and Tassile, 2002), and its depth in the soil profile corresponds to the precipitation gradient, which increases W-E (Iriondo and García, 1993; Manzur, 1997). Thus, zones III and IV, the largest ones, covering together 72.7% of the total of sites, correspond to the plain area of the province and two smaller zones (I, II) correspond to the sierras (mountain) area and the west of the mountain range. The general characteristics of the territory comprising zones I and II include high spatial and vertical heterogeneity due to abrupt changes in relief and highland Pampas patches (Gorgas and Tassile, 2002) as well as the presence of $CaCO_3$ near the surface and a region of salt flats to the NW. Another particular aspect of this area is that it bears the highest proportion of native vegetation (Zak and Cabido, 2002; Cabido et al., 2005).

Zone I exhibited concordance with an area named Bolsón chaqueño, characterized by strong topographic heterogeneity (piedmont and plains), a negative hydrological balance and a marked thermal amplitude (Gorgas and Tassile, 2002). This is the zone with the highest pH, which is partly explained by the origin of the alkalinity associated with the presence of $CaCO_3$ on the surface (Manzur, 1997), It corresponds to a salt flat area, which was originated from a NE-SW geological fault that left the bed of an old sea exposed (Gorgas and Tassile, 2002). It is also characterized by the abundance of sodium chloride, sodium sulfate, which along with other minerals have been originated by cyclical sedimentation and evaporation of waters with high mineral concentration (Bertolino et al., 2000; Gorgas and Tassile, 2002). Zone II covers the mountain region and part of the piedmont in the east and west, being the area of highest altitude of the four defined zones. The positive relationship between TOC and CEC is associated with organic colloids or humified organic matter (Parfitt et al., 1995), suggesting that SOM of zone II would present abundant fresh or barely humified organic matter. Another distinctive trait that reinforces this assumption is that Cu content is the highest of all zones and extractable forms of this element are favored by the increase in pH. However, the dynamics of Cu related to SOM due to the formation of complexes (Mortensen, 1963; McGrath et al., 1988; Burke et al., 1989; Alvarez and Lavado, 1998) needs to be further studied. Zone III was one of the largest zones, geologically corresponding mostly to the so called “loess plains/flatlands” (high, flat, of Altos de Morteros) (Gorgas and Tassile, 2002). It presented several distinctive characteristics, since it is the area of lowest altitude and presented the highest values of pp., clay, K, WHC, and CEC. This set of traits defines a region with agricultural potential, which is indeed the main use (Cabido et al., 2005; Jarsún et al., 2006). Finally, zone IV is the largest (40.3% of the sites) and given that sand content is the most characteristic trait, grain size was clearly the factor that discriminated this portion of the territory from the remaining zones. It corresponds to a group of regions defined earlier in the literature by their grain size as Pampas characterized by dunes and high and flat sandy Pampas (Gorgas and Tassile, 2002). The lowest mean contents of TOC, TN and CEC and several elements as Zn, Mn, and Cu were on average consistent with the grain size traits of zone IV. These results suggest that, despite its edapho-climatic indicators of agricultural potential, this zone that also presents fragile characteristics.

5. Conclusions

The proposed workflow allows us to differentiate homogeneous zones within a large area based on soil and climatic variables with high spatial co-variability. Modeling spatial behavior of each variable allowed us to gather information from different sources. The classification of the sites using spatial principal components of soil, topographic and climatic variables, as input of the fuzzy k-means algorithm, created spatially coherent zones. Spatial covariances between variables enhance the understanding of zoning. Spatial co-variation among site variables allowed us to define four contiguous and homogeneous edaphoclimatic zones in the territory of Cordoba province, Argentina.

Acknowledgements

We thank the Argentinian National Scientific and Technological Promotion Agency (ANPCyT-PICT 2014-1071), Ministry of Science and Technology of Cordoba province (MinCyT-PIODO 2015) and the Argentinian National Scientific and Technical Research Council (CONICET-PIP 2015), for their support of this research.

References

Alvarez, R., Lavado, R.S., 1998. Climate, organic matter and clay content relationships in the Pampa and Chaco soils, Argentina. Geoderma 83 (1), 127–141.
Anderberg, M.R., 1973. Cluster Analysis for Applications. Monographs and Textbooks on Probability and Mathematical Statistics. New York, Academic Press, Inc.
Arrouays, D., Saby, N.P., Thioulouse, J., Jolivet, C., Boulonne, L., Ratié, C., 2011. Large trends in French topsoil characteristics are revealed by spatially constrained multivariate analysis. Geoderma 161 (3), 107–114.
Ball, G.H., Hall, D.J., 1965. ISODATA: A Novel Method of Data Analysis and Pattern Classification. Stanford Research Institute, Menlo Park.
Bertolino, S.R., Poiré, D.G., Carignano, C., 2000. Primer registro de sedimentitas marinas terciarias aflorantes en las Sierras Pampeanas de Córdoba, Argentina. Rev. Asoc. Geol. Argent. 55 (1–2), 121.
Bezdek, J.C., Coray, C., Gunderson, R., Watson, J., 1981. Detection and characterization of cluster substructure i. Linear structure: Fuzzy c-lines. SIAM J. Imag. Sci. 40 (2), 339–357.
Bivand, R., Keitt, T., Rowlingson, B., 2014. rgdal: bindings for the geospatial data abstraction library. R package version 0.8–16. http://CRAN.R-project.org/package=rgdal, Accessed date: 2 February 2017.
Breiman, L., 2001. Random forests. Mach. Learn. 45 (1), 5–32.
Buol, S.W., Hole, F., McCracken, R.J., 1990. Génesis y clasificación de suelos. Trillas, México.
Burke, I.C., Yonker, C.M., Parton, W.J., Cole, C.V., Schimel, D.S., Flach, K., 1989. Texture, climate, and cultivation effects on soil organic matter content in US grassland soils. Soil Sci. Soc. Am. J. 53 (3), 800–805.
Burrough, P.A., Swindell, J., 1997. Optimal Mapping of Site-specific Multivariate Soil Properties. Precision Agriculture: Spatial and Temporal Variability of Environmental Quality. John Wiley & Sons, Ltd., England, pp. 208–220.
Burrough, P.A., Maguire, D.J., Goodchild, M.F., Rhind, D.W., 1991. Soil Information Systems. Geographical Information Systems: Principles and Applications. DJ Maguire MF Goodchild & DW Rhind, Longmans, Harlow, UK, pp. 153–169.
Burrough, P.A., Van Gaans, P.F.M., MacMillan, R.A., 2000. High-resolution landform classification using fuzzy k-means. Fuzzy Sets Syst. 113 (1), 37–52.
Busby, J., 1991. BIOCLIM-a bioclimate analysis and prediction system. Plant Prot. Q. 6 (1), 8–9.
Cabido, M., Zak, M.R., Cingolani, A., Cáceres, D., Díaz, S., 2005. Cambios en la cobertura de la vegetación del centro de Argentina. Factores directos o causas subyacentes. La heterogeneidad de la vegetación de los agroecosistemas. Nacional University of Buenos Aires, pp. 271–300.
Caliński, T., Harabasz, J., 1974. A dendrite method for cluster analysis. Communications in Statistics-theory and Methods 3 (1), 1–27.
Charrad, M., Ghazzali, N., Boiteau, V., Niknafs, A., Charrad, M.M., 2014. Package ‘NbClust’. J. Stat. Softw. 61, 1–36.
Chessel, D., Dufour, A.B., Thioulouse, J., 2004. The ade4 package-I: one-table methods. R News 5–10.
Cosby, B.J., Hornberger, G.M., Clapp, R.B., Ginn, T., 1984. A statistical exploration of the relationships of soil moisture characteristics to the physical properties of soils. Water Resour. Res. 20 (6), 682–690.
Cressie, N., 1985. Fitting variogram models by weighted least squares. J. Int. Assoc. Math. Geol. 17 (5), 563–586.
Cressie, N., 1993. Statistics for Spatial Data: Wiley Series in Probability and Statistics. 15. Wiley-Interscience, New York, pp. 105–209.
Cressie, N., Chan, N.H., 1989. Spatial modeling of regional variables. J. Am. Stat. Assoc. 84 (406), 393–401.
Davies, D.L., Bouldin, D.W., 1979. A Cluster Separation Measure. IEEE Trans. Pattern Anal. Mach. Intell. 1 (2), 224–227.
Development Team, Q.G.I.S., 2014. QGIS Geographic Information System. Open Source Geospatial Foundation Project. http://qgis.osgeo.org, Accessed date: 2 February 2017.
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 (1), 45–56.

Dunn, J., 1974. Well Separated Clusters and Optimal Fuzzy Partitions. J. Cybern. 4 (1), 95–104.
Farr, T.G., Rosen, P.A., Caro, E., Crippen, R., Duren, R., Hensley, S., Seal, D., 2007. The shuttle radar topography mission. Rev. Geophys. 45 (2), 1–33.
Friedman, H.P., Rubin, J., 1967. On Some Invariant Criteria for Grouping Data. J. Am. Stat. Assoc. 62 (320), 1159–1178.
Gorgas, J.A., Tassile, J.L., 2002. Regiones Naturales de la Provincia de Córdoba. Serie C, Publicaciones Técnicas. Agencia Córdoba Ambiente. Ferreyra Editor, Córdoba, Argentina.
Halkidi, M., Vazirgiannis, M., 2001. Clustering Validity Assessment: Finding the Optimal Partitioning of a Data Set. In: ICDM'01 Proceedings of the 2001 IEEE International Conference on Data Mining, pp. 187–194.
Hang, S., Negro, G.J., Becerra, A.M., Rampoldi, A.E., 2015. Suelos de Córdoba: Variabilidad de las propiedades del horizonte superficial. Maita Jorge Omar Editorial, Córdoba, Argentina.
Hartigan, J.A., 1975. Clustering Algorithms. John Wiley & Sons, New York.
Hengl, T., Heuvelink, G.B., Stein, A., 2004. A generic framework for spatial prediction of soil variables based on regression-kriging. Geoderma 120 (1), 75–93.
Hennig, C., 2007. Cluster-wise assessment of cluster stability. Comput. Stat. Data Anal. 52 (1), 258–271.
Hubert, L.J., Levin, J.R., 1976. A General Statistical Framework for Assessing Categorical Clustering in Free Recall. Psychol. Bull. 83 (6), 1072–1080.
Imbellone, P.A., Teruggi, M.E., 1993. Paleosols in loess deposits of the Argentine Pampas. Quat. Int. 17, 49–55.
Iriondo, M.H., García, N.O., 1993. Climatic variations in the Argentine plains during the last 18,000 years. Palaeogeogr. Palaeoclimatol. Palaeoecol. 101 (3–4), 209–220.
Irvin, B.J., Ventura, S.J., Slater, B.K., 1997. Fuzzy and isodata classification of landform elements from digital terrain data in Pleasant Valley, Wisconsin. Geoderma 77 (2–4), 137–154.
Jarsún, B., Gorgas, J., Zamora, E., Bosnero, H., Lovera, E., Ravelo, A., Tassile, J., 2006. Los suelos de Córdoba. Agencia Córdoba Ambiente e Instituto Nacional de Tecnología Agropecuaria. EEA Manfredi, Córdoba, Argentina.
Jenny, H., 1941. Factors of Soil Formation. McGraw-Hill, New York (281 pp).
Johnson, R.A., Wichern, D.W., 1998. Applied Multivariate Statistical Analysis. Printice-Hall. Inc., Upper Saddle River, NJ.
Kohonen, T., 1982. Self-organized formation of topologically correct feature maps. Biol. Cybern. 43 (1), 59–69.
Krzanowski, W.J., Lai, Y.T., 1988. A criterion for determining the number of groups in a data set using sum-of-squares clustering. Biometrics 44, 23–34.
Lark, R.M., 2000. A comparison of some robust estimators of the variogram for use in soil survey. Eur. J. Soil Sci. 51 (1), 137–157.
Long, D.S., 1998. Spatial autoregression modeling of site-specific wheat yield. Geoderma 85, 181–197.
Manzur, A., 1997. Dinámicas evolutivas de suelos en Atum Pampa, sierras pampeanas, Córdoba, Argentina. Multequina 6, 67–83.
McBratney, A.B., Santos, M.M., Minasny, B., 2003. On digital soil mapping. Geoderma 117 (1), 3–52.
McClain, J.O., Rao, V.R., 1975. CLUSTISZ: A Program to Test for The Quality of Clustering of a Set of Objects. J. Mar. Res. 12 (4), 456–460.
McGrath, S.P., Sanders, J.R., Shalaby, M.H., 1988. The effects of soil organic matter levels on soil solution concentrations and extractabilities of manganese zinc and copper. Geoderma 42 (2), 177–188.
Meyer, D., Dimitriadou, E., Hornik, K., Weingessel, A., Leisch, F., 2014. e1071: Misc Functions of the Department of Statistics (e1071), TU Wien. R package version 1.6-3.
Miller, B.A., Schaetzl, R.J., 2016. History of soil geography in the context of scale. Geoderma 264, 284–300.
Milligan, G.W., 1980. An Examination of the Effect of Six Types of Error Perturbation on Fifteen Clustering Algorithms. Psychometrika 45 (3), 325–342.
Milligan G.W., 1981. A Monte Carlo Study of Thirty Internal Criterion Measures for Cluster Analysis. Psychometrika 46 (2), 187–199.
Milligan, G.W., Cooper, M.C., 1985. An Examination of Procedures for Determining the Number of Clusters in a Data Set. Psychometrika 50 (2), 159–179.
Mortensen, J.L., 1963. Complexing of metals by soil organic matter. Soil Sci. Soc. Am. J. 27 (2), 179–186.
Oliver, M.A., Webster, R., 2014. A tutorial guide to geostatistics: computing and modelling variograms and kriging. Catena 113, 56–69.
Parfitt, R.L., Giltrap, D.J., Whitton, J.S., 1995. Contribution of organic matter and clay minerals to the cation exchange capacity of soils. Commun. Soil Sci. Plant Anal. 26 (9–10), 1343–1355.
Pebesma, E.J., 2004. Multivariable geostatistics in S: the gstat package. Comput. Geosci. 30, 683–691.
R Development Core Team, 2016. R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0. http://www.Rproject.org, Accessed date: 2 February 2017.
Ratkowsky, D.A., Lance, G.N., 1978. A Criterion for Determining the Number of Groups in a Classification. Aust. Comput. J. 10 (3), 115–117.
Rocca, R.J., Redolfi, E.R., Terzariol, R.E., 2006. Características geotécnicas de los loess de Argentina. Revista Internacional de Desastres Naturales, Accidentes e Infraestructura Civil 6 (2), 149–166.
Rousseeuw, P., 1987. Silhouettes: A Graphical Aid to the Interpretation and Validation of Cluster Analysis. J. Comput. Appl. Math. 20, 53–65.
Sarle, W.S., 1983. SAS Technical Report A-108. In: Cubic Clustering Criterion. SAS Institute, Cary, NC.
Schabenberger, O., Gotway, C.A., 2004. Statistical Methods for Spatial Data Analysis. CRC Press, Boca Raton, Florida.
Scott, A.J., Symons, M.J., 1971. Clustering Methods Based on Likelihood Ratio Criteria. Biometrics 27 (2), 387–397.
Soil Survey Staff, 1975. Soil Taxonomy: A Basic System of Soil Classification for Making and Interpreting Soil Surveys. Agric. Handbook No. 436. U.S. Govt. Print. Office, Washington, DC.
Theodoridis, S., Koutroumbas, K., 2008. Pattern recognition. IEEE Trans. Neural Netw. 19 (2) 376–376.
Wackernagel, H., 2013. Multivariate Geostatistics: An Introduction with Applications. Springer Science & Business Media, Berlin.
Wartenberg, D., 1985. Multivariate spatial correlation: a method for exploratory geographical analysis. Geogr. Anal. 17, 263–283.
Wu, J., Norvell, W.A., Hopkins, D.G., Smith, D.B., Ulmer, M.G., Welch, R.M., 2003. Improved prediction and mapping of soil copper by kriging with auxiliary data for cation-exchange capacity. Soil Sci. Soc. Am. J. 67 (3), 919–927.
Zak, M.R., Cabido, M., 2002. Spatial patterns of the Chaco vegetation of central Argentina: integration of remote sensing and phytosociology. Appl. Veg. Sci. 5, 213–226.