BIOSYSTEMS ENGINEERING 143 (2016) 95–107

Available online at www.sciencedirect.com

ScienceDirect

journal homepage: www.elsevier.com/locate/issn/15375110

Research Paper

Protocol for multivariate homogeneous zone delineation in precision agriculture

Mariano A. Córdoba [a], Cecilia I. Bruno [a], José L. Costa [b], Nahuel R. Peralta [b], Mónica G. Balzarini [a,*]

[a] Cátedra de Estadística y Biometría, Facultad de Ciencias Agropecuarias, Universidad Nacional de Córdoba and Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET), Ing Agr. Felix Aldo Marrone 746, Ciudad Universitaria, C.C. 509, 5000, Córdoba, Argentina

[b] Instituto Nacional de Tecnología Agropecuaria (INTA), Estación Experimental Balcarce and Facultad de Ciencias Agrarias, Universidad Nacional de Mar del Plata, C.C.276, 7620, Balcarce, Buenos Aires, Argentina

ARTICLE INFO

Article history:
Received 19 August 2015
Received in revised form
9 December 2015
Accepted 12 December 2015
Published online xxx

Keywords:
Spatial data
Spatial autocorrelation
Multivariate classification
Site-specific management
Software specifications

Uniform management of agricultural fields has been increasingly replaced with environmentally based management, which is benefited by the identification of homogeneous zones within crop fields. Such zones are based on the classification of field sites into groups of homogeneous features. Multiple causative agents of variability and the response of agricultural crops should be considered for zoning. Several correlated variables are usually measured and georeferenced for this purpose at multiple sites within the field. This paper presents an approach to promoting the integration of different statistical tools for identifying homogeneous zones based on site covariates. The methodological innovation of this work involves cleaning and re-scaling of spatial data, as well as multivariate and geostatistical analyses in a logical sequence (protocol). Statistical topics for further improvement and protocol applications are noted. The analytical process has been illustrated using a rain-fed wheat crop (60 ha) from the Argentine Pampas, with apparent electrical conductivity, elevation and soil depth as master variables for zoning, and yield, soil organic matter and clay to validate the created management zones; however, it may be applied to other production systems using georeferenced data. The R scripts and the sample file to run the proposed protocol are provided as electronic supplementary material.

1. Introduction

Farm machinery equipped with new technologies provides the opportunity for a more accurate measurement of spatial soil and terrain variability of crop fields. The recorded spatial data is used for multivariate classification of field sites into groups of homogeneous features. The classification is then used to delimit contiguous zones within the field aimed at site-specific management in precision agriculture (PA) (Yao et al., 2014). Management zones (MZs) are usually areas with similar characteristics, such as texture, topography, water status and soil nutrient levels (Moral, Terrón, & Marques Da

* Corresponding author. Tel.: +54 351 4334103x219; fax: +54 351 4334118x114.
E-mail address: mbalzari@agro.unc.edu.ar (M.G. Balzarini).

http://dx.doi.org/10.1016/j.biosystemseng.2015.12.008

Nomenclature

AIC	Akaike information criterion
ECa	Apparent electrical conductivity (mS m^{-1})
ECa30	Apparent electrical conductivity at 30 cm depth (mS m^{-1})
ECa90	Apparent electrical conductivity at 90 cm depth (mS m^{-1})
KM-sPC	Fuzzy k-means clustering from spatial PCA
Elev	Elevation (m)
Sd	Soil depth (m)
MZ	Management zone
MLM	Mixed Linear Model
I_i	Moran's local index
PCA	Principal component analysis
PA	Precision agriculture
sPCA	Spatial principal components analysis
SOM	Soil organic matter (%)

Silva, 2010). Physical and chemical soil properties are the most widely reported variables used for zoning, followed by landscape attributes (Khosla, Westfall, Reich, Mahal, & Gangloff, 2010). In environments in which water and nutrient availability are crop limiting factors, production largely depends on soil type, specifically, on its capacity to retain water and nutrients. In the last years, besides the introduction of yield monitors, there has been an increase in the use of proximal sensors that capture spatial data on apparent electrical conductivity (ECa) as a soil attribute. The use of soil ECa has gained attention as a good surrogate method for detecting spatial variation in soil chemical and physical properties (Arno, Martinez-Casasnovas, Ribes-Dasi, & Rosell, 2011; Corwin & Lesch, 2010; Moral et al., 2010; Peralta, Costa, Balzarini, & Angelini, 2013; Rodríguez-Pérez, Plant, Lambert, & Smart, 2011; Taylor, McBratney, & Whelan, 2007). Data captured by ECa sensors are georeferenced using high-precision GPS. Elevation is another topographic characteristic that influences water movement and soil development within the field, generating yield spatial variability. Effective soil depth is also a useful soil characteristic to delimitate MZs (Peralta, Costa, Balzarini, & Castro Franco, 2013) because this variable affects the water storage capacity and its spatial distribution, generating yield variability.

The hypothesis that the productivity gap of crops is influenced by the interaction of soil and landform characteristics has been present in several works. Pennock (2003) and Kravchenko, Robertson, Thelen, and Harwood (2005) related topography to productivity. Siqueira, Marques, and Pereira (2010) and Sanchez, Marques Jr., Siqueira, Camargo, and Pereira (2013) used multivariate analysis and geostatistics to analyse the potential use of site covariates to predict the variability of important agronomic traits. The combined application of multivariate and geostatistical analysis together has already been proposed, but using different programs and non-spatially restricted multivariate methods. In this work we propose a logical sequence of statistical analyses that can be implemented using a protocol. The protocol application is illustrated using data on ECa, elevation and effective soil depth of a 60-ha wheat field cultivated under PA in the southern Argentine Pampas as master variables for zoning. Agricultural fields in the south-eastern Pampas frequently have multiple soil map units within them, despite their sometimes relatively small size, and wide range of soil textures and properties, causing high soil spatial variability (Peralta, Costa, Balzarini, & Angelini, 2013). The adoption of ECa sensors in Argentina is increasing because the ECa signal provides an integration of several effects at a site. In the southern Argentine Pampas ECa measurements successfully delimited homogeneous soil zones associated with spatial distribution of clay, soil moisture, CEC, SOM content and pH (Peralta, Costa, Balzarini, & Angelini, 2013).

From the analytical point of view, Taylor et al. (2007) set the basis for a protocol to delineate management zones. However, as the computing capability increases, new statistics can be used to handle multivariate spatial data. New statistical techniques are available for handling spatial data at the clean-up step, in the multivariate classification, and in predictive and validation steps.

A first step in any quantitative protocol is the removal of artefacts within data before analysis. Local spatial autocorrelation indices which are particularly useful for data debugging, such as Moran local index (Anselin, 1995) and Getis-Ord index (Getis & Ord, 1992), can integrate analytic protocols given their current availability in free software. On the other hand, the availability of several types of variables requires combining data from different sources of information and usually of different spatial resolutions. Several spatial interpolation techniques (Oliver & Webster, 2014) can be used to perform a re-scaling of original measurements to associate data from the different variables from each site onto a common grid. The issue of different scales (grid sizes) for different variables should be taken into account to proceed with multivariate classification. The spatial correlation structure for each variable is a first guide to follow, but the spatial covariance of variables is also meaningful. Other practical aspects as the minimum area needed for crop management are usual constraints. In choosing the grid size, there is a trade-off between maintaining spatial precision by selecting a fine grid and reducing noise and making the data more manageable by selecting a coarser grid (Long, 1998). Since variability may be studied at any spatial scale, the choice of grid size depends on the aims of the investigation. In making this choice, the investigator is aided by knowledge of how much variability is lost for each variable in moving from one scale to a larger one (Roel & Plant, 2004). Grid sizes of 10- to 50-m range are common for mapping in site-specific agriculture (Peralta et al., 2015; Ping & Dobermann, 2003; Roel & Plant, 2004), and this is because 108 variation in soil properties appears to occur at a much near scale than the 1 ha strata (McBratney & Pringle, 1999). Choosing a coarse resolution (>16 m) for spatial interpolation may result in biased aggregated data sets (Ping & Dobermann, 2003). In practice, the selection of an appropriate sampling cell size requires understanding of the relationships between grid size, yield variability accounted for, and the resulted spatial map fragmentation.

After re-gridding measurements for all variables the next step, in a protocol of analysis, consists of multivariate site classification. Several clustering algorithms (Anderberg, 1973), such as the ISODATA method (Fraisse, Sudduth, & Kitchen, 2001; Guastaferro et al., 2010), non-parametric approaches developed by Aggelopooulou, Castrignanò, Gemtos, and De Benedetto (2013), a hierarchical approach presented by Fleming, Westfall, Wiens, and Brodahl (2000) and the fuzzy k-means method (Bezdek, 1981), have been widely used to identify potential MZs in PA (Boydell & McBratney, 2002; Davatgar, Neishabouri, & Sepaskhah, 2012; Moral et al., 2010). The software FuzME (Minasny & McBratney, 2002) and MZA (Fridgen et al., 2004) were especially designed to perform clustering analysis using the fuzzy k-means algorithm. However, these cluster algorithms do not work with spatial restrictions. They were not developed for spatial data. A high zone fragmentation was observed when the techniques ignore the spatial nature of the data (Frogbrook & Oliver, 2007; Ping & Dobermann, 2003). Córdoba, Bruno, Costa, and Balzarini (2013) developed a method (KM-sPC) based on the use of principal component analysis (PCA) with spatial restriction and the posterior application of the fuzzy k-means algorithm using the spatial principal components as input for site classification. More contiguous classes and reduced fragmentation of the delimited homogeneous zones can be obtained by incorporating spatial information in the protocol analysis. Thus, the multivariate analysis was included in the proposed protocol with the aim of obtaining a site classification taking into account the spatiality in the data. For this reason, the fuzzy k-means algorithm is implemented using the spatial principal components as input variables. Moreover, the interpretation of the coefficients of soil variables at each spatial principal component provides insight into the spatial correlation among input properties.

Different statistical indices, such as partition coefficient and classification entropy (also known as fuzziness performance index-FPI and normalized classification entropy-NCE) (Bezdek, 1981) can be used to determine the optimum number of clusters for site classification. Other indices, such as Xie-Beni (Xie & Beni, 1991), Fukuyama-Sugeno (Fukuyama & Sugeno, 1989), and proportion exponent (Windham, 1981) might also be used to determine this number. It may occur that none of the indices agrees in the optimum class number. A combination of indices will be recommended for a complementary use of the information provided by those indices, thereby obtaining a single index that summarizes four indices (Galarza, Mastaglia, Albornoz, & Martínez, 2013). The use of spatial filters applied on the results of the classification is also recommended to improve zoning contiguity (Lark, 1998; Ping & Dobermann, 2003).

Once zoning is performed, the appropriateness of delineated zones, need to be evaluated which requires determining if there are differences among the zones in terms of yield and other soil variables selected as validation traits. Conventional statistical models like ANOVA are not recommended to evaluate mean differences among zones because independence is not met when the dataset is spatially referenced (Lawes & Bramley, 2012). Instead, Mixed Linear Models (MLM) (West, Welch, & Galecki, 2015) are preferred because they account for spatial correlation in the data. New computational intensive techniques (Falivene, Cabrera, Tolosana-Delgado, & Sáez, 2010; Webster & Oliver, 2007) are available for enhance the crucial step of validation.

Here we discuss the overall process of multivariate homogeneous zone delineation and provide technical specifications for its implementation and particularly to zoning an agricultural field with soil data taken at a fine scale. The whole protocol was developed using the freely available R software (R Core Team, 2015) and provide online access to source code to perform the analysis in R (Appendix A. Supplementary Material).

2. Material and methods

2.1. *Example data*

Data were collected from an agricultural production field (60 ha) under continuous agriculture with annual crop rotation, located in south-eastern Pampas, Argentine. The Argentine Humid Pampas region is one of the world's best regions for grain crop production (Satorre & Slafer, 1999). The major soil types in this region are Typic Argiudolls, with a loam texture at the surface layer, loam to clay loam at subsurface layers, and sandy loam below 110 cm in depth, and Petrocalcic Paleudoll, which presents discontinuous layers of a petrocalcic horizon between 50 and 100 cm and greater clay contents at subsurface layers than Typic Argiudolls (Peralta et al., 2015). According to the Thornthwaite moisture index, the climate is sub-humid/humid (Burgos & Vidal, 1951), with annual rainfall of 880 mm and mean annual temperature of 13.3 °C.

Georeferenced measurements of apparent electrical conductivity (ECa) taken at two depths: 0–30 cm (ECa30) and 0–90 cm (ECa90), as well as elevation (Elev), and soil depth (Sd), were recorded. Soil ECa measurements were taken using a Veris 3100 (Veris 3100, Division of Geoprobe Systems, Salina, KS) calibrated following the manufacturer's instructions. The sensor was pulled across the field in a series of parallel transects spaced at 15–20 m intervals, the appropriate spacing to avoid measurement errors and information loss (Farahani & Flynn, 2007). ECa was simultaneously measured and georeferenced using a Differential Global Positioning System (DGPS) (Trimble R3, Trimble Navigation Limited, USA) with sub-metre measurement accuracy and set up to record position once per second. A total of 6425 sites were measured. Terrain elevation data were processed with a vertical accuracy of 3–5 cm. Soil depth was measured using a hydraulic penetrometer (Gidding) on a 30 × 30-m regular grid. Eight georeferenced soil sampling points were taken within each delineated MZ to validate the process. Each sampling point consisted of three subsamples centred in the MZ. Soil cores were taken to a depth of 90 cm using a 5-cm diameter hydraulically driven soil tube (Giddings Machine Co., Windsor, CO). As soil profile is not uniform through the 0–90 cm depth, soil in each core was carefully mixed to homogenise the sample. To validate the created MZ we used soil organic matter (SOM) content measured from the 0–30 cm stratum (Barbieri, Echeverría, & Sainz Rozas, 2009), particle-size distribution determined using the Bouyoucos method (Dewis & Freitas, 1970) and wheat grain yield recorded using calibrated commercial yield monitors mounted on combines equipped with DGPS. A random sample of 1000 data points was used to validate zoning.

2.2. Protocol

2.2.1. Conversion of spatial coordinates

In geodesy, a datum is a set of reference points on the Earth's surface that serves as the basis for measuring locations, as well as an associated model of the Earth shape (reference ellipsoid) to define the geographic coordinate system. Datums have different radii and central points; hence, a point measured with different datums can have different coordinates. Hundreds of reference datums have been developed for referencing points in certain areas and suitable for those areas. The datum WGS84 (World Geodetic System 84) is commonly used, and is the default standard datum for coordinates in the commercial GPS units. GPS users should check the datum used because an artefact may lead to a coordinate translation of several metres. For analytical purposes, this geodetic coordinate system (expressed in degrees, minutes and seconds) is projected onto another Cartesian coordinate system (from a 3D to a 2D model) known as a typical UTM (universal transverse Mercator) projection system. Thus, the distances between sites within the field can be expressed as absolute distances (metres) rather than relative distances (degrees), therefore facilitating interpretation. In the proposed protocol, we used the spTransform function of the "rgdal" package (Bivand, Keitt, & Rowlingson, 2014) to convert geographic coordinates into UTM Cartesian coordinates.

2.2.2. Removal of outliers

Outliers, or atypical values, are observations that fall outside the general pattern or distribution of the data set. Removing outliers before analysis is essential to ensure correct decision making based on the analysis (Taylor et al., 2007). In our protocol, outliers are easily removed following a process that includes different complementary techniques and theories: 1) the data set is constrained to a natural variation range in which thresholds are obtained from the data distribution; 2) the mean and standard deviation (SD) are calculated for the data set of a variable, and the values that are outside the mean ± 3 SD are identified and removed. According to Chebyshev's theorem (Amidan, Ferryman, & Cooley, 2005) it is inferred that a minimum 89% of the data is within the mean ± 3 SD, regardless the distribution. Even though real data could belong to this interval, the upper and lower limits use to be modified to obtain robust variance estimators for the next protocol steps. This aspect could be necessary because there are occasions where yield monitor data are negatively skewed as the result of harvesting artefacts such as those caused by harvesting with an incompletely-filled header or travelling/turning over harvested areas with the header down (Taylor et al., 2007). A reduction of the lower limit coefficient may be warranted in such cases (e.g. mean-1.5 SD). The need to adjust this value depending on the range of true yield variation was noted by Simbahan, Dobermann, and Ping (2004). Optimization of the SD parameter for a particular field could be achieved by iteratively choosing values and observing the effect on the resulting yield data distribution (Sudduth & Drummond, 2007).

2.2.3. Removal of inliers

Step 2.2 is intended to remove the outer-tails extreme values of the data set, but not the local extreme values (spatial inliers). Inliers are data that differ significantly from their neighbourhood, but lie within the general range of variation of the data set. Some PA software applications, such as Yield Editor (Sudduth & Drummond, 2007) or the one developed by Sun, Whelan, McBratney, and Minasny (2013), were designed for diagnosis and cleaning of data based on combine yield information. Here, we use spatial autocorrelation Moran's local index (I_i) (Anselin, 1995) as a tool for identifying inliers for each recorded variable. Because a data set belongs to different neighbourhoods, I_i is basically a Moran's index applied individually to each neighbourhood and suggests the degree of similarity or difference between the value of an observation and the neighbours' value. A spatial weights matrix should be defined for a local spatial autocorrelation analysis. This matrix can be represented graphically (as neighbourhood graphs), in which the nodes correspond to the field sites and the borders to non-nil spatial weights. There are different methodological options to define the neighbourhoods (Dray, Legendre, & Peres-Neto, 2006). In our protocol, the neighbourhood network was defined using the Euclidean distance.

A positive I_i value corresponds to spatial grouping of similar (high or low) values (positive autocorrelation), whereas a negative I_i value indicates grouping of different values (e.g., a site with a low value of the variable is surrounded by neighbours with high values) (negative autocorrelation). In both situations, the p-value for a given index must be small enough for the value of interest to be considered an inlier. At this step, we suggest adjusting the p-values according to Bonferroni's criterion due to the multiplicity issue in hypothesis testing (Bland & Altman, 1995). The Bonferroni adjustment deflates the $\alpha = 0.05$ applied to each test performed, so the study-wide error rate remains at 0.05. Anselin (1996) proposed visualizing I_i in a scatter plot, a useful visual tool for the exploratory analysis of spatial data that enables the evaluation of the similarity of an observed value with respect to the neighbouring observations. The horizontal axis represents observation values, whereas the vertical axis represents the spatial lag of the analysed variable. In addition, linear or non-linear regression models can be fitted and added to this plot. In the proposed protocol, *localmoran* and *moran.plot* functions of the "spdep" package (Bivand, 2014) are used to calculate I_i and perform the Moran scatter plot to identify inliers. I_i and its statistical significance for each observation are obtained by applying the *localmoran* function. The *moran.plot* function not only performs the scatter plot but also fits a linear regression model and calculates a series of diagnostic statistics. Data that depart from the 45° slope suggest sites with a spatial autocorrelation different from that of their neighbourhood values. The diagnostic criteria are: *Cook Distance*, *Leverage*, *DFFITS*, *DFBETAS* and *COVRATIO* (Draper & Smith, 1998). The *moran.plot* function calculates these indices for each observation and considers an observation as influential if at least one of the diagnostic indices detects it as such. In our protocol, the first inliers to be removed are those detected with Local Moran' Index (data with negative and statistically significant I_i) then a Moran plot is constructed because others potentially influential observations may be identified by at least one regression diagnostic statistics.

2.2.4. *Spatial interpolation*

Spatial interpolation is performed to estimate the values of a variable of interest at sites where that variable was not sampled. Before performing spatial interpolation, it is necessary to generate a regular artificial grid, ideally with similar arrangement on the field for all the information sources. Grid spacing should reflect the required detail level, the capacity for data processing and the statistical ability of the software for the analysis. A 10 $\times$ 10 m grid is recommended to avoid computing problems derived from a grid of a higher point density and to maintain an appropriate resolution for data visualization and further geostatistical analysis in PA (Taylor et al., 2007). The selection of the spatial interpolation method will depend on the number and density of sites from which data were collected. Geostatistical techniques, such as block kriging (Webster & Oliver, 2007), are common in PA. When the database is too large or data density too high, efficient methods of block kriging, such as interpolation by kriging neighbourhood, are preferable due to their low computational requirements (Oliver & Webster, 2014). There are numerous software applications available for performing the gridding and within-field spatial prediction, such as Surfer (Golden Software, Inc.), Vesper (Minasny, McBratney, & Whelan, 2005), and GIS software. Our protocol uses the packages “geoR” (Ribeiro Jr. & Diggle, 2001) and “gstat” (Pebesma, 2004) for geostatistical interpolations by means of block kriging. First, an empirical semi-variogram is adjusted for each variable. Then, the exponential and spherical theoretical models are fitted by weighted least squares, because they are usually the ones of the best fit to model soil variability (Gili, 2013). The residual sum of squares (RSS) of each model is reported as a criterion to be used for selecting the theoretical model to fit the spatial data, and is therefore not a good parameter to assess the precision and accuracy of interpolation. Cross-validation or external validation is implemented to validate spatial interpolation. These computing intensive techniques provide a means of choosing among plausible models for semi-variograms for prediction. Each and every one of the N data points is omitted in turn from the set of data and its value there is predicted by ordinary punctual kriging with the proposed model (Oliver & Webster, 2014). Alternative statistics can be calculated in the validation step. The proposed protocol will include three of them: the mean squared error (MSE), root mean squared (RMSE) and the mean squared deviation ratio (MSDR), which is the mean of the squared errors divided by the corresponding kriging variances. The best-fitting model is that one that minimizes the MSE and RMSE, while the MSDS should be equal to 1.

Site classification from a multivariate perspective requires that the different interpolated information layers (layers of different variables) be included in a single file. For this purpose, individual files must be horizontally concatenated using the columns corresponding to the X and Y coordinates as concatenation criteria. This step will ensure that the data from different sources of information are correctly linked on the geographical space.

2.2.5. *Multivariate site classification*

A fuzzy k-means cluster analysis on the principal components obtained from the spatial principal component algorithm is included in the analytic protocol to be used after the information of the different variables is cleaned and concatenated. We computed the fuzzy k-means cluster analysis with the “e1071” package (Meyer, Dimitriadou, Hornik, Weingessel, & Leisch, 2014) using as input spatial principal components (sPC) obtained by MULTISPATI-PCA (Dray, Saïd, & Débias, 2008). Data are standardized to eliminate the effect of the differential unit of the variables. The same weighting matrix used to obtain the spatial autocorrelation indices was selected to perform MULTISPATI-PCA. Thus, the spatial correlation between the measured soil variables is included in the analysis. The MULTISPATI-PCA is performed with the “ade4” (Chessel, Dufour, & Thioulouse, 2004) and “spdep” packages (Bivand, 2014). An index summarizing four indices to determine the optimum number of sites classes (Galarza et al., 2013) was included in the protocol.

2.2.6. *Smoothing of classification results*

In delineating MZs, spatial filters on the clustering results can be applied to promote contiguous zoning and reduce class fragmentation (Lark, 1998; Ping & Dobermann, 2003). Statistical filters (non-linear spatial filters) are applied (Arce, 2005), which respond according to the ranking of pixels contained in a portion of the image (mask). The masks can be of different sizes: 3 $\times$ 3, 5 $\times$ 5 or 7 $\times$ 7 pixels. The median filter (Gonzalez & Woods, 2008) is widely used and replaces the central pixel value with the median of the values from the neighbourhood of that pixel (the original pixel value is included in the calculation of the median). For the protocol, we implemented a function in R to apply the median filter on the classification obtained in step 2.5.

2.2.7. *Validation of delineated zones*

Conducting a soil and/or yield random stratified sampling is recommended to validate the delineated MZs, using the potential MZs as strata. Three to four samples per MZ are usually sufficient. Taylor et al. (2007) discussed that the soil properties measured at this stage should reflect existing local knowledge on which soil variables are likely to be affecting yield and may differ between cropping regions. MLM can be used to determine between-MZ differences. A MLM with a zone fixed effect and spatially correlated errors is fitted for each dependent variable. Exponential and spherical spatial correlation functions with and without nugget effect are fitted for the model error terms. Akaike information criterion (AIC) (Akaike, 1976) is reported for model selection. Four models plus the independent error model are compared. This analysis is conducted using the function *gls* of “nlme” package (Pinheiro, Bates, DebRoy, Sarkar, & R Core Team, 2014).

2.3. *Protocol outline and implementation*

Figure 1 shows a diagram that illustrates the proposed protocol to analyse multiple variables for the creation of MZs. The R scripts and the sample file for running the proposed protocol are provided as electronic supplementary material. The

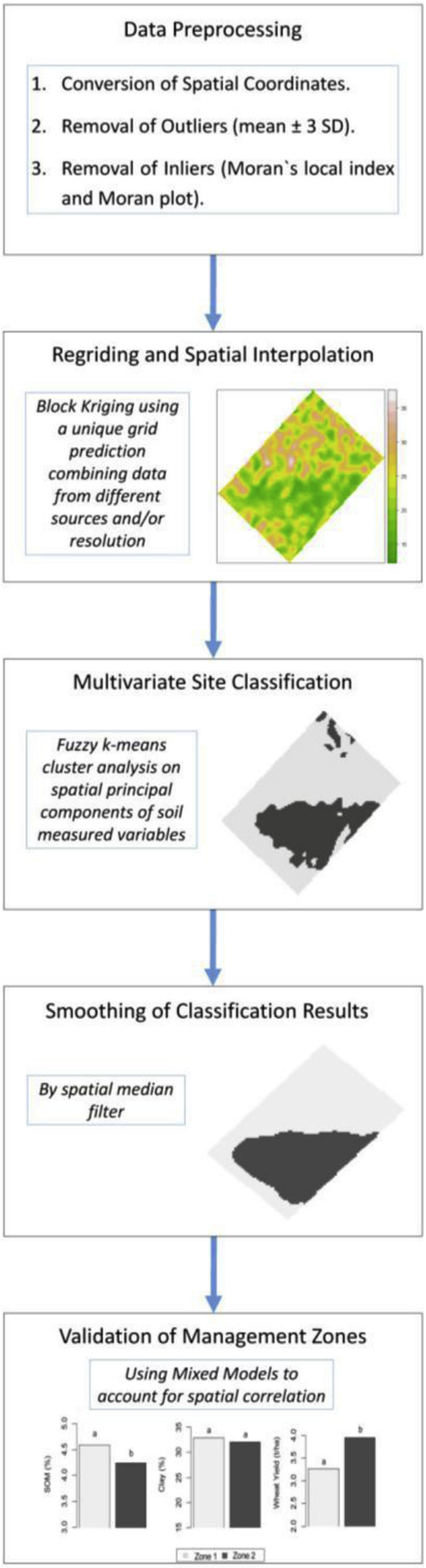

Fig. 1 – Diagram outlining an analytical protocol to delineate MZs from site properties.

following R packages must be installed and loaded for the analyses: “spdep”, “rgdal”, “geoR”, “gstat”, “ade4”, “e1071” and “nlme”. Data from ECa30 are used as an example, in the first part of protocol illustration (3.1–3.2). These steps are repeated for all the measured variables to be used for multivariate classification. The initial data set requires several columns, the first two identify the bidimensional spatial coordinates (x and y) and the others correspond to the measured variables (e.g. ECa30).

3. Results and discussion

3.1. *Data preprocessing*

For conversion of spatial coordinates, it was necessary to specify the band or zone (in this case, Zone 21, south) and the ellipsoid (WGS84); data with the transformed coordinates are shown in Fig. 2. To remove outliers for ECa30 we used its mean (23.8 mS m^{-1}) and its total standard deviation (SD = 7.0 mS m^{-1}). Then, data between the mean ± 3 SD were selected for the next step of the protocol; 48 cases were removed at this clean-up stage, representing 1% of the total of measured sites (Fig. 3). To remove inliers, neighbour points were defined as those contiguous points located at a distance ≤25 m. The Local Moran's Index (I_i) and its statistical significance for $\alpha = 0.05$ (adjusting the p-values according to Bonferroni's criterion) identified 12 potential inliers (Fig. 4). Figure 5 shows the fit of a linear regression on the Moran's plot. The influential regression points are identified using different diagnostic statistics, such as DFBETAS (dfb.1_ for intercept and dfb.x for slope), DFFITS (dffit), Covratio (cov.r), Cook's distance (cook.d) and leverage (hat). A point is regarded as influential if it meets at least one of those statistics (fact denoted by “*” in the output) (Fig. 6). In Fig. 5 influential points were denoted by rhomboidal symbols and are considered inliers. The new data set includes 5910 cases, i.e. 467 cases (7% of the data) were removed with respect to the data set without outliers.

3.2. *Spatial interpolation of data*

To perform geostatistical spatial interpolation, we fit an empirical semivariogram (shown for ECa30 in Fig. 7). The RSS (SSR_exp = 2924423 vs. SSR_sph = 4084724) and cross-validation parameters (Table 1) shown that the best model for spatial geostatistical interpolation in the case study was the exponential one. The maximum semi-variance found between pairs of ECa30 data points was 31.05 (mS m^{-1})2 (sill). This variability, expressed as an SD of 5.57 mS m^{-1}, is important to measure uncertainty in prediction at unsampled locations obtained using kriging. Regarding ECa30, here it represents 24% of the process mean. Spatial interpolation will exhibit lower precision and accuracy, depending on the magnitude of this variance. The range (24.92 m) is the distance at which semivariance stops increasing, interpreted as the spatial lag at which the observations become essentially uncorrelated. In the international literature the scope or range parameter can be interpreted as information useful for defining the sampling plan, indicating which recommended

Case	x	y	ECa30	x	y	ECa30
1	-59.13236	-37.91546	27.8	312558.9	5801421	27.8
2	-59.13241	-37.91550	26.1	312554.9	5801416	26.1
3	-59.13246	-37.91554	22.4	312550.7	5801412	22.4
4	-59.13251	-37.91558	20.0	312546.5	5801407	20.0
5	-59.13256	-37.91562	23.6	312542.2	5801402	23.6

Fig. 2 – Input data in geographic coordinates for the variable electrical conductivity (ECa) (left) and the corresponding output in Cartesian coordinates (UTM WGS84, Zone 21 south) (right) (R software output).

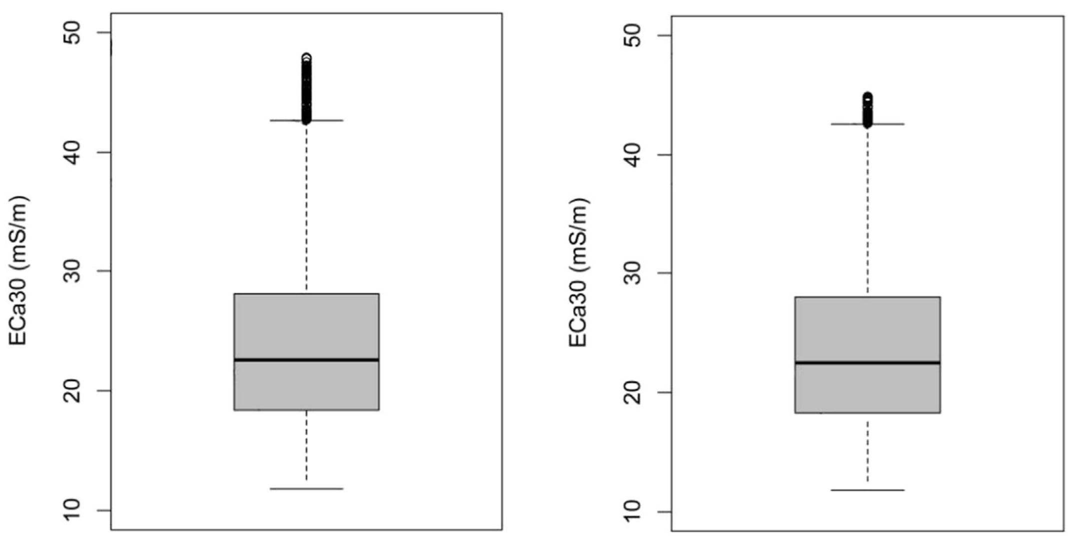

Fig. 3 – Box-plots of apparent electrical conductivity data at 30 cm depth (ECa30) before outlier removal (left) and after outlier removal (right).

spacing to represent the spatial variability in a given region. Thus, can be inferred that the protocol may be used in areas with characteristics (geology, soil type and landscape) similar to southeast in the Argentine pampas using pixel or soil information with resolution up to 25 m. This is possible, provided that the model exhibits a spatial correlation structure (i.e. nugget to sill ratio lesser than 1).

Moreover, this range value could also help to plan other protocol's steps, such as smoothing, preventing very densely packed mascara. The nugget represents the variation at the scale of sampling; it is the variation that remains unresolved including any measurement error; for this example data, nugget was set to 0 $(\mathrm{mS\ m^{-1}})^2$. For block kriging, a data set with the georeferenced points that make up the plot polygon, i.e. points representing the field boundaries, was required. A regular grid (10×10 m) was generated on the polygon, which enables to perform interpolations within the field boundaries (Fig. 8). The map of spatial variability of ECa30 is shown. After all the variables have been processed up to the spatial interpolation with the same prediction grid, the function *cbind* was used for concatenation of predicted values (Fig. 9).

3.3. *Multivariate site classification and zoning*

The graphical display obtained from MULTISPATI-PCA allowed us to study the spatial correlation structure between

Case	Ii	E.Ii	Var.Ii	Z.Ii	Pr(z<0)
1098	-1.401	-0.000157	0.050	-6.276	<0.0001
1962	-3.033	-0.000157	0.055	-12.885	<0.0001
2077	-2.651	-0.000157	0.100	-8.389	<0.0001
5548	-1.187	-0.000157	0.055	-5.045	<0.0001
6362	-0.944	-0.000157	0.077	-3.406	0.0046

Fig. 4 – Local Moran's index values (I_i), expected value (E.Ii), variance (Var.Ii), standardized value (Z.li) and p-value ($Pr(z < 0)$) of the first five inliers detected for ECa30 input data (R software output).

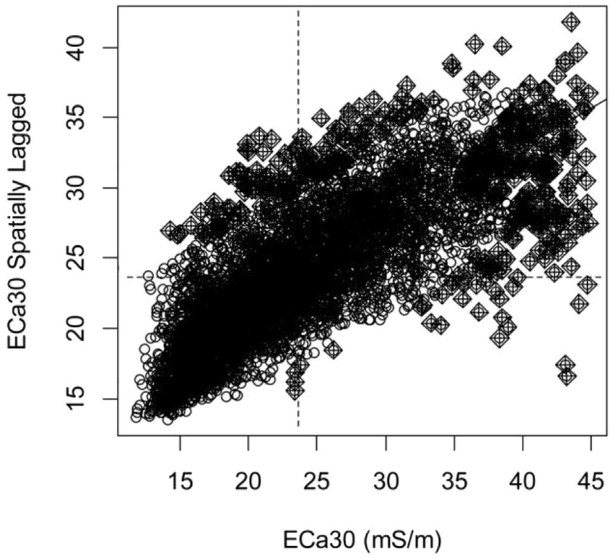

Fig. 5 – Moran scatter plot for the variable apparent electrical conductivity at 30 cm in depth (ECa30). Square dots represent inliers.

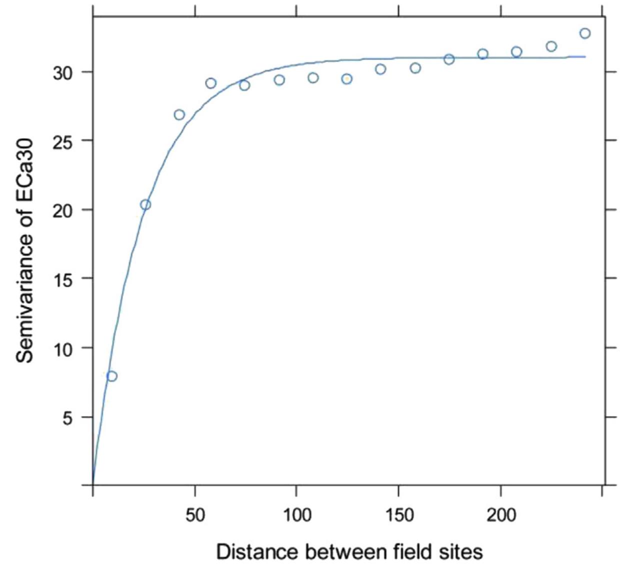

Fig. 7 – Empirical (points) and theoretical (line) semivariogram of the variable apparent electrical conductivity at 30 cm in depth.

the variables used for zone delineation (Fig. 10). The variables ECa30 and Sd were positively correlated and, along with ECa90, are the most important in explaining spatial variability at the first axis level (sPC1, horizontal axis). Previous studies (Fraisse et al., 2001; Sudduth, Hughes, & Drummond, 1995) have reported that zones with higher ECa values correspond to shallower soils where the clay horizon (Bt) is near the soil surface, and lower ECa values correspond to deeper soils where the Bt horizon is also deeper. Similar results were found by Peralta et al. (2015), working on the same soil types of the current work. Since sPC1 is the most important in explaining the total spatial variability, ECa and Sd provide more information for subsequent classification. The variable Elev was negatively correlated with ECa90 and was more important to explain sPC2.

For cluster analysis we selected the columns corresponding to sPC1and sPC 2 such that a large amount of total variation can be explained with the first two axes (≥70%). The Euclidean distance was the similarity distance included in the optimization function of the classification algorithm. The fuzzy partition in this spatial component space was obtained by setting a fuzziness weighting exponent at 1.3, as usual in

Table 1 – Interpolation accuracy of two geostatistical models for the variable electrical conductivity at 30 cm depth, based on mean squared errors (MSE), root mean squared error (RMSE) and mean squared deviation ratio (MSDR).

Model	MSE	RMSE	MSDR
Exponential	4.42	2.10	0.63
Spherical	5.20	2.28	0.57

precision agriculture applications (Fridgen et al., 2004; Guastaferro et al., 2010; Odeh, Chittleborough, & McBratney, 1992). There is no theoretical or computational evidence for distinguishing an optimal fuzziness coefficient; however, because of the smoothing produces in MULTISPATI-PCA, it is supposed to be greater than, or at least equal to, the coefficient used when original soil variables are the classification variables. The fuzzy k-means cluster was performed using the "e1071" package, which also enables to obtain indices for the selection of the number of classes. Here, we evaluate 2, 3 and 4 clusters using those indices. In all of calculated indices, except for partition coefficient, the optimum number of classes is obtained at the lowest index values. For a similar

```
Case  dfb.1_  dfb.x  dffit     cov.r    cook.d  hat
1098  0.10    -0.12  -0.13_*   1.00_*   0.01    0.00
1962  0.19    -0.22  -0.23_*   0.99_*   0.03    0.00_*
2077  0.18    -0.21  -0.22_*   0.99_*   0.02    0.00_*
5548  0.10    -0.12  -0.13_*   1.00_*   0.01    0.00_*
6362  0.09    -0.10  -0.11_*   1.00_*   0.01    0.00
```

Fig. 6 – DFBETAS (dfb.1_ for intercept and dfb.x for slope), DFFITS (dffit), Covratio (cov.r), Cook's distance (cook.d) and leverage (hat) of the first five inliers detected for ECa30 input data (R software output).

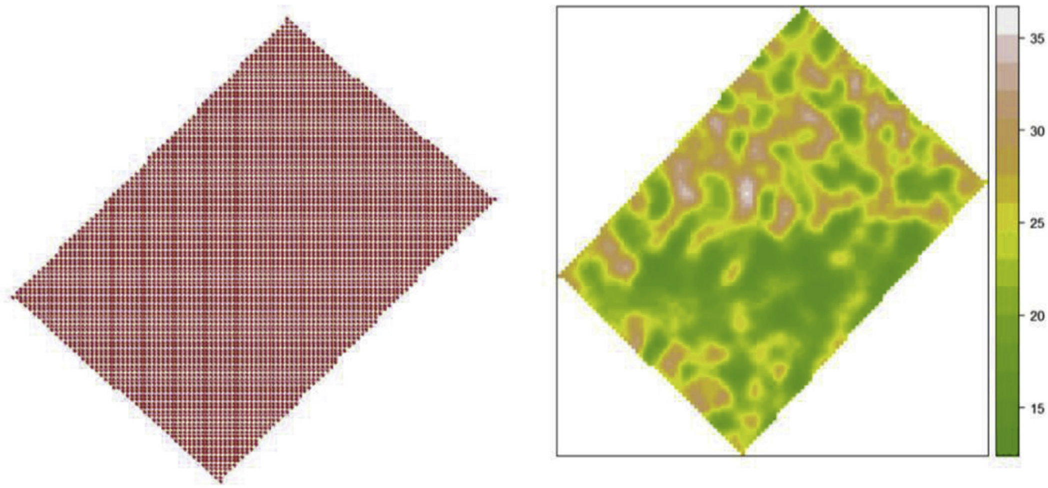

Fig. 8 – **Prediction grid (10 × 10 m) (left) and map of spatial interpolation of the variable apparent electrical conductivity at 30 cm in depth (right).**

Case	x	y	ECa30	ECa90	Elev	Sd
1	312432.78	5800234.17	25.13	28.51	160.41	-78.08
2	312422.78	5800244.17	26.87	28.15	160.42	-77.21
3	312432.78	5800244.17	24.90	27.96	160.43	-78.66
4	312412.78	5800254.17	27.42	27.57	160.42	-75.94
5	312422.78	5800254.17	26.19	27.49	160.43	-76.90

Fig. 9 – **Data set resulting from the application of steps 4.2 to 4.4 to the variables apparent electrical conductivity at 30 (ECa30) and 90 (ECa90) cm in depth, elevation (Elev), and soil depth (Sd). The first five rows are shown (R software output).**

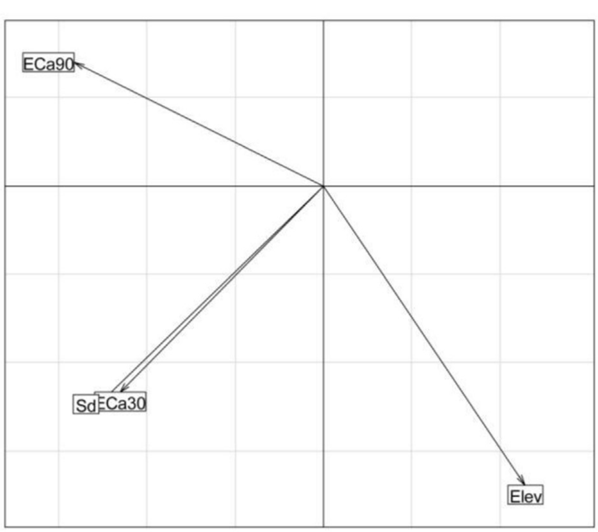

Fig. 10 – **Graphical display of the first two axes of the MULTISPATI-PCA analysis using the following variables: ECa30, electrical conductivity at 30 cm; ECa90, electrical conductivity at 90 cm; Elev, elevation; Sd, soil depth.**

interpretation of partition coefficient to those of the other indices, the value 1 was divided into the partition coefficient index (Fig. 11). In the present example, most of the indices, except for Fukuyama-Sugeno, show that, following a statistical criterion, two classes should be selected. The values of the summary index also show that two management classes should be selected. Likewise, classification with two management classes exhibits large zones with more coherent boundaries than those of classification with 3 and 4 classes, which present several small and irregular zones (Fig. 12). Thus, two classes were further processed to delineate MZs in this field. However, it may happen that the number of MZ indicated by indexes generated as a result several small and irregular areas. The use of calculated indices only provides a statistical metric but does not consider if the output is agronomically sensitive. When selecting the optimum number of management classes to be used, the ability to practically manage the classes needs to be considered (Taylor et al., 2007). For the smoothing of classification results we used the median filter. The masks defining the neighbourhood size (number of pixels) needed to apply this smoothing can be of different sizes. Here we tested masks of 5 × 5, 7 × 7 and 9 × 9 pixels. The 9 × 9 pixel filter (Fig. 13) was more suitable than the others

Index	2 Classes	3 Classes	4 Classes
XieBeni	3.50e-05	6.88e-05	6.33e-05
FukSug	-1.07e+04	-1.16e+04	-1.30e+04
PartCoef	1.06e+00	1.13e+00	1.14e+00
PartEntr	9.81e-02	1.98e-01	2.23e-01

Fig. 11 – Indices used to select the number of classes from fuzzy k-means cluster results. For each index the optimum class number among 2, 3 or 4 classes is suggested by the lowest index value (R software output).

Fig. 12 – Map with two (left), three (centre) and four (right) within-field management classes.

filters to achieve zoning of less abrupt boundaries and with no fragmentation.

3.4. *Comparison of delineated zones*

The comparison of means for the three variables selected for validation indicated statistically significant differences between homogeneous zones in SOM content and wheat yield (Fig. 14). No significant differences were found between zones-means for clay content. In the selection of the spatial correlation model for SOM variable, AIC indicated that the model with exponential spatial correlation function was the best one for these data (Table 2). For clay content the model with spherical spatial correlation function was the best one, whereas for wheat yield the exponential model was the selected one (Table 2).

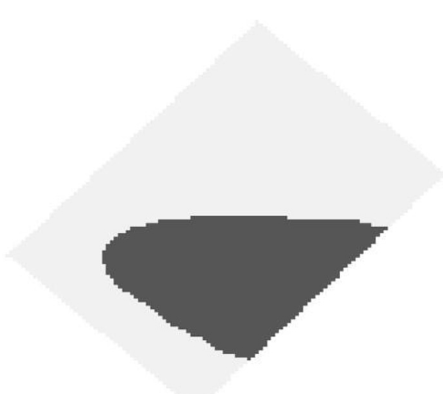

Fig. 13 – Map with two within-field management classes after application of a median filter of 9 × 9 pixels.

We present all of the steps for coming up with zones that are reasonably homogeneous for the soil characteristics that we have used for the classification. Even stating the yields differ, that does not mean that they optimal management decision for these zones differ. That would be a validation that they indeed are management zones. Further studies should be conducted to evaluate these subfield homogeneous zones to better understand the agronomic significance of this classification. In the study case, MZ 1 presents an SOM average of 4.60%, whereas for Zone 2 average SOM value is 4.24%. The MZ 2 was associated with deeper soils and greater yield (Peralta, Costa, Balzarini, & Castro Franco, 2013). In this area, there was possibly an accumulation of eroded material (Buschiazzo, 1986) and higher water accumulation and soil moisture (Kravchenko & Bullock, 2000). In contrast, the MZ 1 showed lower wheat grain yield, which was associated with shallower soils. Following multivariate classification, a field experiment involving variable rate application of inputs was carried out, providing environmental and economic benefits by decreasing fertilisation in the less-productive areas and minimizing the application of chemical substances as a strategy to obtain more cost-effective field management, including less use of agricultural machinery (Peralta et al., 2015).

Temporal stability of the MZs is another important issue in site-specific agriculture, which might require the extension of the current protocol. Further knowledge regarding the yield variation as caused by these soil properties might be gained by adjusting predictive models. The results of the protocol will depend on the input variables. Determining the optimum data layers needed to delineate management zones for a particular field without previous information is difficult. In general, emphasis should be given to sensors that record known or

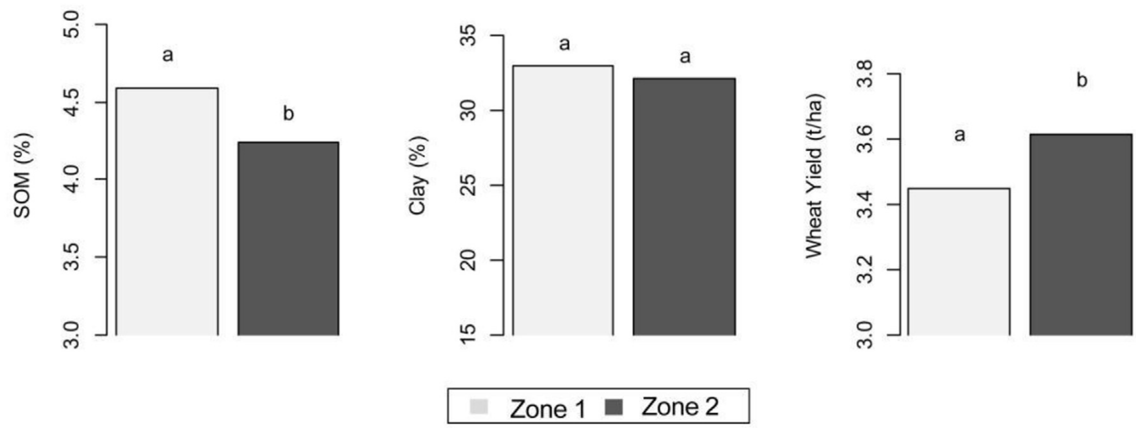

Fig. 14 – Mean differences of soil and yield variables for the delineated management zones. Different letters indicate statistically significant differences ($P < 0.05$) between zones.

expected yield determining factors (Taylor et al., 2007). There is agreement however, that use of multiple layers of data is necessary to adequately describe the spatial variability of a field. Overall, the most promising techniques to delineate management zones use multiple sources of data or layers for the purpose (Khosla et al., 2008). In this paper, MZ delineation was performed using ECa, Elev and Sd data layers. These data were easy accessible and indicative of production potential for the southern Argentine Pampas; hence, the protocol can be easily extrapolated to different cropping systems. However, this does not preclude the use of the statistical methodology with other valid data layers that may be relevant for another specific cropping system. If another variable is used to replace the ECa in soils of other region, the logical sequence of the statistical techniques recommended for multivariate homogeneous zone delineation will be the same as that in the proposed protocol.

4. Concluding remarks

This work presents an analytical protocol to implement a set of statistical techniques for the different stages of multivariate homogeneous zone delineation in PA. In its current state, the protocol integrates a number of statistical methods and models suitable for zoning with multiple variables described by spatial data. The implemented geostatistical and multivariate algorithms have been previously published and validated, but they were not available in a logical sequence of a single protocol. The development of the analytical outline as a pipeline in R, a software package of easy access and free distribution, might facilitate its implementation. The protocol here presented extends that of Taylor et al. (2007) by incorporating specific indices for inlier removal, the spatial principal component analysis to explore the spatial constraints of data during multivariate classification, filters to smooth the classification results and a mixed linear model approach to validate MZ using data with spatial correlation.

Table 2 – AIC (Akaike Information Criterion) for model selection. Mixed Linear Model for differences in soil organic matter, clay content, and wheat yield among delineated management zones.

Model	SOM	Clay	Wheat yield
Exponential	−1.52[a]	49.59[a]	1034.39[a]
Exponential + Nugget	0.48	51.59	1036.39
Spherical	−1.22	47.05	1088.46
Spherical + Nugget	−0.40	49.05	1041.39
Independent error	2.93	51.03	1675.30

[a] A lower AIC value indicates a better model fit.

Acknowledgements

We thank the producers, technical staff from EEA INTA Balcarce for their help in collecting the data for use in the study fields. We are also grateful to the Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET), Universidad Nacional de Córdoba and Ministerio de Ciencia y Tecnología for financial support.

Appendix A. Supplementary data

Supplementary data related to this article can be found at http://dx.doi.org/10.1016/j.biosystemseng.2015.12.008.

REFERENCES

Aggelopooulou, K., Castrignanò, A., Gemtos, T., & De Benedetto, D. (2013). Delineation of management zones in an apple orchard in Greece using a multivariate approach. *Computers and Electronics in Agriculture, 90*, 119–130.

Akaike, H. (1976). An information criterion (AIC). *Mathematical Sciences, 14*(153), 5–9.

Amidan, B., Ferryman, T., & Cooley, S. (2005). Data outlier detection using the Chebyshev theorem. In *IEEE Aerosp. Conf* (pp. 3814–3819).

Anderberg, M. R. (1973). *Cluster analysis for applications*. New York: Academic Press, Inc.

Anselin, L. (1995). Local indicators of spatial association – LISA. *Geographical Analysis, 27*, 93–115.

Anselin, L. (1996). The Moran scatterplot as an ESDA tool to assess local instability in spatial association. In M. Fischer, H. Scholten, & D. Unwin (Eds.), *Spatial analytical perspectives on GIS* (pp. 111–125). London: Taylor and Francis.

Arce, G. R. (2005). *Nonlinear signal processing: A statistical approach*. New Jersey: John Wiley and Son.

Arno, J., Martinez-Casasnovas, J. A., Ribes-Dasi, M., & Rosell, J. R. (2011). Clustering of grape yield maps to delineate site-specific management zones. *Spanish Journal of Agricultural Research, 9*, 721–729.

Barbieri, P., Echeverría, H., & Sainz Rozas, H. (2009). Nitrates in soil at planting or tillering as a diagnostic of the nitrogenated nutrition in wheat in the Southeastern Pampas. *Soil Science, 27*, 41–47.

Bezdek, J. C. (1981). *Pattern recognition with fuzzy objective function algorithms*. New York: Plenum Press.

Bivand, R. (2014). *spdep: Spatial dependence: weighting schemes, statistics and models*. R package version 0.5–75/r559 http://R-Forge.R-project.org/projects/spdep/.

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.

Bland, J. M., & Altman, D. G. (1995). Multiple significance tests: the Bonferroni method. *BMJ, 310*(6973), 170.

Boydell, B., & McBratney, A. B. (2002). Identifying potential within-field management zones from cotton-yield estimates. *Precision Agriculture, 3*, 9–23.

Burgos, J. J., & Vidal, A. L. (1951). The climates of the Argentine Republic according to the new Thornthwaite classification. *Annals of the Association of American Geographers, 41*, 237–263.

Buschiazzo, D. (1986). Estudio sobre la tosca. Parte I: evidencias de un movimiento descendente del carbonato en base a la interpretación de características macro y geomorfológicas. *Ciencia del Suelo, 4*, 55–65.

Chessel, D., Dufour, A. B., & Thioulouse, J. (2004). The ade4 package-I – One-table methods. *R News*, 5–10.

Córdoba, M., Bruno, C., Costa, J. L., & Balzarini, M. (2013). Subfield management class delineation using cluster analysis from spatial principal components of soil variables. *Computers and Electronics in Agriculture, 97*, 6–14.

Corwin, D. L., & Lesch, S. M. (2010). Delineating site-specific management units with proximal sensors. In O. Margaret (Ed.), *Geostatistical applications in precision agriculture* (pp. 139–165). New York: Springer.

Davatgar, N., Neishabouri, M. R., & Sepaskhah, A. R. (2012). Delineation of site specific nutrient management zones for a paddy cultivated area based on soil fertility using fuzzy clustering. *Geoderma, 173–174*, 111–118.

Dewis, J., & Freitas, F. (1970). *Métodos físicos y químicos de análisis de suelos y aguas*. FAO. Boletín sobre Suelos N° 10.

Draper, N. R., & Smith, H. (1998). *Applied regression analysis*. New York, 3rd Edition: John Wiley and Sons.

Dray, S., Legendre, P., & Peres-Neto, P. R. (2006). Spatial modelling: a comprehensive framework for principal coordinate analysis of neighbour matrices (PCNM). *Ecological Modelling, 196*(3), 483–493.

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*, 45–56.

Falivene, O., Cabrera, L., Tolosana-Delgado, R., & Sáez, A. (2010). Interpolation algorithm ranking using cross-validation and the role of smoothing effect. A coal zone example. *Computers & Geosciences, 36*(4), 512–519.

Farahani, H. J., & Flynn, R. L. (2007). Map quality and zone delineation as affected by width of parallel swaths of mobile agricultural sensors. *Biosystems Engineering, 96*, 151–159.

Fleming, K. L., Westfall, D. G., Wiens, D. W., & Brodahl, M. C. (2000). Evaluating farmer defined management zones for variable rate fertilizer application. *Precision Agriculture, 2*, 201–215.

Fraisse, C., Sudduth, K., & Kitchen, N. (2001). Delineation of site-specific management zones by unsupervised classification of topographic attributes and soil electrical conductivity. *Transactions of the ASAE, 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 Journal, 96*, 100–108.

Frogbrook, Z. L., & Oliver, M. A. (2007). Identifying management zones in agricultural fields using spatially :constrained classification of soil and ancillary data. *Soil Use Manage, 23*, 40–51.

Fukuyama, Y., & Sugeno, M. (1989). A new method of choosing the number of clusters for the fuzzy c-means method. In *Proc. 5th fuzzy Syst. Symp* (pp. 247–250).

Galarza, R., Mastaglia, N., Albornoz, E. M., & Martínez, C. E. (2013). Identificación automática de zonas de manejo en lotes productivos agrícolas. In *V Congreso Argentino de Agroinformática (CAI) – 42da. JAIIO*.

Getis, A., & Ord, K. (1992). The analysis of spatial association by use of distance statistics. *Geographical Analysis, 24*, 189–206.

Gili, A. A. (2013). *Modelación de la variación espacial de variables edáficas y su aplicación en el diseño de planes de muestreo de suelos*. PhD Thesis. Argentina: Universidad Nacional de Córdoba.

Gonzalez, R. C., & Woods, R. (2008). *Digital image processing*. Upper Saddle River, New Jersey: Pearson Prentice Hall.

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

Khosla, R., Inman, D., Westfall, D. G., Riech, R., Frasier, W. M., Mzuku, M., et al. (2008). A synthesis of multi-disciplinary research in precision agriculture: site specific management zones in the semi-arid western Great Plains of the USA. *Precision Agriculture, 9*, 85–100.

Khosla, R., Westfall, D. G., Reich, R. M., Mahal, J. S., & Gangloff, W. J. (2010). Spatial variation and site-specific management zones. In O. Margaret (Ed.), *Geostatistical applications in precision agriculture* (pp. 195–219). New York: Springer.

Kravchenko, A. N., & Bullock, D. G. (2000). Correlation of grain yield with topography and soil properties. *Agronomy Journal, 92*, 75–83.

Kravchenko, A. N., Robertson, G. P., Thelen, K. D., & Harwood, R. R. (2005). Management, topographical, and weather effects on spatial variability of crop grain yields. *Agronomy Journal, 97*(2), 514–523.

Lark, R. M. (1998). Forming spatially coherent regions by classification of multi-variate data: an example from the analysis of maps of crop yield. *International Journal of Geographical Information Science, 12*, 83–98.

Lawes, R. A., & Bramley, R. G. V. (2012). A simple method for the analysis of on-farm strip trials. *Agronomy Journal, 104*(2), 371–377.

Long, D. S. (1998). Spatial autoregression modeling of site-specific wheat yield. *Geoderma, 85*, 181–197.

McBratney, A. B., & Pringle, M. J. (1999). Estimating average and proportional variograms of soil properties and their potential use in precision agriculture. *Precision Agriculture, 1*, 212–236.

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 1.6-4 http://CRAN.R-project.org/package=e1071.

Minasny, B., & McBratney, A. B. (2002). *FuzME version 3.0, Australian Centre for Precision Agriculture, McMillan Building A05, The University of Sydney, NSW 2006*.

Minasny, B., McBratney, A. B., & Whelan, B. M. (2005). *VESPER version 1.62. Australian Centre for Precision Agriculture, McMillan Building A05, The University of Sydney, NSW 2006*.

Moral, F. J., Terrón, J. M., & Marques Da Silva, J. R. (2010). Delineation of management zones using mobile measurements of soil apparent electrical conductivity and multivariate geostatistical techniques. *Soil and Tillage Research, 106*, 335–343.

Odeh, I. O. A., Chittleborough, D. J., & McBratney, A. B. (1992). Soil pattern recognition with fuzzy-c-means: application to classification and soil-landform interrelationships. *Soil Science Society of America Journal, 56*, 505.

Oliver, M. A., & Webster, R. (2014). A tutorial guide to geostatistics: computing and modelling variograms and kriging. *Catena, 113*, 56–69.

Pebesma, E. J. (2004). Multivariable geostatistics in S: the gstat package. *Computers & Geosciences, 30*, 683–691.

Pennock, D. J. (2003). Terrain attributes, landform segmentation, and soil redistribution. *Soil and Tillage Research, 69*(1–2), 15–26.

Peralta, N. R., Costa, J. L., Balzarini, M., & Angelini, H. (2013). Delineation of management zones with measurements of soil apparent electrical conductivity in the southeastern pampas. *Canadian Journal of Soil Science, 93*, 205–218.

Peralta, N. R., Costa, J. L., Balzarini, M., & Castro Franco, M. (2013). Delimitación de zonas de manejo con modelos de elevación digital y profundidad de suelo. *Interciencia, 38*(6), 1–7.

Peralta, N. R., Costa, J. L., Balzarini, M., Castro Franco, M., Córdoba, M., & Bullock, D. (2015). Delineation of management zones to improve nitrogen management of wheat. *Computers and Electronics in Agriculture, 110*, 103–113.

Ping, J. L., & Dobermann, A. (2003). Creating spatially contiguous yield classes for site specific management. *Agronomy Journal, 95*, 1121–1131.

Pinheiro, J., Bates, D., DebRoy, S., Sarkar, D., & R. Core Team. (2014). *nlme: Linear and nonlinear mixed effects models*. R package version 3.1–118. URL http://CRAN.R-project.org/package=nlme, 670.

R Core Team. (2015). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. URL http://www.R-project.org/.

Ribeiro Junior, P. J., & Diggle, P. J. (2001). geoR: a package from geostatistical analysis. *R News, 1*(2), 15–18.

Rodríguez-Pérez, J. R., Plant, R. E., Lambert, J. J., & Smart, D. R. (2011). Using apparent soil electrical conductivity (ECa) to characterize vineyard soils of high clay content. *Precision Agriculture, 12*, 775–794.

Roel, A., & Plant, R. E. (2004). Spatiotemporal analysis of rice yield variability in two California fields. *Agronomy Journal, 96*(1), 77–90.

Sanchez, M. G. B., Marques, J., Jr., Siqueira, D. S., Camargo, L. A., & Pereira, G. T. (2013). Delineation of specific management areas for coffee cultivation based on the soil–relief relationship and numerical classification. *Precision Agriculture, 14*(2), 201–214.

Satorre, E. H., & Slafer, G. A. (1999). Wheat production systems of the Pampas. In *Wheat: Ecology and physiology of yield determination* (pp. 333–348). New York, NY: Food Products Press.

Simbahan, G. C., Dobermann, A., & Ping, J. L. (2004). Site-specific management: screening yield monitor data improves grain yield maps. *Agronomy Journal, 96*, 1091–1102.

Siqueira, D. S., Marques, J., Jr., & Pereira, G. T. (2010). The use of landforms to predict the variability of soil and orange attributes. *Geoderma, 155*(1–2), 55–66.

Sudduth, K. A., & Drummond, S. T. (2007). Yield Editor. *Agronomy Journal, 99*(6), 1471–1482.

Sudduth, K. A., Hughes, D. F., & Drummond, S. T. (1995). Electromagnetic induction sensing as an indicator of productivity on claypan soils. In P. C. Robert, et al. (Eds.), *Proc. Int. Conf. on Site-Specific Management for Agricultural Systems, Minneapolis, Madison, USA* (pp. 671–681).

Sun, W., Whelan, B., McBratney, A. B., & Minasny, B. (2013). An integrated framework for software to provide yield data cleaning and estimation of an opportunity index for site-specific crop management. *Precision Agriculture, 14*, 376–391.

Taylor, J. A., McBratney, A. B., & Whelan, B. M. (2007). Establishing management classes for broadacre agricultural production. *Agronomy Journal, 99*, 1366–1376.

Webster, R., & Oliver, M. A. (2007). *Geostatistics for environmental scientists* (2nd ed.). Chichester: John Wiley & Sons.

West, T. B., Welch, K. B., & Galecki, A. T. (2015). *Linear mixed models: A practical guide using statistical software* (2nd ed.). Boca Raton: Chapman & Hall/CRC, 440 pp.

Windham, M. P. (1981). Cluster validity for fuzzy clustering algorithms. *Fuzzy Sets and Systems, 5*, 177–185.

Xie, L. X., & Beni, G. (1991). Validity measure for fuzzy clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 3*(8), 841–847.

Yao, R. J., Yang, J. S., Zhang, T. J., Gao, P., Wang, X. P., Hong, L. Z., et al. (2014). Determination of site-specific management zones using soil physico-chemical properties and crop yields in coastal reclaimed farmland. *Geoderma, 232*, 381–393.