THE INTERNATIONAL SOCIETY OF PRECISION AGRICULTURE PRESENTS THE
14th INTERNATIONAL CONFERENCE ON
PRECISION AGRICULTURE
JUNE 24th-27th 2018
Le Centre Sheraton • Montreal, Quebec • Canada

Creating thematic maps and management zones for agriculture fields

[1]Souza, E. G.; [2]Schenatto, K.; [3]Bazzi, C. L.

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

A paper from the Proceedings of the

14th International Conference on Precision Agriculture

June 24 – June 27, 2018

Montreal, Quebec, Canada

Abstract. *Thematic maps (TMs) are maps that represent not only the land but also a topic associated with it, and they aim to inform through graphic symbols where a specific geographical phenomenon occurs. Development of TMs is linked to data collection, analysis, interpretation, and representation of the information on a map, facilitating the identification of similarities, and enabling the visualization of spatial correlations. Important issues associated with the creation of TMs are: selection of the coordinate system; exploratory data analysis; data interpolation; decision of the number of classes and the method for breaking the data into ranges; choosing an effective color scheme. A special kind of TMs is management zones (MZs), where MZ is a subregion of a field that expresses a functionally homogeneous combination of yield-limiting factors for which a single rate of a specific crop input is appropriate. The use of MZs in sampling is likely to reduce laboratory costs while maintaining the level of reliability, and it has shown to improve the use efficiency of nutrients, maintaining or increasing the yield and potentially reducing the overloading of nutrients into the environment. Several approaches have been developed to define MZs; these are often classified as empirical or clustering according to the technique used. Here the important issues associated with their creation are: selection of the variable to be used in the cluster analysis; necessity of data normalization; decision of method of delineation of MZs; using of indices and ANOVA to evaluate quality of MZs. In this context, the objective of this work was to discuss the particularities behind the creation of TMs and MZs for agriculture fields.*

Keywords. *Contour maps, cluster analysis, Precision agriculture.*

The authors are solely responsible for the content of this paper, which is not a refereed publication.. Citation of this work should state that it is from the Proceedings of the 14th International Conference on Precision Agriculture. EXAMPLE: Lastname, A. B. & Coauthor, C. D. (2018). Title of paper. In Proceedings of the 14th International Conference on Precision Agriculture (unpaginated, online). Monticello, IL: International Society of Precision Agriculture.

Introduction

In addition to representing the terrain, thematic maps (TMs) are used to illustrate themes. Generally, TMs are used to identify different cartographic representations, and they represent not only the land but also associated characteristics. The development of TMs is linked to data collection, analysis, interpretation, and representation of the information on a map. This facilitates the identification of similarities and enables the visualization of spatial correlations. One specific case of TMs is contour maps, which are built by connecting points of the same value, and are applicable to geographical phenomena that show continuity in the geographic space. Contour maps can be constructed from absolute data (elevation, temperature, precipitation, humidity, and atmospheric pressure) or relative data (density, percentages, and indexes). Based on samples collected before, during e after the life period of the culture, TMs are generated in order to identify the variability of properties of the topography, soil, and plants, and to compare with the yield. However, first it is necessary to interpolate the data into a dense and regular grid to provide values for locations that were not sampled. This task is performed with the aid of interpolation methods, and geostatistical analysis is the most used interpolation method. In this analysis the spatial variability is determined through monitoring and measurements, making it possible to create a plan for the correction of any deficiencies, particularly when localized management is intended, in order to improve soil quality, and, consequently, increase production (Davidson, 2014; Mzuku et al., 2005).

Timlin et al. (1998) showed that data on spatial variability and distribution of productivity can be effectively used in localized management of inputs (precision agriculture) with the aim of increasing the efficiency of fertilizers and environmental sustainability, although it is often costly (Khosla et al., 2008).

Normally, soil samples are analyzed to determine the levels of nutrients in the soil. The sampling, therefore, should be dense enough to allow the determination of the variability of nutrients in the soil so that fertilizers can be used profitably and in an environmentally sustainable manner (Ferguson and Hergert, 2009; Franzen et al., 2002). In order to determine the appropriate density of soil sampling in an area, the time and budget available for sampling should be considered.

A management zone (MZ) is a subregion of a field that expresses a functionally homogeneous combination of yield-limiting factors for which a single rate of a specific crop input is appropriate (Doerge, 2000; Moral et al., 2010; Moshia et al., 2014; Bobryk et al., 2016). Although variable-rate application machines could be used, MZs usually involve conventional machinery. After delineation of the MZs, they can be used in smart sampling, and the number of samples needed to delineate the field soil variability can be reduced to one composite sample per zone (according Wollenhaupt, Wolkowski, and Clayton, 1994), the use of use of subsamples collected around georeferenced points ensures superior evaluation of nutrients in the area). This approach (smart sampling) is likely to reduce laboratory costs while maintaining the level of reliability (Ferguson and Hergert, 2009; Mallarino and Wittry, 2004), and it has shown to improve the use efficiency of nutrients, maintaining or increasing the yield and potentially reducing the overloading of nutrients into the environment (Moshia et al., 2014; Khosla et al., 2002). Many studies related to the sampling density have been performed (Journel and Huijbregts, 1978; Demattê et al., 2014; Wollenhaupt, Wolkowski, and Clayton, 1994; Franzen et al., 2002; Ferguson and Hergert, 2009; Doerge, 2000), resulting in a suggested minimum density of 1 sample ha−1 (Ferguson and Hergert, 2009) to 2.5 sample ha−1 (Journel and Huijbregts, 1978; Doerge, 2000), which should be composed of at least eight individual samples (Wollenhaupt, Wolkowski, and Clayton, 1994).

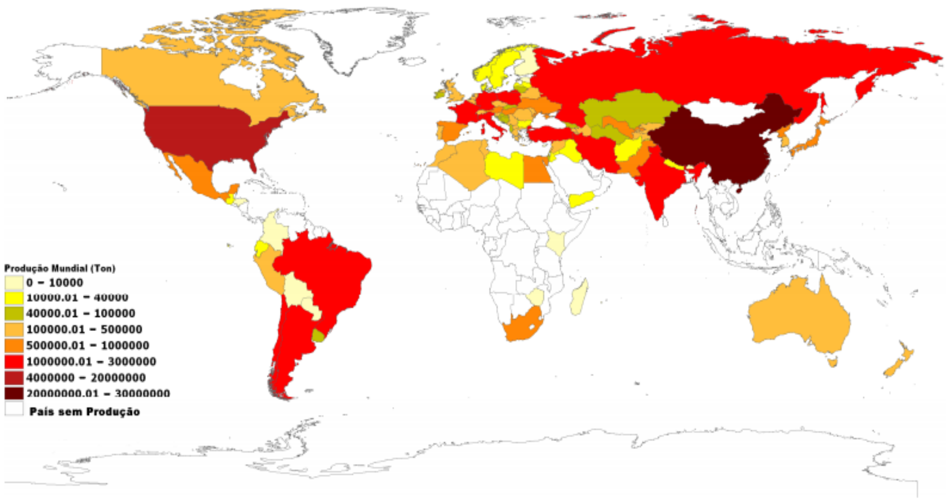
Several kinds of sample data can be used to define MZs; however, it is advantageous to use a set of multivariate data attributes that do not vary significantly over time (topography, electrical conductivity, physical properties of the soil) and that are correlated with target variable (usually yield), thus producing more stable MZs (Buttafuoco et al., 2010; Doerge, 2000). That is important because usually we want to use the MZs for many years. But there are other situations in which the purpose is to use immediately and just the once the MZs. It is the case MZs for agrochemicals applications.

In this context, the objective of this work was to discuss the particularities behind the creation of TMs and MZs for agriculture fields.

Methodology

Thematic Maps

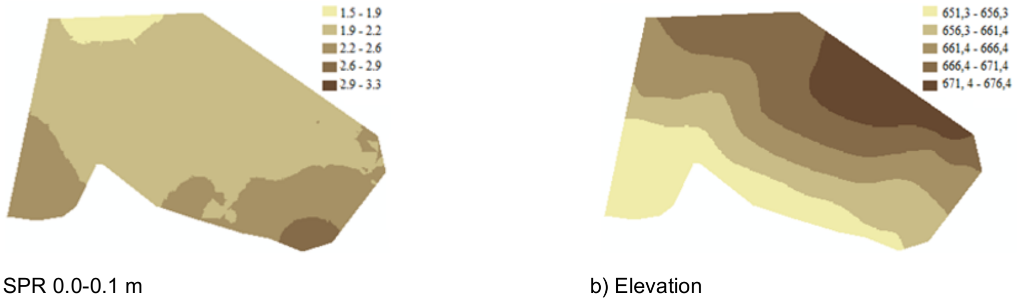
Maps that represent not only the land but also a topic associated with it are called thematic maps (TMs), and they aim to inform through graphic symbols where a specific geographical phenomenon occurs. Development of TMs is linked to data collection, analysis, interpretation, and representation of the information on a map, facilitating the identification of similarities, and enabling the visualization of spatial correlations. The information presented in TMs may include, for example, maximum temperature or maximum precipitation at a given date, amount of calcium and potassium in the soil, and soybean yield at a given agricultural area. Fig. 1 shows a TM of world apple production in 2009.



Source: Carvalho (2011).

Fig. 1. Thematic map of world apple production in 2009.

The combination of visual variables gives rise to different types of TMs (e.g., contour maps, zonal maps, graduated circle maps). Contour maps are built by connecting points of the same value, and are applicable to geographical phenomena that show continuity in the geographic space. They can be constructed from absolute (elevation, temperature, precipitation, humidity, atmospheric pressure) or relative data (density, percentages, or indexes). Fig. 2 shows two examples of contour maps.



a) SPR 0.0-0.1 m

b) Elevation

Fig. 2. Examples of contour maps: a) Soil Penetration Resistance (SPR)(Mpa), b) elevation (m).

In order to construct TMs about attributes collected in agriculture fields, it is necessary to follow a protocol like the presented at Fig. 3.

Coordinate system - A Geographic Information System (GIS Software) is designed to store, retrieve, manage, display, and analyze all types of geographic and spatial data. To construct TMs in 2D we use GIS software and a file with at least three columns representing the X (longitude) and Y (latitude) coordinates and the value of the measured attribute (for 3D, we need one more coordinate, z (altitude)). These coordinates are associated a coordinate system, being the most typical the geographic and UTM (universal transverse Mercator). The geographic coordinate system is associated to model of the Earth shape (reference ellipsoid) called datum. The datum WGS84 (World Geodetic System 84) is most commonly used. In the geodetic coordinate system, the units are in degrees, minutes and seconds. More practical, UTM uses coordinates in meters.

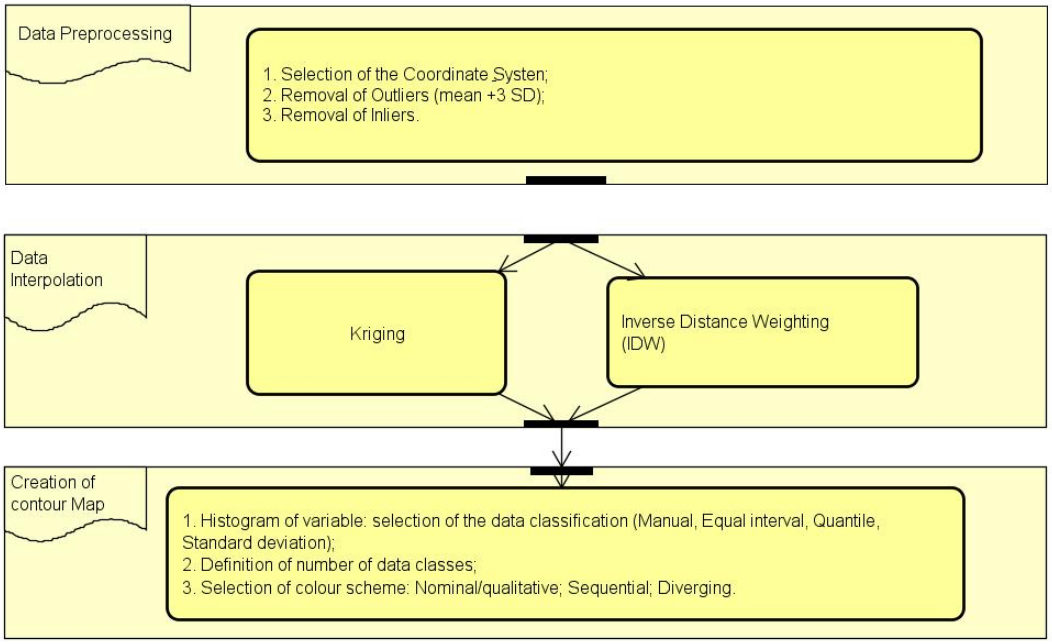

Data Preprocessing
1. Selection of the Coordinate Systen;
2. Removal of Outliers (mean +3 SD);
3. Removal of Inliers.
Data Interpolation
Kriging
Inverse Distance Weighting (IDW)
Creation of contour Map
1. Histogram of variable: selection of the data classification (Manual, Equal interval, Quantile, Standard deviation);
2. Definition of number of data classes;
3. Selection of colour scheme: Nominal/qualitative; Sequential; Diverging.

Fig. 3. Flowchart of the typical protocol do create a thematic map.

Exploratory data analysis (EDA) – is the summarization of the data set through their main characteristics. EDA employs a variety of techniques (mostly graphical) to maximize insight into a data set; uncover underlying structure; extract important variables; detect inliers and outliers (atypical values) and anomalies; test underlying assumptions; develop parsimonious models; and determine optimal factor settings (NIST/SEMATECH, 2013). When constructing TMs, the most important use of EDA is to detect and remove outliers. According Amidan et al. (2005), data outliers can have a significant impact upon data-driven decisions, and in many cases, they do not reflect the true nature of the data and, hence, should not be included in the analyses. They proposed an outlier detection method using Chebyshev's inequality to form a data-driven outlier detection method that is not dependent upon knowing the distribution of the data. According to Córdoba et al. (2016), using Chebyshev's theorem, the values that are outside the mean ± 3 SD are identified as outliers and should be removed (also Haghverdi et al., 2015), They remarked that even though real data could belong to this interval, the upper and lower limits use to be modified to obtain robust variance estimators. Also important are the removal of inliers, data that differ significantly from their neighborhood, but lie within the general range of variation of the data set (Cordoba et. Al, 2016).

Data interpolation – With the objective of generating TMs and MZs that are continuous and smooth, usually the sample data are interpolated in a dense and regular grid to provide values in the places that were not sampled. This task is performed with the aid of interpolation methods.

The inverse distance weighting (IDW) and kriging are the interpolation methods commonly used in precision agriculture (PA) and are differentiated by the way the weights are assigned to the different samples, which can influence the estimated values (REZA et al., 2010). They are many software available for performing data interpolation, such as Surfer (Golden Software, Inc.) and ArcGIS (ESRI, Environmental Systems Research Institute).

Geostatistics analysis is the most used method of interpolation. The best geostatistical model for a series of georeferenced data is selected by comparing theoretical values with those obtained from sampling, and then analyzing the estimation errors and choosing the best model (Arlot and Celisse, 2010; Kohavi, 1995). This technique, called cross-validation, was chosen by Faraco et al. (2008) as the best way to evaluate the adjustment of spatial theoretical models, and was deemed better than Akaike's and Filiben's information criteria and the maximum logarithm value of the likelihood function.

Cross-validation allows for the evaluation of estimation errors by comparing predicted values with values determined through samples. The average error (AE) is calculated as the arithmetic average difference between the original values and those simulated by the interpolation, temporarily discarding any sample taken from the same location where the estimate is made by the interpolator (Isaaks and Srivastava, 1989). Other measures indicating the accuracy of the estimation are then calculated using the reduced average error (RE), standard deviation of the average error (SAE), and standard deviation of the reduced error (SRE) (Cressie, 1993; McBratney and Webster, 1986). According to non-tendentiousness criteria, to choose the best-adjusted model, the values for AE and (RE) should be as close to zero as possible, the value of SAE should be as small as possible, and the value of SRE should be close to 1 (Cressie, 1993; McBratney and Webster, 1986). Because cross-validation makes it possible for ambiguous situations to occur, Souza et al. (2016) propose an estimation called the error comparison index (ECI), which uses (RE) and SRE as interpolators. This allows for the calculation of the standard deviation of the interpolation. ECI is considered a better semivariance model that one using a lower ECI.

$$ECI_i = \frac{ABS(\overline{RE})_i}{\max \left|_{i=1}^{j} [ABS(\overline{RE})]\right.} + \frac{ABS(S_{AE}-1)_i}{\max \left|_{i=1}^{j} [ABS(S_{AE}-1)]\right.} \tag{1}$$

where ECI_i is the error comparison index for model *i*, $ABS(\overline{RE})$ is the absolute value of the reduced average error differing from zero of the cross-validation, S_{AE} is the standard deviation of the reduced average error differing from zero, and $\max \left|_{i=1}^{j}\right.$ is the highest value among the compared *j* semivariograms.

In case of using the deterministic and stochastic methods of interpolation, the best method can be selected using the interpolation selection index (ISI, Equation 2; Bier and Souza, 2017), which assumes a lower value as the interpolator improves:

$$ISI = \left\{ \frac{abs(AE)}{\max \left|_{i=1}^{j} [abs(AE)]\right.} + \frac{\left[S_{AE} - \min \left|_{i=1}^{j} (S_{AE})\right. \right]}{\max \left|_{i=1}^{j} [abs(S_{AE})]\right.} \right\} \tag{2}$$

where:

$$AE = \frac{1}{n} \sum_{i=1}^{n} \left(Z(s_i) - \hat{Z}(s_i) \right) \tag{3}$$

$$S_{AE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(Z(s_i) - \hat{Z}(s_i)\right)^2} \qquad (4)$$

where, AE – average error; S_{AE} – standard deviation of the average errors; n – number of data; $Z(S_i)$ – value observed at point S_i; $\hat{Z}(S_i)$ – value predicted by kriging at point S_i; $abs(AE)$ – module value of the average error of the crossed validation; $min|_{i=1}^{j}$ – lowest value found between the compared j models; $max|_{i=1}^{j}$ – highest value found between the compared j models.

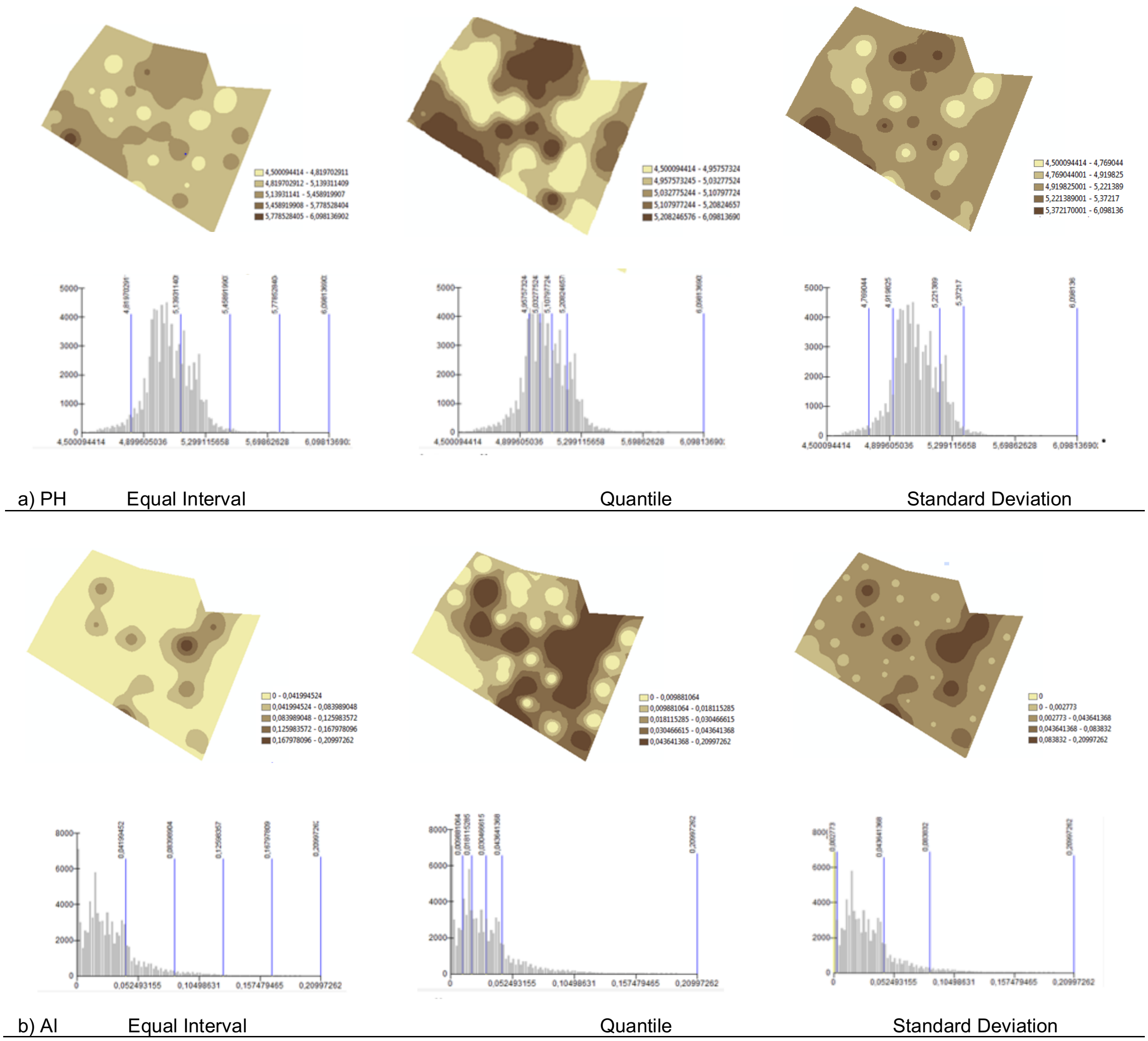
Creation of Contour Maps – after the data interpolation, to draw TMs with our data, we must decide both the number of classes and the method for breaking the data into ranges. The goal is to group together similar observations and split apart observations that are substantially different (Indiemapper, 2016). The first thing to do is looking at the histogram (or scatterplot) to determine the 'form' of your observations. This important step of map creation and how we do that can dramatically change the look of the map, and thus, its message, and it is one of the easiest ways to "lie with maps". There is no escape from the cartographic paradox: to present a useful and truthful picture, an accurate map must tell white lies (Monmonier, 1996). They are many ways to systematically classify data and each GIS software will offer some of them. The most popular are:

- Manual interval: we set manually the one or all of the class breaks. We use this method when the others do not give a good solution. A good way is to start with one of the standard classifications and make adjustments as needed
- Equal interval: we divide the data into equal size classes and works well on data that is generally spread across the entire range. This classification should be avoided if data are skewed to one end or there are one or two really large outlier values.
- Quantile: we divide in classes with an equal number of features and work well on data that is linearly distributed across the entire range. Nevertheless, the resulting map can be misleading, with similar features placed in adjacent classes, or widely different values put in the same class.
- Standard deviation: it a special case of equal interval where the class size is a multiple of standard deviation and work well with data that has a normal distribution. It is god for seeing which features are above or below an average value.

The number of data classes is also an important part of contour map design. Increasing the number of data classes will result in a more revealing map but require more colors. Generally, it is advised not to exceed of seven classes.

Examples of contour maps are presented in the Fig. 4. In each case is presented the map using five classes, classified by equal interval, quantile and standard deviation, and its corresponding histogram. The case of pH (Fig. 4a) we have an attribute with a distribution closes to normal and the equal interval classification looks like the best choice but standard deviation classification is also good. However, with the map of aluminium (Fig. 4b), the distribution is moderately skewed right and then the quantile is visually the best option.

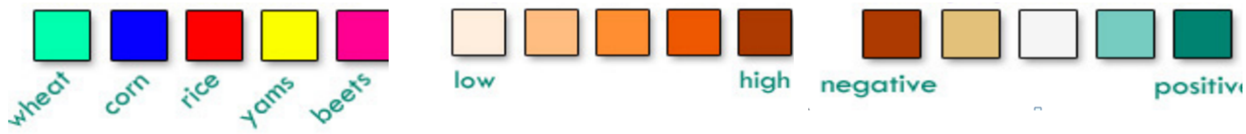
After we selected the way to classify the data it is important to choose an effective color scheme for the TM. A good color scheme needs to be attractive but also support the message of the map and be appropriately matched to the nature of the data (Harrower and Brewer, 2003), being necessary to choose three dimensions of color: hue, lightness, and saturation. There are three kinds of color scheme: **Nominal/qualitative** (unorderable data, like land use, Fig. 5a): different hues that keep lightness, and saturation constant should be used; **Sequential** (orderable, like numerical data (or low/med/high), like yield, Fig. 5b): single or multi-hue with different lightness/saturation should be used; **Diverging** (when there is a mid-point, like zero, or if we want compare with an average, like profit, Fig. 5c). Harrower and Brewer (2003) designed an online tool “ColorBrewer.org” to help users to select appropriate color schemes for their specific mapping needs. Fig. 6 presents some practical examples of application of color scheme.



a) PH Equal Interval Quantile Standard Deviation

b) Al Equal Interval Quantile Standard Deviation

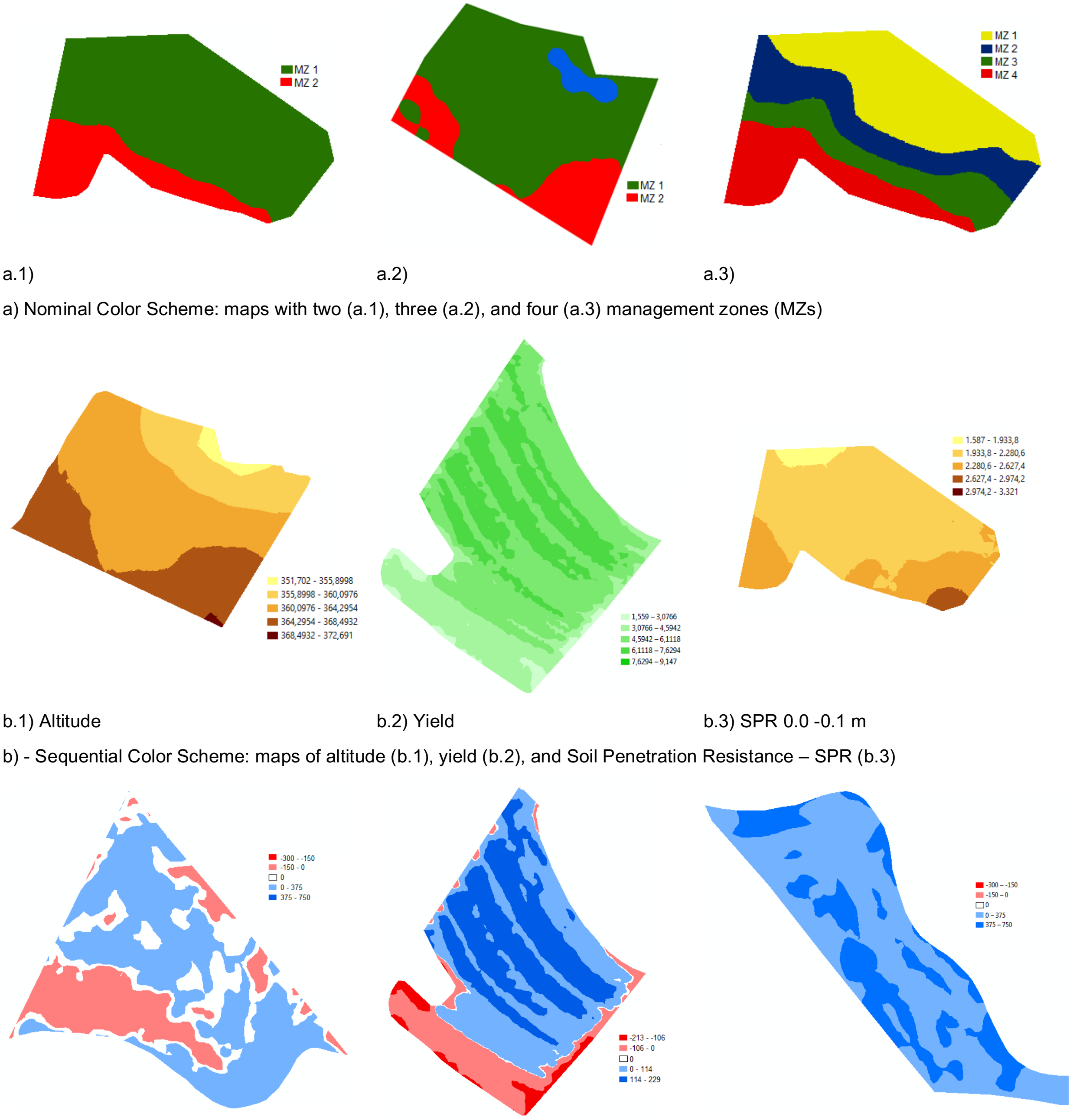
Fig. 4. Thematic Maps for pH (a), and aluminium (b) using three form of classification.



a) Nominal Color Scheme b) Sequential Color Scheme c) Diverging Color Scheme

Fig. 5. Three kinds of color scheme: nominal/qualitative (a), sequential (b), and diverging (c).

Contours maps using continuous scale - despite of most common are using a discrete scale, some people like most continuous scale. The problem with a color ramp is that perception of color intensity is not linear and consequently the user could make a false assumption about what data value it represented. Basso et al. (2009), studying the effects of landscape position and rainfall on spatial variability of wheat yield and protein on a 10-ha field with rolling landscape of Southern Italy, presented an interpolated map of wheat yield (Fig. 7) using a continuous scale.



a.1) a.2) a.3)

a) Nominal Color Scheme: maps with two (a.1), three (a.2), and four (a.3) management zones (MZs)

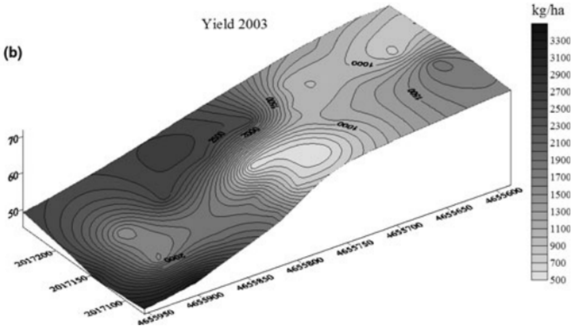
b.1) Altitude b.2) Yield b.3) SPR 0.0 -0.1 m

b) - Sequential Color Scheme: maps of altitude (b.1), yield (b.2), and Soil Penetration Resistance – SPR (b.3)

c.1) c.2) c.3)

c) - Diverging Color Scheme – Profit Maps (US\$ ha^{-1})

Fig. 6. Examples of color scheme: nominal/qualitative (a), sequential (b), and diverging (c).



Source: Basso et al. (2009).

Fig. 7. 3D interpolated map of wheat yield (kg ha^{-1}) for 2003.

Management Zones

Despite of the original concept of an MZ be is a sub-region of a field that expresses a functionally homogeneous combination of yield-limiting factors, the target agriculture variables can be other than yield, like pest & disease infestation, water content, brix, soil resistance to penetration, etc. A MZ can be used for one year or for several years, usually three to five. This fact is very important when we are choosing variables. If we are planning to use only once, that can be the case of weed infestation, we can use variables that are not temporally stable (like weed infestation), to create the MZs. But, in most cases we want to use the MZs for multiple years and we should to use relatively temporally stable variables like topography data (elevation, and slope), and physical data (Bulk density, soil texture, Soil Penetration Resistance – SPR).

There are many approaches presented in literature for the purpose of delineating management zones using yield map. Among those approaches, the two approaches commonly used for delineating MZs using yield maps (Xiang et al., 2007) include: 1) The empirical method, which uses frequency distribution of yield and expertise knowledge to divide the field usually in three or four management zones (Blackmore, 2000); and 2) cluster analysis such as K-means and Fuzzy C-means (Taylor et.al., 2003; Taylor, Mcbratney, and Whelan, 2007; Yan et.al., 2007) and/or iterative self-organizing of data analysis technique (Fridgen, Kitchen, and Sudduth, 2000; Kitchen et al., 2002). While the empirical classification methods are simpler, cluster analysis allows for a greater degree of differentiation between and among management zones or yield classes. Empirical methods are mostly used when the target variable (usually yield) is used to create the MZs. When we use attributes that are correlated to the target variable to create the MZs, usually we use clustering methods.

A typical protocol (Fig. 8) do delineate MZs is:

Data preprocessing – Likewise the first phase of the construction of a TM, we need select the coordinate system and remove the outliers and inliers.

Normalization methods - Data clustering techniques and the Fuzzy C-Means algorithm are the most widely used processes for defining MZs. The most common similarity measurement used is Euclidean distance; however, because the algorithm is sensitive to the range of the input variables, these variables are typically normalized dividing the value by the maximum value, by the mean, or sum of the observations. Schenatto et. al. (2017) analyzed the influence of data normalization methods for defining MZs. Tests were conducted in three experimental fields with 9.9, 15.0, and 19.8 ha, located in Southern Brazil. The variables (attributes) used for defining MZs were selected using spatial correlation statistics and data were normalized using methods of standard score, range, and mean. The MZs were defined using the Fuzzy C-Means algorithm, which generated clusters of two, three, and four classes. It was proved that when the MZs definition uses more than one variable in the clustering process which similarity measure is Euclidian distance, the normalization is required. The range method was considered the overall best normalization method.

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

Three variable selection techniques that can be applied in combination with the Fuzzy C-means algorithm are as follows: spatial correlation analysis (Reich, 2008; Schepers et al., 2004), applied as described by Bazzi et al. (2013) and Schenatto et al. (2016); principal component analysis (PCA) (Hotelling, 1933), used by Fraisse et al. (2001), Li et al. (2007), Moral et al. (2010), and Cohen et al. (2013); and multivariate spatial analysis based on Moran's index PCA (MULTISPATI-PCA) (Dray, Saïd, and Débias, 2008), applied by Córdoba et al. (2013, 2016), and Peralta et al.

(2015).

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

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

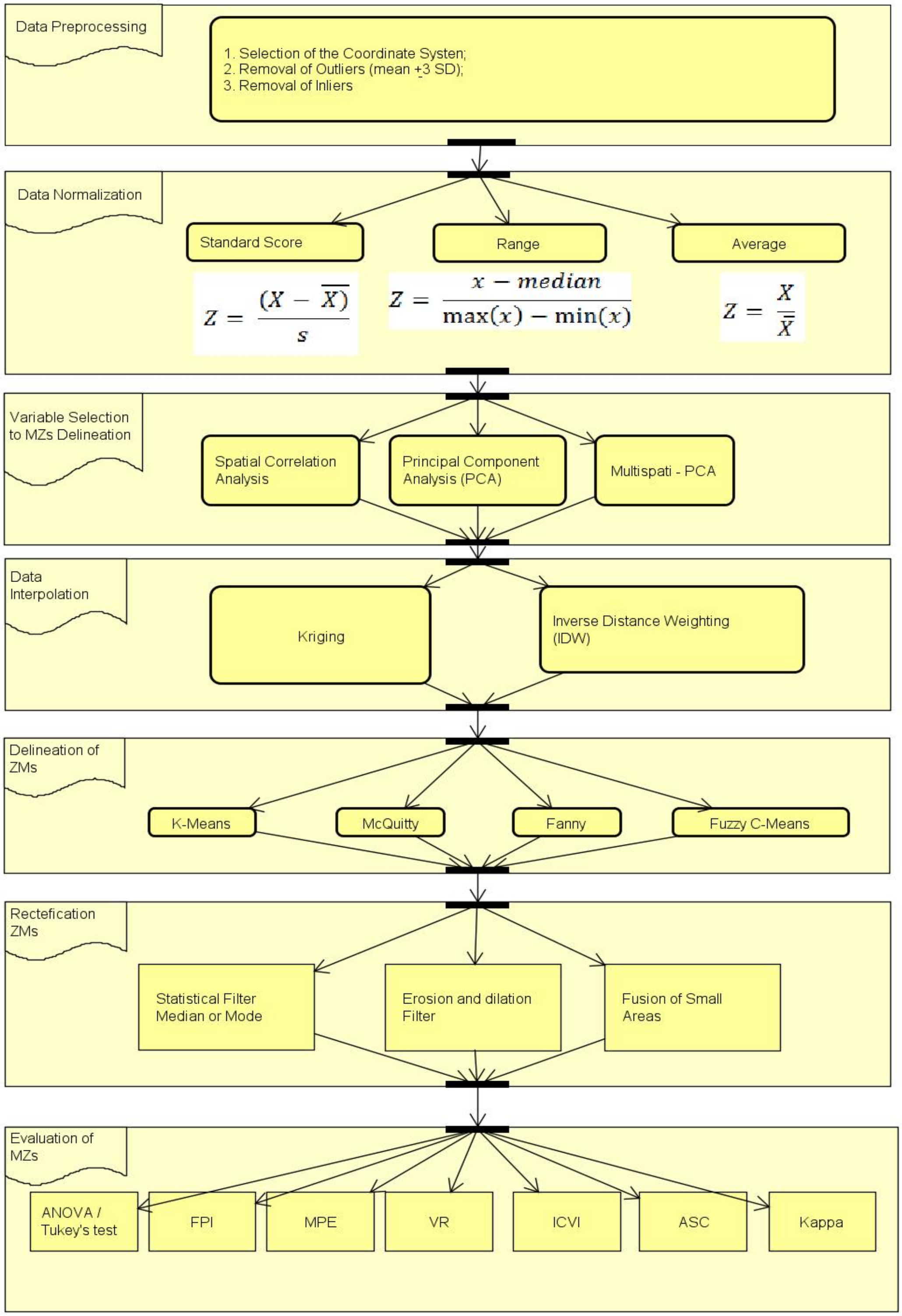

Data Preprocessing
1. Selection of the Coordinate Systen;
2. Removal of Outliers (mean +3 SD);
3. Removal of Inliers
Data Normalization
Standard Score
Range
Average
$Z = \frac{(X - \overline{X})}{s}$
$Z = \frac{x - median}{\max(x) - \min(x)}$
$Z = \frac{X}{\overline{X}}$
Variable Selection to MZs Delineation
Spatial Correlation Analysis
Principal Component Analysis (PCA)
Multispati - PCA
Data Interpolation
Kriging
Inverse Distance Weighting (IDW)
Delineation of ZMs
K-Means
McQuitty
Fanny
Fuzzy C-Means
Rectefication ZMs
Statistical Filter Median or Mode
Erosion and dilation Filter
Fusion of Small Areas
Evaluation of MZs
ANOVA / Tukey's test
FPI
MPE
VR
ICVI
ASC
Kappa

Fig. 8. Flowchart of the typical protocol do delineate management zones.

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

Gavioli et al. (2016) studied the efficiency of each of these three techniques (spatial correlation analysis, PCA MULTISPATI-PCA and a new method proposed by them, named MPCA-SC, based on the combined use of Moran's bivariate spatial autocorrelation statistic and MULTISPATI-PCA. The evaluation was performed by using data collected from 2010 to 2014 from three agricultural areas in Paraná State, Brazil, with corn and soybean crops, generating two, three, and four classes. The delineated MZs were different according to the method used, and MPCA-SC provided the best performance for the Fuzzy C-means algorithm.

Data interpolation – Likewise the first phase of the construction of a TM, we interpolate the data to create MZs that are continuous and smooth. Usually this task is performed with the inverse distance weighting (IDW) or kriging interpolation methods.

Clustering methods - The cluster analysis methods are intended to divide the data points of an agricultural area into classes, which are also termed groups, by employing a similarity evaluation function for this division. In practice, these classes are employed to define the MZs, which can be subsequently delimited in the field (Boydell and McBratney, 2002; Córdoba et al., 2016).

The terms management zone and management class are frequently used in precision agriculture (PA) literature and often as interchangeable terms. However, these terms are not identical. A management class is the area to which a particular treatment may be applied. A management zone is a spatially contiguous area to which a particular treatment may be applied. Thus, a management class may consist of numerous zones whereas a management zone can contain only one management class (Taylor, Mcbratney, and Whelan, 2007).

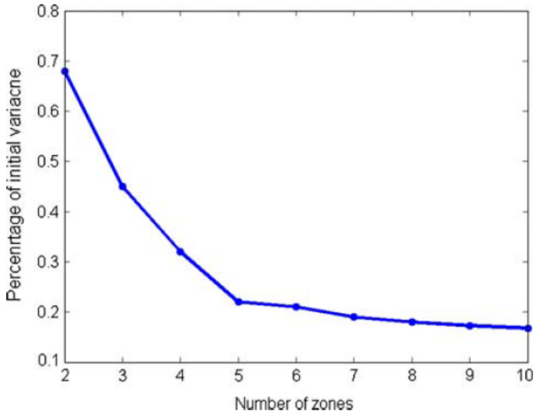
The clustering methods are considered to be more complex than the empirical methods but they enable greater differentiation between classes by less subjective criteria. They employ several variables in the process of MZ definition. Among the many clustering algorithm options described in the literature, two algorithms have been frequently applied in studies related to the generation of MZs: K-means (MacQueen, 1967) and Fuzzy C-means (FCM) (Bezdek, 1981). Examples of specific software for MZ delineation using FCM are Management Zone Analyst (MZA, Fridgen et al., 2004), FuzME (Minasny and McBratney, 2002), Software for Defining Management Zones (SDUM, Bazzi et al, 2013), ZoneMAP (Zhang et al., 2009), and the user-friendly software from Albornoz et al. (2017).

Gavioli et al. (2018) evaluated 20 algorithms for the clustering algorithms. The evaluation was conducted with data obtained between 2010 and 2015 in three commercial agricultural areas cultivated with soybean and corn in the state of Paraná, Brazil. From variables elevation, clay, sand, silt, soil penetration resistance, slope and bulk density, a method based on principal component analysis (PCA) was applied to generate new variables that were employed as inputs for the clustering algorithms. McQuitty's Method and Fanny were considered to be the best algorithm because they produced the largest reductions in the variance of yield in the three areas. These methods generated classes with high internal homogeneity and delimited MZs without fragmentation (suitable for field operations). The classic FCM and K-means generated significantly different subareas in only two areas, in which the obtained results were similar to the results of McQuitty's Method and Fanny.

Rectification - Regardless of the method used to delimit these zones, patches or isolated pixels generally appear. Gonzalez and Woods (2008), Córdoba et al. (2016), Albornoz et al. (2017), and Betzek et al. (2018) used median and dilatation filters and erosion to reduce the fragmentation of MZs.

Evaluation of MZs - Zhang et al. (2009) commented that a method to evaluate classification success is to estimate how much within-zone variability is reduced for a number (n) of zones as compared with n - 1 zones. Generally, the total within-zone variance decreases rapidly initially and then approaches an asymptotic value slowly as the number of zones continues to increase. They proposed a two criteria method to decide the optimal number of zones: (1) overall reduction of variance is >50%; and (2) consecutive reduction of variance is <20%. For the case shown in

Fig. 9, the optimal number of zones would be 5.



Source: Zhang et al. (2009).

Fig. 9. The total within-zone variability as a percentage of initial variance normally decreases with the number of zones.

A more advanced analysis of the performance of the clustering process can be assessed using indices and analysis of variance (ANOVA). The most cited in literature are:

1) Variance Reduction (VR) (Ping and Dobermann, 2003): is calculated for the target variable, with the expectation that the sum of the variances of the data from MZs generated is smaller than the total variance (Equation 4).

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

where *c* is the number of MZs; W_i is the proportion of the area of *i-th* MZ to the total area; V_{mzi} is the data variance of the *i-th* MZ; and V_{field} is the data variance corresponding to the area as a whole.

2) Fuzziness Performance Index (FPI) (Fridgen et al., 2004): it allows determining the degree of separation between the fuzzy *c* groups generated from a data set. FPI varies between 0 and 1, such that the closer this value to 0, the lower is the degree of sharing of elements among the generated groups (Equation 5).

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

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

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

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

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

4) Improved Cluster Validation Index (ICVI) (Gavioli et al., 2016): it was proposed to solve a possible problem when the estimates for FPI, MPE, and VR did not indicate similar methods to the definition of MZs. ICVI (Equation 8) lies between 0 and 1, such that the greater the value of VR and lower the values of the FPI and the MPE, the closer will the ICVI be to 0. In a comparison between *n* clustering methods, the best method is the one with the lowest *ICVI$_i$*.

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

where *FPI$_i$* is the FPI value of the *i-th* variable selection method; *MPE$_i$* is the MPE value of the *i-th* variable selection method; *VR$_i$* is the VR value of the *i-th* variable selection method; and *Max{Index_X}* represents the maximum value of the *Index_X* index among the *n* variable selection methods.

5) Smoothness Index (SI) (Gavioli et al., 2016): it gives the pixel-by-pixel frequency of change of classes in a TM in the horizontal and vertical directions and along the diagonal (Equation 7). It also characterizes the smoothness of the boundary curves of the MZs. If a map has a completely homogeneous area, the result is SI equals to 100% because of lack of changes in class. On the other hand, if the map is completely generated with random values, the SI will have a value close to 0.

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

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

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

7) Average silhouette coefficient (ASC) (Rousseeuw 1987): The ASC coefficient is obtained from the silhouette coefficient (SC) (Rousseeuw 1987), which is an evaluation index that measures the level of satisfactory internal formation and external separation of clusters. The SC value for point *p*, which is denoted by *sc$_p$*, is calculated using the average of the intra-group distances *a$_p$* and the average of the inter-group distances *b$_p$* (Equation 9):

$$sc_p = \frac{b_p - a_p}{Max(a_p, b_p)} \qquad (9)$$

where *a$_p$* is the average of the distances between point *p* and all other points in the same group, and *b$_p$* is the average of the distances between point *p* and all points in the closest group that contains *p*. The group silhouette coefficient (GSC) is obtained by calculating the average of the silhouette coefficients for the points of this group, and the value that corresponds to the ASC coefficient of *k* groups is obtained by calculating the average of the GSC values of the *k* groups.

The ASC values vary from -1 to 1; -1 indicates an incorrect grouping, and 1 indicates groups with the best intra-group formation and the best possible inter-group separation.

8) Kappa coefficient (K) (Cohen, 1960): we want K to compare the agreement of two MZ delineation approach and use the classification proposed by Landis and Koch (1977): $0 < K \leq 0.2$ indicates no agreement, $0.2 < K \leq 0.4$ weak agreement, $0.4 < K \leq 0.6$ moderate agreement, $0.6 < K \leq 0.8$ strong agreement, and $0.8 < K \leq 1$ very strong agreement.

Final Remarks

The use of thematic maps and management zones are essential to analyze the spatial behavior of selected variables but their ability to transmit the correct information depend on the competence and the knowledge of the author of the maps. In this sense, this work aims to help the creation of the referred maps.

Acknowledgements

The authors would like to thank the Coordination for the Improvement of Higher Education Personnel (CAPES), and the National Council for Scientific and Technological Development (CNPq) for funding.

References

Albornoz, E.M., Kemerer, A.C., Galarza, R., Mastaglia, N., Melchiori, R., Martínez, C.E., 2017. Development and evaluation of an automatic software for management zone delineation. Precis. Agric. 1–14. doi:10.1007/s11119-017-9530-9.

Amidan, B., Ferryman, T., & Cooley, S. (2005). Data outlier detection using the Chebyshev theorem. In IEEE Aerosp. Conf (pp. 3814e3819).

Arlot, S., Celisse, A., 2010. A survey of cross-validation procedures for model selection. Statist. Surv. 4, 40–79

Basso, B., Cammarano, D., Chen, D., Cafiero, G., Amato, M., Bitella, G., Rossi, R., Basso, F., 2009. Landscape position and precipitation effects on spatial variability of wheat yield and grain protein in Southern Italy. J. Agron. Crop Sci. 195, 301–312.

Bazzi, C.L., Souza, E.G., Uribe-Opazo, M.A., Nóbrega, L.H.P., Rocha, D.M., 2013. Management zones definition using soil chemical and physical attributes in a soybean area. Engenharia Agrícola 33, 952–964.

Betzek, N., Souza, E.G., Bazzi, C.L., Schenatto, K., Gavioli, A., 2018. Rectification methods for optimization of management zones. Comput. Electron. Agric. 146, 1-11.

Bezdek, J. C. Pattern Recognition with Fuzzy Objective Function Algorithms. New York: Plenum Press, 1981. 256 p.

Bier, V. A.; Souza, E. G., 2017. Interpolation selection index for delineation of thematic maps. Computers and Electronics in Agriculture, 136, 202-209.

Bobryk, C.W., Myers, D.B., Kitchen, N.R., Shanahan, J.F., Sudduth, K.A., Drummond, S.T., Gunzenhauser, B., Gomez Raboteaux, N.N., 2016. Validating a digital soil map with corn yield data for precision agriculture decision support. Agron. J. 108, 957–965

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

Buttafuoco, G., Castrignano, A., Colecchia, A.S., Ricca, N., 2010. Delineation of management zones using soil properties and a multivariate geostatistical approach. Ital. J. Agron. 5, 323–332.

Cohen, J.A., 1960. Coefficient of agreement for nominal scales. Educ. Psychol. Measur. 20, 37–46.

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

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

Córdoba, M., Bruno, C., Costa, J.L., Peralta, N.R., Balzarini, M., 2016. Protocol for multivariate homogeneous zone delineation in precision agriculture. Biosyst. Eng. 143, 95–107.

Cressie, N. 1993. Statistics for spatial data. Revised ed. John Wiley & Sons, New York.

Davidson, D., 2014. Evaluating the quality of your soil. Crop Soils Mag. 4–13.

Demattê, J.A.M., Demattê, J.L.I., Alves, E.R., Barbosa, R.N., Morelli, J.L., 2014. Precision agriculture for sugarcane management: a strategy applied for Brazilian conditions. Acta Sci. Agron. 36, 111.

Doerge, T.A., 2000. Site-Specific Management Guidelines. Potash & Phosphate Institute, Norcross.

Dray, S., Saïd, S., Débias, F., 2008. Spatial ordination of vegetation data using a generalization of Wartenberg's multivariate spatial correlation. J. Veg. Sci. 19, 45–56

Faraco, M. A., Uribe-Opazo, M. A., Da Silva, E. A. A., Johann, J. A. and Borssoi, J. A. 2008. Seleção de modelos de variabilidade espacial para elaboração de mapas temáticos de atributos físicos do solo e produtividade da soja. Revista Brasileira de Ciência do Solo, Viçosa 32:463-476.

Ferguson, R.B., Hergert, G.W., 2009. Soil sampling for precision agriculture. Ext. Precis. Agric. 1–4.

Fraisse, C.W., Sudduth, K.A., Kitchen, N.R., 2001. Delineation of site-specific management zones by unsupervised classification of topographic attributes and soil electrical conductivity. Trans. ASAE 44 (1), 155–166.

Franzen, D.W., Hopkins, D.H., Sweeney, M.D., Ulmer, M.K., Halvorson, A.D., 2002. Evaluation of soil survey scale for zone development of site-specific nitrogen management. Agron. J. 94, 381–389

Fridgen, J. J., Kitchen, N . R., Drummond, K. A. S., Wiebold, S. T., & Fraisse, C. W. (2004). Software Management Zone Analyst (MZA): Software for subfield management zone delineation. Agronomy Journal, 96 (1), 100-108.

Fridgen, J.J., N.R. Kitchen , and K.A. Sudduth, 2000. Variability of soil and landscape attributes within sub-field management zones. In P. C. Roberts, et al. (Eds.), Precision agriculture. Proceedings of the 5th International Conference of the ASA, CSSA, and SSSA, Madison, WI, USA.

Gavioli, A., Souza, E. G.; Bazzi, C. L., Schenatto, K, Betzek, M., 2018. Data clustering methods for definition of management zones. Computers and Electronics in Agriculture (In analysis).

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

Gnanadesikan, R., Kettenring, J., and Tsao, S., 1995. Weighting and selection of variables for cluster analysis. J. Classif. 12, 113–136.

Gonzalez, R.C., Woods, R., 2008. Digital image processing, third ed. Pearson Prentice Hall. New Jersey.

Haghverdi, A., Leib, B.G., Washington-Allen, R.A., Ayers, P.D., Buschermohle, M.J., 2015. Perspectives on delineating management zones for variable rate irrigation. Comput. Electron. Agric. 117, 154–167.

Harrower M, Brewer CA. ColorBrewer.org: an online tool for selecting colour schemes for maps. Cartogr. J. 2003; 40:27-37.

Hotelling, H., 1933. Analysis of a complex of statistical variables into principal components. J. Educ. Psychol. 24, 417–441.

Indiemapper (2016). The basics of data classification. http://indiemapper.com/app/learnmore.php?l=classification.

Isaaks, E.H; Srivastava, R.M. Applied geostatistics. New York: Oxford University Press, 1989.

561p.

Johnson, R.A., Wichern, D.W., 2007. Applied Multivariate Statistical Analysis, sixth ed. Pearson, New Jersey.

Journel, A.G., Huijbregts, C.J., 1978. Mining Geostatistics. London, New York, San Francisco: Academic Press.

Khosla, R., Inman, D., Westfall, D. G., Reich, R. M., Frasier, M., Mzuku, M., ... & Hornung, A. (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(1-2), 85-100.

Khosla, R., K. Fleming, J.A. Delgado, T. Shaver, and D.G. Westfall. 2002. Use of site-specifc management zones to improve nitrogen management for precision agriculture. J. Soil Water Conserv. 57:513–518.

Kitchen, N.R., K.A. Sudduth, S.T. Drummond, J.J. Fridgen, and C.W. Fraisse. 2002. Procedures for evaluating unsupervised classification to derive management zones. p. 330-345. In P.C. Robert et al. (ed.) Proc. 6th Int. Conf. on Precision Agriculture, Minneapolis, MN. [CD-ROM]. 14-17 July 2002. ASA, CSSA, and SSSA, Madi¬son, WI.

Kohavi, R., 1995. A study of cross-validation and bootstrap for accuracy estimation and model selection. In: Mellish, C.S. (Ed.), Proceedings of the 14th International Joint Conference on Artificial Intelligence, pp. 1137–1143.

Li, Y., Shi, Z., Li, F., Li, H.Y., 2007. Delineation of site-specific management zones using fuzzy clustering analysis in a coastal saline land. Comput. Electron. Agric. 56, 174–186.

MacQueen, J. B. (1967). Some methods for classification and analysis of multivariate observations. In Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability (Vol. 1, pp. 281–297). Berkeley: University of California Press.

Mallarino, A.P., Wittry, D.J., 2004. Efficacy of grid and zone soil sampling approaches for site-specific assessment of phosphorus, potassium, pH, and organic matter. Precis. Agric. 5, 131–144.

Mcbratney, A., Webster, R., 1986. Choosing functions for semi-variograms of soil properties and fitting them to sample estimates. J. Soil Sci. 37, 617–639

Minasny, B., Mcbratney, A.B., 2002. FuzME version 3. Australian Centre for Precision Agriculture, The University of Sydney, Sydney.

Monmonier M. (1996). How to Lie with Maps. University of Chicago Press, Chicago, IL.

Moral, F.J., Terrón, J.M., da Silva, J.R.M., 2010. Delineation of management zones using mobile measurements of soil apparent electrical conductivity and multivariate geostatistical techniques. Soil Tillage Res. 106, 335–343.

Moshia, M.E., Khosla, R., Longchamps, L., Reich, R., Davis, J.G., Westfall, D.G., 2014. Precision manure management across site-specific management zones: grain yield and economic analysis. Agron. J. 106, 2146–2156.

Mzuku, M., Khosla, R., Reich, R., Inman, D., Smith, F., MacDonald, L., 2005. Spatial variability of measured soil properties across site-specific management zones. Soil Sci. Soc. Am. J. 69, 1572–1579.

NIST/SEMATECH e-Handbook of Statistical Methods, http://www.itl.nist.gov/div898/handbook/, date 11/21/2016.

Ord, J.K., 1975. Estimation methods for models of spatial interaction. J. Am. Stat. Assoc. 70, 120–126.

Peralta, N.R., Costa, J.L., Balzarini, M., Franco, M.C., Córdoba, M., Bullock, D., 2015. Delineation of management zones to improve nitrogen management of wheat. Comput. Electron. Agric. 110, 103–113.

Ping, J.L., Dobermann, A., 2003. Creating spatially contiguous yield classes for site-specific management. Agron. J. 95, 1121–1131

Reich, R.M., 2008. Spatial Statistical Modeling of Natural Resources. Colorado State University, Fort Collins.

Reza, S.K., Sarkar, D., Daruah, U., Das, T.H., 2010. Evaluation and comparison of ordinary kriging and inverse distance weighting methods for prediction of spatial variability of some chemical parameters of Dhalai district, Tripura. Agropedology 20, 38–48.

Rousseeuw, P. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. Journal of Computational and Applied Mathematics, 20, 53-65.

Schenatto, K., Souza, E.G.; Bazzi, C.L., Gavioli, A.; Beneduzzi, H.M., 2017. Normalization of data for delineating management zones. Computers and Electronics in Agriculture, v. 143, p. 238-248.

Schenatto, K., Souza, E.G., Bazzi, C.L., Bier, V.A., Betzek, N.M., Gavioli, A., 2016. Data interpolation in the definition of management zones. Acta Scientiarum 38, 31–40.

Schepers, A.R., Shanahan, F.J., Liebig, M.A., Schepers, J.S., Johnson, S.H., Luchiari, J.A., 2004. Appropriateness of management zones for characterizing spatial variability of soil properties and irrigated corn yields across years. Agronomy J. 96, 195–203.

Souza, E.G., C.L. Bazzi, R. Khosla, M.A. Uribe-Opazo, and R.M. Reich. 2016. Interpolation type and data computation of crop yield maps is important for precision crop production. Journal of Plant Nutrition. 39:531–538.

Taylor, J.A., Mcbratney A.B., Whelan B.M., 2007. Establishing management classes for broadacre agricultural production. Agron J 99(5), 1366-1376.

Taylor, J.C., G.A. Wood, R. Earl, and R.J. Godwin. 2003. Soil Factors and their Influence on Within-field Crop Variability, Part II: Spatial Analysis and Determination of Management Zones. Biosystems Engineering. 4(84):441-453.

Timlin, D.J., Pachepsky, Y., Snyder, V.A., Bryant, R.B., 1998. Spatial and temporal variability of corn grain yield on a hillslope. Soil Sci. Soc. Am. J. 62, 764.

Wollenhaupt, N.C., Wolkowski, R.P., Clayton, M.K., 1994. Mapping soil test phosphorus and potassium for variable-rate fertilizer application. J. Prod. Agric. 7, 441–448.

Xiang, L.P. Yu-Chun, G. Zhong-Qiang, and Z. Chun-Jiang, 2007. Delineation and Scale Effect of Precision Agriculture management zones using yield monitor data over four years. Agriculture Sciences. 6(2):180-188.

Yan, L., S. Zhou, L. Feng, and L. Hong-yi. 2007. Delineation of site-specific management zones using fuzzy clustering analysis in a coastal saline land. Computers and Electronics in Agriculture. 56:174-186.

Zhang, X., Shi, L.,Jia, X., Seielstad, G., Helgason, C., 2009. Zone mapping application for precision-farming: a decision support tool for variable rate application. Precision Agric. 11, 103–114.