Computers and Electronics in Agriculture 143 (2017) 238–248

Contents lists available at ScienceDirect

Computers and Electronics in Agriculture

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

Normalization of data for delineating management zones

Kelyn Schenatto[a,*], Eduardo Godoy de Souza[b], Claudio Leones Bazzi[c], Alan Gavioli[c], Nelson Miguel Betzek[c], Humberto Martins Beneduzzi[d]

[a] Computer Science Department, Technological Federal University of Paraná, Rua Cerejeira s/n, 85892-000 Santa Helena, Paraná State, Brazil
[b] Technological and Exact Sciences Center, West Paraná State University, Rua Universitária 2069, 85819-110 Cascavel, Paraná State, Brazil
[c] Computer Science Department, Technological Federal University of Paraná, Av. Brasil 4232, 85884-000 Medianeira, Paraná State, Brazil
[d] Analysis and Systems Development Department, Paraná Federal Institute, Av. Araucária 780, 85860-000 Foz do Iguaçu, Paraná, Brazil

ARTICLE INFO

Keywords:
Standardization
Fuzzy C-Means
Precision agriculture
Smoothness index
Variance reduction
Euclidean distance

ABSTRACT

Management zones (MZs) are a viable economic alternative to variable-rate application (VRA) based on prescription maps; however, unlike the latter, MZs can employ conventional machinery. The use of management zones (MZs) is considered an economically viable alternative because of its low initial cost and high return in economic and environmental benefits. Data clustering techniques and the Fuzzy C-Means algorithm are the most widely used processes for delineating 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 standard deviation, maximum value, average, or data set range. The objective of this study was to assess the influence of data normalization methods for delineating MZs. The experiment was conducted in three experimental fields with 9.9, 15.0, and 19.8 ha, located in Southern Brazil between 2010 and 2014. The variables used for delineating MZs were selected using spatial correlation statistics and data were normalized using methods of standard score, range, and average. The MZs were delineated using the Fuzzy C-Means algorithm, which created two, three, and four clusters. The normalization methods were evaluated by five indices (modified partition entropy [MPE], fuzziness performance index [FPI], variance reduction [VR], smoothness index [SI], and kappa), and ANOVA. It was found that when the MZs delineation uses more than one variable with different scales in the clustering process using Euclidean distance, normalization is required. The range method was considered the overall best normalization method.

1. Introduction

The study of the spatial distribution of soil and plant variables is important to the establishment of appropriate management zones (MZs) to be used in application of the fertilizer, soil management, and irrigation. Appropriate MZs may maximize yield, while reducing costs and minimizing potential environmental damage (Tilman et al., 2011; Li et al., 2013; Bansod and Pandey, 2013; Hedley, 2015).

A MZ is defined as a subregion of a field that exhibits similar combinations of yield-limiting factors (Tagarakis et al., 2013). This facilitates the application of precision agriculture (PA) techniques by reducing the costs of its adoption and implementation, since MZs can use constant rate equipment and may reduce the number of samples needed to characterize the soil nutrients availability. Delineating MZs is not a simple task because numerous variables may influence crop yield. Considering that a MZ is often used for several years, the considered variables should be temporally stable (Doerge, 2000) and correlated to the yield. Among the variables identified in the literature good potential to delineate temporally stable MZs are elevation (Bazzi et al., 2015; Fraisse et al., 2001; Jaynes et al., 2005; Peralta and Costa, 2013; Farid et al., 2016; Schepers et al., 2004), soil electrical conductivity (ECa) (Li et al., 2007;Farid et al., 2016), soil penetration resistance (Gavioli et al., 2016), and soil texture (Farid et al., 2016).

Techniques such as principal component analysis (PCA) (Bansod and Pandey, 2013) and the Moran's bivariate spatial autocorrelation statistic proposed by Czaplewski and Reich (1993), and used by Reich et al. (1994) and Bonham et al. (1995) can be used to create (when PCA is used) or select layers for delineation MZs. When there is more than one crop cultivated in the same field during the year, which is a common practice in Brazil, normalizing yield data makes possible to create a more representative variable (Bunselmeyer and Lauer, 2015) to be used in ANOVA and Tukey's test.

Several techniques to delineate MZs are proposed in the literature (Pedroso et al., 2010; Xiang et al., 2007), however the most used is

* Corresponding author at: Computer Science Department, Technological Federal University of Paraná, Rua Cerejeira, s/n – Bairro São Luiz, 85892-000 Santa Helena, Paraná State, Brazil.
E-mail addresses: kschenatto@utfpr.du.br (K. Schenatto), humberto.beneduzzi@ifpr.edu.br (H.M. Beneduzzi).

http://dx.doi.org/10.1016/j.compag.2017.10.017
Received 28 February 2017; Received in revised form 18 October 2017; Accepted 21 October 2017
Available online 04 November 2017

cluster analysis (Li et al., 2007; Iliadis et al., 2010). The most commonly used clustering methods to delineate MZs are the K-means algorithm (Rodrigues Jr. et al., 2011; Ortega and Santibañez, 2007) and fuzzy C-means (Li et al., 2007,2013; Fu et al., 2010; Zhang et al., 2013; Moral et al., 2010). This algorithm, that incorporates the theory of fuzzy logic in the division algorithm, uses a weighting exponent to control the degree of sharing between classes (Bezdek, 1981), allowing individuals to exhibit partial adhesion in each of the classes, which is important when dealing with the continuous variability of natural phenomena (Burrough, 1989). Before a dataset can be formed, it is necessary to establish an appropriate measure of similarity. Euclidean distance is most regularly used; this measure gives equal weight to all measured variables and is sensitive to correlated variables (Bezdek, 1981). In geometrical terms, the Euclidean distance creates agglomerates having a spherical shape, which rarely occur in a soil (Odeh et al., 1992). Fridgen et al. (2004) reports that Euclidean distance should be used only for statistically independent variables demonstrating equal variances. In this sense, when the Euclidean distance is used to clustering, the normalization data can be very important step before creating MZs.

The normalization methods such as Standard score or Z-score method (Eq. (1)) has been used by many researchers for delineation of MZs (Anderberg, 1973; Romesburg, 1984; Larscheid and Blackmore 1996; Stafford et al., 1996; Molin, 2002; Kitchen et al., 2005). This method is used for transforming normal variables to standard score where the transformed variable will have a mean of 0.0 and a variance of 1.00.

$$Z = \frac{(X-\overline{X})}{s} \tag{1}$$

where X is the original data value; $\overline{X}$ is the sample average; and s is the standard deviation.

Several researches reported the use of the average method (Eq. (2)) for delineation MZs (Stafford et al., 1996; Molin, 2002; Kitchen et al., 2005) with the assumption that the average represents the dataset well; however, the average is sensitive, can be modified by adding any constant, and can easily change the distribution of the normalized data (Anderberg, 1973).

$$Z = \frac{X}{\overline{X}} \tag{2}$$

Good results were also reported by Milligan and Cooper (1988), Bazzi et al. (2013), Gavioli et al. (2016), and Schenatto et al. (2016) using the Range (Eq. (3)) normalization method. This method is bounded by 0.0 and 1.0 with at least one observed value at each of these end points. The Min(X) value used in Eq. (3) can be changed for Median(X) (Mielke and Berry, 2007) and have the same behavior because Min(X) and Median(X) are constants and not change the data distribution.

$$Z = \frac{X-Min(X)}{Max(X)-Min(X)} \tag{3}$$

The goal of this study was to evaluate the performance of these methods, frequently used in the data clustering process by the Fuzzy C-Means algorithm to delineate MZs.

2. Materials and methods

A step-by-step flowchart (Fig. 1) was created to show the methodology used.

2.1. Datasets

This research was conducted in three fields (Fig. 2) located in Paraná State, Brazil: Field A (15 ha), located in the municipality of Céu Azul (central geographical location of 25°06′32″S, 53°49′55″W, and average elevation of 460 m). Field B (9.9 ha) located in the municipality of Serranópolis do Iguaçu (central geographic location of 25°24′28″S, 54°00′17″W, and average elevation of 355 m) and Field C (19.8 ha) located in the municipality of Cascavel (central geographic location of 24°57′08″S, 53°33′59″W, and average elevation of 650 m).

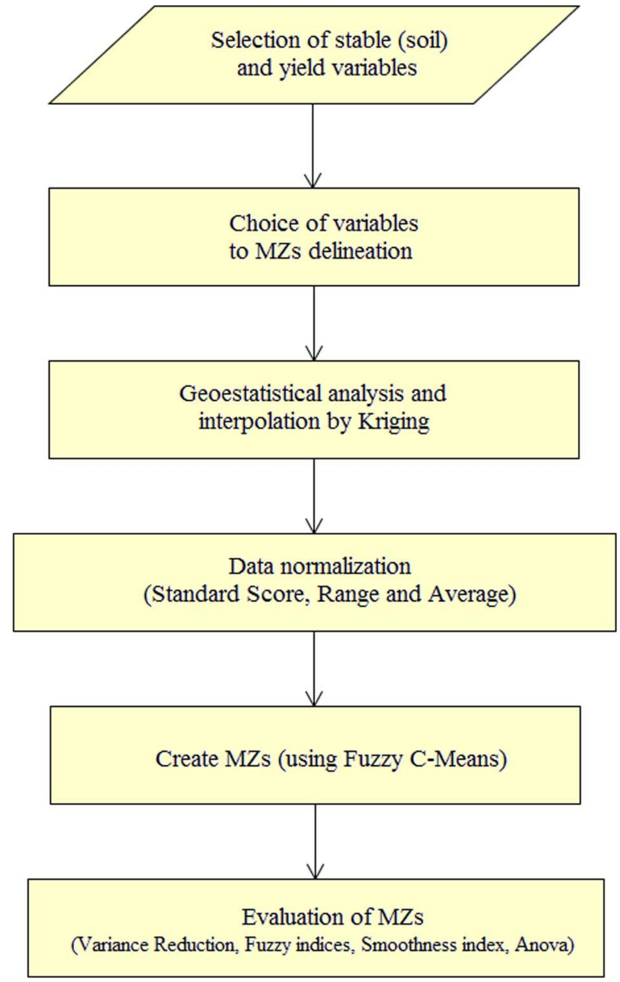

Fig. 1. Step-by-step flowchart of the methodology used to evaluate the normalization methods for delineation of MZ's.

For the delineation of MZs, only variables considered temporally stable collected between 2010 and 2014 (Table 1) were used, to meet the recommendation of Doerge (2000). To meet the constraints of geostatistical analysis (Journel and Huijbregts, 1978) in terms of the minimum number of pairs (30) to calculate the semivariances of the semivariogram, a dense sampling grid (Table 1) was used, with 2.7 points ha^{-1} for Field A, 4.2 points ha^{-1} for Field B, and 3.4 points ha^{-1} for Field C. The irregular sampling grids were defined taking into account an imaginary central line between the elevation contour lines of each field (Fig. 2).

Elevation was determined with a total station (Topcon GPT-7505, Topcon Corporation, Tokyo, Japan), and soil penetration resistance (SPR) was determined with a soil penetrometer (penetroLOG PGL1020, Falker Automação Agrícola, Porto Alegre, Brazil). Soil samples were collected at a depth of 0–0.2 m and sent to the laboratory for analysis. Soybean yield for Field A was determined with a yield monitor (AFS