Geoderma 310 (2018) 170–177

Contents lists available at ScienceDirect

Geoderma

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

Enhancing edaphoclimatic zoning by adding multivariate spatial statistics to regional data

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

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

ARTICLE INFO

Handling Editor: A.B. McBratney

Keywords:
Multivariate zoning
Spatial principal components
Fuzzy k-means

ABSTRACT

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

1. Introduction

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

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

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

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

http://dx.doi.org/10.1016/j.geoderma.2017.09.011
Received 10 February 2017; Received in revised form 9 August 2017; Accepted 10 September 2017
0016-7061/ © 2017 Elsevier B.V. All rights reserved.

variability, as classical non-spatial PCA does.

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

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

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

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

2. Material and methods

2.1. Study area

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

2.2. Database

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

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

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

2.3. Data preprocessing

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

2.4. Protocol sequence

2.4.1. Step 1: deriving spatial synthetic variables

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

Download English Version:

https://daneshyari.com/en/article/5770371

Download Persian Version:

https://daneshyari.com/article/5770371

Daneshyari.com