Generalized linear models and multivariate analysis applied to predict soil spatial distribution in south Brazil.

Alexandre ten Caten[A], **Ricardo Simão Diniz Dalmolin[B]**, Fabrício de Araújo Pedron[C]

[A]Instituto Federal Farroupilha Campus Júlio de Castilhos, Júlio de Castilhos, RS, Brasil, Email acaten@yahoo.com.br
[B]Universidade Federal de Santa Maria – Departamento de Solos, Santa Maria, RS, Brasil, Email dalmolinrsd@gmail.com
[C]Universidade Federal de Santa Maria – Departamento de Solos, Santa Maria, RS, Brasil, Email fapedron@ymail.com

Abstract

Digital Soil Mapping (DSM) is an interdisciplinary science involving soil, statistics, mathematics, and geomatics knowledge applied to generate soil spatial information. This study aimed to use Principal Component (PC) as covariates in logistic models for the prediction of soil classes in the south of Brazil. Principal Component Analysis (PCA) was applied to nine terrain attributes: elevation, slope, distance to the nearest stream; planar curvature, profile curvature, radiation index, natural logarithm of contributing area, topographic wetness index and sediment transport capacity. The retained PC was used as explanatory covariates in Multiple Logistic Regressions (MLR), which were trained with soil information provided by an available soil map on 1:50.000 scale. The three retained components explained 65.57% of the variability in the original data. The logistic models reproduce the original map in 58.20% (kappa index), and the predictive ability of the models was 48.53%. Soil units with the smallest areas were not properly spatialized, and logistic models were not able to distinguish the soil classes too close on the landscape. MLR need further investigation, since they have a huge applying potential in the immense not mapped areas of the Brazilian territory.

Key Words

Soil map, polytomic, Shuttle Radar Topography Mission.

Introduction

Digital Soil Map (DSM) aims the creation an population of spatial soil information systems by numerical models inferring the spatial and temporal variations of soil types and soil properties from soil observation and knowledge and from related environmental variables (Lagacherie and McBratney 2007). The availability of technologies which generate data related to processes and factors of soil formation, allied to the use of statistical and mathematical techniques, makes possible the use of DSM to cater the rising demand for soil spatial information.

The Principal Component Analysis (PCA) is a multivariate method that allows the change of a set of initial variables, correlated among them, into another set of non-correlated variables, the so called, Principal Components (PC) (Johnson and Wichern 1992). It is a mathematical procedure, not statistical. Does not require the assumption of normality distribution and leads to no statistical test of significance (Webster 2001). The PCA, being based on a linear model, has a positive applicability on studies that relate soil and environmental predictors, once that, rarely exists a non-linear relation between them (Gaussian) (Odeh *et al.* 1991).

In cases where the result of an inference can be given considering many categories or polytomic (soil classes), an alternative is to work with the probability of occurrence in each one of the categories. To do so, it is applied the Multiple Logistic Regression (MLR), which is a flexible technique, because, it does not present any requirement for its application concerning the explicative variable distribution. There is no need for a normal distribution, linear correlation and measures using the same scale or homogeneity of variance. The explicative variables can be a mixture of binary data with discrete and continuous data (Chatterjee and Hadi 2006).

This study aimed at generating a soil map from a training area using predictive principal components as covariates in multiple logistic regressions.

Methods

Caracterization of the study area

The study area is in São Pedro do Sul, located in the central region of Rio Grande do Sul – Brazil. This area has a surface of 873 km^{2,} ,being comprehended between the coordinates 29º46' to 29º26' south latitude and 54º30' to 53º56' west longitude. It comprehends a transitory region between the physiographic regions of

Medium Plateau and Central Depression in Rio Grande do Sul State – Brazil. This area was chosen because of a available semi-detailed soil map in 1:50000 scale (Klamt *et al.* 2001).

Extraction of the principal components

The terrain attributes ELEV (Elevation), SLOP (Slope), DIST (Distance to Nearest Stream), PLNC (Plan Curvature), PRFC (Profile Curvature), RADI (Radiation Index), LNCA (Natural Logarithm of Contributing Area), TWI (Topographic Wetness Index) e SPI (Stream Power Index), were generated according to Wilson & Gallant (2000) from a DEM / SRTM. A set of 70000 points were randomly created to sample the Information Plans (IP) [layers] of the terrain attributes, these charted information in the form of a text (ASCII) were processed to PCA. It was verified the adequacy of the samples through the individual test Measure of Sample Adequacy (MAS) and general Kaiser Meyer Olkin (KMO) aiming at verifying the correlation degree among the variables and the PCA justification. The eigenvalue numbers retained were conditioned to the ones that had the minimum value equal one. The rotated eigenvector (VARIMAX), resulting from PCA, was used to calculate the new variables, which are not correlated.

Predicted Map from the existing soil map

The MLR were generated using the PC as explicative variables and the soil classes in the existing soil map, to the level of order (1st Level of the Brazilian System of soil classification), as predicted variables. For adjusting the MLR models was just considered the significant parameters to the level of 5% (Wald test). Each logit function generated a probability map about the existence of a certain soil class in the landscape. These values were placed together in only one IP, with higher value among the plans defining the predicted class in that pixel. The quality of the generated maps was evaluated concerning its capacity to reproduce the original map. The capacity to predict the soil classes in an area where not data were used to generate the models was evaluated as well.

Results

Principal Components

For the PC analysis application in the correlation matrix of the attributes, it was first verified the data adequacy by the individual test MAS and general KMO. MAS values under 0.5 indicate that the variable is not proper for the PCA application, as it was mentioned in the literature. Among the terrain attributes PRFC and LNCA obtained values respectively of 0.58 and 0.56, which can be considered a low value for the application of these variables in PCA. However, being the number of attributes only nine; it was chosen to keep all the variables. The KMO value of the quality set was only 0.66, being considered a low value for the PCA application, but, the application is still possible in this case.

After the PCA application to the nine terrain attributes it was generated nine PC, each one concentrating a decreasing percentage of variability of the initial data (Figure 1). The three first PC have an eigenvalue higher than one, and they were kept in this study. Keeping just the three first components mean the loss of one-third of the data variability piled up in the new variables of the fourth and ninth component. Although there is a significant loss on the data variability pattern, there is a gain with the simplification in the number of variables.

Predicted Soil Map

The generated soil map did not displayed the spatially among the Leptosols, Plinthosols and Nitisols classes. However, the Cambisols, Lixisols and the Planosols were visually displayed in a similar way to the one already shown in the existing map and in the relation soil-landscape of the study area. The Planosols were displayed in the lower parts of the landscape and the Cambisols in the declivities and hill tops, at the same time, the Lixisols were distributed in the small hills and the Cambisols were attributed to the regions where the Leptosols were found. The reason for not displaying the Leptosols, Plinthosols and Nitisols classes was their small representativeness in the total of the used samples in the logistic models. These classes correspond only to 5%, 3% and 1% of the total in the 70.000 samples selected at random for the generation of the models.

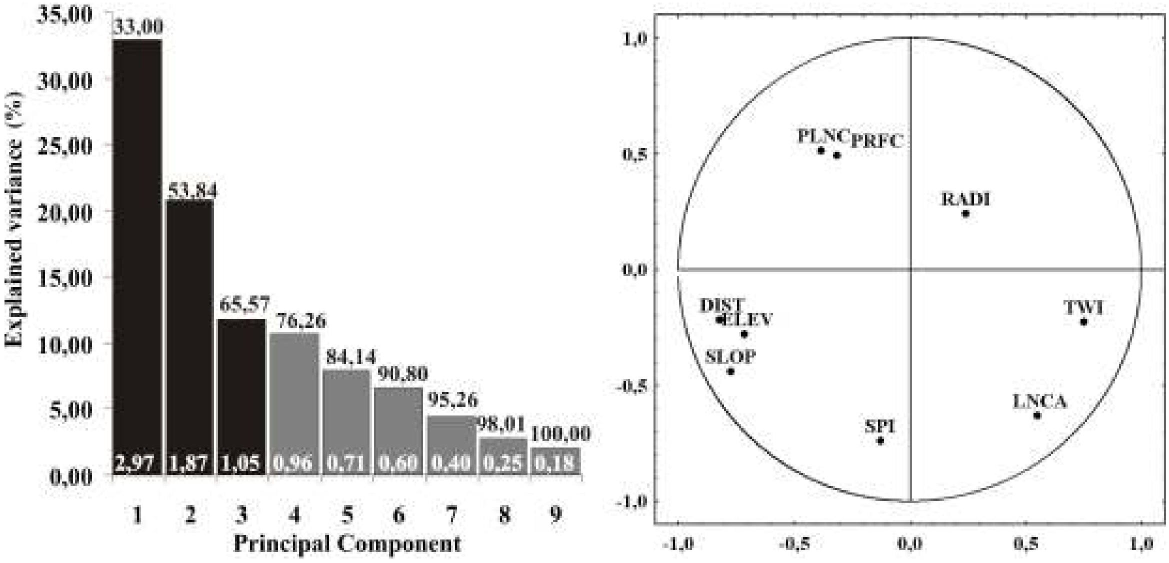

Figure1. (Left) Variability explained by each one of the nine principal components (PC) generated. The eigenvalue of each component are in blank space on the base and in the interior of each bar. The three first bars in bold have eigenvalue higher than the unit. The value over the bars indicates the accumulated percentage variability. (Right) Circular Dispersion Diagram indicates the correlations between the first (1) and second (2) PC variables. ELEV (Elevation), SLOP (Slope), DIST (Distance to Nearest Stream), PLNC (Plan Curvature), PRFC (Profile Curvature), RADI (Radiation index), LNCA (Natural Logarithm of Contributing Area), TWI (Topographic Wetness Index) e SPI (Stream Power Index).

The most common mistakes when mapping were the ones among the classes spatially close related regarding the map outlining. The soils of the wetland, near to river border were mistaken for Lixisols, these ones were mistakenly displayed in the positions of the Cambisols which were mistaken for Leptosols. These errors can have their origin in the borders of each soil class. The inference of the true class, from the analysis of the existing map, can be really difficult for the models due to problems in the outlining of the coropletic map that served as training. The slight difference among the terrain attributes, which may not present any type of gradient in the borders of the soil class polygons, was another difficulty.

Boruvka e Penizek (2007) used neural networks for the soil class spatialization and verified that classes, which are very similar considering the constitution processes, tend to be mistaken by the models. The classes need to be well defined and distinct, to an efficient spatialization to be possible. According to the authors, the use of any methodology should consider the categorical level to be predicted due to the local heterogeneity. The available data to generate the models (number of profiles or area) should be considered as well.

The general accuracy (GA) of the predicted map was 74.3%. This value can be considered positive even that does not consider the correct mapped points by chance. To change this, it is considered the kappa index (K) 58.20% as a more realistic measure for the predicted map quality.

To verify the predictive capacity of the models, it was carried out an accuracy test in an area where no data was used to generate the model. This procedure validates the model, or, in another way, tests its real inference capacity or predictive capacity. The GA was higher than the general accuracy of the reference area, where data were used to train the models, reaching the value of 79.4%. However, a more realistic measure about the map quality in this region, shows that the map accuracy was smaller, with the K index of 48.53%, ten points less than the area where data were used to train the models.

The results of this study show the PCA potential to reduce the number of variables to be applied on the models. It also allows the visualization of the correlations among the original variables and produces new non-correlated ones. Although the PCA application implies loss of original data variability, using the three first PC allowed the logistic models to reproduce the soil classes with an accuracy of 60%, in relation to the original map. This accuracy is compatible with literature data. The use of PCA will probably be increased, with a higher number of predictive variables and with better values of MAS and KMO.

The MLR can be more effective for the soil class spatialization if these classes have a higher relative representativeness among them. Sustaining the Hengl *et al*. (2007) data, this study showed that the soils classes with smaller relative area are not spatially adequate by the logistic models.

Conclusion

The use of principal components in multiple logistic regressions implies on simplified models, comparatively, in the use of all original covariates. However, when explaining the new variables, this simplification can be associated to power reduction because of a smaller retained variance, and also, the new variables cannot have a physical, chemical or biological meaning to constitute the soil.

The logistic models presented lower quality when used outside the training area. This way, the predictive capacity of the models is associated to the model generation in similar areas; areas where the models will be applied considering the processes and factors of soil formation.

Soil classes spatially close to each other in the landscape are mistaken by predictive models.

Soil classes with smaller relative proportion data, used for model training, tend to be inappropriately predicted.

References

Boruvka L, Penizek V (2007) A test of an artificial neural network allocation procedure using the Czech Soil Survey of Agricultural Land data. In 'Digital soil mapping: an introductory perspective'. (Eds P Lagacherie, A Mcbratney, M Voltz) pp. 415-424. (Elsevier)

Chatterjee S, Hadi AS (2006) 'Regression analysis by example'. (John Willey & Sons)

Hengl T (2007) Methods to interpolate soil categorical variables from profile observations: Lessons from Iran. *Geoderma*, **140,** 417-427.

Johnson RA, Wichern DW (1992) 'Applied multivariate statistical analysis. (New Jarsey: Prentice-Hall)

Klamt E, Flores CA, Cabral DR (2001) Solos do Município de São Pedro do Sul. Departamento. de Solos/CCR/UFSM. Santa Maria, 96 p.

Lagacherie P, McBratney A (2007) Spatial soil information systems and spatial soil inference systems: perspectives for digital soil mapping. In 'Digital soil mapping: an introductory perspective.' (Eds. P Lagacherie, A Mcbratney, M Voltz). pp. 3-22. (Elsevier).

Odeh IOA, Chittleborough DJ, McBratney A (1991) Elucidation of soil-landform interrelationships by canonical ordination analysis. *Geoderma* **49**, 1-32.

Webster R (2001) Statistics to support soil research and their presentation. *European Journal of Soil Science*, **52**, 331-340.

Wilson JP, Gallant JC (2000) Digital terrain analysis. In 'Terrain analysis: principles and applications.' (Eds JP Wilson, JC Gallant) pp. 1-27. (Wiley & Sons)