MAPEAMENTO PEDOLÓGICO DIGITAL COM BASE NO RELEVO E TREINAMENTO POR AMOSTRAGEM DE SOLOS DESENVOLVIDOS DE ARENITOS

LAURA MILANI DA SILVA DIAS[1]
RICARDO MARQUES COELHO[2]
STANLEY ROBSON DE MEDEIROS OLIVEIRA[3]
FLÁVIO MARGARITO MARTINS DE BARROS[4]
1 Centro de Solos/Instituto Agronômico
laurads5@yahoo.com.br
2 Centro de Solos/Instituto Agronômico
rmcoelho@iac.sp.gov.br
3 Embrapa Informática Agropecuária
stanley.oliveira@embrapa.br
4 Feagri/Unicamp
flaviomargarito@gmail.com

Resumo

O entendimento de associações solo-relevo pode contribuir para o mapeamento digital de solos. Por ser estratégia de mapeamento em desenvolvimento, algoritmos de mineração de dados, base de dados para treinamento dos modelos e escalas de trabalho ainda necessitam ser avaliados. Para treinar modelos de classificação a partir de observações pontuais em campo, foram produzidos mapas pedológicos digitais em escala grande de bacia hidrográfica, em Botucatu (SP), em que predominam arenitos. Do modelo digital de elevação da bacia foram geradas sete variáveis morfométricas. A classificação dos solos para treinamento e validação dos modelos foi realizada em campo até o nível de subgrupo mais grupamento textural. Foram testados três algoritmos de mineração de dados. A pertinência de grupos de atributos de relevo às classes taxonômicas foi verificada por análise de agrupamento. Apesar do melhor desempenho do algoritmo MLP (redes neurais), este foi considerado pouco confiável, já que não classificou nenhum exemplo da classe GXbdt, com apenas dois exemplos para treinamento. Os classificadores J48 e Random Forest apresentaram acurácia equivalente na classificação dos solos a partir de dados de relevo, com índice kappa ligeiramente superior para o J48 (0,42). A combinação da extensão da área de estudo com o grau de detalhe das variáveis geomorfométricas produziu uma variabilidade de atributos preditivos difícil de representar no conjunto de treinamento criado por amostragem em campo. A presença de classes de solo representativas e distintas pela textura no mesmo grupo de atributos de relevo criado pela análise de agrupamento indicou que relevo não é fator preponderante na diferenciação textural dos solos, principal atributo diferencial dos solos da área de estudo.

Palavras-chave: Mineração de dados de solo. Classificação supervisionada. Predição de classes de solo

Abstract

Digital soil mapping can benefit from soil-relief relationships background knowledge. Due to being a developing approach, data mining algorithms, type of data base for algorithm training, and working scale

are elements of the strategy still to be evaluated. Aiming the production of large-scale digital soil class maps and to train classification models on soil classes identified *in situ* by field and laboratory analysis we studied a watershed located in Botucatu, state of São Paulo, Brasil. Soils of the watershed were mainly developed from sandstones. Seven terrain attributes were generated from a digital elevation model (DEM). Soil classification for model training and validation were generated by field soil sampling and by laboratory analysis to the subgroup level plus texture group of the Brazilian System of Soil Classification. Three data mining algorithms were evaluated. Soil taxonomic classes were compared to terrain attributes clusters. Despite MLP better performance, this neural network algorithm was considered less reliable due to omitting one soil class with few training instances from the prediction set. J48 and Random Forest classifiers showed equivalent accuracy, but with slightly greater kappa index (0,42) for J48. The combination of a large study area with the great detail of terrain attributes generated from the 30-m resolution DEM produced large variability of terrain attributes, difficult to represent in the training set only by field sampling. Occurrence of representative soil classes distinguished fundamentally by particle-size groups in the same terrain attributes cluster suggests relief is not a prevalent factor to soil particle-size class differentiation, the main soil *diferentia* of the studied area.

1. Introdução

Algoritmos que relacionam a ocorrência dos solos na paisagem com seus fatores de formação e que utilizam como apoio Sistemas de Informação Geográfica (SIG's) e técnicas de sensoriamento remoto fazem parte da estratégia do Mapeamento Digital de Solos (MDS). Dentre as covariáveis que vêm sendo aplicadas no MDS, destacam-se os atributos do relevo derivados de modelos digitais de elevação (MDE) (Arruda et al., 2013). A extrapolação das associações solo-paisagem de uma área mapeada, a área de referência, para áreas onde estas associações ainda não são conhecidas é frequentemente feita por treinamento em mapas. O treinamento por observações pontuais do solo em campo é pouco utilizado, apesar de prescindir da existência de áreas de referência e pontualmente corresponder mais fielmente à verdade de campo.

O MDS aplicado à predição de classes de solos no Brasil tem sido, em sua maioria, em escalas pequenas (1:50.000 ou menor). Justificativas para isto estão na pequena disponibilidade de mapas de solos para serem usados como áreas de referência, bem como à falta de cartas topográficas em menor escala (ten Caten et al., 2012). Em níveis generalizados, a legenda admite unidades de mapeamento na forma de associações e frequentemente há simplificação de legenda.

Os objetivos do trabalho foram (i) produzir um mapa pedológico digital para a bacia do córrego Águas da Lúcia no município de Botucatu-SP em escala grande e (ii) treinar modelos de classificação a partir de observações pontuais em campo.

2. Metodologia de Trabalho

O trabalho foi desenvolvido na microbacia do córrego Águas da Lúcia, no município de Botucatu, SP. A bacia tem 1894 ha de extensão e localiza-se entre as coordenadas UTM 771.253 e 776.535 m E e 7.473.291 e 7.479.874 m N, zona 22 S E. Na classificação de Koeppen, o clima da região é do tipo Cwa. A litologia é de arenitos finos a médios com matriz síltico-argilosa e estratificação cruzada de médio a grande porte da formação Pirambóia e arenitos eólicos da formação Botucatu, ambas do grupo São Bento. A partir de curvas de nível de 5 metros foi elaborado um MDE com resolução de 30 m no programa ArcGIS 10 e derivadas as seguintes variáveis geomorfométricas: altitude, declividade, orientação das vertentes, curvatura em planta, curvatura em perfil, distância diagonal e índice topográfico de umidade (Wilson & Gallant, 2000) em formato raster, também com resolução de 30 m. Nesta resolução espacial, a matriz de variáveis geomorfométricas totalizou 20.772 linhas.

Os locais de observação do solo em campo tanto para treinamento dos modelos quanto para sua validação foram determinados em amostragem por hipercubo latino, usando as variáveis do relevo para aleatorizar os locais de amostragem. Foram identificados 75 locais de amostragem para treinamento e 25 para validação. Os solos foram caracterizados no campo em mini-trincheiras (0,7x0,7x0,7m) e com trado até 2 m e coletadas amostras para análise em laboratório. Dos 75 pontos usados para treinamento, em 12 foram abertas trincheiras até 2 m para caracterização completa. Os solos da bacia foram classificados de acordo com o Sistema Brasileiro de Classificação de Solos até o 4º nível categórico mais grupamento textural (Embrapa, 2006).

Para diminuir a fragmentação do mapa digital produzido e a quantidade de classes com poucos exemplos, foram usados apenas 62 pontos de observação de solo para o treinamento dos algoritmos, excluindo-se do treinamento 13 pontos considerados inclusões de solo. Como a proporção de exemplos classificados era pequena (0,3%) para o total da base de dados de relevo, expandiu-se o conjunto de treino,

considerando-se a combinação de variáveis preditivas (relevo). Assim, para cada exemplo Y não classificado, com a mesma combinação de variáveis de um exemplo X com solo classificado, identificou-se Y com a mesma classe de solo de X. Assim, o conjunto de treino de 62 exemplos foi expandido para 2.454 (12 % da base de dados).

Foram testados três diferentes classificadores: J48, uma implementação de árvores de decisão baseada na abordagem de aprendizado de máquina supervisionado (Hall et al., 2009); redes neurais do tipo *Multi-layer Perception* (MLP) calibradas através de retropropagação (Kanellopoulus & Wilkinson, 1970) e o *Random Forest* (Breiman, 2001), que combina classificações feitas por diversas árvores de decisão. O treinamento dos modelos foi realizado no programa *Waikato Environment for Knowledge Analysis* (Weka) (Witten & Frank, 2005). A exatidão dos mapas preditos foi avaliada por meio da "acurácia global"; da "exatidão do produtor"; da "exatidão do usuário" e do coeficiente Kappa (Congalton, 1991). Todas as determinações de exatidão foram calculadas a partir de matrizes de erro, onde os 25 pontos classificados em campo foram considerados verdade e confrontados com a classe predita pelos algoritmos. A base de dados foi também analisada pelo algoritmo K-Means, que realiza a tarefa de agrupamento e visa identificar e aproximar os registros similares, para k=6 e k=10 (Camilo et al., 2009).

3. Resultados

Tabela 1: Acurácia global e índice kappa para os algoritmos testados

Algoritmo	Acurácia (%)	Kappa
J48	56	0.42
Random Forest	56	0.39
MLP	60	0.46

Tabela 2: Resultados da análise de agrupamento pelo algoritmo K-means para K= 6.

	Agrupamento					
Classe de solo	**1**	**2**	**3**	**4**	**5**	**6**
LVAdt	0	0	47	0	50	0
LVdt	188	99	123	0	39	0
PVdabrup	0	0	0	23	50	0
RQot	406	19	163	11	485	680
Outros	0	0	0	67	4	0

Tabela 3: Resultados da análise de agrupamento pelo algoritmo K-means para K= 10.

	Agrupamento									
Classe de solo	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
LVAdt	0	47	0	0	50	0	0	0	0	0
LVdt	0	0	123	0	131	0	0	0	0	195
PVdabrup	15	0	0	0	0	0	50	8	0	0
RQot	0	57	87	849	326	122	0	0	20	303
Outros	58	0	0	0	0	0	0	9	4	0

Figura 1: Mapas pedológicos digitais da bacia do córrego Àguas da Lúcia gerados através dos algoritmos (a) J48, (b) *Random Forest* e (c) MLP.

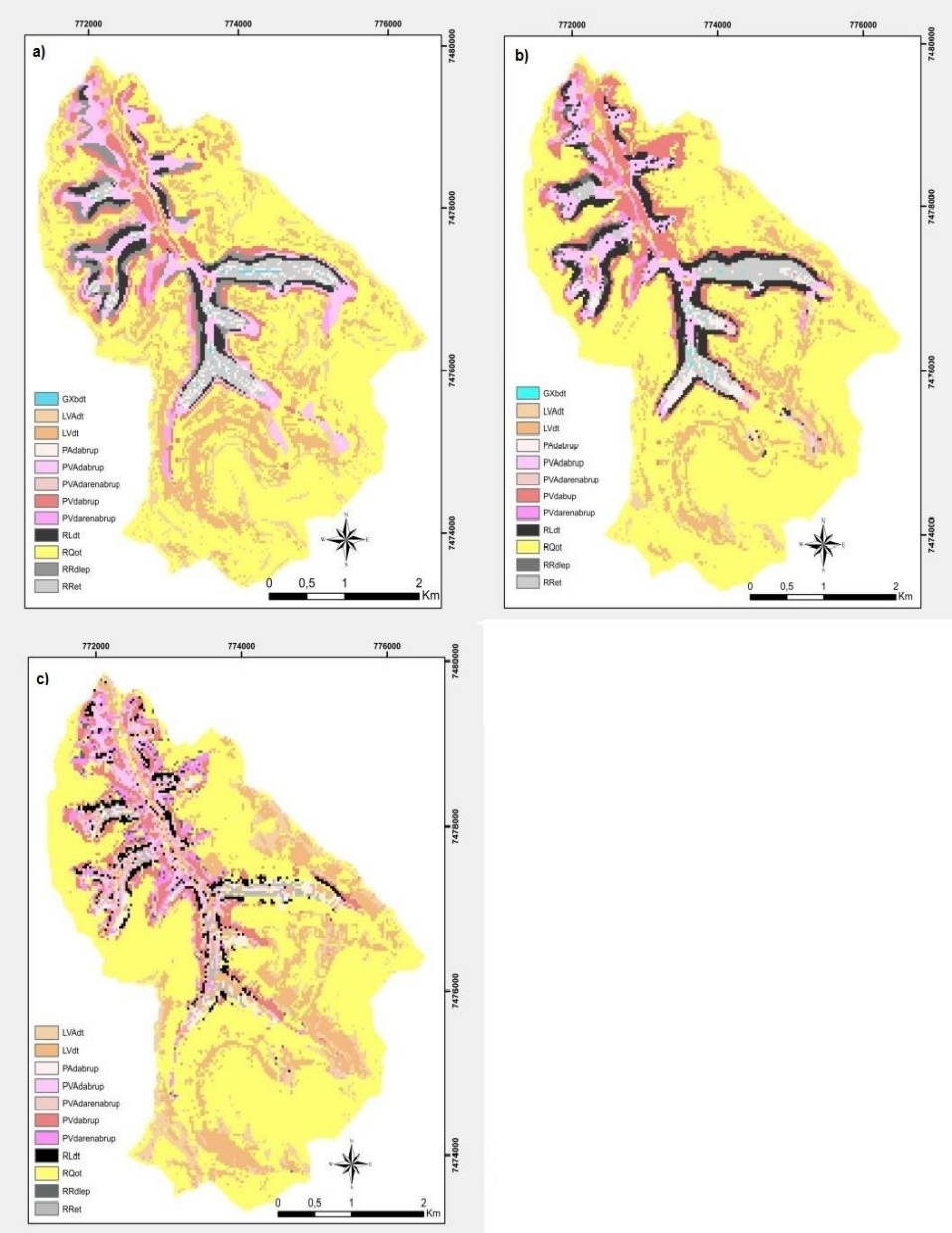

Tabela 3: Matriz de erro com as observações de solo classificado no campo e as preditas pelo algoritmo J48, bacia do Águas da Lúcia, Botucatu, SP.

	Preditas														
Reais	**LV Adt**	**LV dt**	**PAda brup**	**PA dt**	**PVAda brup**	**PVAdare nabrup**	**PV Adt**	**PVda brup**	**PVdaren abrup**	**PVe lat**	**RL dt**	**RQ ht**	**RQ ot**	**To tal**	**Exati dão do usuá rio (%)**
LVAdt	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
LVdt	0	3	0	0	0	0	0	0	0	0	0	0	1	4	75
PAdabru p	0	0	1	0	0	0	0	0	0	0	0	0	0	1	100
PAdt	0	0	0	0	0	0	0	0	0	0	1	0	0	1	0
PVAdabr up	0	0	0	0	1	0	0	0	0	0	0	0	0	1	100
PVAdare nabrup	1	0	0	0	0	0	0	0	0	0	0	0	0	1	0
PVAdt	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0
PVdabru p	0	1	0	0	0	0	0	1	0	0	0	0	0	2	50
PVdarena brup	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
PVelat	0	0	0	0	0	0	0	1	0	0	0	0	0	1	0
RLdt	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
RQht	0	0	0	0	1	0	0	0	0	0	0	0	1	2	0
RQot	0	0	0	0	0	0	0	1	1	0	1	0	8	11	72.7
Total	1	4	1	0	2	0	0	3	1	0	2	0	11	25	
Exatidão do produtor (%)	0	75	100	0	50	0	0	33.3	0	0	0	0	72.7		
Acurácia Global	0.56														
Kappa	0.42														

4. Considerações Finais

Mesmo com quatro classes de solo identificadas na validação (25 observações) não constarem do conjunto de treinamento, os classificadores J48 e *Random Forest* apresentaram acurácia global de 56 %. Todavia, melhor desempenho foi atribuído ao J48 por este apresentar maior kappa (0,42) que o *Random Forest* (0,39).

Apesar do melhor desempenho geral do algoritmo MLP (kappa 0,56 e acurácia global 60 %), ele não classificou nenhum exemplo da classe GXbdt aren, com apenas dois exemplos para treinamento, sugerindo baixa confiabilidade do classificador para mapeamentos em maior detalhe, em que espera-se predição também de classes com menor frequência de ocorrência..

O tamanho da área de estudo e o grau de detalhe de análise das variáveis geomorfométricas geraram uma variabilidade de atributos preditivos das classes de solo não representada pelo conjunto de treinamento, que foi adicionalmente reduzido devido à exclusão das classes com poucos exemplos.

Mais de um grupo de variáveis de relevo criado pela análise de *cluster* selecionou diferentes classes de solos de ocorrência majoritária (classes representativas em extensão) e distintas

taxonomicamente por diferenciação de textura, o que indicou que relevo não é fator preponderante para a variabilidade textural de solos da área.

A elevada demanda operacional para o treinamento de modelos preditivos por pontos de observação de solo em campo cria empecilho à realização dos MDS devido a baixa representatividade do universo amostrado nos levantamentos de nível detalhado, que abrangem áreas pequenas.

Referências

ARRUDA, P.G. et al. Mapeamento digital de sols por redes neurais artificiais com bases nas relações solo – paisagem. Revista Brasileira de Ciência do Solo. Viçosa, v.37, n.2, p. 327-338, 2013.

BREIMAN, L. Random Forests. Journal of Machine Learning, v.45, p.5-32, 2001.

CAMILO, et al. Mineração de dados: Conceitos, tarefas,métodos e ferramentas. Relatório técnico. Instituto de informática. Universidade Federal de Goiás, 2009.

CONGALTON, R. G. A review of assessing the accuracy of classifications of remotely sensed data. Remote Sensing Environment, v.37, p.35-46, 1991.

EMPRESA BRASILEIRA DE PESQUISA AGROPECUÁRIA - EMBRAPA. Centro Nacional de Pesquisa de Solos. Sistema brasileiro de classificação de solos. 2.ed. Rio de Janeiro, 2006. 306p.

HALL, M. A. et al. The WEKA Data Mining Software: Na Update. SIGKDD Explorations, v. 11, n. 1, p. 10 – 18, 2009.

KANELLOPOULOS, I.; WILKINSON, G.G. Strategies and best pratice for neural network image classification. International Journal of Remote Sensing, v. 18, n.4, p.711–725, 1997.

TEN CATEN, A. et al. Mapeamento digital de classes de solos: características da abordagem brasileira. Ciência Rural, v.42, n.11, 2012.

WILSON, J.P. & GALLANT, J.C. Digital terrain analysis. p.1-27. In: WILSON, J.P. & GALLANT, J.C., eds. Terrain analysis: Principles and applications. New York, Wiley & Sons, 2000.

WITTEN, I.H. & FRANK, E. Data mining: practical machine learning tools and techniques. 2.ed. San Francisco, Morgan Kaufmann, 2005. 524p.