Predição de classes de solo por mineração de dados em área da bacia sedimentar do São Francisco

Laura Milani da Silva Dias[1], Ricardo Marques Coelho[2], Gustavo Souza Valladares[3], Ana Carolina Cunha de Assis[2], Edilene Pereira Ferreira[2] e Rafael Cipriano da Silva[4]

[1]Universidade Estadual de Campinas, Rua Pandiá Calógeras, nº 51, Cidade Universitária, CEP 13083-870 Campinas, SP, Brasil. E-mail: laurads5@yahoo.com.br [2]Instituto Agronômico, Avenida Barão de Itapura, nº 1.481, Jardim Guanabara, CEP 13012-970 Campinas, SP, Brasil. E-mail: rmcoelho@iac.sp.gov.br, assisacc@gmail.com, edilene_agro@yahoo.com.br [3]Universidade Federal do Piauí, Campus Universitário Ministro Petrônio Portella, Ininga, CEP 64049-550 Teresina, PI, Brasil. E-mail: valladares@ufpi.edu.br [4]Universidade de São Paulo, Escola Superior de Agricultura Luiz de Queiroz, Avenida Pádua Dias, nº 11, Vila Independência, CEP 13418-260 Piracicaba, SP, Brasil. E-mail: ciprianors@usp.br

Resumo – O objetivo deste trabalho foi avaliar diferentes estratégias para a predição da distribuição de classes de solo em mapas pedológicos digitais de áreas sem dados de referência, na bacia sedimentar do São Francisco, no Norte de Minas Gerais. As estratégias incluíram: o detalhamento da legenda, o treinamento por observações em campo, a ampliação do conjunto de treinamento e o uso de diferentes algoritmos de mineração de dados. Foram elaboradas quatro matrizes, diferenciadas pelo volume de dados, para o aprendizado dos algoritmos, e pelo nível taxonômico das classes de solo a serem preditas. Avaliou-se o desempenho dos algoritmos de aprendizado de máquina – Random Forest, J48 e MLP –, associados a procedimentos de discretização, balanceamento de classes, seleção de variáveis e expansão do conjunto de treinamento. O balanceamento de classes, a discretização de variáveis por frequências iguais e o algoritmo Random Forest apresentaram os melhores desempenhos. A extensão da representatividade das observações em campo, que presume uma área de treinamento mais ampla, não trouxe ganho preditivo. A generalização taxonômica para subordem diminui a fragmentação dos polígonos mapeados e aumenta a acurácia dos mapas pedológicos digitais. Quando são produzidos após treinamento por observações de solo in situ, na área de mapeamento, os mapas pedológicos digitais têm valores de acurácia equivalentes aos dos treinados em mapas preexistentes.

Termos para indexação: acurácia de mapas pedológicos, algoritmos de classificação, mapa digital de solos, variáveis preditivas do meio físico.

Soil class prediction by data mining in an area of the sedimentary São Francisco basin

Abstract – The objective of this work was to evaluate different strategies for the prediction of soil class distribution on digital soil maps of areas without reference data, in the sedimentary basin of San Francisco, in the north of the state of Minas Gerais, Brazil. The strategies included: taxonomic generalization, training by field observations, training set expansion, and the use of different data mining algorithms. Four matrices were developed, differentiated by the volume of data for machine learning and by soil taxonomic levels to be predicted. The performance of the machine learning algorithms – Random Forest, J48, and MLP –, associated with discretization, class balancing, variable selection, and expansion of the training set was evaluated. Class balancing, variable discretization by equal frequencies, and the Random Forest algorithm showed the best performances. The representativeness extension of field observations, that assumes a larger training area, brought no predictive gain. Soil taxonomic generalization to the suborder level reduces the fragmentation of mapped polygons and improves the accuracy of digital soil maps. When generated by training on in situ soil observations at the mapping area, digital soil maps are as accurate as those trained on preexistent maps.

Index terms: soil map accuracy, classification algorithms, digital soil map, predictive variables of the terrain.

Introdução

Em face da demanda permanente por mapeamento pedológico para planejamento da gestão e ocupação racional das terras no Brasil, as limitações para a aquisição de dados de solos têm levado ao estudo de técnicas de mapeamento digital de solos (MDS) (Mendonça-Santos & Santos, 2003; Chagas et al., 2010), cuja principal abordagem é a predição das classes ou propriedades de solos por meio de modelos

matemáticos e seu mapeamento digital de forma contínua e espacial (McBratney et al., 2003).

O nível de detalhamento dos mapas pedológicos e da legenda está relacionado, entre outros fatores, à finalidade à qual eles se destinam. A necessidade de maior detalhe do mapa pedológico requer legendas de solos com maior detalhamento taxonômico. Esta relação "detalhe do mapa/detalhe taxonômico" está bem estabelecida nos mapeamentos convencionais (Manual técnico de pedologia, 2007), mas não nos mapas digitais, especialmente pela escassa validação destes últimos, o que impede o conhecimento do nível de confiança dos mapas digitais (Ten Caten et al., 2012). Embora a predição de classes de solo venha sendo realizada em distintos níveis hierárquicos do Sistema Brasileiro de Classificação de Solos (SiBCS), grande parte dos trabalhos usam os níveis de ordem e subordem (Carvalho Junior et at., 2011; Giasson et al., 2011), ou mesmo associações de classes de solo nesses níveis hierárquicos (Giasson et al., 2006; Ten Caten et al., 2011). No 3.º e 4.º níveis, os solos são frequentemente classificados a partir de características de mais difícil associação com processos morfogenéticos do solo (Ten Caten et al., 2012) e, assim, de difícil predição por métodos que utilizem apenas variáveis como as do relevo, obtidas de modelos digitais de elevação de baixa resolução (por exemplo, 30 m).

Como informação básica para o treinamento de modelos preditivos de classes ou propriedades de solos, têm prevalecido os mapas convencionais preexistentes, com extrapolação das relações solo-paisagem de áreas de referência adjacentes e fisiograficamente semelhantes (Grinand et al., 2008; Sarmento et al., 2012; Silva et al., 2013; Höfig et al., 2014). Na ausência destas informações, há possibilidade de treinar os modelos preditivos em pontos de observação e amostragem de campo, na própria área de mapeamento; esta alternativa metodológica não tem registro na literatura consultada. Assim, espera-se que o mapa alcance acurácia comparável aos elaborados com outras técnicas, se o número de pontos de treinamento for equivalente ao usado para treinamento em áreas de referência (Ten Caten et al., 2012), já que o treinamento não depende da qualidade dos mapas preexistentes ou da impureza decorrente da generalização cartográfica (Sarmento et al., 2014).

A capacidade preditiva dos modelos apoiados na mineração de dados também depende de bases de treinamento robustas, que captem ao máximo a variação dos atributos preditivos e dos solos (Teske et al., 2015). O treinamento por pontos de observação e coleta em campo cria elevada demanda por estes pontos para treinamento dos modelos, o que pode tornar o método pouco operacional. Todavia, a ampliação dos conjuntos de treinamento pode ser feita, estendendo-se a classificação do solo de locais com classe de solo conhecida para locais sem observação do solo in situ, mas com conjunto similar de variáveis preditivas. A efetividade deste procedimento pode ser avaliada pela qualidade dos mapas resultantes.

O objetivo deste trabalho foi avaliar diferentes estratégias que incluem o detalhamento da legenda, o treinamento por observações em campo, a ampliação do conjunto de treinamento e o uso de diferentes algoritmos de mineração de dados, para a predição da distribuição de classes de solo em mapas pedológicos digitais de áreas sem dados de referência, na bacia sedimentar do São Francisco, no norte de Minas Gerais.

Material e Métodos

A área estudada está inserida na Microrregião de Montes Claros, norte do Estado de Minas Gerais, com o Município de Capitão Enéas em posição central, e tem 110.289 ha, delimitada entre 16º02'09" e 16º40'14"S e 43º56'51" e 43º36'54"W. Segundo a classificação de Köppen-Geiger, o clima é Aw – tropical de savana, com estação seca de inverno, com médias anuais de precipitação e temperatura de 1.060 mm e 24,2ºC, respectivamente.

Com base no relevo e geologia, a área do estudo é representativa regionalmente. A área se insere nos Patamares e Depressão do Alto-Médio Rio São Francisco, unidades de relevo mais extensas da bacia sedimentar do São Francisco, no domínio geomorfológico do Cráton Neoproterozoico do Nordeste (Mapa de unidades de relevo do Brasil, 2006). Pela compilação da carta geológica em escala 1.100.000 (Carta geológica, 2011) e do mapa metalogenético em escala 1:250.000 (Feboli, 1985), ambos da folha Montes Claros, verifica-se o predomínio de rochas sedimentares pelíticas ao norte da área, onde o relevo é bastante aplainado, e de rochas metapelíticas ao sul, em relevo suave-ondulado e ondulado, das formações Lagoa do Jacaré e Serra de Santa Helena, amplamente distribuídas no norte de Minas Gerais

(Carta geológica do Brasil ao milionésimo, 2004). Também são encontrados depósitos aluvionares nas cotas baixas da região, ao longo do vale do Rio Verde Grande, compostos por sedimentos grossos e depósitos colúvio-eluviais residuais de sedimentos areno-siltosos, em testemunhos distintos e subnivelados altimetricamente em toda a área do estudo.

A elaboração das matrizes de dados, seu pré-processamento, mineração e a avaliação dos mapas pedológicos digitais estão descritos nas etapas a seguir, como: variáveis preditivas; classes de solo; matrizes de dados; treinamento dos modelos e predição das classes de solo; validação de dados.

Variáveis preditivas – a partir de dados de hipsometria SRTM, com resolução espacial de 90 m, foi gerado um modelo digital de elevação (MDE) e foram derivadas as variáveis geomorfométricas altitude, declividade, curvatura planar, curvatura em perfil e orientação das vertentes, em ambiente ArcGIS 10, e a distância da drenagem e o índice topográfico de umidade, pelo módulo Terrain Analysis do SAGA GIS (System for automated geoscientific analyses). A variável litologia foi obtida da compilação dos mapas geológicos descritos anteriormente (Feboli, 1985; Carta geológica, 2011). Dados do sensor OLI (operational land imager) do Landsat 8, obtidos em agosto de 2014, foram utilizados para derivar os índices clay minerals (CMI) e iron oxides (IOI) (Sabins, 1997). O primeiro foi gerado por divisão da banda 6 (1,57–1,65 μm) pela 7 (2,11–2,29 μm), e o segundo pela divisão da banda 4 (0,64–0,67 μm) pela 2 (0,45–0,51 μm). A esses índices acrescentou-se o índice de vegetação NDVI (normalized difference vegetation index), todos obtidos com a utilização do programa ENVI 4.7.

Classes de Solo – a seleção dos locais de observação de solos foi realizada pelo programa de amostragem aleatória estratificada cLHS (Wyss & Jorgensen, 1998), tendo como condicionantes as variáveis de relevo, litologia e os índices derivados da imagem Landsat 8. A caracterização de solos em campo foi em minitrincheiras, com sondagens de trado na base destas, ou em cortes de estrada, com caracterização morfológica de solo e coleta de amostras para análise. Caracterizaram-se 26 perfis completos (caracterização morfológica e analítica, em todos os horizontes pedogenéticos) e 226 cortes de barranco ou minitrincheiras (caracterização morfológica parcial ou coleta parcial de horizontes). Coletaram-se amostras de solo de todos os pontos de observação, nos seguintes horizontes pedogenéticos: em todos os horizontes; em dois horizontes (superficial e subsuperficial); no horizonte superficial ou no horizonte subsuperficial. Com as amostras, realizaram-se análises granulométricas e químicas de rotina pedológica (Claessen, 1997) e a classificação dos perfis, pelo Sistema Brasileiro de Classificação de Solos (SiBCS), até o quarto nível hierárquico e grupamento textural (Santos et al., 2013). Das 252 observações de solo, 190 foram usadas para o aprendizado dos modelos e 62 para posterior validação dos mapas.

Matrizes de dados – as variáveis geomorfométricas, os índices derivados da imagem Landsat 8 e a litologia foram considerados preditivos das classes de solo avaliadas em dois níveis hierárquicos do SiBCS: subgrupo (4º nível), acrescido do grupamento textural; e subordem (2º nível). As variáveis preditivas em formato raster, com pixels de 90 x 90 m compuseram duas matrizes de dados – 1 e 3 (Tabela 1), de 136.149 linhas cada. Destas, apenas 190 linhas (pixels) (0,14%) tinham classe de solo determinada. A classe dos solos das demais linhas foi predita pelos algoritmos de mineração de dados. Na tentativa de expandir o conjunto de treinamento e a representatividade do ponto classificado em campo para além do pixel de 90 x 90 m, atribuiu-se – para cada pixel com solo classificado – a mesma classe de solo para os oito pixels a ele adjacentes. Assim, cada observação de solo que, inicialmente, era representada por uma área de 90 x 90 m (tamanho do pixel), passou a ser representada por uma área de 270x270 m. Desta forma, duas novas matrizes de dados foram elaboradas – 2 e 4 (Tabela 1), e o conjunto original composto por 190 linhas classificadas passou a ter 1.710 linhas (1,25% do total).

Treinamento dos modelos e predição das classes de solo – o treinamento dos modelos preditivos foi realizado no programa de mineração de dados Weka 3.6.6. Verificou-se o desempenho de três algoritmos: J48, uma implementação para a técnica de árvores

Tabela 1. Descrição das matrizes de dados.

Matriz	Generalização taxonômica	Conjunto de treinamento
1	4.º nível + grupamento textural	Original
2	4.º nível + grupamento textural	Expandido
3	2.º nível	Original
4	2.º nível	Expandido

de decisão (Hall et al., 2009); Random Forest, uma extensão da técnica de árvores de decisão, que combina a predição de diversas árvores (Breiman, 2001); e MLP, que emprega redes neurais. Para cada um dos algoritmos, testaram-se técnicas de pré-processamento dos dados, como discretização, balanceamento de classes e ordenamento e seleção de variáveis. O procedimento de discretização dividiu as variáveis – originalmente numéricas e contínuas – em cinco intervalos, por dois critérios distintos: intervalos de mesmo tamanho; e intervalos de mesma frequência (mesmo número de amostras). Os balanceamentos de classes foram calibrados em 0, 0,5 e 1, que representam, respectivamente: a distribuição original dos dados; a distribuição com subamostragem das classes com maior número de observações; e a distribuição com reamostragem das classes, considerando-se igual quantidade de observações em todas as classes. Para o ordenamento e a seleção das variáveis, aplicaram-se os algoritmos ChiSquared e GainRatio baseados, respectivamente, no teste de qui-quadrado e na entropia, que ordenam as variáveis considerando sua capacidade preditiva. Os modelos, constituídos pelo algoritmo e pela sequência de técnicas de pré-processamento que resultaram na maior acurácia, foram aplicados às matrizes para a predição das classes de solo faltantes. As matrizes foram extraídas do programa Weka e inseridas no ArcGIS 10, em formato raster, para gerar os mapas pedológicos digitais.

Validação dos mapas – em matrizes de erro, as 62 observações em campo separadas para validação foram confrontadas com os mapas preditos, tendo-se obtido a acurácia global, que é a proporção de observações corretamente classificadas, e o índice kappa.

Resultados e Discussão

Os diferentes procedimentos de pré-processamento dos dados, combinados aos três algoritmos classificadores, resultaram no desempenho distinto de cada uma das matrizes de dados minerados (Tabela 2). O algoritmo de classificação Random Forest, com raro registro de uso em mapeamento digital de solos (Stum et al., 2010), gerou o modelo de maior acurácia para as quatro matrizes de dados e teve desempenho superior ao dos algoritmos J48 e MLP.

As matrizes de dados discretizados por frequências iguais mostraram maior acurácia do que as matrizes de dados contínuos e do que aquelas com dados discretizados por intervalos iguais. Valores contínuos dificultam o processo de identificação de padrões para a modelagem, pois, algumas variáveis podem conter tantos valores distintos que o algoritmo encontra dificuldade de identificar os padrões nos dados para os quais quer criar o modelo (Han & Kamber, 2006). Por sua vez, a divisão em intervalos de igual tamanho desconsidera padrões naturais de ocorrência das variáveis preditivas e elimina distinções úteis, quando reúne na mesma classe diferentes padrões de ocorrência natural ou separa um mesmo padrão em classes diferentes (Han & Kamber, 2006).

O balanceamento de classes com índice 1 teve desempenho superior na grande maioria dos modelos. O fato de a área de estudo ter ampla distribuição de extensão territorial das classes de solo pode ter contribuído para o bom desempenho, já que esta calibração considera igual quantidade de observações, para todas as classes de solo. O índice 0,5 é uma estratégia de subamostragem, que elimina aleatoriamente exemplos das classes majoritárias, e pode causar a perda de informação útil ao modelo, enquanto a ausência de balanceamento (índice zero) supre o modelo com pouca informação em classes de pequena extensão de ocorrência (Chawla, 2010).

Na seleção de variáveis por poder preditivo (algoritmos Chi-Squared e GainRatio), a altitude e

Tabela 2. Acurácia dos três algoritmos e procedimentos de pré-processamento para as quatro matrizes de dados[(1)].

Matriz	Algoritmo	Pré-processamento			Acurácia
		Discretização	Balanceamento de classes	Seleção de variáveis	
1	RF	DFI	1	ECP	69,1
	MLP	DFI	0,5	E_NDVI	68,6
	J48	DFI	1	ECP	55,8
2	RF	DFI	1	Completo	67,9
	MLP	DFI	1	E_NDVI	61,0
	J48	DFI	1	E_NDVI	56,8
3	RF	Contínua	1	E_NDVI	76,0
	MLP	Contínua	0,5	E_NDVI	69,6
	J48	Contínua	0,5	E_NDVI	72,8
4	RF	DFI	1	E_NDVI	73,3
	MLP	DFI	1	E_NDVI	64,4
	J48	DFI	1	E_NDVI	67

[(1)]RF, Random Forest; DFI, discreta por frequências iguais; ECP, exclui curvatura planar; E_NDVI, exclui NDVI; .

a litologia se destacaram. O controle litológico na distribuição dos solos é fato bastante frequente (Ten Caten et al., 2012). Por sua vez, a altitude na área está relacionada ao formato do relevo e declividade: nas cotas mais elevadas, a topografia é mais ondulada do que nas cotas inferiores, e a topografia condiciona a diferenciação dos solos.

O uso de um maior número de variáveis preditivas deveria fornecer ao modelo maior poder de discriminação das classes, porém, isto não acontece quando há variáveis irrelevantes ou redundantes (Lee et al., 2006). Por este motivo, para as matrizes de dados 1, 3 e 4, houve aumento da acurácia, quando as variáveis curvatura planar e NDVI foram retiradas. Cerca de 75% da área de estudo apresenta curvatura planar plana, prevalência justificada pela predominância (86%) de relevo plano e suave-ondulado. Tal homogeneidade da variável não contribuiu para a discriminação das classes de solo e, consequentemente, para o modelo preditivo. Quanto ao NDVI, o índice é mais sensível a variações na cobertura vegetal do que nas propriedades espectrais dos solos, o que é importante quando a cobertura vegetal é heterogênea. Isto ocorre na área mapeada, especialmente nas áreas mais planas ao norte, com maior fragmentação do uso da terra, o que promove variação do NDVI desvinculada de variação das propriedades do solo.

Quando aplicados às matrizes de dados para predição das classes de solo, os modelos preditivos resultaram em mapas com diferentes distribuições e extensões espaciais das classes de solo, e em distintos valores medidos de acurácia global e índice kappa (Tabela 3).

O mapa gerado a partir do modelo da matriz de dados 1 é bastante fragmentado (Figura 1), com quatorze classes de solo, identificadas nas observações em campo e classificadas em nível de subgrupo e grupamento textural (Tabela 4). O detalhamento taxonômico (4.º nível + grupamento textural) da classificação dos solos desta matriz gerou mais classes de solo, cada uma associada a um conjunto próprio de variáveis preditivas. O maior número de instâncias de "variáveis preditivas + classes de solo" para treinamento é a causa mais provável para a maior fragmentação dos mapas gerados com essa matriz de dados.

Tabela 3. Validação dos mapas dos quatro modelos de maior acurácia por matriz de dados.

Matriz	Acurácia global	Índice kappa	
		Valor	Concordância
1	54,8	0,50	Moderada
2	48,3	0,44	Moderada
3	58,0	0,44	Moderada
4	53,2	0,41	Moderada

As classes Cambissolo Háplico distrófico típico textura argilosa e Latossolo Vermelho distrófico típico textura argilosa representam a maior área relativa no mapa (40%), mas os Cambissolos aparecem mais bem distribuídos por toda sua extensão. Latossolos Vermelhos distróficos típicos textura média (14,5%) e Latossolos Vermelhos eutróficos típicos textura argilosa (10%), ambos localizados ao norte da área, e os Nitossolos Vermelhos eutróficos típicos textura argilosa (9%), localizados ao centro, também têm áreas representativas (Figura 1). A acurácia global do mapa foi de 54,8%, e o índice kappa, de 0,50 (concordância moderada) (Tabela 3).

A matriz de dados 2 tem o mesmo nível taxonômico da matriz de dados 1, mas o procedimento que expandiu o conjunto de treinamento originou mapa com proporção e distribuição das classes de solo diferentes (Figura 1). A classe Latossolos Vermelhos distróficos típicos argilosos permanece extensa (19,7%), porém, mais concentrada ao norte da área, em relevo plano. Os Latossolos Vermelho-Amarelos distróficos típicos textura média, localizados principalmente no centro e ao norte da área, têm extensão muito maior (42%, Figura 1 B), em comparação aos do mapa da matriz 1 (16,5%, Figura 1 A). As duas classes de solo da ordem dos Cambissolos, distintas pela saturação por bases, têm sua área reduzida em 12%. Ao contrário do que se esperava pelo maior volume de dados de treinamento (Sarmento et al., 2012; Ten Caten et al., 2013), tanto a acurácia global (48,3%) quanto o índice kappa (0,44) foram inferiores aos do mapa da matriz de dados 1 (Tabela 3).

O mapa pedológico digital produto da matriz de dados 3 tem acurácia global de 58% e índice kappa 0,44 (Tabela 3). Tem menos polígonos, em consequência do nível hierárquico taxonômico mais elevado das classes de solos distribuídos em nove classes (Figura 1 C; Tabela 4). A distribuição dos Latossolos Vermelhos (50% da área), concentrada ao norte da área, e dos Cambissolos Háplicos (26%), ao sul da área, em relevo mais ondulado, permanece.

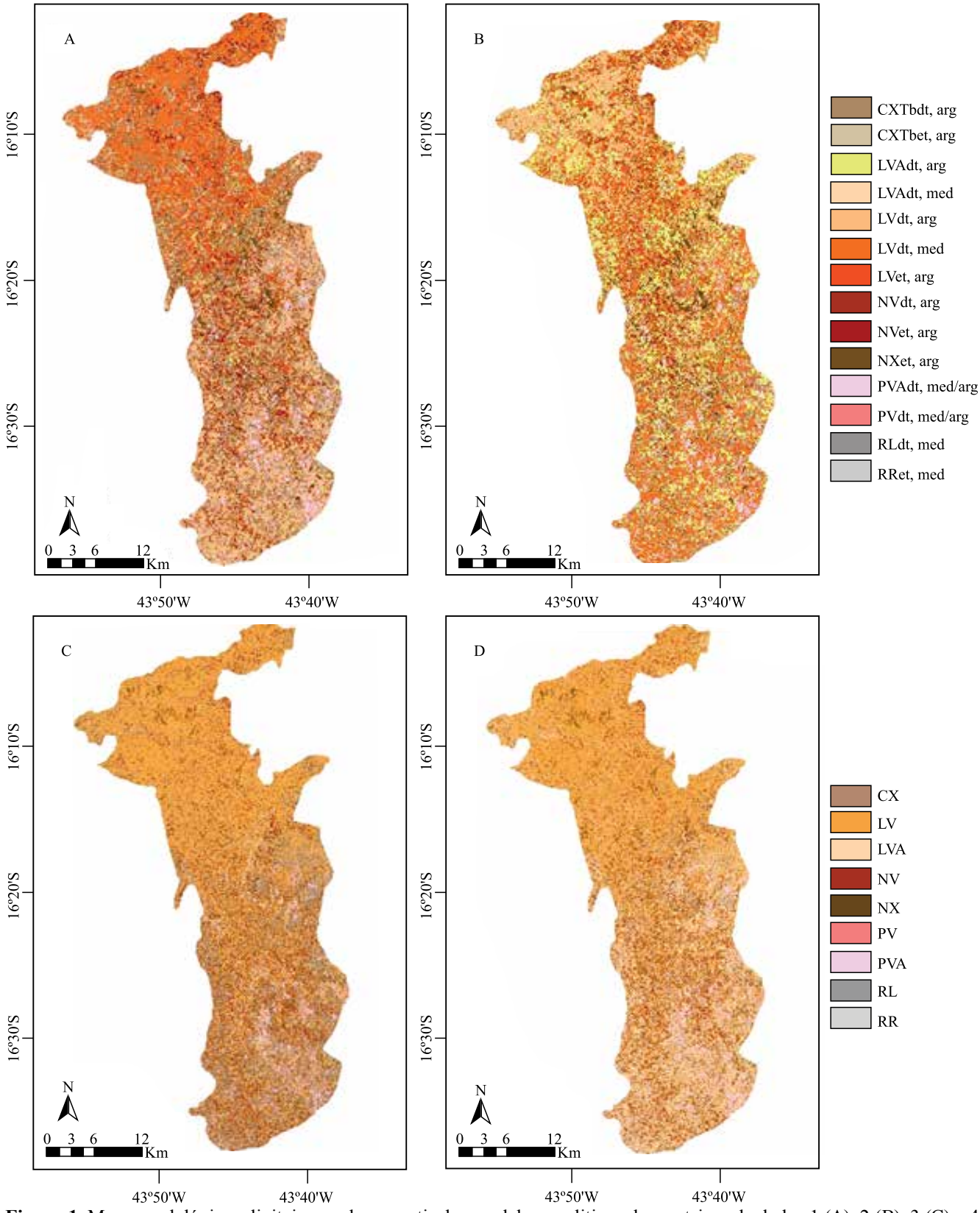

Figura 1. Mapas pedológicos digitais gerados a partir dos modelos preditivos das matrizes de dados 1 (A), 2 (B), 3 (C) e 4 (D), com o algoritmo de maior acurácia (Random Forest).

O mapa gerado a partir da matriz de dados 4, por sua vez, foi predominantemente classificado como Latossolo Vermelho e Latossolo Vermelho-Amarelo que, juntos, representam cerca de 70% da área do mapa (Figura 1 D). Os valores de acurácia (53,2%) e índice kappa (0,41) do mapa com a base ampliada (Tabela 3) não superaram os do mapa gerado com o conjunto original (Figura 1 C).

Atribuir a classificação do solo de um pixel para os pixels a ele adjacentes ampliou a base de treinamento de 190 para 1710 observações. Esta ampliação do conjunto de treinamento deveria contribuir para modelos mais robustos, mas a acurácia global dos mapas gerados após a ampliação mostrou-se inferior àquela com o conjunto de treinamento original (Tabela 3). Entende-se, assim, que não ocorre a homogeneidade (taxonômica) de solos presumida, pela extensão da classificação do solo de cada observação para os pixels adjacentes, ou seja, estender cada observação de um (8.100 m^2) para nove (72.900 m^2) pixels inclui classes de solo distintas da identificada no pixel central (observação inicial de campo). Por esse motivo, o conjunto de treinamento expandido (matrizes 2 e 4) apresentou instâncias (pixels) com associações solo-variáveis preditivas espúrias, o que reduziu a acurácia dos mapas.

Tabela 4. Símbolo e classes de solo nas matrizes de dados.

Símbolo	Classes de solo
	Matrizes 1 e 2
CXTbd tip, arg	Cambissolo Háplico Tb distrófico típico, argilosa
CXTbe tip, arg	Cambissolo Háplico Tb eutrófico típico, argilosa
LVAd tip, arg	Latossolo Vermelho-Amarelo distrófico típico, argilosa
LVAd tip, med	Latossolo Vermelho-Amarelo distrófico típico, média
LVd tip, arg	Latossolo Vermelho distrófico típico, argilosa
LVd tip, med	Latossolo Vermelho distrófico típico, média
LVe tip, arg	Latossolo Vermelho eutrófico típico, argilosa
NVd tip, arg	Nitossolo Vermelho distrófico típico, argilosa
NVe tip, arg	Nitossolo Vermelho eutrófico típico, argilosa
NXe tip, arg	Nitossolo Háplico eutrófico típico, argilosa
PVAd tip, med/arg	Argissolo Vermelho-Amarelo distrófico típico, média/argilosa
PVd tip, arg	Argissolo Vermelho distrófico típico, argilosa
RLe tip, med	Neossolo Litólico eutrófico típico, média
RRe tip, med	Neossolo Regolítico eutrófico típico, média
	Matrizes 3 e 4
CX	Cambissolo Háplico
LVA	Latossolo Vermelho-Amarelo
LV	Latossolo Vermelho
NV	Nitossolo Vermelho
NX	Nitossolo Háplico
PVA	Argissolo Vermelho-Amarelo
PV	Argissolo Vermelho
RL	Neossolo Litólico
RR	Neossolo Regolítico

Os resultados também mostraram que a acurácia global dos mapas com classes de solo em nível de subordem foi superior em pelo menos 5% à dos mapas com classificação em nível de subgrupo. O acréscimo dos valores de acurácia foi verificado tanto para as matrizes com conjunto de treinamento original quanto para as que tiveram o conjunto ampliado (Tabela 3). O aumento da acurácia global não se repetiu para o índice kappa, em razão da maior chance de acertos ao acaso (Congalton, 1991) com a generalização taxonômica. Os mapas com classes de solo mais generalizadas (nível hierárquico superior) têm menor detalhamento e menor fragmentação do mapa.

Os valores de acurácia global e índice kappa, alcançados com o treinamento dos modelos em pontos de observação de solo no campo (Tabela 3), mostraram-se comparáveis aos de estudos anteriores, treinados em mapas convencionais de referência, como os de Bui & Moran (2003), com acurácia global média de 66% e índices kappa entre 0,33 a 0,74. Outros estudos com treinamento em mapas convencionais – como os de Giasson et al. (2011) com acurácia global 68% e índice kappa 0,54), Teske et al. (2015) com acurácia global 63% e índice kappa 0,46), e Silva et al. (2013) com acurácia global 52 % e índice kappa 0,41 – também são comparáveis aos maiores valores de acurácia global (58%) e índice kappa (0,50) obtidos no presente estudo.

Observa-se, ainda, que o valor de acurácia dos modelos preditivos (Tabela 2) é superior ao valor de acurácia dos mapas validados por meio das matrizes de erros, tendo-se a classificação do solo em campo como referência (Tabela 3). Assim, a acurácia do modelo da matriz de dados 1 foi de 69,1%, após a retirada da variável curvatura planar. No entanto, a acurácia do mapa gerado a partir dessa matriz foi de 54,8%. Esta tendência se repete para as quatro matrizes de dados avaliadas. A acurácia do modelo é consequência da qualidade das variáveis preditivas utilizadas, dos procedimentos de pré-processamento empregados e da capacidade do algoritmo de classificação em interpretar, extrair os padrões das matrizes de dados e associá-los às diferentes classes de solo. Além dos fatores que influenciam o desempenho do modelo, a acurácia do mapa ainda está relacionada ao quanto a variabilidade de solos da área de estudo está sendo representada pelo conjunto de treinamento. Neste

caso, a grande extensão adotada para a área de estudo – 110.289 ha e resolução de 90x90 m –geraram um conjunto de combinações de variáveis preditivas dos solos muito grande (136.149 linhas), para toda a área, e difícil de representar com o conjunto de treinamento por observações de solos em campo (190 observações), o que reduziu a acurácia dos mapas em relação à dos modelos preditivos.

Conclusões

1. O algoritmo Random Forest é eficaz na predição de classes de solo a partir de variáveis de relevo, geologia e sensoriamento remoto, e supera os algoritmos J48 e o MLP.

2. A extensão da representatividade das observações em campo de treinamento para uma área mais abrangente (de 0,81 ha para 7,9 ha) reduz a acurácia dos mapas, por presumir maior homogeneidade pedológica do que a existente na área de estudo.

3. Quando se consideram informações de solos mais pormenorizadas (classes de solo em nível de subgrupo), os mapas digitais de solos apresentam informação espacial muito fragmentada.

4. Mapas pedológicos digitais, produzidos com treinamento em observações de solo in situ na área de mapeamento, têm valores de acurácia equivalentes aos dos treinados em mapas preexistentes.

Agradecimentos

À Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (Capes), pela concessão de bolsa; à Petrobrás/PBIO (Rede de Sistemas de Produção de Mamona, Girassol e Pinhão-Manso), pelo financiamento da pesquisa.

Referências

BREIMAN, L. Random forests. **Machine Learning**, v.45, p.5-32, 2001. DOI: 10.1023/A:1010933404324.

BUI, E.N.; MORAN, C.J. A strategy to fill gaps in soil survey over large spatial extents: an example from Murray-Darling basin of Australia. **Geoderma**, v.111, p.21-44, 2003. DOI: 10.1016/S0016-7061(02)00238-0.

CARTA geológica do Brasil ao milionésimo: Belo Horizonte: folha SE-23. Brasília, DF: Companhia de Pesquisas e Recursos Minerais, 2004. 1 mapa. Escala 1:1.000.000. Programa Geologia do Brasil.

CARTA geológica: Montes Claros: folha SE-23-X-A-XI. Rio de Janeiro: Companhia de Pesquisas e Recursos Minerais; Universidade Federal de Minas Gerais, 2011. 1 mapa. Escala 1:100.000.

CARVALHO JUNIOR, W. de; CHAGAS, C. da S.; FERNANDES FILHO, E.I.; VIEIRA, C.A.O.; SCHAEFER, C.E.G.; BHERING, S.B.; FRANCELINO, M.R. Digital soilscape mapping of tropical hillslope areas by neural networks. **Scientia Agricola**, v.68, p.691-696, 2011. DOI: 10.1590/S0103-90162011000600014.

CLAESSEN, M.E.C. (Org.). **Manual de métodos de análise de solo**. 2.ed. rev. atual. Rio de Janeiro: Embrapa-CNPS, 1997. 212p. (Embrapa-CNPS. Documentos, 1).

CHAGAS, C. da S.; FERNANDES FILHO, E.I.; VIEIRA, C.A.O.; SCHAEFER, C.E.G; CARVALHO JUNIOR, W. de. Atributos topográficos e dados do Landsat7 no mapeamento digital de solos com uso de redes neurais. **Pesquisa Agropecuária Brasileira**, v.45, p.497-507, 2010. DOI: 10.1590/S0100-204X2010000500009.

CHAWLA, N. V. Data mining for imbalanced datasets: an overview. In: MAIMON, O.; ROKACH, L. (Ed.). **Data mining and knowledge discovery handbook**. 2nd ed. New York: Springer, 2010. p.875-886. DOI: 10.1007/978-0-387-09823-4.

CONGALTON, R.G. A review of assessing the accuracy of classifications of remotely sensed data. **Remote Sensing Environment**, v.37, p.35-46, 1991. DOI: 10.1016/0034-4257(91)90048-B.

FEBOLI, W.L. **Projeto mapas metalogenéticos e de previsão de recursos minerais**: Montes Claros: folha SE.23-X-A. Montes Claros: Companhia de Pesquisa de Recursos Minerais, 1985. 1 mapa. Escala 1:250.000.

GIASSON, E.; CLARKE, R.T.; INDA JUNIOR, A.V.; MERTEN, G.H.; TORNQUIST, C.G. Digital soil mapping using multiple logistic regression on terrain parameters in southern Brazil. **Scientia Agricola**, v.63, p.262-268, 2006. DOI: 10.1590/S0103-90162006000300008.

GIASSON, E.; SARMENTO, E.C.; WEBER, E.; FLORES, C.A.; HASENACK, H. Decision trees for digital soil mapping on subtropical basalt steep lands. **Scientia Agricola**, v.68, p.167-174, 2011. DOI: 10.1590/S0103-90162011000200006.

GRINAND, C.; ARROUAYS, D.; LAROCHE, B.; MARTIN, M.P. Extrapolating regional soil landscapes from an existing soil map: sampling intensity, validation procedures, and integration of spatial context. **Geoderma**, v.143, p.180-190, 2008. DOI: 10.1016/j.geoderma.2007.11.004.

HALL, M.; FRANK, E.; HOLMES, G.; PFAHRINGER, B.; REUTEMANN, P.; WITTEN, I.H. The WEKA data mining software: an update. **SIGKDD Explorations**, v.11, p.10-18, 2009. DOI: 10.1145/1656274.1656278.

HAN, J.; KAMBER, M. **Data mining**: concepts and techniques. 2nd ed. San Francisco: Morgan Kaufmann, 2006. 770p.

HÖFIG, P.; GIASSON, E.; VENDRAME, P.R.S. Mapeamento digital de solos com base na extrapolação de mapas entre áreas fisiograficamente semelhantes. **Pesquisa Agropecuária Brasileira**, v.49, p.958-966, 2014. DOI: 10.1590/S0100-204X2014001200006.

LEE, H.D.; MONARD, M.C.; VOLTOLINI, R.F.; PRATI, R.C.; CHUNG, W.F. A simple evaluation model for feature subset selection algorithms. **Inteligencia Artificial**, v.10, n. 32, p.9-17, 2006. DOI: 10.4114/ia.v10i32.923.

MANUAL técnico de pedologia. 2.ed. Rio de Janeiro: Instituto Brasileiro de Geografia e Estatística, 2007. 323p. (IBGE. Manuais técnicos em geociências, 4).

MAPA de unidades de relevo do Brasil. Rio de Janeiro: Instituto Brasileiro de Geografia e Estatística, 2006. 1 mapa. Escala 1:5.000.000.

MCBRATNEY, A.B.; MENDONÇA-SANTOS, M.L.; MINASNY, B. On digital soil mapping. **Geoderma**, v.117, p.3-52, 2003. DOI: 10.1016/S0016-7061(03)00223-4.

MENDONÇA-SANTOS, M. de L.; SANTOS, H.G. dos. **Mapeamento digital de classes e atributos de solos**: métodos, paradigmas e novas técnicas. Rio de Janeiro: Embrapa Solos, 2003. 19p. (Embrapa Solos. Documentos, 55).

SABINS, F.F. **Remote sensing**: principles and interpretation. 3rd ed. New York: Waveland, 1997. 432p.

SANTOS, H.G. dos; JACOMINE, P.T.K.; ANJOS, L.H.C. dos; OLIVEIRA, V.A. de; LUMBRERAS, J.F.; COELHO, M.R.; ALMEIDA, J.A. de; CUNHA, T.J.F.; OLIVEIRA, J.B. de. **Sistema brasileiro de classificação de solos**. 3.ed. rev. e ampl. Rio de Janeiro: Embrapa Solos, 2013. 353p.

SARMENTO, E.C.; GIASSON, E.; WEBER, E.; FLORES, C.A.; HASENACK, H. Prediction of soil orders with high spatial resolution: response of different classifiers to sampling density. **Pesquisa Agropecuária Brasileira**, v.47, p.1395-1403, 2012. DOI: 10.1590/S0100-204X2012000900025.

SARMENTO, E.C.; GIASSON, E.; WEBER, E.J.; FLORES, C.A.; ROSSITER, D.G.; HASENACK, H. Caracterização de mapas legados de solos: uso de indicadores em mapas com diferentes escalas no Rio Grande do Sul. **Revista Brasileira de Ciência do Solo**, v.38, p.1672-1680, 2014. DOI: 10.1590/S0100-06832014000600002.

SILVA, C.C. da; COELHO, R.M.; OLIVEIRA, S.R. de M.; ADAMI, S.F. Mapeamento pedológico digital da folha Botucatu (SF-22-Z-B-VI-3): treinamento de dados em mapa tradicional e validação de campo. **Revista Brasileira de Ciência do Solo**, v.37, p.846-857, 2013. DOI: 10.1590/S0100-06832013000400003.

Stum, A.K.; Boettinger, J.L.; White, M.A.; Ramsey, R.D. Random forests applied as a soil spatial predictive model in arid Utah. In: BOETTINGER, J.L.; HOWELL, D.W.; MOORE, A.C.; HARTEMINK, A.E.; KIENAST-BROWN, S. (Ed.). **Digital soil mapping**: bridging research, environmental application, and operation. New York: Springer, 2010. p.179-190. DOI: 10.1007/978-90-481-8863-5_15.

TEN CATEN, A.; DALMOLIN, R.S.D.; MENDONÇA-SANTOS, M. de L.; GIASSON, E. Mapeamento digital de solos: características da abordagem brasileira. **Ciência Rural**, v.43, p.1989-1997, 2012. DOI: 10.1590/S0103-84782012001100013.

TEN CATEN, A.; DALMOLIN, R.S.D.; PEDRON, F. de A.; MENDONÇA-SANTOS, M. de L. Extrapolação das relações solo-paisagem a partir de uma área de referência. **Ciência Rural**, v.41, p.812-816, 2011. DOI: 10.1590/S0103-84782011000500012.

TEN CATEN, A.; DALMOLIN, R.S.D.; PEDRON, F. de A.; RUIZ, L.F.C.; SILVA, C.A. da. An appropriate data set size for digital soil mapping in Erechim, Rio Grande do Sul, Brazil. **Revista Brasileira de Ciência do Solo**, v.37, p-359-366, 2013. DOI: 10.1590/S0100-06832013000200007.

TESKE, R.; GIASSON, E.; BAGATINI, T. Comparação de esquemas de amostragem para treinamento dos modelos preditores no mapeamento digital de classes de solo. **Revista Brasileira de Ciência do Solo**, v.19. p.14-20, 2015. DOI: 10.1590/01000683rbcs20150344.

WYSS, G.D.; JORGENSEN, K.H. **A user's guide to LHS**: Sandia's Latin Hypercube Sampling Software. Albuquerque: Sandia National Laboratories, 1998. 138p. DOI: 10.2172/573301.

Recebido em 31 de agosto de 2015 e aprovado em 27 de janeiro de 2016