Mineração de dados para inferência de relações solo-paisagem em mapeamentos digitais de solo

Rafael Castro Crivelenti[1], Ricardo Marques Coelho[1], Samuel Fernando Adami[1] e Stanley Robson de Medeiros Oliveira[2]

[1]Instituto Agronômico, Avenida Barão de Itapura, nº 1.481, Jardim Guanabara, CEP 13012-970 Campinas, SP. E-mail: grilasso@hotmail.com, rmcoelho@iac.sp.gov.br, samuel@iac.sp.gov.br [2]Embrapa Informática Agropecuária, Avenida André Tosello, nº 209, CEP 13083-886 Barão Geraldo, Campinas, SP. E-mail: stanley@cnptia.embrapa.br

Resumo – O objetivo deste trabalho foi desenvolver uma metodologia para mapeamento digital de solos na escala 1:100.000 com a aplicação de técnicas de mineração de dados a descritores de relevo e a dados de mapas geológico e pedológico preexistentes. Foi criada uma base de dados digitais a partir de cartas topográficas e temáticas, que permitiu elaboração do modelo digital de elevação (MDE) da folha Dois Córregos, SP (escala 1:50.000). A partir do MDE, foram calculados os parâmetros geomorfométricos declividade, curvaturas em planta e perfil, área de contribuição e distância diagonal de drenagem. A matriz que associou esses dados georreferenciados foi analisada por meio de árvores de decisão, no ambiente de aprendizado de máquina Weka, o que gerou um modelo de predição de unidades de mapeamento de solos. A acurácia geral do modelo aumentou de 54 para 61% com a eliminação das classes com probabilidade nula de ocorrência. A associação da mineração de dados com sistemas de informações geográficas permite a elaboração de mapas digitais passíveis de uso em estudos que requeiram menor detalhamento que aqueles realizados com o mapa original.

Termos para indexação: árvores de decisão, levantamento pedológico, parâmetros geomorfométricos, sistemas de informação geográfica.

Data mining to infer soil-landscape relationships in digital soil mapping

Abstract – The objective of this work was to develop a methodology for digital soil mapping at a 1:100,000 scale by applying data mining techniques to preexisting relief descriptors and data from pedological and geological maps. A digital database was created from topographic and thematic maps, and allowed the generation of a digital elevation model (DEM) of the Dois Córregos (SP, Brazil) sheet (1:50,000 scale). The slope gradient, slope profile, contour profile, basin contributing area, and diagonal distance to drainage geomorphometric parameters were extracted from the DEM. The matrix which associated this georeferred data was analyzed by means of decision trees within the Weka machine-learning environment, and a model for soil mapping unit prediction was generated. The overall model accuracy increased from 54 to 61% when soil classes with no chances of being predicted were excluded. The association of data mining techniques with geographical information systems produced digital soil maps feasible to be used in studies requiring less detail than those made with the original reference soil maps.

Index terms: decision trees, soil survey, geomorphometric parameters, geographic information system.

Introdução

A longa duração e o elevado custo dos meios tradicionais motivam o estudo de métodos alternativos de levantamento pedológico (McBratney et al., 2003). O mapeamento digital de solos apresenta vantagens em relação ao método tradicional por ser uma alternativa rápida e econômica (Mendonça-Santos et al., 2008). Ele pode ser definido como a criação de sistemas de informação espacial com uso de modelos numéricos para a inferência das variações espaciais e temporais nos tipos de solos e em suas propriedades, a partir de observações e do conhecimento dos solos e de variáveis ambientais correlacionadas (Lagacherie & McBratney, 2007), como as que se baseiam nos fatores de formação dos solos, por exemplo, material de origem, clima, organismos, relevo e tempo. Uma das vantagens do mapeamento digital com base no conhecimento dos padrões regionais de solos é a possibilidade de prever a ocorrência de tipos de solos em áreas não mapeadas, com uso de informações geradas previamente em áreas de referência (Lagacherie & Voltz, 2000).

Entre as técnicas utilizadas para inferência dessas variações espaciais do solo, pode-se utilizar a mineração de dados, que é a principal etapa do processo de descoberta de conhecimento em banco de dados (Han & Kamber, 2006) e tem como objetivo encontrar padrões em dados armazenados nesses bancos. A tarefa de classificação (Han & Kamber, 2006) tem por objetivo inferir uma variável dependente a partir de um conjunto de dados que contêm atributos relacionados a essa variável.

Entre as várias técnicas de mineração de dados, os algoritmos de árvores de decisão são utilizados para classificação e predição das amostras desconhecidas por meio de aprendizado de máquina, ou seja, com base em registros conhecidos, cria-se um conjunto de treinamento do qual uma árvore de decisão é montada. Dessa árvore, pode-se classificar a amostra desconhecida sem necessariamente testar todos os valores dos seus atributos. A árvore de decisão consiste de uma hierarquia de nós conectados por ramos. O nó interno, também conhecido como decisório ou nó intermediário, é a unidade de tomada de decisão que avalia, por meio de teste lógico, qual será o próximo nó descendente, ou filho. Em contrapartida, um nó externo que não tem nó descendente, também conhecido como folha ou nó terminal, está associado a um rótulo ou valor. Assim, apresenta-se um conjunto de dados ao nó inicial da árvore e, dependendo do resultado do teste lógico usado pelo nó, a árvore ramifica-se para um dos nós filhos. Esse procedimento é repetido até que um nó terminal seja alcançado. A repetição desse procedimento caracteriza a recursividade da árvore de decisão (Breiman et al., 1984). O teste lógico que determina a ramificação da árvore é o coeficiente de incerteza da informação, que é baseado na entropia (Onoda, 2001). A entropia mede a pureza ou impureza dos dados: em um conjunto de dados, a entropia é uma medida da falta de homogeneidade dos dados de entrada em relação à sua classificação, ou seja, quanto menor a entropia, menor a aleatoriedade da variável objetivo.

No caso de mapeamentos, a acurácia geral obtida da análise de árvores de decisão é uma medida simples do total de concordância do mapa predito em relação ao mapa de referência, e a acurácia de classe é essa concordância apenas em relação à classe em consideração.

Embora a mineração de dados ambientais tenha sido usada na predição de variáveis individuais de solo com bom desempenho (Henderson et al., 2005), quando se trata da predição de classes de solos, espera-se desempenho inferior, pois essas classes não são caracterizadas apenas por uma variável, mas por conjuntos de variáveis de solos (Santos et al., 2006). São raros os relatos de técnicas de mineração de dados usadas no mapeamento digital de solos, no Brasil ou no exterior. Para mapeamento digital de solos na região de Toowoomba, Austrália, Bui et al. (1999) extraíram de um modelo digital de elevação (MDE) com 250 m de resolução espacial os parâmetros: declividade; aspecto; curvaturas em perfil, em planta e tangencial; e área de contribuição. Em outro estudo, Chagas (2006) utilizou redes neurais artificiais e de máxima verossimilhança como técnicas de mineração de dados para a predição de classes de solos na região de domínio de mar de morros e de alinhamentos serranos do noroeste do Estado do Rio de Janeiro, com base em conceitos de associações solo-paisagem. A comparação a pontos de observação coletados no campo mostrou que o mapa produzido por redes neurais teve concordância superior (71%) à dos produzidos pela abordagem convencional (53%) e por máxima verossimilhança (51%), o que levou os autores a concluir que a quantificação de atributos do terreno e sua classificação por redes neurais podem aumentar a confiabilidade dos mapeamentos de solos. Zhu (2000) observou que mapas de solos produzidos com dados ambientais classificados por redes neurais artificiais apresentam maior detalhe espacial e melhor qualidade que mapas convencionais.

Comparativamente às redes neurais, árvores de decisão se destacam por: estabelecerem regras de decisão que podem ser visualizadas, o que permite sua comparação às regras não explícitas usadas durante o mapeamento tradicional; permitirem conhecer o poder preditivo de cada uma das variáveis, bem como realizar ajustes para aumentar o poder de predição (Scull et al., 2003); e terem algoritmos muito mais rápidos que os das redes neurais.

O objetivo deste trabalho foi desenvolver uma metodologia para mapeamento digital de solos na escala 1:100.000 com a aplicação de técnicas de mineração de dados a descritores de relevo e a dados de mapas geológico e pedológico preexistentes.

Material e Métodos

Para a seleção da área de estudo, estabeleceram-se alguns pressupostos: selecionar quadrícula do Estado de São Paulo com mapa de solos elaborado, para que o modelo de mapeamento digital pudesse ser testado; selecionar áreas com grande variabilidade de ambientes

formadores de solo, especialmente em termos de geologia e geomorfologia; e selecionar áreas com a menor variação climática possível, pois o clima, juntamente com tempo e organismos, foram fatores de formação do solo considerados homogêneos neste estudo.

Assim, pelas informações das diferentes quadrículas do levantamento de solos do Estado de São Paulo (escala 1:100.000) e dos mapas geológico (escala 1:500.000) (Instituto de Pesquisas Tecnológicas, 1981a) e geomorfológico (escala 1:1.000.000) (Instituto de Pesquisas Tecnológicas, 1981b), chegou-se à folha topográfica Dois Córregos (SF-22-Z-B-III-3), área objeto deste estudo, que é uma das quatro cartas na escala 1:50.000 que compõem a quadrícula Brotas, de escala 1:100.000, que já possui mapa pedológico (Almeida et al., 1981). Essa folha localiza-se na região central do Estado de São Paulo, delimitada pelas coordenadas 48°30'–48°15'W e 22°15'–22°30'S, e apresenta dois tipos climáticos predominantes: Aw, tropical, com estação seca de inverno; e Cwa, tropical de altitude, com inverno seco e verão quente (Cepagri, 2008). O relevo representa o de três províncias geomorfológicas distintas: Planalto Ocidental, colinoso, que ocupa a maior parte da folha; Depressão Periférica, colinoso, com morros testemunhos e pequenas "cuestas", encontrada principalmente na porção sudeste da folha; e Cuestas Basálticas, com escarpas íngremes na sua parte frontal e declive suave em seu reverso, presentes no contato entre Planalto Ocidental e Depressão Periférica (Instituto de Pesquisas Tecnológicas, 1981b). A geologia da folha (Instituto de Pesquisas Tecnológicas, 1981a) é composta, em grande parte (52%), pela formação Itaqueri, de arenitos com intercalações de folhelhos e conglomerados; pela formação Serra Geral (28%), de basaltos, nas escarpas das cuestas e no reverso delas, onde houve dessecamento da cobertura arenítica do Itaqueri pela drenagem; pela formação Pirambóia (15%), predominante na porção sudeste da folha, de arenitos; e pelos arenitos da formação Botucatu (5%), nas escarpas das cuestas.

A construção do banco de dados, a análise desses dados e a elaboração do mapa de solos digital podem ser resumidas em 13 passos, descritos a seguir.

Passo 1 – vetorização e edição dos planos de informação de curvas de nível, pontos cotados, malha viária e rede hidrográfica da folha topográfica 1:50.000 Dois Córregos, no sistema de informação geográfica (SIG) ArcGis 9.2 (Environmental Systems Research Institute, 2004).

Passo 2 – elaboração do modelo digital de elevação (MDE), com análise e interpolação de valores das curvas de nível e dos pontos cotados no módulo Geostatistical Analyst do ArcGIS (Lark, 1999).

Passo 3 – obtenção dos parâmetros morfométricos curvatura em planta, curvatura em perfil, área de contribuição da bacia e declividade, a partir do MDE, com uso do programa Ilwis Academic (Faculty for Geo-Information Science and Earth Observation, 2001). O parâmetro de distância diagonal da drenagem foi obtido a partir de uma macro (Valeriano, 1999) no programa Idrisi Andes (Clark Labs, 2006), que utilizou como dados de entrada os limites geográficos da folha e o MDE.

Passo 4 – estabelecimento de classes discretas nos mapas de cada parâmetro gerado, com base na literatura (Quinn et al., 1991; Gallant & Wilson, 2000; Valeriano, 2003), e atribuição de nomes a cada uma dessas classes quantitativas, transformando-as em variáveis qualitativas (ou nominais), o que é uma exigência do algoritmo de classificação.

Passo 5 – digitalização, georreferenciamento e vetorização do mapa geológico em escala 1:500.000 (Instituto de Pesquisas Tecnológicas, 1981a).

Passo 6 – produção de um novo mapa pedológico a partir do mapa pedológico da quadrícula Brotas já vetorizado, com atualização da legenda de acordo com o Sistema Brasileiro de Classificação de Solos (Santos et al., 2006) no terceiro nível categórico e unificação de polígonos do mapa de solos original da área.

Passo 7 – cruzamento dos mapas de parâmetros morfométricos, de geologia e de solos, e formação de uma matriz de dados nominais onde cada linha corresponde a um pixel.

Passo 8 – retirada das inconsistências da matriz de dados e obtenção do pré-processamento dos dados em 794.273 linhas.

Passo 9 – análise por meio de árvores de decisão (Han & Kamber, 2006), com a retirada aleatória de 10% dos dados da matriz para validação do modelo, enquanto que o modelo de aprendizado de máquina era gerado com os 90% dos dados restantes.

Passo 10 – teste para avaliar o poder preditivo da metodologia nos 10% dos dados retirados no início, com base na acurácia geral, na acurácia de cada unidade de mapeamento e no índice kappa (Jensen, 1996).

Passo 11 – ordenamento dos atributos pelo ganho de informação com base na entropia (H), que caracteriza a pureza/impureza dos dados (Onoda, 2001):
$H(x, y) = - \Sigma p_{ij} \ln p_{ij}$, em que p é a probabilidade de ocorrência de uma unidade de mapeamento de solo, ln é o logaritmo natural e i e j são valores numéricos das unidades de mapeamento reais (x) e preditas (y). Para o ordenamento dos atributos e para a geração do modelo de predição de solos, foram utilizados os algoritmos do software Weka, versão 3.4.4 (Witten & Frank, 2005), que contém uma coleção de algoritmos de aprendizado de máquina usados em problemas de mineração de dados e descoberta do conhecimento.

Passo 12 – realização da pré-poda da árvore, para eliminar determinadas regras que poderiam não contribuir com o modelo gerado (Batista, 2003). As podas foram realizadas com 20, 50 e 100 pixels, sendo estes os números mínimos de pixels necessários para que uma unidade de mapeamento fosse estabelecida.

Passo 13 – transcrição das regras geradas pela árvore de decisão para o SIG Ilwis com uso de ferramentas de modelagem cartográfica, o que permitiu a obtenção do mapa digital de solos da folha Dois Córregos.

Determinaram-se também a acurácia geral, que é computada pelo somatório da diagonal principal (acertos) dividido pelo total de pixels da matriz de erros, e a acurácia de classe, também chamada de precisão da classe, que é avaliada dividindo-se o número de pixels atribuídos corretamente à classe pelo número total de pixels da categoria nos dados de referência (Jensen, 1996). Na determinação das acurácias geral e de classes do modelo, foi utilizada a técnica de balanceamento de classes, que tem por finalidade aumentar a proporção de amostragem nas classes com menor área de ocorrência e reduzir a proporção nas unidades com maior área, ou seja, elevar a representatividade das classes com menor representatividade e diminuir a das classes com maior representatividade (Batista, 2003). Os balanceamentos das classes utilizados foram de: 0, que representa os dados brutos, sem balanceamento de classes; 0,5, em que as classes são balanceadas de modo intermediário entre zero e um; e 1, situação na qual todas as classes apresentam a mesma distribuição na folha.

A ordem de influência de variáveis relacionadas a fatores de formação do solo (formação geológica, distância diagonal da drenagem, declividade, curvatura em perfil, curvatura em planta, área de contribuição da bacia) foi determinada pela análise de entropia (passo 11) (Onoda, 2001).

Resultados e Discussão

Após a simplificação da legenda, os solos da folha Dois Córregos ficaram distribuídos em quatro ordens do Sistema Brasileiro de Classificação de Solos (Santos et al., 2006): Latossolos, Argissolos, Nitossolos e Neossolos (Tabela 1). De acordo com a folha, os Latossolos ocupam a maior parte da área (64,3%), com destaque para o Latossolo Vermelho-Amarelo distrófico textura média (40,5%). Os Argissolos Vermelho-Amarelos também estão presentes em grande parte da folha (21,6%), e os Neossolos (5,5%) e Nitossolos (8,2%) têm menor expressão na área.

Tabela 1. Proporção de ocorrência e acurácia (%) individual com o balanceamento de classes das unidades de mapeamento na folha Dois Córregos, após simplificação da legenda.

Unidades de mapeamento[(1)]	Área	Balanceamento de classes		
	(%)	0	0,5	1
LVAd textura média	40,5	63,0	63,4	68,2
LVd textura argilosa	3,20	0,0	12,3	3,8
LVd textura média	12,3	18,6	18,8	21,2
LVdf textura argilosa ou muito argilosa	3,90	0,0	8,3	11,6
LVef textura argilosa ou muito argilosa	4,45	20,1	20,2	18,2
NVe ou NVd textura argilosa ou muito argilosa	8,15	31,7	34,3	37,5
PVAd textura média ou arenosa/média	3,70	0,0	13,3	8,2
PVAe textura arenosa/média ou média/argilosa	17,90	63,0	65,3	65,0
RLe ou RLd textura média	4,47	57,9	54,5	52,4
RQd	1,00	34,5	18,9	8,9
Urbana	0,43	-	-	-

[(1)] LVAd, Latossolo Vermelho-Amarelo distrófico; LVd, Latossolo Vermelho distrófico; LVdf, Latossolo Vermelho distroférrico; LVef, Latossolo Vermelho eutroférrico; NVe, Nitossolo Vermelho eutrófico; NVd, Nitossolo Vermelho distrófico; PVAd, Argissolo Vermelho-Amarelo distrófico; PVAe, Argissolo Vermelho-Amarelo distrófico; RLe, Neossolo Litólico eutrófico; RLd, Neossolo Litólico distrófico; RQd, Neossolo Quartzarênico distrófico.

A acurácia das predições por árvores de decisão, nos diferentes balanceamentos de classes, estão apresentadas nas Tabelas 1 e 2. Observa-se que três das cinco unidades de mapeamento de menor extensão no mapa original – Latossolo Vermelho distrófico textura argilosa, Latossolo Vermelho distroférrico textura argilosa ou muito argilosa e Argissolo Vermelho-Amarelo distrófico textura média ou arenosa/média – não foram preditas no modelo quando o balanceamento de classes foi zero. Com balanceamento de classes 0,5, que aumenta a amostragem nas classes pouco extensas e diminui a amostragem nas classes já abundantemente amostradas, essas três unidades são preditas com acurácia em torno de 10%. Em contrapartida, os Neossolos Quartzarênicos, com acurácia de 34,5% no balanceamento de classes zero, têm a acurácia reduzida no balanceamento 0,5. Essa unidade de mapeamento é a de menor área na folha (1%) e, ao contrário das três unidades citadas anteriormente, foi predita com acurácia relativamente elevada, em relação à sua extensão, no balanceamento zero (34,5%). No mapa original, observa-se que a localização do único polígono de Neossolo Quartzarênico é concentrada no limite norte da folha (Figura 1 A). Imagina-se que essa situação, aliada ao sistema de amostragem do algoritmo para treinamento do modelo, seja a causa desse contraste de acurácia em relação às outras classes de pequena extensão, com menor acurácia no balanceamento zero.

As unidades de mapeamento com maior distribuição na folha, como Latossolo Vermelho-Amarelo distrófico textura média e Argissolo Vermelho-Amarelo eutrófico, quando foram subamostradas no treinamento do modelo (balanceamentos de classes 0,5 e 1), não tiveram sua acurácia reduzida. Isso se deve à grande representatividade dessas unidades na folha, o que minimiza efeitos de redução de amostragem. O elevado número de exemplos para treinamento da árvore, superior a 714.000 (90% do total de registros), certamente contribuiu para que uma redução na amostragem das classes mais representativas não alterasse a acurácia dessas classes. À exceção dos Neossolos Quartzarênicos (RQd) e dos Neossolos Litólicos (RLe ou RLd), o balanceamento de classes 0,5 apresentou maior acurácia para as classes individualmente que o balanceamento nulo. Essa não alteração da acurácia com o aumento do balanceamento de classes ocorre em casos de treinamento em grande número de registros, em razão do elevado número de exemplos para treinamento (Batista, 2003). Apesar de o balanceamento 1 ter proporcionado resultados semelhantes entre as classes mais representativas individualmente, a acurácia geral do modelo neste balanceamento reduziu-se de 53 para 36% (Tabela 2). Dessa forma, o balanceamento de classes 0,5 foi o que proporcionou maior acurácia ao mapeamento dos solos, considerando-se o conjunto das unidades de mapeamento.

Tabela 2. Efeito do balanceamento de classes na acurácia (%) geral do modelo gerado para a folha Dois Córregos.

Parâmetro do modelo	Balanceamento de classes		
	0	0,5	1
Regras geradas	172	294	418
Acurácia geral	54,24	53	36,13
Coeficiente kappa	0,37	0,36	0,25

As diferenças obtidas na acurácia do mapa predito sugerem que o balanceamento de classes pode se adequar à finalidade do mapa. Dessa forma, em estudos para estimativa apenas da distribuição das unidades de mapeamento com maior extensão, não se aplica balanceamento de classes e, em estudos em que a distribuição de solos com baixa representatividade também é importante, aplica-se balanceamento próximo a 0,5.

Ao eliminarem-se 11% dos dados das três unidades de mapeamento com probabilidade nula de ocorrência (0% de acurácia geral) no balanceamento de classes zero (Tabela 1), houve ganho de 6% sobre a acurácia geral (Tabela 2), tomando-se a média dos três balanceamentos de classes. A maior acurácia (61%) ocorreu no balanceamento zero seguida do 0,5 (59%) e 1 (43%). Com a eliminação das três unidades, houve ganho também na acurácia das classes individualmente (Tabela 3), com exceção das unidades LVd textura média e RQd. Esse aumento da acurácia geral deve-se a uma redistribuição da amostragem de treinamento de modo mais homogêneo entre as demais unidades. Nesse caso, a mudança no balanceamento de classes para treinamento de 0 para 0,5 praticamente não altera a acurácia, enquanto a mudança no balanceamento de 0 para 1 reduz 18% da acurácia de predição do modelo, o que ocorre por aumento de classificações incorretas por subamostragem de classes de maior área (Batista, 2003).

O número de regras gerado pelo modelo foi elevado mesmo após eliminação das unidades com acurácia zero (98 regras) (Tabela 4). Para diminuir o número de regras e aumentar a acurácia das estimativas de erro, que são avaliadas em amostras cada vez menores

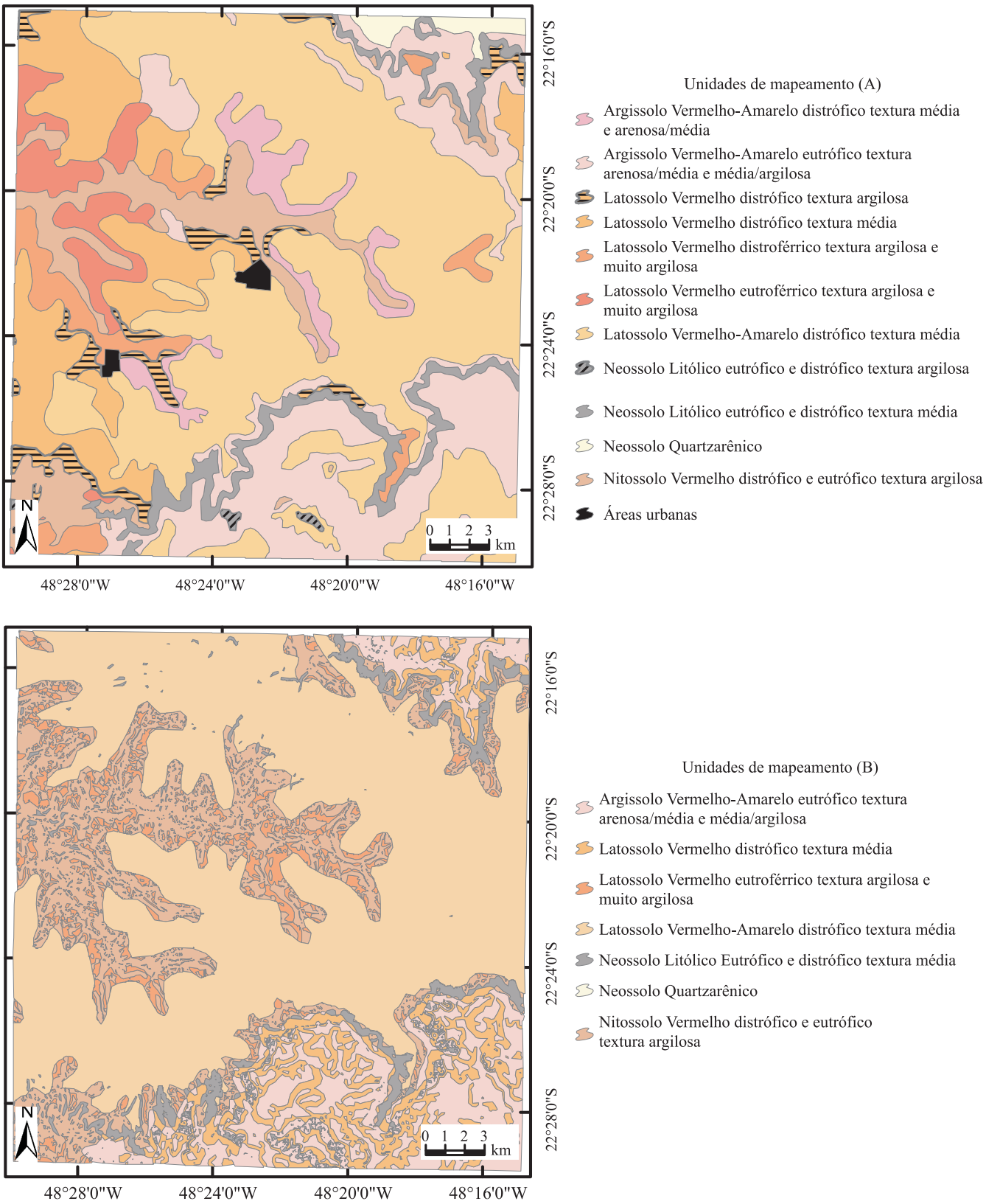

Figura 1. Mapa de solos da folha Dois Córregos obtido por (A) métodos tradicionais (Almeida et al., 1981) após simplificação e unificação de legenda e (B) aplicação das regras desenvolvidas na análise por árvores de decisão.

com o crescimento da árvore, realizou-se a pré-poda da árvore (Batista, 2003). As classes de pré-poda aplicadas (20, 50 e 100) representam o número mínimo de pixels que as regras devem considerar para definir uma unidade de mapeamento de solo (folha da árvore). A acurácia do modelo não diferiu entre as três classes de pré-poda, mantendo-se em 61% (Tabela 5), igual à acurácia obtida sem aplicação de pré-poda. A pré-poda 100 contribuiu para a diminuição do número de regras, de 98 para 86. Isso evidencia que há combinações de variáveis (regras) geomorfométricas e geológicas pouco relevantes na inferência das unidades de mapeamento de solos (Batista, 2003) e, assim, podem ser descartadas sem prejuízo da precisão cartográfica do mapa digital pedológico. Considerando a resolução de 30 m da base de dados utilizada e a área mínima mapeável como 0,6x0,6 cm na superfície representada (Santos et al., 1995), conclui-se que a classe de pré-poda 100 representa a área mínima mapeável na escala 1:50.000 (9 ha). Assim, a manutenção da acurácia, mesmo na pré-poda 100, mostra que ainda é possível aumentá-la (maior número de pixels) sem prejuízo da precisão cartográfica do mapa digital e com maior redução do número de regras de classificação. Teoricamente, poder-se-ia aumentar o número de pré-poda para 400 pixels, valor correspondente à área mínima mapeável na escala 1:100.000 (36 ha), escala de publicação original, sem prejuízo para a precisão cartográfica do mapa digital. Isso está diretamente relacionado com a acurácia do mapa original (Drohan et al., 2003) mas, ainda assim, indica que a pré-poda da árvore de decisão, em modelos de predição de mapas pedológicos, pode orientar-se pelo nível de detalhe do mapa de treinamento.

Tabela 3. Acurácia (%) individual das unidades de mapeamento de solos, retiradas aquelas com probabilidade nula de ocorrência no balanceamento de classes igual a zero.

Unidade de mapeamento[(1)]	Balanceamento de classes		
	0	0,5	1
LVd textura média	14,3	0,0	23,4
LVAd textura média	68,8	68,9	73,2
PVAe textura arenosa/média ou média/argilosa	63,7	65,0	65,1
RLe ou RLd textura média	62,3	56,0	55,6
RQd	61,1	17,6	10,9
NVd ou NVe textura argilosa	41,3	46,0	46,1
LVef textura argilosa ou muito argilosa	27,3	26,2	23,9

[(1)]LVd, Latossolo Vermelho distrófico; LVAd, Latossolo Vermelho-Amarelo distrófico; PVAe, Argissolo Vermelho-Amarelo distrófico; RLe, Neossolo Litólico eutrófico; RLd, Neossolo Litólico distrófico; RQd, Neossolo Quartzarênico distrófico; NVd, Nitossolo Vermelho distrófico; NVe, Nitossolo Vermelho eutrófico; LVef, Latossolo Vermelho eutroférrico.

A ordem de influência das variáveis na determinação das unidades de mapeamento de solos, em função da sua entropia (Onoda, 2001), foi a seguinte: formação geológica > distância diagonal da drenagem > declividade > curvatura em perfil > curvatura em planta > área de contribuição da bacia. Acredita-se que a importância da geologia como elemento preditivo das unidades de mapeamento de solos na folha Dois Córregos esteja associada à diferenciação da ordem de solos mais extensa que ocorre na folha (Latossolos), predominantemente por grupamento de textura de solo. Isso, conjugado à diferenciação das formações geológicas locais (arenitos e basalto) por fatores determinantes dos constituintes granulométricos do material de sua alteração, contribui para conferir maior poder preditivo da variável formação geológica. As variáveis distância da drenagem e declividade, por sua vez, são associadas ao acúmulo de água e aos fluxos hídricos, que têm grande influência na formação dos solos. Isso explica o ordenamento delas como importantes variáveis preditivas dos solos.

A possibilidade de ordenamento das variáveis em termos da sua contribuição preditiva dos solos é uma informação original fornecida pela análise por árvores de decisão (Batista, 2003). Por exemplo, apesar de ser amplamente aceito que o material de origem, aqui representado pela formação geológica, é um fator fundamental de formação dos solos (Buol et al., 2003),

Tabela 4. Regras geradas e índices de acurácia nos diversos balanceamentos de classes, depois da retirada das unidades de mapeamento de solos com probabilidade nula de ocorrência no modelo inicial da folha Dois Córregos.

Parâmetro do modelo	Balanceamento de classes		
	0	0,5	1
Regras geradas	98	156	214
Acurácia geral do modelo (%)	60,88	58,77	43,0
Coeficiente kappa	0,43	0,41	0,3

Tabela 5. Diferentes classes de pré-poda da árvore de decisão, aplicadas aos dados de melhor acurácia geral do modelo.

Poda da árvore de decisão	Nº de regras geradas	Acurácia (%)
20 pixels	98	60,75
50 pixels	92	60,75
100 pixels	86	60,75

a ordenação da contribuição desse fator em relação a outros fatores analisados em determinado local ou região não é conhecida, muito menos mostrada de forma quantitativa, como é feito pela análise da entropia (Onoda, 2001).

O coeficiente kappa obtido no mapa da folha Dois Córregos foi relativamente baixo: 0,43 (Tabela 4), o que indica uma concordância moderada (Landis & Koch, 1977) dos resultados preditos com os observados. Chagas (2006) obteve índices kappa superiores a 0,80 e acurácias global e de classe superiores a 80%, em média, com diferentes arquiteturas de redes neurais testadas para o mapeamento de solos em região do domínio de mares de morros e geologia do Pré-Cambriano no Estado do Rio de Janeiro. Ao aplicar árvores de decisão a um conjunto de variáveis semelhantes ao utilizado no presente trabalho, Bui et al. (1999) obtiveram coeficiente kappa para as classes individuais de 0,23 a 0,89, e coeficiente kappa geral do mapa de 0,64, indicando concordância substancial dos resultados do modelo, segundo os referenciais de Landis & Koch (1977). Bui et al. (1999) também obtiveram coeficiente de incerteza de 0,48 para o mapa predito, e concordância entre mapa predito e original de 69%. Acredita-se que a principal causa da baixa acurácia obtida no presente trabalho tenha sido a simplificação da legenda do mapa original, que reuniu, em uma mesma classe, domínios de relevo distintos. Pode-se também citar outras possíveis causas, como: não inclusão de outras variáveis determinantes da variabilidade dos solos do local; baixa associação dos parâmetros de relevo escolhidos com os solos; problemas de precisão cartográfica ou de exatidão taxonômica do mapa pedológico original; ou problemas de precisão ou de exatidão na base de dados de relevo.

Na comparação do mapa digital de solos elaborado com as regras geradas pela árvore de decisão (Figura 1 B) ao mapa de solos de Almeida et al. (1981) com legenda simplificada (Figura 1 A), a principal diferença foi a maior fragmentação das unidades de mapeamento do mapa digital. Há considerável descontinuidade dos polígonos mapeados no mapa digital que, associada com a eliminação ou deslocamento espacial de algumas unidades de mapeamento, mostra nítida diferença entre mapa original e predito. Isso reflete o baixo poder de predição do modelo, conforme já mostrado pelos índices de acurácia. Essa descontinuidade pode resultar da metodologia empregada, que não considera as relações de vizinhança entre os pixels. Contudo, a incompatibilidade entre menor detalhe do mapa pedológico (original em escala 1:100.000), simplificado e com legenda unificada, e o maior detalhe da carta topográfica (escala 1:50.000) podem fazer com que possíveis variações de solo, eventualmente associadas ao relevo, não sejam representadas no mapa pedológico original. Há, assim, a possibilidade de que essa fragmentação seja parcialmente verídica, ou seja, de que as diferenciações no relevo da folha se reflitam em diferenciações dos solos na paisagem, não identificadas no mapeamento original devido à sua pequena escala. A comprovação dessa hipótese, mediante identificação e amostragem dos solos em campo em desenho amostral específico, deve ser testada em outros trabalhos.

Conclusões

1. A análise por árvores de decisão, associada a sistemas de informações geográficas, permite a elaboração de mapas digitais que representam aproximações dos mapas de solos elaborados por métodos tradicionais.

2. O balanceamento de classes no treinamento da árvore permite adequar o mapa predito à sua finalidade, de forma a predizer maior número de unidades de mapeamento ou apenas as mais representativas.

3. Para redução do número de regras geradas pela árvore de decisão sem prejuízo da acurácia do mapa predito, a pré-poda pode se orientar pela área mínima mapeável do mapa original, considerando sua escala de publicação.

4. A ordenação das variáveis por importância decrescente na predição das unidades de mapeamento de solos foi formação geológica > distância diagonal de drenagem > declividade > curvatura do perfil > curvatura da planta > área de contribuição.

5. O mapeamento digital produz elevada fragmentação dos polígonos de classes de solo, inexistente no mapa original.

Referências

ALMEIDA, C.L.F. de; OLIVEIRA, J.B. de; PRADO, H. do. **Levantamento pedológico semidetalhado do Estado de São Paulo**: quadrícula de Brotas. I. Mapas de solos. Campinas: Instituto Agronômico, 1981. Mapa. Escala 1:100.000.

BATISTA, G.E. de A.P.A. **Pré-processamento de dados em aprendizado de máquinas supervisionado**. 2003. 204p. Tese (Doutorado) - Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos.

BREIMAN, L.; FRIEDMAN, J.H.; OLSHEN, R.A.; STONE, C.J. **Classification and regression trees**. Monterey: Wadsworth and Brooks, 1984. 358p.

BUI, E.N.; LOUGHHEAD, A.; CORNER, R. Extracting soil-landscape rules from previous soil surveys. **Australian Journal of Soil Research**, v.37, p.495-508, 1999.

BUOL, S.W.; SOUTHARD, R.J.; GRAHAM, R.C.; MCDANIEL, P.A. **Soil genesis and classification**. 5th ed. Ames: Iowa State Press, 2003. 494p.

CHAGAS, C.S. **Mapeamento digital de solos por correlação ambiental e redes neurais em uma bacia hidrográfica no domínio de mar de morros**. 2006. 223p. Tese (Doutorado) - Universidade Federal de Viçosa, Viçosa.

CLARK LABS. **IDRISI Andes**. Worcester: Clark Labs, 2006. 327p.

DROHAN, P.J.; CIOLKOSZ, E.J.; PETERSEN, G.W. Soil survey mapping unit accuracy in forested field plots in northern Pennsylvania. **Soil Science Society of America Journal**, v.67, p.208-214, 2003.

ENVIRONMENTAL SYSTEMS RESEARCH INSTITUTE. **ArcGIS 9**: getting started with ArcGIS. Redlands: Esri, 2004. 272p.

FACULTY FOR GEO-INFORMATION SCIENCE AND EARTH OBSERVATION. **ILWIS 3.3**: user's guide. Enschede: ITC, 2001. 530p.

GALLANT, J.C.; WILSON, J.P. Primary topographic attributes. In: WILSON, J.P.; GALLANT, J.C. (Ed.). **Terrain analysis**: principles and applications. New York: John Wiley, 2000. p.51-85.

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

HENDERSON, B.L.; BUI, E.N.; MORAN, C.J.; SIMON, D.A.P. Australia-wide predictions of soil properties using decision trees. **Geoderma**, v.124, p.383-398, 2005.

INSTITUTO DE PESQUISAS TECNOLÓGICAS. **Mapa geológico do Estado de São Paulo**. São Paulo: Instituto de Pesquisas Tecnológicas, 1981a. v.1: nota explicativa; v.2: mapas. (IPT. Série monografias, 6).

INSTITUTO DE PESQUISAS TECNOLÓGICAS. **Mapa geomorfológico do Estado de São Paulo**. São Paulo: Instituto de Pesquisas Tecnológicas, 1981b. v.1: nota explicativa; v.2: mapas. (IPT. Série monografias, 5).

JENSEN, J.R. **Introductory digital image processing**: a remote sensing perspective. 2nd ed. Upper Saddle River: Prentice Hall, 1996. 318p.

LAGACHERIE, P.; MCBRATNEY, A.B. Spatial soil information systems and spatial soil inference systems: perspectives for digital soil mapping. In: LAGACHERIE, P.; MCBRATNEY, A.B.; VOLTZ, M. (Ed.). **Digital soil mapping**: an introductory perspective. Amsterdam: Elsevier, 2007. p.3-24.

LAGACHERIE, P.; VOLTZ, M. Predicting soil properties over a region using sample information from a mapped reference area and digital elevation data: a conditional probability approach. **Geoderma**, v.97, p.187-208, 2000.

LANDIS, J.R.; KOCH, G.G. The measurement of observer agreement for categorical data. **Biometrics**, v.33, p.159-174, 1977.

LARK, R.M. Soil-landform relationships at within-field scales: an investigation using continuous classification. **Geoderma**, v.92, p.141-165, 1999.

MCBRATNEY, A.B.; MENDONÇA-SANTOS, M. de L.; MINASNY, B. On digital soil mapping. **Geoderma**, v.117, p.3-52, 2003.

MENDONÇA-SANTOS, M. de L.; SANTOS, H.G.; DART, R.O.; PARES, J.G. Digital mapping of soil classes in Rio de Janeiro State, Brazil: data, modelling and prediction. In: HARTEMINK, A.E.; MCBRATNEY, A.; MENDONÇA-SANTOS, M. de L. (Org.). **Digital soil mapping with limited data**. Amsterdam: Elsevier, 2008. p.381-398.

ONODA, M. **Estudo sobre um algoritmo de árvores de decisão acoplado a um sistema de banco de dados relacional**. 2001. 110p. Dissertação (Mestrado) - Universidade Federal do Rio de Janeiro, Rio de Janeiro.

QUINN, P.F.; BEVEN, K.J.; CHEVALLIER, P.; PLANCHON, O. The prediction of hillslope flow paths for distributed hydrological modeling using digital terrain models. **Hydrological Processes**, v.5, p.59-79, 1991.

SANTOS, H.G. dos; HOCHMÜLLER, D.P.; CAVALVANTI, A.C.; RÊGO, R.S.; KER, J.C.; PANOSO, L.A.; AMARAL, J.A.M. do. **Procedimentos normativos de levantamentos pedológicos**. Brasília: Embrapa-SPI, 1995. 116p.

SANTOS, H.G. dos; JACOMINE, P.K.T.; ANJOS, L.H.C. dos; OLIVEIRA, V.A. de; OLIVEIRA, J.B. de; COELHO, M.R.; LUMBRERAS, J.F.; CUNHA, T.J.F. (Ed.). **Sistema brasileiro de classificação de solos**. 2.ed. Rio de Janeiro: Embrapa Solos, 2006. 316p.

SCULL, P.; FRANKLIN, J.; CHADWICK, O.A.; MCARTHUR, D. Predictive soil mapping: a review. **Progress in Physical Geography**, v.27, p.171-197, 2003.

VALERIANO, M. de M. Curvatura vertical de vertentes em microbacias pela análise de modelos digitais de elevação. **Revista Brasileira de Engenharia Agrícola e Ambiental**, v.7, p.539-546, 2003.

VALERIANO, M. de M. **Estimativa de variáveis topográficas para modelagem da perda de solos por geoprocessamento**. 1999. 172p. Tese (Doutorado) - Universidade Estadual Paulista Júlio de Mesquita Filho, Rio Claro.

WITTEN, I.H.; FRANK, E. **Data mining**: practical machine learning tools and techniques. 2nd ed. San Francisco: Morgan Kaufmann, 2005. 525p.

ZHU, A.X. Mapping soil landscape as spatial continua: the neural network approach. **Water Resources Research**, v.36, p.663-677, 2000.

Recebido em 20 de março de 2009 e aprovado em 21 de novembro de 2009