DIVISÃO 1 - SOLO NO ESPAÇO E NO TEMPO

Comissão 1.3 - Pedometria

COMPARAÇÃO DE ESQUEMAS DE AMOSTRAGEM PARA TREINAMENTO DE MODELOS PREDITORES NO MAPEAMENTO DIGITAL DE CLASSES DE SOLOS

Rodrigo Teske[1], Elvio Giasson[2]* e Tatiane Bagatini[1]

[1] Universidade Federal do Rio Grande do Sul, Faculdade de Agronomia, Programa de Pós-graduação em Ciência do Solo, Porto Alegre, Rio Grande do Sul, Brasil.

[2] Universidade Federal do Rio Grande do Sul, Faculdade de Agronomia, Departamento de Solos, Porto Alegre, Rio Grande do Sul, Brasil.

* Autor correspondente.
E-mail: giasson@ufrgs.br

RESUMO

Os modelos preditores usados no mapeamento digital de solos (MDS) precisam ser treinados com dados que captem ao máximo a variação dos atributos do terreno e dos solos, a fim de gerar correlações adequadas entre as variáveis ambientais e a ocorrência dos solos. Para avaliar a acurácia desses modelos, tem sido constatado o uso de diferentes métodos de avaliação da acurácia no MDS. Os objetivos deste estudo foram comparar o uso de três esquemas de amostragem para treinar algoritmo de árvore de classificação (CART) e avaliar a capacidade de predição dos modelos gerados por meio de quatro métodos. Foram utilizados os esquemas de amostragem: aleatório simples; proporcional à área de cada unidade de mapeamento de solos (UM); e estratificado pelo número de UM. Os métodos de avaliação testados foram: aparente, divisão percentual, validação cruzada com 10 subconjuntos e reamostragem com sete conjuntos de dados independentes. As acurácias dos modelos estimadas pelos métodos foram comparadas com as acurácias mensuradas obtidas pela comparação dos mapas gerados, a partir de cada esquema de amostragem, com o mapa convencional de solos na escala 1:50.000. Os esquemas de amostragem influenciaram na quantidade de UMs preditas e na acurácia dos modelos e dos mapas gerados. Os esquemas de amostragem proporcional e estratificada resultaram mapas digitais menos acurados, e a acurácia dos modelos variou conforme o método de avaliação empregado. A amostragem aleatória resultou no mapa digital mais acurado e apresentou valores da acurácia semelhantes para todos os métodos de avaliação testados.

Recebido para publicação em 13 de junho de 2014 e aprovado em 24 de setembro de 2014.
DOI: 10.1590/01000683rbcs20150344

Palavras-chave: amostragem proporcional, amostragem aleatória, amostragem estratificada, avaliação da acurácia, árvore de classificação.

ABSTRACT: *COMPARISON OF SAMPLING PROCEDURES FOR TRAINING PREDICTIVE MODELS IN DIGITAL SOIL CLASS MAPPING*

The predictive models used in digital soil mapping (DSM) need to be trained with data that most fully capture the variation of terrain and soil properties in order to generate adequate correlations between environmental variables and the occurrence of soil unities. Several methods have been used in DSM to evaluate the accuracy of these models. The aims of this study were to compare the use of three sampling procedures for training a classification and regression tree (CART) algorithm, and evaluate the predictive capacity of the models generated using four methods. The sampling procedures used were: simple random; proportional to the area of each soil mapping unit (MU), and stratified by the number of MUs. The evaluation methods tested were: apparent, percentage division, cross-validation with 10 subsets, and resampling with seven independent data sets. The accuracies of the models estimated by the methods were compared with the measured accuracies. This was achieved by comparing the maps generated, based on each sampling procedure, with the conventional soil map at the scale of 1:50,000. The sampling procedures influenced the number of MUs predicted and the accuracy of the models and of the maps generated. The proportional and stratified sampling procedures resulted in less accurate digital soil maps, and the accuracies of the models varied according to the evaluation method adopted. Random sampling resulted in the most accurate digital soil map and presented accuracy values that were similar for all the evaluation methods tested.

Keywords: proportional sampling, random sampling, stratified sampling, accuracy evaluation, classification tree.

INTRODUÇÃO

O mapeamento digital de solo (MDS) visa correlacionar os solos com os atributos do terreno de forma mais quantitativa que os levantamentos convencionais e utilizando modelos numéricos ou estatísticos para inferir as variações espaciais dos solos (Lagacherie e McBratney, 2007). Para isso, os modelos precisam ser treinados e validados com dados que capturem ao máximo a variação espacial dos atributos do terreno e dos solos, e é necessário o uso de estratégias de amostragem estatisticamente robustas para diminuir os erros da predição (Minasny e McBratney, 2007; Brungard e Boetinger, 2010).

Segundo Brus e Gruijter (1997), quando realizada a predição de ocorrência de classes ou propriedades de solo por correlação ambiental entre mapas auxiliares (variáveis preditoras) e mapas de referência de solo, a amostragem deve ser fundamentada na teoria da probabilidade. Assim, a amostragem aleatória tem sido comumente utilizada no MDS para amostrar os dados para treinamento, porque elimina a subjetividade e permite a reprodutibilidade simples (Hengl et al., 2003). Todavia, alguns autores têm relatado que classes e UMs pouco extensas e pouco representativas de uma área não são preditas pelos modelos, quando utilizada a amostragem aleatória (Giasson et al., 2011; Ten Caten et al., 2012). Em trabalho realizado por Grinand et al. (2008), os autores utilizaram a amostragem proporcional à área de cada UM, como recomendado por Moran e Bui (2002); dessa forma, as UMs menos representativas não foram subamostradas, como pode ocorrer na amostragem aleatória simples. A amostragem estratificada também tem sido recomendada para amostrar as UMs pouco representativas (Hengl et al., 2003; Stehman, 2008).

Os modelos de predição mais acurados possibilitam gerar mapas digitais de solo mais precisos e, semelhante aos mapeamentos convencionais de solo, os mapas gerados também apresentam erros que devem ser identificados e quantificados (McBratney et al., 2003; Brus et al., 2011). Para isso, Rossiter (2004) abordou diferentes métodos para obter a acurácia de mapas temáticos, sendo a comparação do mapa gerado com um mapa de referência um método comumente empregado no MDS para medir a acurácia, como realizado por Bui e Moran (2003) e Giasson et al. (2011). Além disso, as estimativas da acurácia podem ser obtidas por métodos avaliação de modelos preditores, que fornecem ao usuário estimativa da acurácia e dos erros da predição ao final da construção do modelo (Chatfield, 1995; Steyerberg, 2009) e em etapa anterior à geração dos mapas digitais (Brus et al., 2011).

A avaliação da acurácia com uso de dados independentes e distintos daqueles usados para o treinamento é indicada para estudos preditivos de diversas áreas do conhecimento (Grinand et al., 2008;

Steyerberg, 2009; Brus et al., 2011). Todavia, no MDS tem sido encontrado o uso de diferentes métodos de avaliação de modelos preditores (Grunwald, 2009). Entre os métodos de avaliação de modelos disponíveis no programa Weka (Hall et al., 2009), que tem sido comumente utilizado em trabalhos com MDS no Brasil, estão: a validação cruzada e a validação com divisão percentual, que repartem o conjunto de dados inicialmente amostrados para treinamento em dois subconjuntos, sendo um para treinamento e outro para avaliação; o método de validação aparente, que utiliza todo o conjunto de dados usados no treinamento para a avaliação; e o método de validação com dados independentes e distintos daqueles usados para treinamento.

Nesse sentido, tanto o esquema de amostragem dos dados para treinamento como o método de estimativa da acurácia usado podem influenciar na seleção de modelos preditores de ocorrência de classes de solos. Os objetivos deste estudo foram avaliar e comparar os resultados das predições de ocorrência de classes de solos geradas por árvore de classificação com dados oriundos de três esquemas de amostragem.

MATERIAL E MÉTODOS

O estudo foi realizado na bacia do Rio Santo Cristo, região noroeste do Estado do Rio Grande do Sul, com área de 898 km². O clima da região é subtropical úmido, tipo Cfa de Köppen, e o material de origem da região é basalto da Formação Serra Geral. O mapa de solos utilizado como referência, na escala de 1:50.000 (Figura 1a), cujas unidades de mapeamento são descritas no quadro 1, faz parte do Levantamento Pedológico e Análise Qualitativa do Potencial de Uso dos Solos para o Descarte de Dejetos Suínos da Microbacia do Rio Santo Cristo (Kämpf et al., 2004).

Os atributos do terreno usados para caracterizar a paisagem foram gerados em ambiente de sistemas de informação geográfica (SIG), com o programa ArcGis 9.3 (ESRI, 2009). Foram gerados seis atributos do terreno (declividade, direção do fluxo, acúmulo do fluxo, comprimento do fluxo, curvatura e índice de umidade topográfica), a partir do modelo digital de elevação (MDE) ASTER-GDEM v2 com resolução espacial de 30 m (Meyer et al., 2012). A partir do arquivo vetorial de hidrografia da base contínua do Rio Grande do Sul (Hasenack e Weber, 2010), foi gerada a variável distância dos rios. Todos os mapas foram rasterizados com resolução espacial de 30 m.

As correlações entre os atributos do terreno e a distribuição espacial dos solos foram geradas por árvore de classificação (Breiman et al., 1984) com o algoritmo *SimpleCart* implementado no

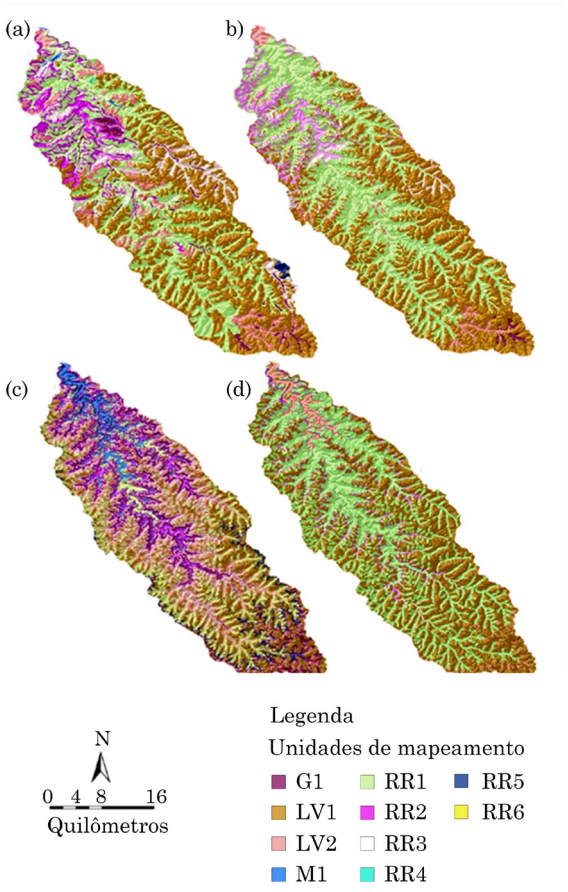

Figura 1. (a) Mapa convencional de solos da bacia do Rio Santo Cristo (Kämpf et al., 2004); (b) mapa digital de solos com amostragem aleatória simples; (c) mapa digital de solos com amostragem estratificada); e (d) mapa digital de solos com amostragem proporcional à área de cada unidade de mapeamento.

programa de mineração de dados Weka 3.6.3 (Hall et al., 2009). As árvores de classificação foram utilizadas porque são robustas e permitem o uso de variáveis preditoras tanto nominais como numéricas, além de usar os dados em qualquer escala e com valores discrepantes (Scull et al., 2003; Witten et al., 2011). Para proceder ao treinamento dos modelos preditores, as informações dos atributos do terreno e das unidades de mapeamento (UM) foram amostradas em 45000 pontos (1 ponto a cada 2 ha), usando três esquemas de amostragem: aleatória simples; aleatória proporcional à área ocupada por cada UM, em que os pontos amostrais foram distribuídos aleatoriamente sobre o mapa obedecendo o critério da proporcionalidade; e aleatória estratificada pelo número de UM, em que a mesma quantidade de pontos amostrais (4500) foi distribuída aleatoriamente dentro da área de cada uma das 10 UMs.

Quadro 1. Unidades de mapeamento de solos ocorrentes na bacia do Rio Santo Cristo, RS

UM[1]	Descrição taxonômica[2]	Proporção dos componentes	Inclusão	Área
		%		%
G1	Gleissolo Háplico	-		2,50
LV1	Latossolo Vermelho distroférrico	-	RR, CX	38,30
LV2	Associação Latossolo Vermelho + Neossolo Regolítico	60 e 40	CX	7,90
M1	Chernossolo Háplico	-	CX	0,20
RR1	Associação Neossolo Regolítico + Cambissolo Háplico	60 e 40	RL, CX	34,80
RR2	Complexo Neossolo Regolítico + Neossolo Litólico	50 e 50	CX	8,10
RR3	Associação Neossolo Regolítico + Latossolo Vermelho	60 e 40	CX	7,90
RR4	Associação Neossolo Regolítico + Neossolo Litólico	70 e 30		0,03
RR5	Associação Neossolo Regolítico + Cambissolo Háplico + Latossolo Vermelho	50, 30 e 20		0,40
RR6	Associação Neossolo Regolítico + Chernossolo Háplico	60 e 40	CX	0,01

[1] UM: unidades de mapeamento de solos (Kämpf et al., 2004); [2] De acordo com Santos et al. (2013). RR: Neossolo Regolítico; CX: Cambissolo Háplico; e RL: Neossolo Litólico.

Os modelos preditores gerados para cada esquema de amostragem tiveram suas acurácias estimadas pelos seguintes métodos de avaliação: validação aparente, que se utiliza da totalidade do conjunto de dados usados para treinamento; divisão percentual, sendo 66 % dos dados usados para treinamento e 34 % para avaliação; validação cruzada com 10 subconjuntos, sendo nove subconjuntos para treinamento e um para avaliação, sendo intercalados até que todos os conjuntos sejam usados para treino e para teste (Steyerberg, 2009); validação com dados independentes, utilizando sete conjuntos de dados distintos daqueles usados para treinamento, sendo amostrados aleatoriamente sobre a área de estudo, obedecendo, assim, ao critério de amostragem probabilística (Grinand et al., 2008; Brus et al., 2011). Os dados independentes foram amostrados em diferentes proporções (0,75; 1,5; 3,0; 4,5; 6,0; 7,5; e 9,0 %) em relação ao tamanho total da microbacia, resultando em conjuntos de dados com 7500, 15000, 30000, 45000, 60000, 75000 e 90000 pontos amostrais.

As estimativas da acurácia obtidas pelos métodos de avaliação foram comparadas à acurácia mensurada, obtida pela comparação dos mapas gerados a partir de cada esquema de amostragem com o mapa convencional de solos. Para isso, cada modelo preditor foi transcrito em regras de classificação, que foram implementadas em ambiente de SIG para obter os mapas preditores de ocorrência de solos, sendo um mapa digital para cada esquema de amostragem. Com auxílio da função *Tabulate Area* no ArcGis 9.3, cada mapa digital foi comparado *pixel* a *pixel* com o mapa convencional de solos, usando a matriz de confusão de Congalton (1991) para alcançar os valores da acurácia geral, que é a proporção de instâncias corretamente classificadas, e do índice kappa (Cohen, 1960), que mede as concordâncias compensando o acaso.

RESULTADOS E DISCUSSÃO

Os esquemas de amostragem resultaram na geração de diferentes modelos preditores de ocorrência de solos, com distintos valores estimados e medidos da acurácia que estão apresentados no quadro 2. A observação das árvores de classificação geradas permitiu identificar que os atributos do terreno mais importantes para explicar a distribuição espacial dos solos na paisagem foram a elevação, a declividade, o comprimento de fluxo, a distância de rios e o índice de umidade topográfica, os quais têm sido comumente relatados como importantes variáveis preditoras para o MDS (Behrens et al., 2010; Ten Caten et al., 2012; Giasson et al., 2013).

Quando o modelo foi gerado com os dados oriundos da amostragem aleatória, a elevação foi a principal variável preditora de ocorrência de solos, sendo utilizada como nó raiz e em diversos nós internos da árvore de classificação. O comprimento de fluxo, a distância de rios e a declividade também foram variáveis importantes no modelo gerado com dados aleatórios, sendo utilizadas como principais nós internos ao separar os dados em subconjuntos mais homogêneos pelo índice *Gini* do algoritmo *SimpleCart*. Ao utilizar os dados da amostragem estratificada, a variável elevação também foi usada como a principal variável preditora, enquanto para os nós internos do modelo foram utilizadas as variáveis declividade, acúmulo de fluxo e índice de umidade topográfica. Com uso dos dados da amostragem proporcional, foi gerada a árvore com menor tamanho, que apresentou a variável distância de rios como nó raiz; para os demais nós internos foram usadas apenas as variáveis elevação e declividade.

Os distintos modelos resultaram diferentes distribuições espaciais dos solos, como apresentadas

Quadro 2. Resultados obtidos pelos métodos de estimativas da acurácia dos modelos e pela medição da acurácia *pixel* a *pixel* entre cada mapa gerado e mapa convencional de solos

Método de avaliação	Aleatória simples		Estratificada aleatória		Proporcional aleatória	
	Acurácia geral	kappa	Acurácia geral	kappa	Acurácia geral	kappa
	%		%		%	
			Avaliação Interna			
VC-10[(1)]	62,1	0,45	66,0	0,62	55,6	0,33
DP[(2)]	61,3	0,44	65,9	0,63	55,8	0,33
VA[(3)]	64,0	0,46	67,1	0,64	57,0	0,35
Média	62,5	0,45	66,3	0,63	56,1	0,34
			Avaliação com dados independentes			
7500	63,0	0,46	29,6	0,14	51,4	0,27
15000	62,3	0,45	29,0	0,13	50,1	0,26
30000	62,9	0,46	29,4	0,14	51,7	0,27
45000	63,5	0,46	29,0	0,13	51,3	0,26
60000	62,6	0,45	29,5	0,14	51,3	0,27
75000	62,6	0,45	29,5	0,14	51,3	0,27
90000	63,0	0,46	30,1	0,15	51,6	0,27
Média	62,8	0,46	29,4	0,14	51,2	0,27
			Acurácia mensurada			
	63,0	0,46	29,2	0,14	51,0	0,26

[(1)] VC-10: validação cruzada com 10 subconjuntos; [(2)] DP: divisão percentual; e [(3)] VA: validação aparente.

nos mapas digitais (Figura 1). No mapa digital gerado a partir da predição com dados aleatórios (Figura 1b), foram estimadas as seis UMs mais extensas e representativas da microbacia do Rio Santo Cristo (G1, LV1, LV2, RR1, RR2 e RR3). Como constatada em diversos trabalhos com MDS, a predição de ocorrência de UMs pouco representativas é prejudicada quando utilizada a amostragem aleatória (Giasson et al., 2011; Ten Caten et al., 2012). Todavia, a distribuição espacial das UMs foi a mais semelhante em relação ao mapa convencional de solos, resultando na maior acurácia mensurada, com concordância com o mapa convencional de 63,0 % e índice kappa de 0,46.

Os mapas digitais gerados a partir de dados amostrados proporcionalmente à área de cada UM e de forma estratificada resultaram nos menores valores da acurácia mensurada. Ao utilizar os dados da amostragem proporcional, que embora garanta a amostragem de todas as UMs, resultou num mapa digital (Figura 1d) com as 6 UMs mais extensas da microbacia com concordância geral de 51,0 % e índice kappa de 0,26. A partir dos dados amostrados de forma estratificada, o mapa preditor de solos estimou todas as 10 UMs (Figura 1c); porém, com os menores valores de concordância geral (29,2 %) e índice kappa (0,14) em razão de as UMs menos representativas terem sido preditas para áreas maiores que aquelas originalmente mapeadas no mapa convencional. Esses resultados são comparáveis e concordantes com estudos anteriores com MDS, como o realizado por Giasson et al. (2011), os quais encontraram valores de acurácia geral de 68,0 % e índice kappa de 0,54, ao compararem o mapa gerado com o mapa convencional de solos; e o de Bui e Moran (2003), que encontraram valores de acurácia geral entre 53,0 e 79,0 % e índice kappa 0,33 a 0,74, conforme a escala do mapa convencional e das sub-regiões daquela área de estudo.

Os resultados da avaliação da capacidade preditiva dos modelos preditores estão dispostos no quadro 2. Os valores da acurácia geral e do índice kappa foram diferenciados para cada modelo gerado pelas combinações de esquema de amostragem e dos métodos de avaliação aplicados. A estimativa da acurácia dos modelos preditores com uso dos dados independentes e distintos resultou valores de índice kappa e da acurácia geral muito semelhantes, indiferentemente do esquema de amostragem e do tamanho do conjunto de dados independentes. Esses resultados estão em conformidade com a indicação da avaliação com uso de dados independentes e distintos daqueles usados para treinamento dos algoritmos (Grinand et al., 2008; Steyerberg, 2009; Brus et al., 2011).

Ao realizar a avaliação com os métodos de validação aparente, cruzada e por divisão percentual, a acurácia dos modelos gerados com dados oriundos

da amostragem estratificada e proporcional foi superestimada, enquanto a avaliação da acurácia do modelo preditor gerado com os dados da amostragem aleatória resultou valores semelhantes aos obtidos pela acurácia mensurada e pelo uso de dados independentes. Assim, o uso da amostragem aleatória se mostrou mais robusta e vantajosa para a predição de ocorrência de UM por gerar correlações solo-paisagem mais acuradas, cujos valores foram semelhantes para quaisquer dos métodos de avaliação testados. Esses resultados diferem em parte dos obtidos por Hengl et al. (2003), os quais concluíram que a amostragem estratificada foi o método mais apropriado para predizer a ocorrência de classes de solos usando mapas auxiliares e regressão logística.

Os resultados deste trabalho revelaram que o uso de dados independentes é o mais indicado para validar modelos preditores (Grinand et al., 2008; Brus et al., 2011). Adicionalmente, ao utilizar a amostragem aleatória, a avaliação do modelo pode ser realizada com diferentes métodos de avaliação sem que se obtenham falsos valores (super ou subestimados) da acurácia. Isso pode estar associado ao fato da acurácia obtida pela validação cruzada ser calculada a partir de uma matriz de confusão com base no uso de subconjuntos; sempre que um subconjunto for utilizado para a avaliação, esse não será utilizado para treinamento do classificador numa mesma rodada, sendo alternadamente trocados até que todos os subconjuntos sejam utilizados para avaliação. Por isso, a estimativa da acurácia obtida pela validação cruzada resulta de uma estimativa da média das classificações e, por isso, é considerada como indicador confiável para estimar o desempenho da predição de algoritmos supervisionados quando a amostragem for aleatória (Elkan, 2012). Adicionalmente, Steyerberg (2009) citou que a validação cruzada é com base no método da divisão percentual e, como a subdivisão das amostras é realizada de forma aleatória, ambos os métodos de avaliação garantem independência dos dados, permitindo obter valores da acurácia semelhantes à acurácia mensurada.

Na amostragem aleatória, os valores da acurácia geral e do índice kappa avaliados pelo método de validação aparente foram muito semelhantes aos obtidos pela avaliação com dados independentes e pela acurácia mensurada. Todavia, o uso do método de validação aparente não é recomendado para avaliar a acurácia de modelos preditores (Steyerberg, 2009), porque esse método retorna estimativas da acurácia superestimadas, uma vez que todas as comparações das predições são realizadas exatamente sobre as mesmas informações (das variáveis preditoras e dos solos) utilizadas para a geração do modelo. Dessa forma, podem resultar em maior quantidade de acertos de classificação, porém apenas para os dados em que foram treinados (Steyerberg, 2009). Portanto, sua aplicação deve ser restrita apenas em situações em que o banco de dados seja muito pequeno a ponto de inviabilizar a partição dos dados em subconjuntos, como os métodos de divisão percentual e validação cruzada (Elkan, 2012), fornecendo alguma estimativa da acurácia da predição.

Com base nos resultados encontrados, esta pesquisa demonstrou que o esquema de amostragem totalmente aleatória resultou modelos e mapas preditores de ocorrência de solos mais acurados do que os esquemas de amostragem estratificada e proporcional à área de cada UM. Adicionalmente, os modelos gerados com dados aleatórios podem ser avaliados por quaisquer dos métodos de estimativa da acurácia testados (validação aparente, divisão percentual, cruzada e dados independentes). Em relação às demais estratégias de amostragem, os resultados indicaram que os esquemas de amostragem influenciaram na capacidade preditiva dos modelos; o método de estimativa da acurácia usado para avaliar os modelos interfere na obtenção dos resultados estimados da acurácia. Dessa forma, a avaliação de modelos preditores deve ser realizada sempre que possível, com reamostragem de dados independentes e que não foram usados para o treinamento dos modelos, como indicado por Grinand et al. (2008) e Brus et al. (2011). Em situações que o modelo preditor venha a ser gerado com dados totalmente aleatórios e que uma reamostragem para avaliação com dados independentes seja inviável, a avaliação do modelo pode ser realizada com o método de validação cruzada ou divisão percentual.

CONCLUSÕES

Os esquemas de amostragem influenciaram na quantidade de UMs preditas e na acurácia dos modelos preditores; o modelo preditor de ocorrência de solos gerado com dados do esquema de amostragem aleatório simples foi o mais acurado.

A acurácia dos modelos preditores gerados com dados dos esquemas estratificado e proporcional é superestimada quando estimada pelos métodos de validação aparente, validação cruzada e com divisão percentual.

A avaliação de modelos preditores com dados independentes garante obter estimativas da acurácia semelhantes à acurácia mensurada para todos os esquemas de amostragem testados (aleatória, estratificada e proporcional).

Não houve influência do tamanho dos conjuntos de dados independentes usados na estimativa da acurácia de modelos preditores.

AGRADECIMENTOS

À Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), pela bolsa de doutorado

para o primeiro autor. Ao Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), pela bolsa de Produtividade em Pesquisa concedida ao segundo autor e pela bolsa de doutorado para a terceira autora.

REFERÊNCIAS

Behrens T, Zhu AX, Schmidt K, Scholten T. Multi-scale digital terrain analysis and feature selection in digital soil mapping. Geoderma. 2010;155:175-85.

Breiman L, Friedman JH, Olshen RA, Stone CJ. Classification and regression trees. Pacific Grove: Cole Advanced Books and Software; 1984.

Brungard CW, Boettinger JL. Application of conditioned Latin hypercube sampling in arid Rangelands in Utah, USA. In: Boettinger JL, Howell DW, Moore AC, Hartemink AE, Kienast-Brown S, editors. Digital soil mapping: Bridging research, environmental application, and operation. Dordrecht: Springer; 2010. p.67-75.

Brus DJ, Gruijter JJ. Random sampling or geostatistical modelling? Choosing between design-based and model-based sampling strategies for soil (with discussion). Geoderma. 1997;80:1-59.

Brus DJ, Kempen B, Heuvelink GBM. Sampling for validation of digital soil maps. Eur J Soil Sci. 2011;62:394-407.

Bui EN, Moran CJ. A strategy to fill gaps in soil survey over large spatial extents: An example from the Murray-Darling basin of Australia. Geoderma. 2003;111:21-44.

Chatfield C. Model uncertainty, data mining, and statistical inference (with discussion) J Royal Stat Soc. 1995;158:419-66.

Cohen J. A coefficient of agreement for nominal scales. J Educ Measur. 1960;20:37-46.

Congalton RG. A review of assessing the accuracy of classification of remotely sensed data. Rem. Sens Environ. 1991;37:35-46.

Elkan, C. Evaluating classifiers. San Diego: University of California; 2012.

Environmental Systems Research Institute - ESRI. ArcGIS 9.3.1. [CD-ROM]. Redlands: Environmental Systems Research Institute; 2009.

Giasson E, Hartemink AE, Tornquist CG, Teske R, Bagatini T. Avaliação de cinco algoritmos de árvores de decisão e três tipos de modelos digitais de elevação para mapeamento digital de solos a nível semidetalhado na Bacia do Lageado Grande, RS, Brasil. Ci Rural. 2013;43:61-7.

Giasson E, Sarmento EC, Weber E, Flores CA, Hasenack H. Decision trees for digital soil mapping on subtropical basaltic steeplands. Sci Agric. 2011;68:167-74.

Grinand C, Arrouays D, Laroche B, Martin MP. Extrapolating regional soil landscapes from an existing soil map: Sampling intensity, validation procedures, and integration of spatial context. Geoderma. 2008;143:180-90.

Grunwald S. Multi-criteria characterization of recent digital soil mapping and modeling approaches. Geoderma. 2009;152:195-207.

Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P, Witten IH. The WEKA data mining software: an update. SIGKDD Explorations. 2009;11:10-8.

Hasenack H, Weber E. Base cartográfica vetorial contínua do Rio Grande do Sul [DVD-ROM]. Porto Alegre: Universidade Federal do Rio Grande do Sul; 2010. Escala 1:50.000. (Série Geoprocessamento, 3).

Hengl T, Rossiter DG, Stein A. Soil sampling strategies for spatial prediction by correlation with auxiliary maps. Aust J Soil Res. 2003;41:1403-22.

Kämpf N, Giasson E, Streck EV. Levantamento pedológico e análise qualitativa do potencial de uso dos solos para o descarte de dejetos suínos da bacia do Rio Santo Cristo [relatório]. Porto Alegre: Secretaria do Meio Ambiente do Rio Grande do Sul; 2004.

Lagacherie P, McBratney AB. Spatial soil information systems and spatial soil inference systems: Perspectives for digital soil mapping. In: Lagacherie P, McBratney AB, Voltz M., editors. Digital soil mapping: An introductory perspective. Amsterdam: Elsevier; 2007. p.3-24.

McBratney AB, Mendonça-Santos ML, Minasny B. On digital soil mapping. Geoderma. 2003;117:3-52.

Meyer DJ, Tachikawa T, Abrams M, Crippen R, Krieger T, Gesch D, et al. Summary of the validation of the second version of the ASTER GDEM. Int Arch Photogram Remote Sens Spatial Inf Sci. 2012;39-B4:291-3.

Minasny B, McBratney AB. Incorporating taxonomic distance into spatial prediction and digital mapping of soil classes. Geoderma. 2007;142:285-93.

Moran CJ, Bui EN. Spatial data mining for enhanced soil map modelling. Int J Geogr Inf Sci. 2002;16:533-49.

Rossiter DG. Technical note: Statistical methods for accuracy assessment of classified thematic maps. Enschede: International Institute for Geo-information Science and Earth Observation; 2004.

Santos HG, Jacomine PKT, Anjos LHC, Oliveira VA, Lumbreras JF, Coelho RM, Almeida JÁ, Cunha TJF, Oliveira JB, editores. Sistema brasileiro de classificação de solos. 2ª ed. Rio de Janeiro: Empresa Brasileira de Pesquisa Agropecuária; 2013.

Scull P, Franklin J, Chadwick OA, McArthur D. Predictive soil mapping: A review. Progr Phys Geogr. 2003;27:171-97.

Stehman SV. Sampling designs for assessing map accuracy. In: Proceedings of the 8th International Symposium on Spatial Accuracy Assessment in Natural Resources and Environmental Sciences; Jun 25-27 2008; Shanghai, China. Shanghai: World Academic Press; 2008. v.2, p.8-15.

Steyerberg EW. Validation of prediction models. In: Steyerberg EW, editor. Clinical prediction models. New York: Springer; 2009. p.299-311.

Ten Caten A, Dalmolin RSD, Pedron FA, Mendonca-Santos ML. Spatial resolution of a digital elevation model defined by the wavelet function. Pesq Agropec Bras. 2012;47:449-57.

Witten IH, Frank E, Hall MA. Data mining: Practical machine learning tools and techniques. 3th ed. San Francisco: Morgan Kaufmann: 2011.