XXXIV CONGRESSO BRASILEIRO DE CIÊNCIA DO SOLO
28 de julho a 2 de agosto de 2013 | Costão do Santinho Resort | Florianópolis | SC

Teste de densidade de amostragem para treinamento de modelos de árvore de decisão[(1)].

Tatiane Bagatini[(2)]; Elvio Giasson[(3)]; Rodrigo Teske[(4)].

[(1)] Trabalho executado com recursos do Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq);
[(2)] Doutoranda, Bolsista CNPq; Universidade Federal do Rio Grande do Sul (UFRGS), Programa de Pós-graduação em Ciência do Solo (PPGCS); Porto Alegre – RS; e-mail: tatibagatini@yahoo.com.br; [(3)] Professor do Departamento de Solos, Bolsista de Produtividade em Pesquisa do CNPq; UFRGS; Porto Alegre – RS; e-mail: giasson@ufrgs.br
[(4)] Doutorando, Bolsista Capes; UFRGS - PPGCS; Porto Alegre – RS; e-mail: rodrigoteske.agr@gmail.com.

RESUMO: Nas últimas décadas o mapeamento digital de solos está ganhando espaço devido aos aumento ao acesso de dados numéricos e com o desenvolvimento de novas ferramentas de processamento de informação. Todavia, algumas metodologias ainda precisam ser melhor definidas, dentre elas, a densidade de amostras para treinamento dos modelos. Assim sendo, o objetivo desse trabalho foi determinar o efeito da densidade de amostragem na acurácia dos modelos de árvores de decisão em duas bacias hidrográficas no noroeste do RS. As áreas utilizadas são as bacias do Rio Santo Cristo e do Lageado Grande. Os mapas de solos utilizados encontram-se na escala de 1:50.000. Em SIG, a partir do MDE ASTER-GDEM foram gerados sete mapas de variáveis preditoras dos solos na paisagem. As densidades amostrais foram geradas de forma aleatória na proporção de 0,1 a 4 pontos por hectare. O treinamento dos modelos foi realizado no programa Weka e as acurácias foram calculada a partir da matriz de erros. Das árvores de decisão geradas selecionou-se as árvores mais complexas e árvores com tamanho possível de ser aplicadas manualmente em SIG. O aumento da densidade de amostragem resultou no aumento da acurácia geral e no aumento do número de unidades de mapeamento de solos preditas, nas árvores de decisão maiores. Nas árvores de decisão menores o aumento da densidade de amostragem não influenciou a ácurácia geral e influenciou muito pouco no número de unidades de mapeamento de solos.

Termos de indexação: pedologia, pedometria, mapeamento digital.

INTRODUÇÃO

A utilização de mapeamento digital de solos (MDS) com base em sistemas de informação geográfica (SIG), estatística e pedologica está aumentando continuamente nas últimas décadas devido ao aumento de fontes de dados numéricos, tais como aqueles fornecidos pelos modelos digitais de elevação (MDE), combinado com o desenvolvimento de novas ferramentas de processamento de informação (McBratney et al., 2003). A predição de classes de solos na paisagem a partir desta técnica consiste na utilização de modelos matemáticos que conseguem descrever essas relações. Dentre esses modelos, os que vem ganhando destaque são os algoritmos de aprendizagem de máquinas. Neste sentido Qi e Zhu (2003), comparando três desses algoritmos, descobriram que os algoritmos de árvore de decisão (AD) são os mais adequados para extrair e representar o conhecimento de solo.

A utilização de AD é uma técnica eficiente, pois permite processar uma grande quantidade de dados sem interferência humana (Henderson et al., 2005). São geralmente simples e fáceis de interpretar e de discutir (Xu et al., 2005), possibilitam o uso de dados de diferentes tipos de fontes e possuem tempo relativamente reduzido de treinamento e de processamento (Miller e Franklin, 2001).

Contudo, apesar da viabilidade do uso de ADs no mapeamento digital de solos, algumas metodologias ainda precisam ser definidas. Dentre essas metodologias pode-se citar a densidade de amostras para treinamento dos modelos (McBratney et al., 2003). Isso ocorre porque existem várias técnicas de amostragens de pontos (de Gruijter et al., 2006) e, consequentemente, cada uma responde diferentemente à densidade de pontos amostrados.

A questão da densidade de amostragem é importante pois o tamanho de amostras pode afetar significativamente a capacidade de predição dos algoritmos, bem como sua acurácia (Hjort, 2008). Neste sentido, Zhu (2000) sugere adotar como número mínino de amostras pelo menos 30 vezes o número de classes existentes ou a serem preditas. Moran & Bui (2002) apontam que o mais apropriado seria a utilização da totalidade dos dados. Por outro lado, Grinand et al., (2008) apontaram que a densidade de um ponto a cada 10 hectares é suficiente para captar a variabilidade de classes de solos. Considerando o exposto, o objetivo desse trabalho foi determinar o efeito da densidade de amostragem na acurácia geral e no número de unidades de mapeamento de slo (UMS) preditas em duas bacias hidrográficas no noroeste do RS.

XXXIV CONGRESSO BRASILEIRO DE CIÊNCIA DO SOLO
28 de julho a 2 de agosto de 2013 | Costão do Santinho Resort | Florianópolis | SC

MATERIAL E MÉTODOS

As áreas de estudo são as bacias hidrográficas do Rio Santo Cristo e do Rio Lageado Grande. Estas bacias estão inseridas na Bacia Hidrográfica U30 e apresentam áreas de 898 km^2 e 500 km^2, respectivamente. O clima da região é subtropical úmido, tipo Cfa de Köppen, com precipitação média anual de 1.778 mm e temperatura média anual de 18,5 °C. O material de origem da região é basalto da Formação Serra Geral.

Os mapas de solos utilizados encontram-se na escala de 1:50.000 e fazem parte dos levantamentos pedológicos e análise qualitativa do potencial de uso dos solos para o descarte de dejetos suínos da bacia do Rio Santo Cristo (Kämpf et al., 2004a) e da bacia do Rio Lageado Grande (Kämpf et al., 2004b). A bacia de Santo Cristo é composta por 10 UMS e a Lageado Grande por 15 UMS. No programa ArcGis 9.3 (Esri, 2006) a partir do MDE ASTER-GDEM (Abrams et al., 1999) foram gerados mapas de variáveis preditoras dos solos na paisagem, tais como declividade, direção do fluxo, acúmulo do fluxo, comprimento do fluxo, curvatura do declive e índice de umidade topográfica (Beven & Kirkby, 1979). Para geração da variável distância dos rios foi utilizado o arquivo vetorial de hidrografia da base continua do Rio Grande do Sul (Hasenack & Weber, 2010). Todos os mapas citados foram gerados com resolução espacial de 30 metros.

Os arquivos com diferentes densidades amostrais foram gerados de forma aleatória utilizando-se a função *random* do ArcGis 9.3 na proporção de 0,1; 0,3; 1; 1,5; 2; 3 e 4 pontos por hectare. Essas densidades amostrais correspondem a 9; 30; 90; 135; 180; 270 e 360 mil pontos na bacia do Rio Santo Cristo e, 5; 15; 50; 75; 100; 150 e 200 mil pontos para a bacia do Rio Lageado Grande. A partir dessas amostragens coletaram-se as informações de todas as variáveis preditoras e das UMS utilizando-se função *Sample*. Os dados foram exportados para proceder ao treinamento dos modelos com o algoritmo Simple Cart no programa Weka 3.6.3 (Hall et al., 2009). A seleção do tamanho das ADs se deu em função do número de elementos no nó final (M), sendo que, quando M é igual a 2 ocorre a geração das maiores ADs. Para a seleção das ADs menores foram testados sete valores de M, dos quais selecionou-se os que resultaram ADs com tamanho possível de ser aplicado manualmente (de 150 a 200 folhas), dado que a implementação das mesmas, em SIG, é feita manualmente. As acurácias gerais foram calculada a partir da matriz de erros de Congalton (1991).

RESULTADOS E DISCUSSÃO

Na **figura 1** são apresentados os resultados das acurácias, o número de UMS preditas e o tamanho das ADs, resultantes das árvores mais complexas da Bacia Santo Cristo. Em relação à acurácia, observa-se que o modelo não conseguiu prever a totalidade das classes (10) quando da utilização de menos de um ponto por hectare, entretanto, mesmo com a predição de somente seis UMS a acurácia geral foi de 60%. Vale salientar que a área total ocupada por essas seis UMS equivalem a 99,3% da área portanto, as demais UMS provavelmente não foram preditas devido à baixa representatividade das mesmas na área. A acurácia variou de 60 a 76% da menor para a maior densidade de pontos amostrais. Observa-se também que a maior diferença ocorreu entre as densidades de 0,3 pontos por ha para a densidade de 1 ponto por ha. A partir de 1 ponto por hectare observa-se que o tamanho das árvores foram aumentando, porém não refletiu no aumento proporcional das acurácias. Entretanto, mesmo havendo um pequeno ganho na acurácia geral dos modelos com o aumento da densidade de pontos, sem o auxílio de alguma ferramenta que processe esse montante de informação, a implementação dessas árvores torna-se impraticável, dado que a implementação das ADs nos SIGs é manual (a partir da geração dos modelos a montagem das equações, bem como a implementação em SIG de cada regra de decisão é realizada manualmente utilizando o *map calculator*). Vale ressaltar que, nas densidades menores que 1 ponto por hectare, mesmo com M=2, o tamanho das ADs não foi superior a 200 folhas.

Por outro, lado quando selecionou-se árvores com tamanhos menores, entre 150 e 200 folhas (**Figura 2**), não se observou grandes diferenças nos valores de acurácia (60 a 63%) e nem diferenças no número de UMS preditas (seis) entre as diferentes densidades de amostragem. Vale salientar que as seis UMS preditas abrangem aproximadamente 99% da área em estudo, portanto, o modelo não conseguiu predizer as demais devido a baixa representatividade das mesmas.

Comparando-se os dados da **fiura 1** com os da **figura 2**, percebe-se que há um aumento na acurácia com o aumento da densidade de pontos amostrais somente nas ADs maiores e, a partir de 1 ponto por ha. Entretanto, devido ao tamanho das mesmas, a implementação manual em SIG, torna-se impraticável.

Na **figura 3** são apresentados os dados de acurácia, o número de UMS e o tamanho das ADs referentes as árvores maiores geradas na Bacia do Rio Lageado Grande. Analisando os dados observa-se que somente a partir de 2 pontos por hectare o modelo conseguiu predizer todas as 15 UMS. Observa-se também que na menor densidade de amostragem o modelo conseguiu prever somente

XXXIV CONGRESSO BRASILEIRO DE CIÊNCIA DO SOLO
28 de julho a 2 de agosto de 2013 | Costão do Santinho Resort | Florianópolis | SC

cinco UMS. Isso pode ter ocorrido porque as 10 UMS não preditas ocupam somente 7,5% da área. Assim sendo, quando da utilização de baixa densidade de pontos, o modelo não conseguiu captar a variabilidade presente (Moran & Bui, 2002). Em relação a acurácia, observa-se que a mesma variou de 59% na menor densidade de amostragem a 74% na maior densidade. Observa-se também que, igualmente aos dados da bacia anterior, quanto maior a densidade de amostragem, maior foi o tamanho da árvore de decisão.

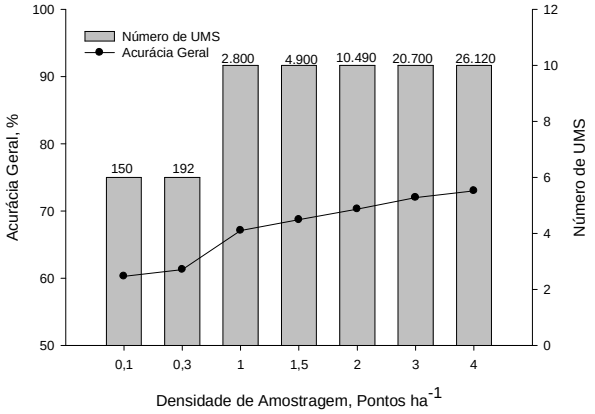

Figura 1 – Acurácia e número de unidades de mapeamento dos solos (UMS) preditas com diferentes densidades de amostragens na Bacia do Santo Cristo, nas árvores de decisão com M=2 (números sobre as barras correspondem ao tamanho das árvores de decisão, em folhas).

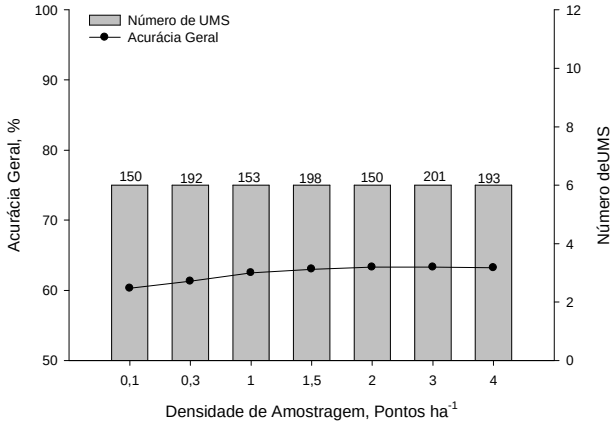

Figura 2 – Acurácia e número unidades de mapeamento (UMS) com diferentes densidades de amostragens na Bacia Santo Cristo, nas árvores de decisão menores (números sobre as barras correspondem ao tamanho das árvores de decisão, em folhas).

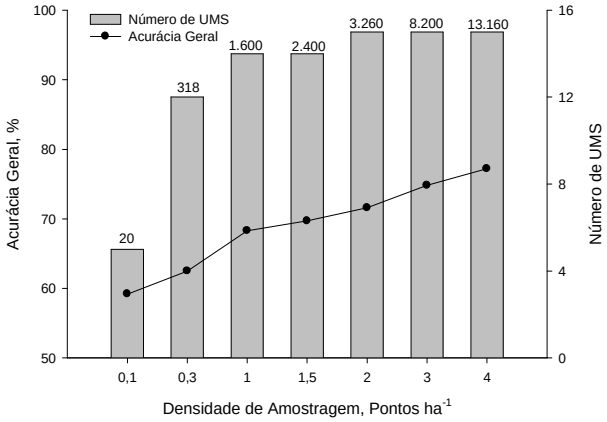

Figura 3 – Acurácia e número de unidades de mapeamento dos solos (UMS) preditas com diferentes densidades de amostragens na Bacia Lageado Grande, nas árvores de decisão com M=2 (números sobre as barras correspondem ao tamanho das árvores de decisão, em folhas).

Em relação aos dados obtidos com tamanho de árvores menores (**Figura 4**) observa-se que, igualmente aos dados da bacia anterior não houve aumento na acurácia geral nas diferentes densidades de amostragem, porém, ocorreu um aumento do número de UMS preditas. Na menor densidade de amostragem foram preditas somente cinco UMS enquanto que nas maiores foram preditas 13. Essa diferença de predição entre as densidades de amostragem pode ter ocorrido devido à existência de várias UMS com baixa representatividade e, com o aumento da densidade de pontos o modelo conseguiu modelar a ocorrência de um maior número de UMS.

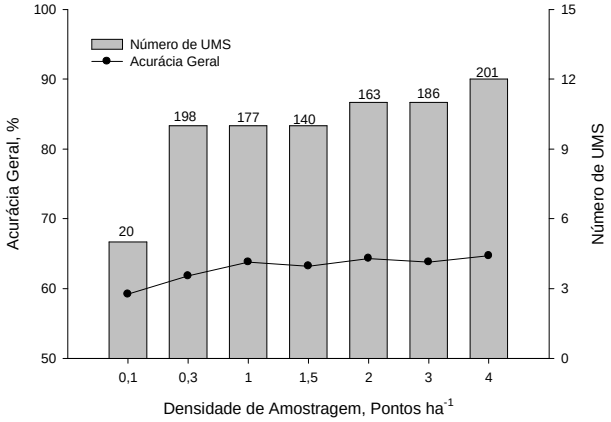

Figura 4 – Acurácia e e número de unidades de mapeamento (UMS) com diferentes densidades de amostragens na Bacia Lageado Grande, nas árvores de decisão menores (números sobre as barras correspondem ao tamanho das árvores de decisão, em folhas).

Assim sendo, percebe-se que, nas duas bacias, o aumento da densidade de amostragem influenciou na acurácia e no número de UMS preditas. Contudo quando selecionou-se as árvores menores, percebe-se que, quando há uma dominância de classes (bacia Santo Cristo), a acurácia geral e o número de UMS preditas não foi influenciadas pela densidade de pontos. Entretanto, quando ocorre o aumento de número de UMS com menor representatividade (bacia Lageado Grande), o aumento da densidade de pontos proporcionou um aumento no número de UMS preditas, porém sem aumento da acurácia.

Portanto, analisando os dados das duas bacias, observa-se que, se houver a possibilidade da aplicação automatizada em SIG dos modelos mais complexos recomenda-se a utilização da densidade de amostras de 1 ponto por ha, por outro lado, se a aplicação for manual, recomenda-se a densidade de 0,3 pontos por ha dado que com essa densidade de amostragem é possível obter valores de acurácia geral maiores que 60% e ao mesmo tempo ter ADs com tamanho possível de ser aplicado manualmente.

CONCLUSÕES

O aumento da densidade de amostragem resultou no aumento da acurácia geral e no aumento do número de unidades de mapeamento de solos preditas nas árvores de decisão maiores.

Nas árvores de decisão menores, o aumento da densidade de amostragem não influenciou a ácurácia geral e, influenciou muito pouco no número de unidades de mapeamento de solos.

REFERÊNCIAS

ABRAMS, M. et al. ASTER users handbook. Pasadena: JPL, 1999. 93p.

BEVEN, K.; KIRKBY, N. A physically based variable contributing area model of basin hydrology. Hydrological Sciences. Bulletin des Sciences Hydrologiques, 24 : 43-69, 1979.

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

de Gruijter JJ, Brus DJ, Bierkens MFP, Knotters M. Sampling for Natural Resource Monitoring. Springer: New York, 2006. 326p.

ESRI. Environmental Systems Research Institute, Inc. (ESRI). ArcGIS. Professional GIS for the desktop, versão 9.3.1 CA. 2009.

HALL, M.; FRANK, E.; HOLMES, G.; PFAHRINGER, B.; REUTEMANN, P.; WITTEN, I. H. The WEKA Data Mining Software: An Update; SIGKDD Explorations, 11, 2009.

HASENACK, H.; WEBER, E. (org.). Base cartográfica vetorial contínua do Rio Grande do Sul - escala 1:50.000. Porto Alegre, UFRGS-IB-Centro de Ecologia. 2010. 1 DVD-ROM (Série Geoprocessamento, 3).

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

HJORT, J.; MARMION, M. Effects of sample size on the accuracy of geomorphological models. Geomorphology, 102: 341 – 350, 2008.

KÄMPF, N. GIASSON, E. STRECK, E.V. Levantamento pedológico e análise qualitativa do potencial de uso dos solos para o descarte de dejetos suínos da microbacia do Rio Santo Cristo. SEMA RS, Programa Nacional do Meio Ambiente II. Relatório final. 2004a.

KÄMPF, N. GIASSON, E. STRECK, E.V. Levantamento pedológico e análise qualitativa do potencial de uso dos solos para o descarte de dejetos suínos da microbacia do Rio Lageado Grande. SEMA RS, Programa Nacional do Meio Ambiente II. Relatório final. 2004b.

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

MILLER, J. FRANKLIN, J. Modeling the distribution of four vegetation alliances using generalized linear models and classification trees with spatial dependence. Ecological Modeling, 57: 227 – 247, 2001.

MORAN, C. J.; BUI, E. Spatial data mining for enhanced soil map modelling. International Journal of Geographical Information Science, 16: 533– 549, 2002

QI, F. Knowledge discovery from area-class resource maps: data preprocessing for noise reduction. Transactions in GIS, 8: 297 – 308, 2004.

QI, F.; ZHU, A. X. Knowledge discovery from soil maps using inductive learning. International Journal of Geographical Information Science, 17: 771 – 795, 2003.

XU, M.; WATANACHATURAPORN, P.; VARSHNEY, P. K.; ARORA, M. K. Decision tree regression for soft classification of remote sensing data. Remote Sensing of Environment, 97: 322 – 336, 2005.

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