XXXV Congresso Brasileiro de Ciência do Solo
CENTRO DE CONVENÇÕES - NATAL / RN
O SOLO E SUAS MÚLTIPLAS FUNÇÕES
02 a 07 DE AGOSTO DE 2015

Avaliação de quatro algoritmos de árvores de decisão usando diferentes densidades de amostragem[(1)].

José Janderson Ferreira Costa[(2)]; Israel Rosa Machado[(3)]; Elvio Giasson[(4)]; Alcinei Ribeiro Campos[(3)]; Elisângela Benedet da Silva[(3)].

[(1)] Trabalho executado com recursos da Epagri, CNPq e da Universidade Federal do Rio Grande do Sul (UFRGS).
[(2)] Mestrando do Programa de Pós-Graduação em Ciência do Solo da UFRGS; janderson.costa@ufrgs.br; [(3)] Aluno do Programa de Pós-Graduação em Ciência do Solo da UFRGS; [(4)] Professor Associado; UFRGS.

RESUMO: No mapeamento digital de solos alguns aspectos metodológicos necessitam ser melhor pesquisados. Dentre esses destaca-se a densidade de amostragem para o treinamento dos modelos de Árvores de Decisão. Este estudo teve como objetivo realizar a comparação de quatro algoritmos de AD, utilizando diferentes densidades de amostragem. Foram derivadas nove variáveis de terreno a partir do Modelo Digital de Elevação, usando seis densidades de amostragem na proporção de 0,2; 0,5; 1; 1,5; 2 e 2,5 pontos amostrais por hectare. As variáveis foram derivadas utilizando o System for Automated Geoscientific Analyses (SAGA GIS) e os dados tabulados no aplicativo computacional ArcGis 9.3. As ADs foram construídas no programa Weka 3.6.3. Obteve-se as melhores porcentagens de acurácia geral nas densidades 1,5; 2 e 2,5. Todos os modelos de AD conseguiram predizer o total de 6 Unidades de Mapeamento. Os algoritmos *Simple Chart*, *BR Tree* e *J48* apresentaram os melhores valores de acurácia geral média, 56,7%, 55,5% e 55,1%, respectivamente.

Termos de indexação: pedometria, classificadores, mapeamento digital de solos.

INTRODUÇÃO

O conhecimento das propriedades e da distribuição na paisagem das diferentes classes de solos é indispensável para o correto uso desse recurso natural. No entanto, apesar da importância são poucas as áreas no Brasil que possuem mapas de solos em escala adequada ao planejamento do uso e ocupação de forma sustentável. Essa carência de informação a respeito de levantamentos dos solos pode ser atribuída aos métodos e técnicas utilizadas para realização desses levantamentos que em grande parte foram feitos de forma tradicional.

Por exigir muito tempo e ser oneroso o levantamento tradicional de solos torna-se impraticável quando existe uma demanda por mapas em escalas de grande detalhamento para áreas extensas. Frente a essas dificuldades McBratney et al., (2003) formularam o conceito de mapeamento digital de solos (MDS) e propuseram essa técnica como uma alternativa econômica e rápida para geração de informações dos solos a partir de modelos numéricos.

No MDS utiliza-se vários métodos para gerar as regras de classificação que posteriormente serão aplicadas para produção dos mapas digitas de solos, dentre esses métodos podemos destacar os modelos de regressões logísticas múltiplas multinominais (Giasson et al., 2006), modelos de redes neurais artificiais (Sirtoli, 2008) modelos logísticos com aplicação de componentes principais (ten Caten et al., 2009) e o modelo de árvores de decisão (Giasson et al., 2013).

Dentre esses modelos, estudos tem evidenciado que o uso de árvores de decisão apresentam bons resultados no MDS. Estudos utilizando esse modelo tem demonstrado bom desempenho na capacidade para descriminação de classes de solos (Coelho & Giasson, 2010; Giasson et al., 2011).

O método de árvores de decisão faz uso de diversos algoritmos para geração das regras de classificação, sendo que cada um deles podem apresentar resultados diferentes em relação a sua capacidade de predição das classes de solos, ocasionando variações na acurácia das regras de classificação e no índice Kappa. Portanto, são necessários estudos que tenham por objetivos selecionar e recomendar os melhores algoritmos para uso no MDS (Coelho & Giasson, 2010). Outro aspecto importante, é definir o número de amostras para treinamento dos modelos (McBratney et al., 2003), pois a densidade de amostragem afeta a capacidade de predição dos algoritmos e influência nos valores de acurácia (Hjort, 2008).

O objetivo deste estudo foi realizar a comparação de quatro algoritmos de árvores de decisão, utilizando diferentes densidades de amostragem.

MATERIAL E MÉTODOS

A área de estudo está localizada no município de Lontras na zona agroecológica 2A – Alto Vale do

Rio Itajaí, no estado de Santa Catarina. A microbacia do rio concórdia apresenta área de 38,80 km^2 e pertence à região hidrográfica 7 (vale do Itajaí). O clima é classificado com Cfa, segundo Köeppen, caracterizado por ser constantemente úmido, com precipitação pluviométrica média anual de 1.480mm e temperatura média anual variando de 17,0 a 19,1°C. A microbacia possui um mapa de Unidade Fisiografica na escala de 1:25.000. Nesse mapa estão espacializados os solos dominantes ocorrentes nas subpaisagens e suas associações (EPAGRI, 2006). A partir do modelo digital de elevação (MDE) com resolução de 1 metro foram derivadas as variáveis do terreno com uso do System for Automated Geoscientific Analyses (SAGA GIS). Foram obtidas nove variáveis, são elas: a) declividade, b) curvatura, c) perfil de curvatura, d) curvatura planar, e) direção de fluxo, f) acúmulo de fluxo e g) índice de umidade topográfica (Beven & Kirkby, 1979). Além destas, foram geradas as variáveis aspecto e a forma sombreada do terreno. Os mapas do estudo foram gerados com resolução espacial de 5m.

Foram testadas diferentes densidades de amostragem na proporção de 0,2; 0,5; 1; 1,5; 2; e 2,5 pontos por hectare. O limiar mínimo foi na proporção de 0,2 pontos para mapas com escala de 1:25.000, recomendado pelo manual técnico de pedologia (FILHO, 2007). Através do ArcGis 9.3 (ESRI, 2009) utilizando-se a ferramenta *Create Random Point* foram criados os pontos amostrais correspondentes a 1; 2; 4; 6; 8 e 10 mil pontos. Com a ferramenta *Sample* realizou-se a amostragem das nove variáveis em cada ponto e em seguida foram exportados para o programa Weka 3.6.3 (Hall et al., 2009). Nesse programa os dados foram inicialmente analisados para gerar as regras de classificação pelo método de árvores de decisão (AD). Os algoritmos testados foram: *J48*, *Simple Chart*, *Random Tree* e *REP Tree*.

Os modelos foram validados com o função cross-validation e avaliados pela acurácia geral (AG), índice Kappa e tamanho da árvore final.

RESULTADOS E DISCUSSÃO

A Tabela 1 apresenta os resultados obtidos com diferentes densidades de amostragem. Os indicadores de desempenho dos algoritmos de árvores de decisão demonstraram que as densidades nas proporções de 1,5; 2 e 2,5 pontos, apresentaram os melhores resultados (Figura 1). No entanto, nessas densidades os algoritmos utilizados conseguiram predizer a ocorrência das unidades de mapeamento (UM) de solos de forma semelhante às demais densidades do estudo.

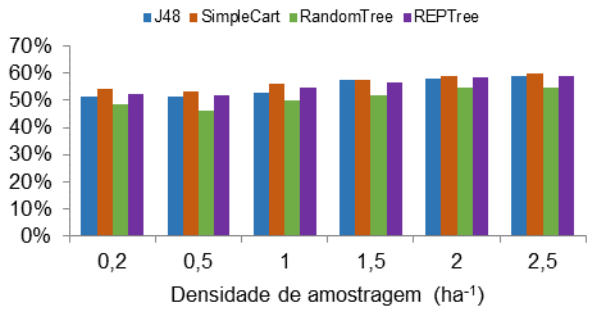

Figura1. Diferentes densidades de amostragem na microbacia do rio concórdia e acurácia geral determinadas pelos algoritmos de árvores de decisão.

Como observado, com o aumento do número de amostras houve melhor desempenho dos algoritmos de aprendizagem de máquina, resultando no aumento de acurácia geral. Esses resultados corroboram com os realizados por Sarmento et al. (2012) e Bagatini et al. (2013), demostrando a importância de se determinar os valores de densidade de amostragem no sentido de melhorar a acurácia geral. O índice Kappa e o tamanho da AD também sofreram alterações positivas com o aumento do número de amostras, exceto na densidade de 0,5 que a média dos resultados de AG foram inferiores a densidade de 0,2 pontos por hectare (Tabela 1).

O algoritmo *Simple Chart* apresentou melhor desempenho, uma vez que alcançou o maior valor de índice Kappa e menor tamanho de AD. O tamanho da AD é um importante critério a ser avaliado, uma vez que cada regra de classificação tem que ser implantada individualmente em software de Sistema de Informação Geográfica - SIG, assim quanto menor a AD menos complexa será sua implementação.

Nos resultados de acurácia, todos os modelos de AD conseguiram predizer o total de 6 UM. O algoritmo *Simple Chart* na densidade de 2,5 pontos atingiu acurácia geral de 60%, mostrando-se ser um bom classificador. Observa-se que todas as UM foram preditas pelos algoritmos de AD. Esse resultado pode estar relacionado ao modelo digital utilizado, que possui alta resolução espacial. Da mesma forma, Sarmento et al. (2012) testaram o MDE de alta resolução com aumento do número de amostragem, obtendo melhores resultados na

predição de solos.

Os valores de acurácia variaram de 46,4% a 60% na densidade de 0,5 a 2,5 pontos/ha, respectivamente. Os métodos usando algoritmo *J48*, *Simple Chart* e *BF Tree* originaram árvores de decisão com capacidade levemente superior aos demais classificadores, apresentando os melhores valores de AG média (56,7% para *Simple Chart*, 55,5% para *BR Tree* e 55,1% *J48*). Esses algoritmos foram utilizados no trabalho de Giasson et al. (2013) e produziram modelos de AD capazes de originar mapas de solos semelhantes ao mapa original, mostrando-se serem bons classificadores e recomendados para futuros estudos com MDS.

CONCLUSÕES

A densidade de amostragem influencia no desempenho dos algoritmos de aprendizagem e o maior número de amostras provoca ligeiro aumento no valor de acurácia geral.

O maior valor de AG e índice Kappa foram obtidos na densidade de amostragem 2,5 pontos com o algoritmo *Simple Chart*.

Os modelos de árvores de decisão *J48*, *Simple Chart* e *BR Tree* obtiveram os melhores valores de acurácia geral, na densidade de 2,5 pontos por hectare.

AGRADECIMENTOS

Ao Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) e à Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) pela concessão das bolsas de estudo dos autores.

REFERÊNCIAS

BAGATINI, T.; GIASSON, E. & TESKE, R. Teste de densidade de amostragem para treinamento de modelos de árvore de decisão. In: XXXIV CONGRESSO BRASILEIRO DE CIÊNCIA DO SOLO. Anais. Florianópolis, SC: Sociedade Brasileira de Ciência do Solo, 2013.

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

COELHO, F.F.; GIASSON, E. Métodos para mapeamento digital de solos com utilização de sistema de informação geográfica. Revista Ciência Rural, Santa Maria, RS, v. 40, n. 10, p. 2099-2106, out. 2010.

EPAGRI, Empresa de Pesquisa Agropecuária e Extensão Rural de Santa Catarina S.A. Inventário de Terras – Microbacia do rio concórdia, município de Lontras – SC, 2006.

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

FILHO, C.J.M., (Coord.) Manual técnico de pedologia. 2. ed. Rio de Janeiro, IBGE, Coordenação de Recursos Naturais e Estudos Ambientais, 2007. 316p. (Manuais Técnicos em Geociências, n.4).

GIASSON, E. et al. Digital soil mapping using multiple logistic regressions on terrain parameters in Southern Brazil. Scientia Agricola, 63:262-268, 2006.

GIASSON, E. et al. Decision trees for digital soil mapping on subtropical basaltic steeplands. Scientia Agrícola, v.68, p.167-174, 2011.

GIASSON, E. et al. 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. Ciência Rural, Santa Maria, v.43, n.11, nov., 2013.

HALL, M. A. Correlation-based feature subset selection for machine learning. Hamilton, New Zealand. 1998. In: HALL, M.; FRANK, E.; HOLMES, G.; PFAHRINGER, B.; REUTEMANN, P.; WITTEN, I.H. 2009. The WEKA Data Mining Software: An Update; SIGKDD Explorations, Volume 11, Issue 1.

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

McBRATNEY, A.B.; MENDONÇA-SANTOS, M.L.; MINASNY, B. On Digital Soil Mapping. Geoderma, Amsterdam, v. 117, p. 3-52, 2003.

SARMENTO, E. C. et al. Predição de ordens de solos com alta resolução espacial: resposta de diferentes classificadores à densidade de amostragem. Pesquisa Agropecuária Brasileira, 47:1395-1403, 2012.

SIRTOLI, A.E. Mapeamento de solos com auxílio da geologia, atributos do terreno e índices espectrais integrados por redes neurais artificiais. 2008. 96f. Tese (Doutorado em Geologia) - Curso de Pós-graduação em Geologia, Universidade Federal do Paraná, PR.

TEN CATEN, A.; DALMOLIN, R. S. D.; RUIZ, L. F. C.; SEBEM, E.; PEREIRA, R. S. P. Mapeamento digital de solos através da aplicação de componentes principais em modelos logísticos. In: Simpósio Brasileiro de Sensoriamento Remoto, Natal, RN, 2009. Anais XIV. INPE:, 2009. p. 7677-7684.

Tabela 1 – Resultados da análise de árvores de decisão obtidos pelos quatro algoritmos classificadores usando difetentes densidades de amostragem.

	Ds 0,2 pontos/ha				Ds 0,5 pontos/ha				Ds 1 ponto/ha			
Método	AG	K	t	UM	AG	K	t	UM	AG	K	t	UM
J48	51,6	0,35	160	6	51,3	0,36	296	6	53	0,37	604	6
Simple chart	54,4	0,38	34	6	53,2	0,38	27	6	56	0,4	198	6
Random tree	48,5	0,31	777	6	46,4	0,3	1471	6	50	0,33	2857	6
REP tree	52,4	0,35	79	6	51,9	0,36	195	6	54,6	0,39	285	6
Média	51,7	0,35	262,5	6	50,7	0,35	497,2	6	53,4	0,37	986	6

	Ds 1,5 pontos/ha				Ds 2 pontos/ha				Ds 2,5 pontos/ha			
Método	AG	K	t	UM	AG	K	t	UM	AG	K	T	UM
J48	57,6	0,44	867	6	58	0,45	1102	6	59,2	0,46	1325	6
Simple chart	57,5	0,43	105	6	59,1	0,45	166	6	60	0,47	173	6
Random tree	51,8	0,36	3931	6	54,5	0,4	5397	6	54,5	0,4	6465	6
REP tree	56,4	0,42	451	6	58,4	0,45	607	6	59,2	0,46	743	6
Média	55,8	0,41	1338	6	57,5	0,44	1818	6	58,2	0,45	2176	6

AG = acurácia geral; K = índice Kappa; t = tamanho da árvore de decisão expresso em número de folhas finais; UM = unidades de mapeamento de solos que o algoritmo foi capaz de prever.