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

Modelos baseados em árvores para predição de classes de solo[(1)]

Helena Saraiva Koenow Pinheiro[(2)]; Lúcia Helena Cunha dos Anjos[(3)]; Waldir de Carvalho Junior[(4)]; César da Silva Chagas[(4)]

[(1)] Trabalho executado com apoio do Curso de Pós-Graduação em Agronomia - Ciência do Solo (CPGA-CS) da Universidade Federal Rural do Rio de Janeiro (UFRRJ), Embrapa Solos - RJ, CAPES e FAPERJ. [(2)] Doutoranda do CPGA-CS (UFRRJ). Seropédica, RJ. lenask@gmail.com; [(3)] Professora Associada IV; Universidade Federal Rural do Rio de Janeiro [(4)] Pesquisador A; Embrapa Solos, Rio de Janeiro.

RESUMO: Modelos solo-paisagem quantitativos representam uma nova tendência nos levantamentos de solos. Neste sentido, diferentes técnicas de mapeamento digital são aplicadas para prever os padrões naturais de ocorrência de classes de solo. O objetivo deste trabalho foi predizer unidades de mapeamento de solos em uma bacia hidrográfica, no Estado do Rio de Janeiro, que apresenta grande variação de condições de paisagem. A abordagem foi baseada em conhecimento pedológico tácito, culminando na escolha de atributos da paisagem que representem a variabilidade dos fatores de formação de solos na região. Na construção do modelo solo-paisagem foram gerados atributos relacionados à pedogênese tais como - altimetria, declividade, curvatura, índice topográfico composto, distância euclidiana de hidrografia, *clay minerals, iron oxide* e índice de vegetação da diferença normalizada (NDVI), geologia e *geomorphons*. Os solos predominantes na bacia foram: Latossolos, Argissolos, Gleissolos, Cambissolos, Neossolos Flúvicos e Neossolos Litólicos. Os algoritmos utilizados foram árvores de decisão e *random forest*. O desempenho dos algoritmos foi avaliado realizado através de índices estatísticos e generalização das unidades de mapeamento. O melhor desempenho foi observado para o modelo *random forest* que apresentou valor superior para os índices estatísticos e melhor generalização das unidades de mapeamento.

Termos de indexação: pedometria, levantamento de solos, mapeamento digital de solos

INTRODUÇÃO

Mapas de solos servem como infra-estrutura básica para gestão de terras e proteção dos recursos naturais. A ciência do solo enfrenta um grande desafio para atender a demanda da sociedade por informações pedológicas, em escala e detalhamento apropriado como suporte à tomada de decisões sobre o uso da terra. Neste contexto, autores como Hengl et al. (2004), Minasny et al. (2003), McBratney et al. (2003), propõem que a direção das mudanças nos levantamentos de solo aborde questões como algoritmos de predicão, modelagem dinâmica, integração de sistemas de informação geográfica (SIG) ferramentas de geoestatística e uso de imagens de alta resolução. Na última década estudos como os de Chagas et al. (2011); Carvalho Júnior et al. (2011); Ten Caten (2011) mostraram a aplicação de técnicas de mapeamento digital de solos no Brasil.

O mapeamento digital de solos une conceitos clássicos de pedologia com modernas ferramentas de análise de dados e modelagem, visando aperfeiçoar as informações fornecidas pelos levantamentos de solos. De acordo com esta perspectiva, o objetivo do estudo foi avaliar métodos baseados em árvores, para a predicao de classes de solo em uma bacia hidrográfica no Estado do Rio de Janeiro. Os métodos testados foram árvores de decisão e *random forest*, e a seleção do mapa para representação das unidades de mapeamento foi baseada nos índices estatísticos e na generalização de classes de solo.

MATERIAL E MÉTODOS

Área de estudo

A bacia hidrográfica do rio Guapi-Macacu, no Estado do Rio de Janeiro, tem área correspondente a 1.250,78 km², perímetro de 199,2 km e 72,68 km de extensão (Ecologus-Agrar, 2003). O clima é classificado como tropical chuvoso com inverno seco (Aw) de acordo com Köppen (1948). A temperatura média é de 23 °C; apresentando, ocasionalmente, baixas temperaturas no inverno. A precipitação média anual é superior a 1.200 milímetros e pode chegar a 2.600 milímetros nas partes mais altas da bacia (Projeto Macacu, 2010).

Co-variaveis preditoras

O modelo digital de elevação (MDE) foi gerado por interpolação de dados primarios de elevação e rede de drenagem, e foi restrito aos limites da bacia hidrográfica. O modelo de elevação foi gerado pela ferramenta "*TopotoRaster*", com base no algoritmo ANUDEM desenvolvido por Hutchinson (1993). Depois de gerar o MDE as células "*sink*" foram preenchidas almejando um modelo sem falhas de interpolação.

Atributos da paisagem foram gerados a partir do MDE para compor o conjunto de variáveis preditoras

usadas como entrada para os modelos preditivos. Os atributos derivados do DEM foram: Altimetria, *Slope*, Curvatura, Indice Topografico Composto (CTI, sigla em ingles), Distância Euclidiana de Hidrografia e *Geomorphons*. Para completar as informações utilizadas como entrada para os modelos preditivos foram gerados três índices espectrais a partir dos dados do Landsat 5 (TM), são eles o Índice de Vegetação por Diferença Normalizada (NDVI), *Clay Minerals* e *Iron Oxide*. Este procedimento foi efetuado no ERDAS Imagine v.9.1.

Os atributos do terreno derivados do DEM, foram obtidos em ArcGIS Desktop v.10, através das ferramentas em *Spatial Analyst* Toolbox (ESRI, 2010). O mapa de formas da superficie foi criado no programa GRASS (*Geographc Resources Analysis Support System*), através do algoritmo *Geomorphon*, que classifica superfícies em dez formas comuns, a partir do MDE e da distância do raio de busca (*lookup distance*), a qual foi pré-definida em 45 células. Informações sobre este método podem ser obtidas em Jasiewicz & Stępiński (2013).

Mapeamento digital de solos

O levantamento de campo para amostragem e descrição das classes de solo, contou com 100 pontos de amostragem definidos com apoio do algoritmo do hipercubo latino condicionado, atraves do software cLHS - *conditioned Latin Hypercube Sampling* (Mathworks, 2005). Este método considera a representatividade das condições da paisagem em função da acessibilidade em campo (Roudier et al., 2012; Minasny & McBratney, 2006), que foi pré-definida como de 100 metros para cada lado das estradas. Os procedimentos empregados no levantamento de campo respeitaram as normas da EMBRAPA. As classes de saída correspondentes às nove unidades de mapeamento de solos (UMs) são identificadas na **Tabela 1**.

Com base nas observações de campo e análise dos padrões de terreno sobre os tipos de solo, foi coletado um conjunto de amostras de 500 *pixels* para cada unidade de mapeamento, 4.500 amostras no total. Tal conjunto e usado como *input* para ambos os modelos preditivos. Após o treinamento dos algoritmos foi feita a validação cruzada para obter a matriz de confusão e índices estatísticos (*kappa, overall* e variância), que servem de base para a avaliação dos classificadores.

Tabela 1. Descrição das unidades de mapeamento na (UMs) na bacia hidrográfica do rio Guapi-Macacu (RJ)

UMs	Descrição
PA	Argissolo Amarelo + Planossolo Háplico
PVA	Argissolo Vermelho-Amarelo
CX	Cambissolo Háplico
LA	Latossolo Amarelo
LVA	Latossolo Vermelho-Amarelo
GX	Gleissolo Háplico + Gleissolo Melânico
GS	Gleissolo Sálico + Gleissolo Tiomórfico
RY	Neossolo Flúvico
RL	Neossolo Litólico + Afloramento Rochoso

Os algoritmos testados foram árvores de decisão e *random forest*, e todos os procedimentos de classificação foram executados através de linhas de comando (*scripts*), o que requer a instalação de pacotes específicos para analises estatísticas no software R (R Development Core Team, 2013).

A predição pelo método *random forest* (RF) requer a instalação do pacote de ferramentas "randomForest" (Liaw & Wiener, 2002), e por sua vez para gerar as Árvores de Decisão (DT) é necessário instalar o pacote "rpart" (RDevelopment Core Team, 2008).

RESULTADOS E DISCUSSÃO

Análise das variáveis do terreno

O sucesso da modelagem do solo está diretamente relacionado com a qualidade dos dados de entrada e a coerencia das variáveis preditoras (Zhu, 2001; Minasny et al., 2003). Tendo isso em mente, a análise detalhada dos atributos do terreno é útil para compreender as relações solo-paisagem e para escolher um conjunto de variáveis relevante. A **figura 1** mostra o comportamento das variáveis por unidade de mapeamento.

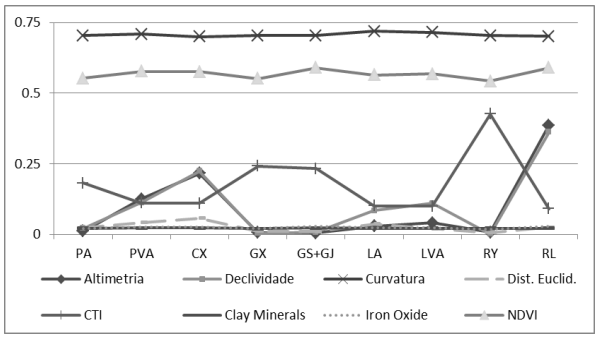

Figura 1. Padrões de variação dos atributos por UM

Com base nos gráficos e observações in situ, e possivel inferir que os Argissolos Amarelos ocorrem

comumente em encostas suaves, de baixa altitude; diferentemente dos Argissolos Vermelho-Amarelos, em cotas mais elevadas e com declives diversos. Os Latossolos Vermelho-Amarelos têm ampla ocorrência na area da bacia, assim como o espectro de condicoes de ocorrência, mas nota-se que predominam em formas de superfícies convexas e apresentam valores reduzidos para CTI. Padrão semelhante é observado nos Argissolos Vermelho-Amarelos, embora estejam intimamente associados a presença de rochas alcalinas. Os Cambissolos Háplicos dominam em formas côncavas, encostas íngremes e grandes altitudes, associados com Neossolos Litólicos e afloramentos de rocha.

Gleissolos Háplicos têm ampla distribuição na bacia hidrográfica, predominando em áreas planas de baixa altitude e declives suaves. Inclusões de Gleissolos Melanicos foram observadas em condições favoráveis ao aumento da matéria orgânica, como vegetação nativa ou depressões naturais onde o nível de lençol freático permanece elevado por longo período no ano. A outra unidade de Gleissolos ocorre sob influência de depósitos fluvio-marinhos, com caráter sálico no horizonte subsuperficial e algumas vezes o caráter tiomórfico. As principais diferenças entre estas duas unidades devem-se a declividade e o índice topográfico combinado, pois os Gleissolos Sálicos apresentam valores mais elevados para CTI e inferiores para declividade. Os Neossolos Flúvicos também mostram valores elevados para CTI e menor distância euclidiana da hidrografia, uma vez esta unidade ocorreu principalmente no limiar dos grandes drenos, como os rios Macacu e Guapi-Açu. Os índices derivados de sensoriamento remoto apresentaram diferencas para as classes de solo, principalmente por causa dos distintos tipos de usos da terra, conforme observado no campo.

Avaliação dos modelos de predição

A avaliação dos resultados é executada em primeiro lugar através de uma matriz de confusão. A comparação da matriz de confusão gerada para ambos os modelos é apresentada na **Tabela 2**.

Os melhores resultados para predição por ambos os métodos foram para as classes PVA e GX. Isso pode estar relacionado com as condições particulares de material de origem que ocorrem (rochas alcalinas e depósitos estuarinos, respectivamente) mostrando a importância da geologia como co-variável preditora.

Tabela 2. Matriz de confusao dos modelos AD, RF.

Arvore de Decisão - AD									
	CX	GS	GX	LA	LVA	PA	PVA	RL	RY
CX	393	0	0	10	60	0	0	37	0
GS	0	499	0	0	0	1	0	0	0
GX	0	0	383	0	0	41	0	0	76
LA	0	0	0	393	107	0	0	0	0
LVA	23	0	0	0	477	0	0	0	0
PA	0	0	72	57	40	278	0	0	53
PVA	0	0	0	0	0	2	498	0	0
RL	58	0	0	0	0	0	0	442	0
RY	0	0	0	23	0	1	0	0	476
Random Forest - RF									
	CX	GS	GX	LA	LVA	PA	PVA	RL	RY
CX	468	0	0	6	13	0	0	13	0
GS	0	499	1	0	0	0	0	0	0
GX	0	0	481	0	0	4	0	0	15
LA	0	0	0	466	31	2	0	0	1
LVA	2	0	0	3	495	0	0	0	0
PA	1	0	13	19	4	454	0	0	9
PVA	0	0	0	0	0	0	500	0	0
RL	10	0	0	0	0	0	0	490	0
RY	0	0	1	2	0	2	0	0	495

Os Argissolos e Latossolos Amarelos apresentaram, em geral, superposição, o que é comum até para pedólogos experientes, uma vez que podem ocorrer nas mesmas condições de paisagem, diferindo apenas na razão do incremento em argila. Nos modelos de árvores de decisão os Argissolos Amarelos tambem mostraram confusão com os Gleissolos Háplicos, onde podem ocorrer associados em areas de pedimento de colinas, embora os valores encontrados para CTI sejam maiores para a segunda classe.

Os Gleissolos Háplicos mostram confusão especialmente com Neossolos Flúvicos, pois ambos são influenciados por processos hidromórficos. Cambissolos Háplicos apresentam confusão com Neossolos Litólicos, ambos ocorrendo nos topos de encostas em declives acentuados. Os Latossolos Amarelos e Vermelho-Amarelos apresentaram confusão entre si, sendo o ultimo de ampla distribuição e variabilidade de condições de ocorrência na área de estudo. Estes solos tambem apresentam confusão com Cambissolos Háplicos, sendo que ambos podem ocorrem em relevo movimentado nos maciços costeiros.

A avaliação do desempenho dos algoritmos baseia-se nos valores para o índice *kappa* e *overall*. Melhores valores foram encontrados para o classificador RF (*overall*= 0,9662222; *Kappa*= 0,96200), em comparação com o modelo DT (*overall*= 0,8531111; *Kappa*= 0,83475). Embora os índices obtidos para ambos os classificadores sejam considerados excelentes (Landis & Koch, 1977; Monserud & Leemans, 1992). Sendo assim, foi selecionado o produto gerado pelo classificador

random forest para representar as unidades de mapeamento de solo.

CONCLUSÕES

A análise das co-variáveis ambientais com base nas observações de campo permitiu identificar relações consistentes entre as caracteristicas da paisagem e ocorrência de tipos de solo.

Com base nos conceitos clássicos de formação do solo, através da escolha de variáveis preditoras relevantes pedogeneticamente e ferramentas modernas de modelagem, a abordagem adotada possibilitou a criacao de mapa de solos em plataforma SIG, com erro de predicao conhecido.

O mapa escolhido para representar os solos na bacia hidrográfica foi o inferido pelo classificador *random forest*, que apresentou melhor desempenho para os índices estatísticos.

REFERÊNCIAS

CARVALHO JÚNIOR, W, CHAGAS, C. S., FERNANDES, E. I., VIEIRA, C. E., SCHAEFER, C. E. G., BHERING, S. B., FRANCELINO, M. R. Digital soilscape mapping of tropical hillslope areas by neural networks. Scientia Agricola. (Piracicaba) 68(6): 691-696, 2011.

CHAGAS, C. S., CARVALHO JÚNIOR, W., BHERING, S. B. Integração de dados do Quickbird e atributos do terreno no mapeamento digital de solos por redes neurais artificiais. R. Bras. Ci. Solo. 35:693-704. 2011.

ECOLOGUS- AGRAR. Plano Diretor dos Recursos Hídricos da Região Hidrográfica da Baía de Guanabara. Rio de Janeiro. RJ. 2003. 3087 p. CD-ROM.

GEOGRAPHIC RESOURCES ANALYSIS SUPPORT SYSTEM -GRASS. 1999-2013. Disponível em: grass.osgeo.org/home/copyright (acesso: Maio 2014).

HENGL, T. E., HEUVELINK, G. B. M. New Challenges for Predictive Soil Mapping. In: Global Workshop on Digital Soil Mapping. Montpellier, France. 2004. 14-17.

HUTCHINSON, M. F. Development of continent-wide DEM with applications to terrain and climate analysis. In: M.F. Goodchild, editor, Environmental Modeling with GIS. New York: Oxford University Press. 392-399, 1993.

JASIEWICZ, J., T.F. STEPINSKI. Geomorphons - a pattern recognition approach to classification and mapping of landforms. Geomorphology 182:147-156, 2013.

KÖPPEN, W. Climatologia: con un estudio de los climas de la tierra. Fondo Cultura Econômica. Panuco, México. 1948. 479p.

LANDIS, J. R., KOCH, G. G. The measurement of observer agreement for categorical data. Biometrics 33:159-174. 1997.

LIAW A., WIENER, M. Classification and regression by randomForest. R News. 2(3):18–22, 2002.

MATHWORKS. Matlab Releases 14. The mathwork Inc. Natick. MA. 2005. Disponivel em: www.mathworks.com/matlabcentral/fileexchange/4352-latin-hypercube-sampling (Acesso em 13 Fevereiro 2011).

MINASNY. B., MCBRATNEY, A. B., SANTOS, M. L., SANTOS, H. G.. Revisão sobre funções de pedotransferência (PTFs) e novos métodos de predição de classes de solos e atributos do solo. (In Portuguese, with English Abstract). (Embrapa Solos. Documentos n. 45). Rio de Janeiro, RJ. 2003. 50 p.

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

MINASNY. B., MCBRATNEY, A. B. A conditioned Latin hypercube method for sampling in the presence of ancillary information. Computers & Geosciences, 32:1378-1388, 2006.

MONSERUD, R. A., LEEMANS, R. Comparing global vegetation maps with the Kappa statistic. Ecological Modelling, 62: 275-293, 1992.

PROJETO MACACU. Planejamento Estratégico da Região Hidrográfica dos Rios Guapi-Macacu e Caceribu-Macacu. Niteroi. RJ: UFF/FEC. 2010. 544p.

R DEVELOPMENT CORE TEAM. R: A language and environment for statistical computing, R Foundation for Statistical Computing. Vienna, Austria. 2013. Disponivel em: http://www.r-project.org (Acesso em 15 Junho 2014).

ROUDIER, P., D.E. BEAUDETTE, A.E. HEWITT. A conditioned Latin hypercube sampling algorithm incorporating operational constraints. Digital Soil Assessments and Beyond: Proceedings of the 5th Global Workshop on Digital Soil Mapping. Sydney, Australia. 10-13. 2012.

TEN CATEN, A. Mapeamento digital de solos: metodologias para attender a demanda por informacao especial em solos. Tese de Doutorado. Universidade Federal de Santa Maria, RS. 2011. 106p.

ZHU, A. X. Soil mapping using GIS, expert knowledge and fuzzy logic. Soil Science Society of American Journal, 65:1463-1472. 2001.


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