Regressão linear múltipla e modelo Random Forest para estimar a densidade do solo em áreas montanhosas

Waldir de Carvalho Junior[1], Braz Calderano Filho[1], César da Silva Chagas[1], Silvio Barge Bhering[1], Nilson Rendeiro Pereira[1] e Helena Saraiva Koenow Pinheiro[2]

[1]Embrapa Solos, Rua Jardim Botânico, nº1.024, Jardim Botânico, CEP 22460-000 Rio de Janeiro, RJ, Brasil. E-mail: waldir.carvalho@embrapa.br, braz.calderano@embrapa.br, cesar.chagas@embrapa.br, silvio.bhering@embrapa.br, nilson.rendeiro@embrapa.br [2]Universidade Federal Rural do Rio de Janeiro, Departamento de Solos, BR-465, Km 47, CEP 23890-000 Seropédica, RJ, Brasil. E-mail: lenask@gmail.com

Resumo – O objetivo deste trabalho foi o desenvolvimento de modelos com diferentes conjuntos de dados, para estimar a densidade de solos de regiões tropicais montanhosas, a partir de atributos de solos comumente encontrados nas análises de perfis de solos descritos nos levantamentos regionais. O conjunto total de dados compõe-se de 163 amostras e foi dividido em seis grupamentos, dos quais três com 73 amostras, com o máximo de 32 covariáveis, e três com 163 amostras, com o máximo de 18 covariáveis. Testaram-se modelos de regressão linear múltipla (RLM) e randomForest (RF). A menor incerteza entre os modelos foi alcançada pelo RLM2, com R^2 de 0,56, 13 covariáveis e 73 amostras. Nos grupamentos com 163 amostras, os melhores modelos foram os RF, com R^2 médio de 0,48. A raiz quadrada da média do erro ao quadrado variou entre 0,09 e 0,14. As covariáveis mais importantes no modelo RF foram: carbono orgânico, hidrogênio, areia fina e grossa, saturação por bases e capacidade de troca catiônica. Pelo método "stepwise regression", as variáveis mais importantes foram: a relação silte/argila; areia grossa e fina; carbono orgânico; saturação por bases; e potássio.

Termos para indexação: estoque de carbono, funções de pedotransferência, modelos dirigidos pelos dados, stepwise.

Multiple linear regression and Random Forest model to estimate soil bulk density in mountainous regions

Abstract – The objective of this work was the development of models with different sets of data for estimating soil bulk density in tropical mountainous regions, from soil attributes commonly found in the analyses of soil profiles described in regional surveys. The complete dataset is composed of 163 samples and it was divided into six groups, of which three groups have 73 samples and the maximum of 32 covariables, and three have 163 samples and the maximum of 18 covariables. The linear regression (RLM) and randomForest (RF) models were tested. The lowest uncertainty between the models was achieved by RLM2, with R^2 of 0.56, 13 covariables, and 73 samples. Considering the groups with 163 samples, the best models were the RFs with mean R^2 of 0.48. The root mean squared error ranged between 0.09 and 0.14. The most important covariables in the RF model were: organic carbon, hydrogen, fine and coarse sand, base saturation, and cation exchange capacity. By the stepwise backward regression, the main covariables were: silt and clay relation; fine and coarse sand; organic carbon; base saturation; and potassium.

Index terms: carbon stock, pedotransfer functions, data-driven models, stepwise.

Introdução

A informação espacial de solos, inclusive classes e atributos, é fundamental na formulação de políticas agrícolas sustentáveis e no monitoramento de impactos causados pelo uso inadequado deste recurso, considerado patrimônio natural. A falta desta informação pode levar à adoção de políticas inadequadas e insustentáveis, aumentando o risco de degradação ambiental e perda de biodiversidade (Mulder et al., 2011).

A Convenção-Quadro das Nações Unidas sobre Mudança do Clima (UNFCCC), criada na Conferência das Nações Unidas sobre Meio Ambiente e Desenvolvimento ocorrida no Rio de Janeiro em 1992, aborda o comprometimento dos países de prover inventários nacionais de emissão de gases do efeito estufa e, para o setor agropecuário, inventários de estoque de carbono (Milne et al., 2007).

A densidade do solo (Dsol) é necessária para converter o conteúdo de carbono orgânico do solo (COS) em massa de carbono orgânico por unidade de

área e é, portanto, obrigatória para comparar diferentes tipos de uso do solo, quanto ao potencial de sequestro de carbono (Brahim et al., 2012). Além disso, Ellert et al. (2002) mostraram que, para medir, monitorar e verificar o sequestro de carbono pelo solo, é necessário calcular o estoque de carbono em uma equivalência de massa, o qual leva em consideração a densidade do solo. Entretanto, a análise da densidade do solo é trabalhosa, onerosa, envolve infraestrutura laboratorial e recursos humanos capacitados, particularmente quando a investigação de solos envolve grandes regiões. A importância da medição da densidade do solo não se restringe apenas à avaliação de estoque de carbono, é importante também para a estimativa de estoque de elementos no solo em geral, como os macronutrientes. Ademais, a Dsol varia conforme o tipo de solo, manejo, cobertura vegetal, etc.

O uso de modelos estatísticos para a estimativa de atributos do solo vem sendo amplamente utilizado, com o intuito de prever um valor a partir da correlação que existe entre a variável e outras variáveis ou covariáveis (Benites et al., 2007; Suuster et al., 2011; Brahim et al., 2012; Carvalho Junior et al., 2013; Nanko et al., 2014; Nasri et al., 2015; Rodríguez-Lado et al., 2015). Assim, os modelos estatísticos têm sido amplamente empregados para estimar o valor de algum atributo, ao invés de realizar a análise propriamente dita. Isto otimiza o processo de coleta e análise e os recursos materiais, humanos e econômicos.

Os modelos estatísticos usados para estimar atributos dos solos são chamados de funções de pedotransferência (FPT), e vêm sendo cada vez mais utilizados para cobrir a falta de informações sobre determinadas propriedades dos solos. Segundo Minasny & Hartemink (2011), é praticamente impossível medir continuamente as propriedades dos solos, em cada local no globo terrestre, o que torna necessário o uso de sistemas robustos para estimar o valor de atributos não determinados em diferentes localidades, principalmente em regiões tropicais, onde a falta de informações pedológicas é grande. Isso também se aplica quanto à estimação de propriedades dos solos a profundidades não amostradas, que denotam funções de profundidade.

O conceito de FPT vem sendo usado para estimar propriedades dos solos que são difíceis de determinar, em razão do custo e tempo necessário, ou mesmo para diminuir o número de análises laboratoriais. A estimativa é feita a partir de atributos do solo comumente encontrados nos bancos de dados de solos. Assim, estudos focados neste tema vêm sendo disponibilizados pela sociedade científica desde o 7º Congresso Internacional de Ciência do Solo, em que De Leenheer & Van (1960), citado por Minasny & Hartemink (2011), questionaram se seria possível predizer alguma característica física do solo, a partir de seus componentes dos solos. Segundo Benites et al. (2007), a medição da densidade do solo é essencial para estimar a reserva de carbono do solo. Entretanto, a amostragem de campo, especialmente a profundidades variadas, e a medição direta da densidade requerem trabalho intenso, demorado e, muitas vezes, tornam-se impraticáveis.

Segundo De Vos et al. (2005), no entanto, as FPT mostram grandes diferenças de performance, quando aplicadas a ambientes diferentes daqueles em que foram calibradas, o que reforça a necessidade de construir FPTs para cada atributo, em ambientes específicos.

O objetivo deste trabalho foi o desenvolvimento de modelos com diferentes conjuntos de dados, para estimar a densidade de solos de regiões tropicais montanhosas, a partir de atributos de solos comumente encontrados nas análises de perfis de solos descritos nos levantamentos regionais.

Material e Métodos

O presente estudo foi desenvolvido para a região de Mar de Morros, especificamente para o Município de Bom Jardim, RJ, na Mesorregião Centro Fluminense, Microrregião de Nova Friburgo, entre 22º 06' e 22º 18' S e 42º 12' e 42º 30' W (Figura 1). Sua área total é de 385,04 km^2, com relevo forte-ondulado, altitudes de 405 a 1.630 m e declividade média de 38%.

A densidade do solo é definida como a relação entre a massa de sólidos secos do solo e seu volume total, conforme a equação $D_s = m_s/V_s$, em que: D_s é a densidade do solo; m_s é a massa do solo seco; e V_s é o volume do solo; e a unidade de medida grama por centímetro cúbico.

A determinação da densidade é obtida pela medição direta dessas duas variáveis (massa e volume). O método usual para a determinação da densidade envolve a obtenção de uma amostra de volume, por meio de anéis volumétricos inseridos no solo com o uso de equipamento apropriado. A massa da amostra é obtida por pesagem em balança analítica, após a remoção da umidade em estufa a 105°C, até a obtenção

de massa constante. Este método, denominado "método do anel volumétrico", é o mais usado em trabalhos de avaliação da densidade de solo (Blake & Hartge, 1986; Claessen, 1997).

Para compor o banco de dados de solos, coletaram-se amostras de solos de 125 locais, conforme o SIBCS (Santos et al., 2005), no total de 579 amostras. Entre estas, 163 amostras com determinação de densidade do solo foram empregadas na elaboração dos modelos, e as demais 416, como teste na aplicação dos modelos, avaliação qualitativa das estimativas e predição final.

O conjunto de dados com 163 amostras foi dividido em dois subconjuntos, um com as 163 amostras e o máximo de 18 covariáveis, e outro com 73 amostras e o máximo de 32 covariáveis (Tabela 1). Cada um destes dois subconjuntos foi subdividido em três grupos e testados pelos modelos de regressão linear e randomForest, no total de seis grupos de dados. Esses grupos de dados diferenciam-se quanto ao número de amostras e de covariáveis envolvidas. Os três primeiros grupos contam com 73 amostras e 32 covariáveis (todas as possíveis), ou 12 covariáveis selecionadas pela regressão linear stepwise backward, conforme adotado por Vasques et al. (2008), Poggio et al. (2013) e Samuel-Rosa et al. (2015), entre outros, ou 13 covariáveis definidas pelo limiar menor ou igual a 0,05 para o valor p da correlação. Estes grupos foram validados pela validação cruzada "leave one out". Os outros três grupos são compostos de 163 amostras e 18 covariáveis, ou 6 covariáveis – seleção regressão linear "stepwise backward" –, ou 7 covariáveis (valor p da correlação menor ou igual a 0,05); estes grupos foram divididos para o ajuste dos modelos (131 amostras) e a validação (32 amostras).

Inicialmente, foram obtidos os valores de estatística básica da densidade, representados por média, desvio padrão, máximo, mínimo, mediana e valores dos 1.º e 3.º quartis e foi observada a distribuição de frequência.

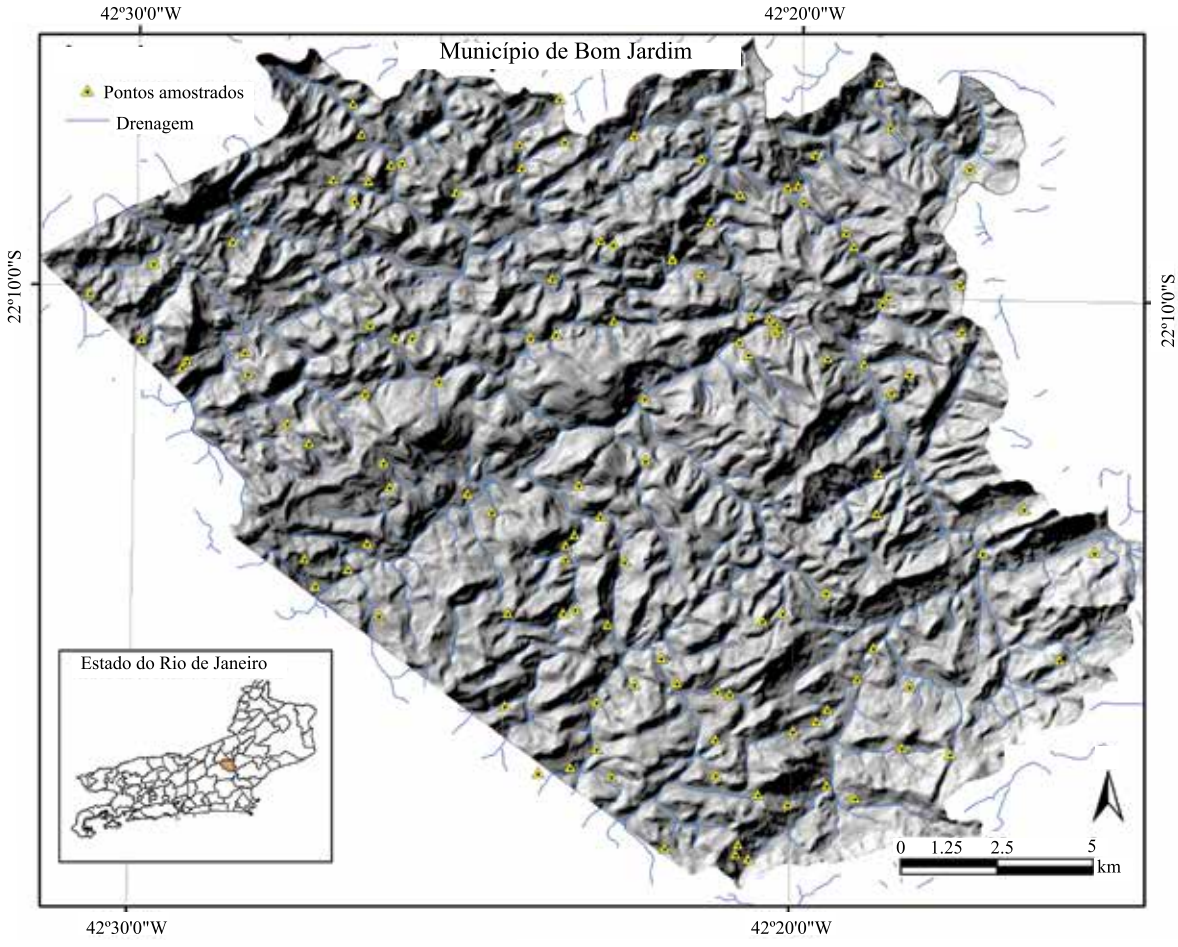

Figura 1. Localização da área de estudo e de pontos amostrados, no Município de Bom Jardim, RJ.

Em seguida, foram elaborados 12 modelos. Destes, seis modelos foram feitos com os subconjuntos de 73 amostras, conforme a seguir: regressão linear múltipla (RLM1) e Random Forest (RF1), com todas as covariáveis; RLM2 e RF2, uso de "stepwise", com a opção "direction=backward"; e RLM3 e RF3, com as covariáveis com valor p do teste de correlação e a densidade menor ou igual a 0,05. Os outros seis modelos foram elaborados com os grupamentos do conjunto total, conforme a seguir: RLM4 e RF4, com todas as covariáveis; RLM5 e RF5, uso de stepwise com a opção "direction=backward"; e RLM6 e RF6, com as covariáveis que possuem valor p do teste de correlação, com a densidade menor ou igual a 0,05 (Tabela 1).

A RLM é um método clássico largamente usado para o desenvolvimento de FPTs, que permite entender a relação entre as variáveis e explorar as formas dessas relações. A RLM foi implementada no R (The R Foundation, 2013), através da função lm, com uso de seleção passo a passo ("backward stepwise"), a 95% de confiança. A correlação entre as covariáveis e a densidade do solo foi avaliada pela função cor.test.

Desenvolvida por Breiman (2001) como extensão do programa CART (Classification and Regression Trees), a RF é uma técnica não paramétrica, que busca melhorar o desempenho de predição deste modelo. Trata-se de uma combinação de muitas árvores preditoras (floresta), em que cada árvore é gerada a partir de um vetor aleatório, amostrado de forma independente e com a mesma distribuição para todas as árvores na floresta. As subdivisões dentro de cada árvore são determinadas com base em um subconjunto de covariáveis, escolhido aleatoriamente a partir do total existente. No caso de RF para regressão, o resultado final consiste da média dos resultados de todas as árvores (Breiman, 2001; Cutler et al., 2007).

As RFs foram implementadas pelo pacote Random Forest do R (The R Foundation, 2013). Para a utilização de uma RF, os seguintes três parâmetros precisam ser definidos: o número de árvores (ntree), o número mínimo de dados em cada nó terminal (nodesize) e o número de variáveis utilizadas em cada árvore (mtry) (Liaw & Wiener, 2002). O padrão para ntree definido no sistema é de 500. Embora resultados mais estáveis possam ser alcançados com um número maior (Grimm et al., 2008), testes preliminares mostraram que o aumento do ntree não melhora o desempenho do modelo; assim, optou-se por utilizar valores de 70, 120 e 500, conforme os testes preliminares da performance dos modelos. Para o valor de "nodesize", utilizou-se o padrão de cinco para cada nó terminal. Quanto ao mtry, para problemas de regressão, o valor-padrão estipulado é de um terço do número total de variáveis preditoras (Liaw & Wiener, 2002); assim, utilizaram-se valores de mtry iguais a 10, para as variáveis com 73 amostras, e 6 e 3 para aquelas com 131 amostras.

O modelo RF fornece estimativas confiáveis dos erros, utilizando dados conhecidos como "out-of-bag" (OOB), que é um subconjunto aleatório dos dados não

Tabela 1. Subconjuntos de dados usados para desenvolver as funções de pedotransferência.

Modelo	Número de amostras	Número de covariáveis	Covariáveis
RLM1 e RF1	73	32	Calhaus, cascalho, terra fina, areia grossa, areia fina, silte, argila, argila dispersa, grau de floculação, relação silte/argila, pH em água, pH em KCl, Ca+Mg, potássio, sódio, soma de bases, alumínio, hidrogênio, capacidade de troca catiônica (T), saturação por bases (V), saturação de alumínio, fósforo, carbono orgânico, nitrogênio, relação C/N, SiO_2, Al_2O_3, Fe_2O_3, TiO_2, Ki, Kr, Al_2O_3/ Fe_2O_3.
RLM2 e RF2	73	12	Calhaus, silte, relação silte/argila, pH em água, pH em KCl, V, fósforo, carbono orgânico, SiO_2, Al_2O_3, TiO_2, Ki.
RLM3 e RF3	73	13	Areia grossa, areia fina, silte, argila, argila dispersa, alumínio, carbono orgânico, nitrogênio, relação C/N, Al_2O_3, TiO_2, Ki, Kr.
RLM4 e RF4	163	18	Areia grossa, areia fina, silte, argila, relação silte/argila, pH em água, pH em KCl, Ca+Mg, potássio, sódio, soma de bases, alumínio, hidrogênio, T, V, saturação por alumínio, fósforo, carbono orgânico.
RLM5 e RF5	163	6	Areia grossa, areia fina, relação silte/argila, potássio, V, carbono orgânico.
RLM6 e RF6	163	7	Areia grossa, areia fina, silte, alumínio, hidrogênio, T, carbono orgânico.

RLM, regressão linear múltipla; RF, Random Forest. Covariáveis em (g kg^{-1}): calhaus, cascalho, terra fina, areia grossa, areia fina, silte, argila, argila dispersa em água, C orgânico, N, SiO_2, Al_2O_3, Fe_2O_3, TiO_2; em ($cmol_c$ kg^{-1}): Ca+Mg, K, Na, valor S, Al, H e T; em (%): grau de floculação, V e saturação por Al; adimensionais: relação C/N, relação silte/argila, pH em água e em KCl, Ki, Kr, e relação Al_2O_3/Fe_2O_3; e P (mg kg^{-1}).

utilizado pelo algoritmo para construção das árvores. A partir das predições OOB de todas as árvores na floresta, calcula-se o erro quadrado médio (MSE_{OOB}), conforme Liaw & Wiener (2002), pela seguinte equação:

$$MSE_{OOB} = n^{-1}\sum_{i=1}^{n}(z_i - \hat{z}_i^{OOB})^2,$$

em que: zi é o valor medido da variável; e $\hat{z}_i^{OOB}$ é a média de todas as predições OOB. No entanto, como MSE é dependente da escala de medida da variável, não pode ser usado para comparação da performance de diferentes modelos; assim, calcula-se a percentagem de variância explicada pelo modelo (Varex), conforme Liaw & Wiener (2002), pela seguinte equação: $Var_{ex} = 1 - (MSE_{OOB}/Var_z)$; em que Var_z é a variância total da variável.

Os coeficientes de determinação (R^2) e a raiz quadrada da média do erro ao quadrado (RMEQ) foram utilizados para avaliar os modelos. Este mesmo procedimento de avaliação foi adotado por Taalab et al. (2015), Aitkenhead & Coull (2016), Jiang et al. (2016), Malone et al. (2016) e Vågen et al. (2016), entre outros. Os modelos de RLM desenvolvidos com 73 amostras foram avaliados pelo R^2 do modelo e por validação cruzada "leave one out". Os modelos RF com 73 amostras foram avaliados pela Var_{ex}. Para os modelos desenvolvidos com os grupamentos do conjunto total de amostras, criou-se, aleatoriamente, um subconjunto de 32 amostras para validação, com o pacote estatístico R (The R Foundation, 2013).

Resultados e Discussão

O histograma da distribuição de frequência mostrou equivalência entre os dois subconjuntos de dados (Figura 2), com maior expressão de valores entre 1,2 e 1,3 g cm^{-3}. O teste t de Student e o de Kolmogorov-Smirnov mostraram valor p de 0,09 e 0,16, respectivamente, o que indica que não há diferença significativa entre os grupamentos a 95% de confiança. Apesar da similaridade dos grupamentos de amostras, os modelos desenvolvidos apresentaram resultados distintos, possivelmente inerentes aos modelos dirigidos pelos dados. A Tabela 2 mostra os valores da estatística básica para os dados de densidade dos dois subconjuntos.

A primeira função estatística de regressão linear múltipla (RLM1) utilizou todas as 32 covariáveis e apresentou um coeficiente de determinação (R^2) de ajuste do modelo igual a 0,75. A partir da RLM1, utilizou-se a opção "stepwise backward" para selecionar as variáveis mais importantes e criar a RLM2 com 12 covariáveis (Tabela 1). Para a RLM2, o valor de R^2 de ajuste do modelo foi de 0,72. Estas duas regressões não apresentaram diferença significativa pela análise de variância entre as regressões. A RLM3 desenvolvida com 13 covariáveis apresentou R^2 de ajuste do modelo de 0,60 (Tabela 3). As funções RLM4, RLM5 e RLM6, que usaram os grupamentos do conjunto total de amostras, tiveram valores R^2 inferiores e diferentes estatisticamente daqueles modelos de regressão com 73 amostras, quando se compara aquelas de mesma quantidade e qualidade de covariáveis.

Os resultados dos testes de validação das funções de regressão linear, tanto por validação cruzada

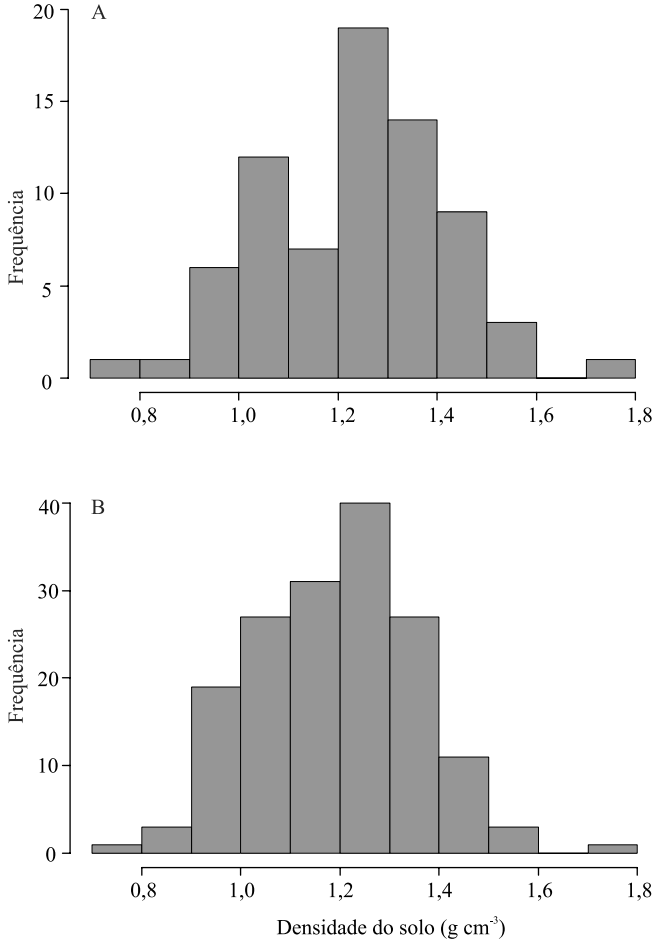

Figura 2. Distribuição de frequência dos valores de densidade do solo para os dois subconjuntos de dados de 163 (A) e 73 (B) amostras.

quanto com amostras independentes, apresentaram R^2 inferiores ao do ajuste dos modelos e significativamente diferentes destes, o que é indicativo de que os melhores resultados de validação foram os que utilizaram as covariáveis selecionadas pela opção "stepwise backward", que foram de 0,56 para RLM2 e 0,25 para RLM5 (Tabela 3). Para estes modelos, as variáveis da fração granulométrica (com exceção da argila), o valor V e o carbono orgânico mostraram-se importantes na estimativa da Dsol. Associados a essas variáveis, SiO_2, Al_2O_3, TiO_2, Ki, P e K também contribuíram com um ou outro modelo.

Além disso, destaca-se a pouca variação entre os modelos desenvolvidos com randomForest, em razão dos valores de R^2 da validação, provavelmente em consequência do próprio método, que faz a estimativa pela média dos nós terminais. Nota-se, entretanto, que os modelos RF4, RF5 e RF6 apresentaram valores de R^2 de validação superiores e estatisticamente diferentes daqueles do ajuste dos modelos.

Os valores de R^2 de ajuste do modelo, comparados aos de validação, para todos os modelos avaliados, mostram que os modelos RLM apresentaram, para todos os conjuntos de dados, valores de R^2 da validação inferiores aos do ajuste (Figura 3). No entanto, os modelos RF mostraram valores de R^2 da validação superiores aos do ajuste. Além disso, os valores R^2 da validação dos modelos randomForest foram superiores também àqueles dos modelos de regressão linear, com exceção do modelo RLM2.

Os resultados do coeficiente de determinação para a validação mostraram que a função que apresentou a menor incerteza foi a RLM2, seguida pelos modelos RF4, RF5 e RF6, os quais apresentaram resultados semelhantes. Observa-se, entre os valores observados e estimados por estes quatro modelos, menor dispersão da estimativa feita pela regressão linear RLM2 (Figura 4). Estes resultados mostram que usar um grande número de covariáveis para modelos de regressão linear não significa uma melhoria do modelo ou baixa incerteza.

Os valores de RMEQ foram ligeiramente superiores nos modelos RF, o que indica que os modelos de regressão linear apresentam menor erro na estimativa dos valores de densidade do solo (Figura 5). Para os modelos randomForest, os valores de RMEQ tiveram uma variação muito pequena – entre 0,13 e 0,14 –, o que mostra que este tipo de modelo é pouco suscetível à variação da quantidade de amostras e, mesmo, da quantidade e qualidade das covariáveis.

Tabela 2. Estatísticas descritivas dos dados analíticos de densidade do solo (g cm^{-3}), para os dois conjuntos de dados de 73 e 163 amostras de solo.

Conjunto	Mínimo	1° Quartil	Mediana	Média	3° Quartil	Máximo	Desvio-padrão
73	0,71	1,095	1,26	1,232	1,35	1,72	0,181
163	0,71	1,07	1,21	1,189	1,304	1,72	0,163

Tabela 3. Resultados da avaliação dos modelos RLM e RF, com o coeficiente de determinação e o RMEQ.

Modelo	n	R^2 modelo	R^2 validação	N.º de covariáveis	RMEQ
RLM1	73	0,75	0,33[(1)]	32	0,090
RLM2	73	0,72	0,56[(1)]	12	0,096
RLM3	73	0,60	0,43[(1)]	13	0,115
RLM4	131	0,41	0,18[(2)]	18	0,114
RLM5	131	0,38	0,25[(2)]	6	0,116
RLM6	131	0,37	0,22[(2)]	7	0,118
RF1	73	0,45	0,45[(1)]	32	0,134
RF2	73	0,44	0,44[(1)]	12	0,135
RF3	73	0,42	0,42[(1)]	13	0,137
RF4	131	0,23	0,48[(2)]	18	0,138
RF5	131	0,28	0,48[(2)]	6	0,134
RF6	131	0,24	0,46[(2)]	7	0,137

[(1)]Validação cruzada. [(2)]Validação com conjunto de amostras independentes. RLM, regressão linear múltipla; RF, randomForest; N, número de amostras; e RMEQ, raiz quadrada da média do erro ao quadrado.

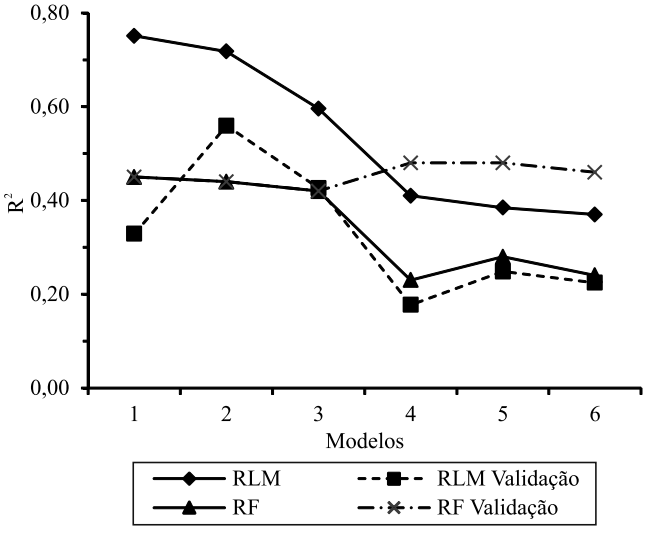

Figura 3. Visualização gráfica dos valores de R^2 dos modelos de regressão linear múltipla (RLM) e randomForest (RF) e das validações.

Para o conjunto de amostras em que a estimativa da densidade do solo não foi analisada, realizou-se o teste final quanto a esta propriedade, com todos os modelos (Tabela 4). Os valores mínimos negativos foram estimados por dois modelos de regressão linear (RLM1 e RLM2). Todos os modelos estimaram valores máximos inferiores àqueles dos conjuntos de dados empregados para os ajustes dos modelos.

Nos quatro modelos com menor incerteza, RLM2, RF4, RF5 e RF6, nota-se a presença de valores negativos para a estimativa pela função de regressão linear, enquanto os modelos randomForest apresentam estimativas com amplitudes dentro da faixa dos dados originais. Estes achados são inerentes aos modelos adotados (Figura 6).

A importância das covariáveis preditoras é um dos resultados fornecidos por seu próprio ranqueamento nos modelos Random Forest. Já as covariáveis mais importantes, nos modelos de regressão linear, são selecionadas pela função "stepwise backward", sem, entretanto, classificá-las. No ranqueamento das covariáveis em função de sua importância, as covariáveis carbono orgânico, areia fina, valor V, valor T, argila e hidrogênio foram as mais importantes nos modelos Random Forest, em que se destaca o carbono orgânico no modelo RF5. O hidrogênio foi a covariável mais importante para os modelos RF4 e RF6 (Figura 7).

Os resultados alcançados para as estimativas dos 416 horizontes corroboram aqueles alcançados por Suuster et al. (2011), quando os modelos de regressão linear múltipla subestimaram valores mais baixos de densidade. No entanto, os modelos RF superestimaram os menores valores e subestimaram os maiores.

Os valores de R^2 da validação para os modelos com menor incerteza, em torno de 0,5, estão abaixo dos encontrados por Nanko et al. (2014), que foram em torno de 0,60, que usou no máximo três covariáveis explanatórias, mas elaboradas por conjunto de dados

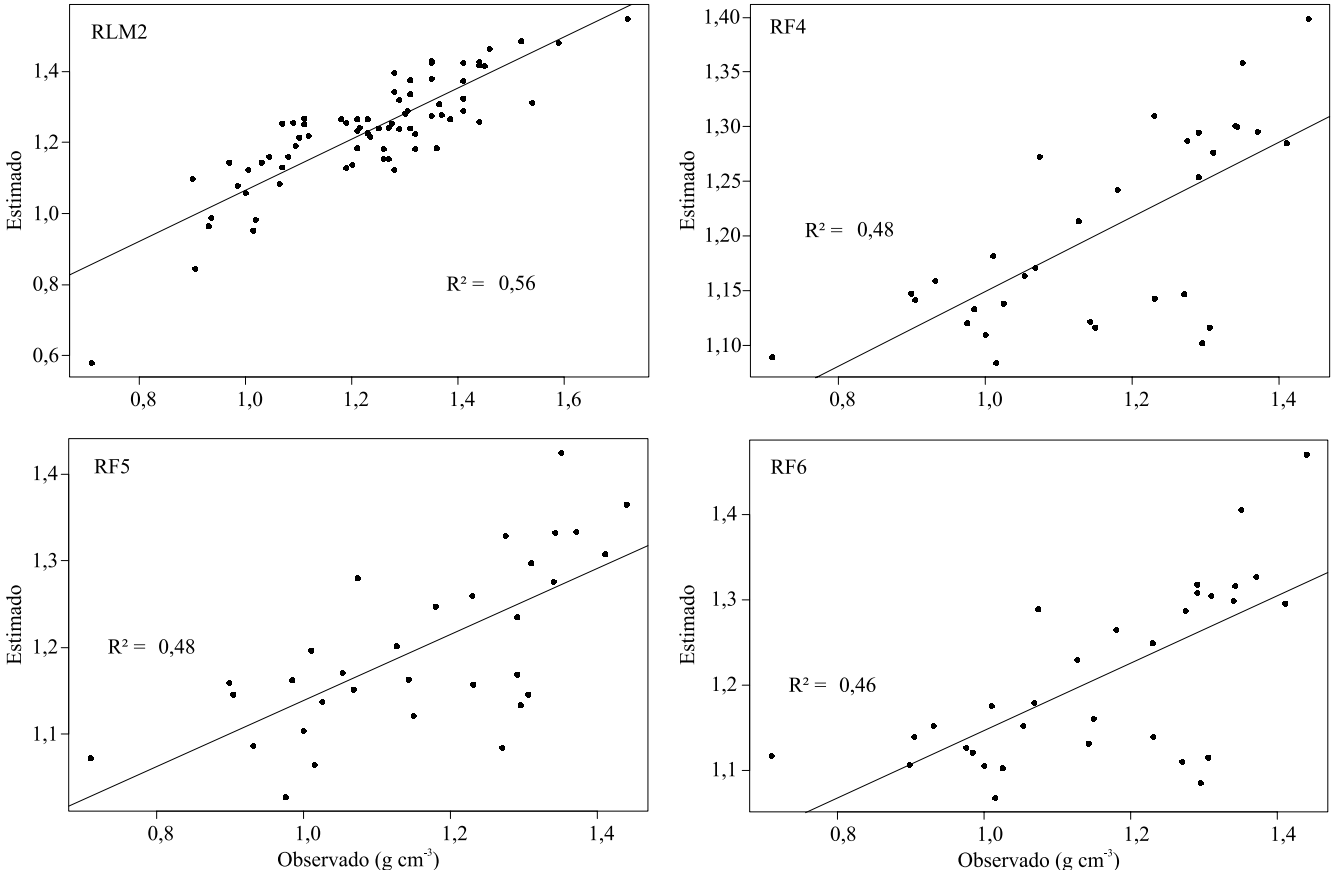

Figura 4. Dispersão entre os valores estimados e observados, para os modelos com melhor R^2 da validação.

com maior número de amostras e em condições ambientais distintas da tropicais.

Rodríguez-Lado et al. (2015) testaram três modelos diferentes (RLM, RF e redes neurais), em um conjunto com 115 observações, e encontraram melhores resultados para o modelo RF, com valores de R^2 de 0,90 e RMEQ de 0,14. No presente estudo, os valores de R^2 para os modelos RF4, RF5 e RF6 estão em torno de 0,47, com RMEQ de 0,09. Assim, por um lado, o R^2 da estimativa da densidade dos solos de Bom Jardim (RJ) foi a metade daquele obtido por Rodríguez-Lado et al. (2015), mas, por outro lado, o RMEQ representou 2/3 daquele obtido por Rodríguez-Lado et al. (2015). Destacam-se também as condições climáticas diferenciadas dos dois estudos.

Brahim et al. (2012) encontraram resultados semelhantes, em trabalho com 707 horizontes de solos da Tunísia e, com modelos de RLM, obtiveram R^2 igual 0,55. Adicionalmente, as principais covariáveis preditoras foram carbono orgânico, argila, areia grossa e pH. Ressaltando o caráter local dos modelos, os autores relatam que o RLM desenvolvido pode ser aplicado para regiões áridas e semiáridas.

Benites et al. (2007) encontraram valores de R^2 de 0,71 e 0,66 para modelos RLM, com base em mais de mil amostras distribuídas aleatoriamente pelo território brasileiro. Segundo estes autores, as covariáveis preditoras mais importantes foram nitrogênio, carbono orgânico, argila e soma de bases, corroborando o fato de que o carbono orgânico está presente na maioria dos modelos. O RMEQ encontrado por Benites et al. (2007) foi de 0,19, acima, portanto, do encontrado no presente trabalho, que variou de 0,09 a 0,14.

As principais covariáveis preditoras apontadas pelos modelos RF foram carbono orgânico, hidrogênio, areia

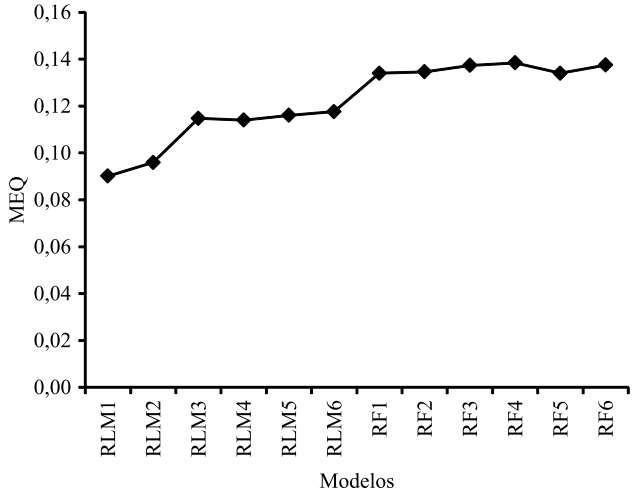

Figura 5. Representação gráfica da evolução da raiz quadrada da média do erro ao quadrado (RMEQ) dos modelos testados.

Tabela 4. Estatísticas básicas descritivas das estimativas de densidade do solo (g cm^{-3}), para o Município de Bom Jardim, RJ.

Modelo	Mínimo	1.º Quartil	Mediana	Média	3.º Quartil	Máximo	DP
73 amostras	0,710	1,095	1,260	1,232	1,350	1,720	0,18
163 amostras	0,710	1,070	1,210	1,189	1,304	1,720	0,16
RLM1	-3,681	1,184	1,291	1,261	1,393	1,675	0,49
RLM2	-0,528	1,211	1,294	1,272	1,385	1,631	0,26
RLM3	0,658	1,227	1,285	1,2760	1,336	1,499	0,17
RLM4	0,733	1,193	1,251	1,243	1,30	1,422	0,14
RLM5	0,757	1,189	1,25	1,238	1,292	1,426	0,10
RLM6	0,837	1,205	1,255	1,242	1,287	1,427	0,10
RF1	0,855	1,225	1,278	1,278	1,327	1,527	0,10
RF2	0,874	1,231	1,277	1,277	1,323	1,525	0,09
RF3	0,867	1,227	1,281	1,282	1,336	1,514	0,10
RF4	1,036	1,204	1,266	1,265	1,324	1,487	0,09
RF5	1,027	1,198	1,270	1,265	1,318	1,483	0,09
RF6	1,041	1,206	1,263	1,271	1,333	1,540	0,10

RLM, regressão linear múltipla; e RF, randomForest.

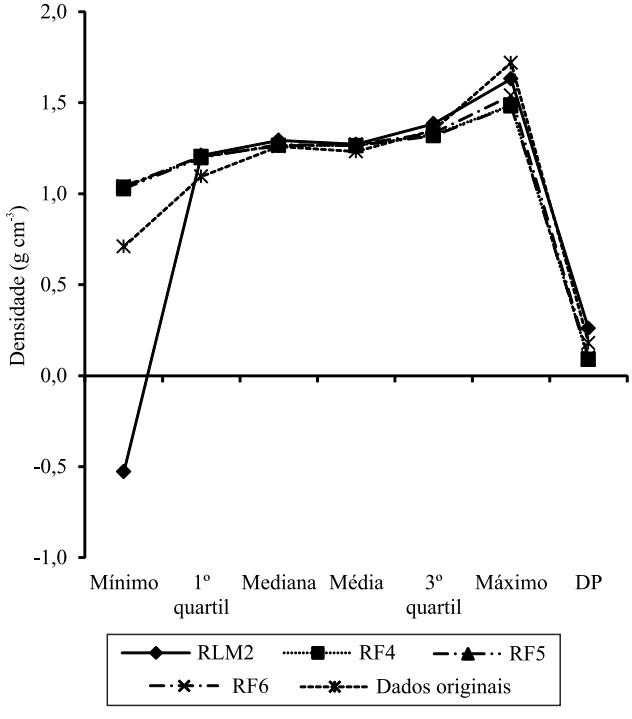

Figura 6. Valores de estatísticas básicas da estimativa dos quatro modelos com menor incerteza.

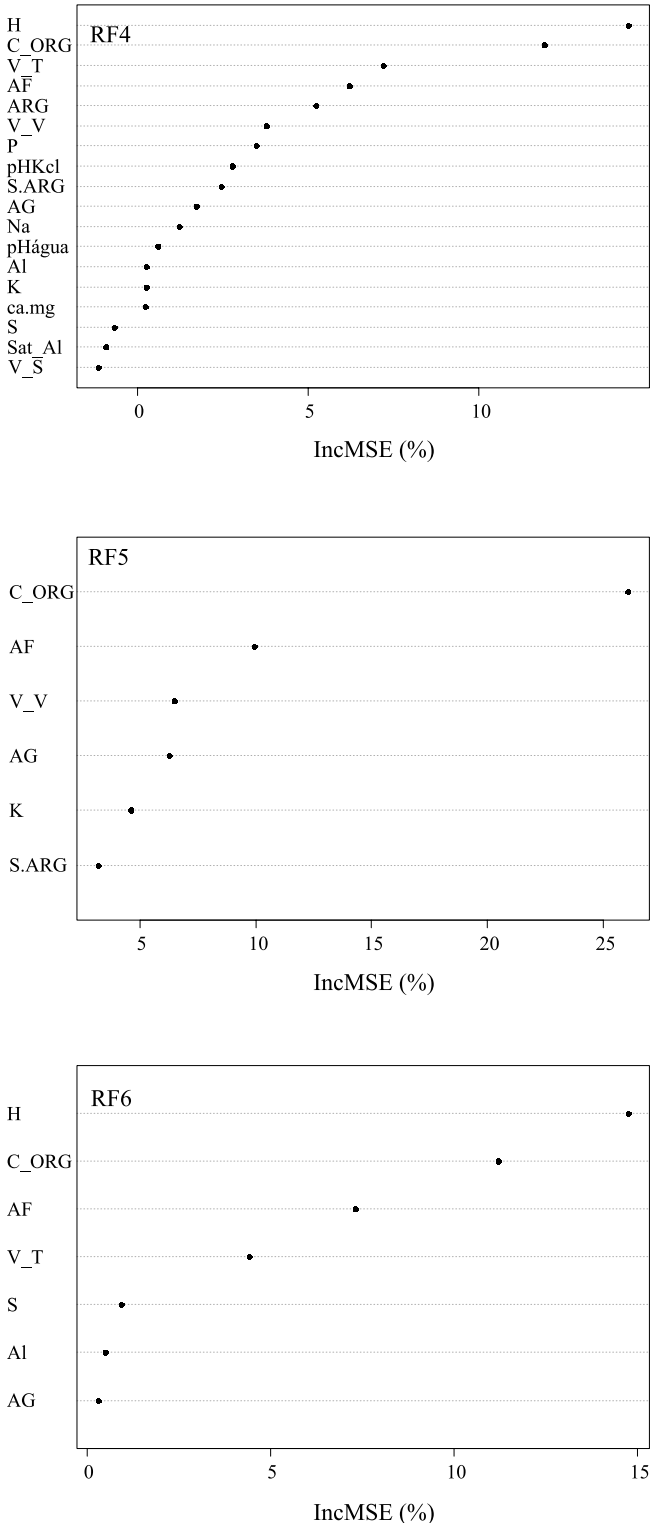

Figura 7. Importância das covariáveis preditoras, nos modelos RF4, RF5 e RF6. AG, areia grossa; AF, areia fina; S, silte; ARG, argila; S.ARG, relação silte/argila; ca.mg, Ca+Mg; V_S, soma de bases; V_T, valor T; V_V, valor V; Sat_Al, saturação por alumínio; C_ORG, carbono orgânico.

fina, areia grossa, valor V e valor T. Pelo método de regressão linear, com uso de "stepwise backward", as variáveis mais importantes foram, sem ordem de importância, a relação silte/argila; areia grossa e fina, carbono orgânico, valor V e potássio.

Conclusões

1. As covariáveis mais promissoras, no método randomForest, para estimar densidade de solo de regiões tropicais montanhosas, foram carbono orgânico, hidrogênio, areia fina e grossa, valor V e valor T; e no método de regressão linear múltipla com "stepwise regression", as covariáveis mais importantes foram relação silte/argila, areia grossa e fina, carbono orgânico, valor V e potássio.

2. O carbono orgânico e frações granulométricas estão presentes em todos os modelos do presente trabalho, o que reforça sua importância e correlação com a densidade do solo.

Referências

AITKENHEAD, M.J.; COULL, M.C. Mapping soil carbon stocks across Scotland using a neural network model. **Geoderma**, v.262, p.187-198, 2016. DOI: 10.1016/j.geoderma.2015.08.034.

BENITES, V.M.; MACHADO, P.L.O.A.; FIDALGO, E.C.C.; COELHO, M.R.; MADARI B.E. Pedotransfer functions for estimating soil bulk density from existing soil survey reports in Brazil. **Geoderma**, v.139, p.90-97, 2007. DOI: 10.1016/j.geoderma.2007.01.005.

BLAKE, G.R.; HARTGE, K.H. Bulk density. In: KLUTE, A. (Ed.). **Methods of soil analysis**: part 1 – physical and mineralogical methods. 2nd ed. Madison: Soil Science Society of America, American Society of Agronomy, 1986. p.363-375. (SSSA book series, 5).

BRAHIM, N.; BERNOUX, M.; GALLALI, T. Pedotransfer functions to estimate soil bulk density for northern Africa: Tunisia case. **Journal of Arid Environments**, v.81, p.77-83, 2012. DOI: 10.1016/j.jaridenv.2012.01.012.

BREIMAN, L. **Random forests**. 2001. Disponível em: <https://www.stat.berkeley.edu/~breiman/randomforest2001.pdf>. Acesso em: 28 dez. 2014.

CARVALHO JUNIOR, W. de; CHAGAS, C. da S.; CALDERANO FILHO, B.; BHERING, S.B. Funções de pedotransferência para estimativa da densidade dos solos de áreas tropicais montanhosas. In: CONGRESSO BRASILEIRO DE CIÊNCIA DO SOLO, 34., 2013, Florianópolis. **Anais**. Viçosa: Sociedade Brasileira da Ciência do Solo, 2013. 4p.

CLAESSEN, M.E.C. (Org.). **Manual de métodos de análise de solo**. 2.ed. rev. e atual. Rio de Janeiro: Embrapa-CNPS, 1997. 212p. (Embrapa-CNPS. Documentos, 1).

CUTLER, D.R.; EDWARDS JR., T.C.; BEARD, K.H.; CUTLER, A.; HESS, K.T.; GIBSON, J.; LAWLER, J.J. Random forests for classification in ecology. **Ecology**, v.88, p.2783-2792, 2007. DOI: 10.1890/07-0539.1.

DE VOS, B.; VAN MEIRVENNE, M.; QUATAERT, P.; DECKERS, J.; MUYS, B. Predictive quality of pedotransfer functions for estimating bulk density of forest soils. **Soil Science Society of America Journal**, v.69, p.500-510, 2005. DOI:10.2136/sssaj2005.0500.

ELLERT, B.H.; JANZEN, H.H.; ENTZ, T. Assessment of a method to measure temporal change in soil carbon storage. **Soil Science Society of America Journal**, v.66, p.1687-1695, 2002. DOI: 10.2136/sssaj2002.1687.

GRIMM, R.; BEHRENS, T.; MÄRKER, M.; ELSENBEER, H. Soil organic carbon concentrations and stocks on Barro Colorado Island - Digital soil mapping using Random Forests analysis. **Geoderma**, v.146, p.102-113, 2008. DOI: 10.1016/j.geoderma.2008.05.008.

JIANG, J.; ZHU, A.-X.; QIN, C.-Z.; ZHU, T.; LIU, J.; DU, F.; LIU, J.; ZHANG, G.; AN, Y. CyberSoLIM: a cyber platform for digital soil mapping. **Geoderma**, v.263, p.234-243, 2016. DOI: 10.1016/j.geoderma.2015.04.018.

LIAW, A.; WIENER, M. Classification and regression by random forest. **R News**, v.2, p.18-22, 2002.

MALONE, B.P.; JHA, S.K.; MINASNY, B.; MCBRATNEY, A.B. Comparing regression-based digital soil mapping and multiple-point geostatistics for the spatial extrapolation of soil data. **Geoderma**, v.262, p.243-253, 2016. DOI: 10.1016/j.geoderma.2015.08.037.

MILNE, E.; Al-ADAMAT, R.; BATJES, N.H.; BERNOUX, M.; BHATTACHARYYA, T.; CERRI, C.C.; CERRI, C.E.P.; COLEMAN, K.; EASTER, M.; FALLOON, P.; FELLER, C.; GICHERU, P.; KAMONI, P.; KILLIAN, K.; PAL, D.K.; PAUSTIAN, K.; POWLSON, D.; RAWAJFIH, Z.; SESSAY, M.; WILLIAMS, S.; WOKABI, S. National and sub-national assessments of soil organic carbon stocks and changes: The GEFSOC modelling system. **Agriculture, Ecosystems and Environment**, v.122, p.3-12, 2007. DOI: 10.1016/j.agee.2007.01.002.

MINASNY, B.; HARTEMINK, A.E. Predicting soil properties in the tropics. **Earth-Science Reviews**, v.106, p.52-62, 2011. DOI: 10.1016/j.earscirev.2011.01.005.

MULDER, V.L.; BRUIN, S. de; SCHAEPMAN, M.E.; MAYR, T.R. The use of remote sensing in soil and terrain mapping – a review. **Geoderma**, v.162, p.1-19, 2011. DOI: 10.1016/j.geoderma.2010.12.018.

NANKO, K.; UGAWA, S.; HASHIMOTO, S.; IMAYA, A.; KOBAYASHI, M.; SAKAI, H.; ISHIZUKA, S.; MIURA, S.; TANAKA, N.; TAKAHASHI, M.; KANEKO, S. A pedotransfer function for estimating bulk density of forest soil in Japan affected by volcanic ash. **Geoderma**, v.213, p.36-45, 2014. DOI: 10.1016/j.geoderma.2013.07.025.

NASRI, B.; FOUCHÉ, O.; TORRI, D. Coupling published pedotransfer functions for the estimation of bulk density and saturated hydraulic conductivity in stony soils. **Catena**, v.131, p.99-108, 2015. DOI: 10.1016/j.catena.2015.03.018.

POGGIO, L.; GIMONA, A.; BREWER, M.J. Regional scale mapping of soil properties and their uncertainty with a large number of satellite-derived covariates. **Geoderma**, v.209/210, p.1-14, 2013. DOI: 10.1016/j.geoderma.2013.05.029.

RODRÍGUEZ-LADO, L.; RIAL, M.; TABOADA, T.; CORTIZAS, A.M. A pedotransfer function to map soil bulk density from limited data. **Procedia Environmental Sciences**, v.27, p.45-48, 2015. DOI: 10.1016/j.proenv.2015.07.112.

SAMUEL-ROSA, A.; HEUVELINK, G.B.M.; VASQUES, G.M.; ANJOS, L.H.C. Do more detailed environmental covariates deliver more accurate soil maps? **Geoderma**, v.243-244, p.214-227, 2015. DOI: 10.1016/j.geoderma.2014.12.017.

SANTOS, R.D. dos; LEMOS, R.C. de; SANTOS, H.G. dos; KER, J.C.; ANJOS, L.H.C. dos; SHIMIZU, S.H. **Manual de descrição e coleta de solo no campo**. 5.ed. rev. e ampl. Viçosa: Sociedade Brasileira de Ciência do Solo, 2005. 100p.

SUUSTER, E.; RITZ, C.; ROOSTALU, H.; REINTAM, E.; KÕLLI, R.; ASTOVER, A. Soil bulk density pedotransfer functions of the humus horizon in arable soils. **Geoderma**, v.163, p.74-82, 2011. DOI: 10.1016/j.geoderma.2011.04.005.

TAALAB, K.; CORSTANJE, R.; ZAWADZKA, J.; MAYR, T.; WHELAN, M.J.; HANNAM, J.A.; CREAMER, R. On the application of Bayesian networks in digital soil mapping. **Geoderma**, v.259/260, p.134-148, 2015. DOI: 10.1016/j.geoderma.2015.05.014.

THE R FOUNDATION. **R**: the R project for statistical computing. Vienna: The R Foundation, 2013.

VÅGEN, T.-G.; WINOWIECKI, L.A.; TONDOH, J.E.; DESTA, L.T.; GUMBRICHT, T. Mapping of soil properties and land degradation risk in Africa using MODIS reflectance. **Geoderma**, v.263, p.216-225, 2016. DOI: 10.1016/j.geoderma.2015.06.023.

VASQUES, G.M.; GRUNWALD, S.; SICKMAN, J.O. Comparison of multivariate methods for inferential modeling of soil carbon using visible/near-infrared spectra. **Geoderma**, v.146, p.14-25, 2008. DOI: 10.1016/j.geoderma.2008.04.007.

Recebido em 28 de agosto de 2015 e aprovado em 8 de dezembro de 2016