Mapeamento digital de areia, argila e carbono orgânico por modelos Random Forest sob diferentes resoluções espaciais

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

[1]Embrapa Solos, Rua Jardim Botânico, nº 1.024, Jardim Botânico, CEP 22460-000 Rio de Janeiro, RJ, Brasil. E-mail: silvio.bhering@embrapa.br, cesar.chagas@embrapa.br, waldir.carvalho@embrapa.br, nilson.pereira@embrapa.br, braz.calderano@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 avaliar a influência da resolução espacial do modelo digital de elevação e da eficiência de modelos Random Forest sobre a predição dos teores de areia, argila e carbono orgânico, com uso de número reduzido de amostras. O trabalho foi realizado em área de Cerrado com diversidade litológica, no Estado do Mato Grosso do Sul, tendo-se utilizado atributos morfométricos, dados do sensor TM Landsat 5 e litologia como covariáveis preditoras. Dados da camada superficial (0,0–0,2 m) de 175 perfis de solos (0,009 perfis km^{-2}) e de 26 covariáveis preditoras foram utilizados com resolução espacial de 30 (conjunto 1) e 90 m (conjunto 2). A análise realizada pelo Random Forest mostrou que as covariáveis de nível de base do canal de drenagem, da elevação e da litologia foram as mais importantes para explicar a variabilidade. A validação dos modelos apresentou similaridade entre os conjuntos quanto à predição de areia, argila e carbono orgânico, o que explica os seguintes valores de variabilidade espacial, respectivamente: 44, 40 e 33%, para a resolução de 30 m; e de 45, 46 e 33%, para a resolução de 90 m. A resolução espacial das covariáveis preditoras tem pouca influência sobre a predição dos atributos, e a abordagem por Random Forest apresenta potencial de utilização para estimar atributos do solo.

Termos para indexação: modelo digital de elevação, morfometria, pedometria, SRTM.

Digital mapping of sand, clay, and soil carbon by Random Forest models under different spatial resolutions

Abstract – The objective of this work was to evaluate the effect of the digital elevation model spatial resolution and of the efficiency of Random Forest models on the prediction of sand, clay, and organic carbon contents, using few soil samples. The study was carried out in a Cerrado area with lithological diversity, in the state of Mato Grosso do Sul, Brazil, using morphometric attributes, TM Landsat 5 sensor data, and lithology as predictive covariates. The surface layer data (0.0–0.2 m) of 175 soil profiles (0,009 profiles km^{-2}) and of 26 predictor covariates were used with 30 (set 1) and 90-m (set 2) spatial resolutions. The performed analysis by Random Forest models showed that channel base level, elevation, and lithology were the most important ones to explain the variability. The validation of the models showed similarity among sets for the prediction of sand, clay, and organic carbon contents, which explains the following values of spatial variability, respectively: 44, 40, and 33%, for the spatial resolution of 30 m; and 45, 46, and 33%, for the spatial resolution of 90 m. The spatial resolution of the predictive covariates has little effect on attribute predictions, and the Random Forest approach has potential use for estimating soil properties.

Index terms: digital elevation model, morphometrics, pedometrics, SRTM.

Introdução

O conhecimento da distribuição espacial das propriedades físicas e químicas é muito importante para a modelagem ambiental e o manejo adequado dos solos. Por essa razão, existe atualmente uma crescente demanda por informações quantitativas dessas propriedades, especialmente em escalas nacional e regional (Carvalho Junior et al., 2014a).

Dessas propriedades, a textura do solo e o carbono orgânico estão entre as mais importantes para o manejo do solo. A textura por afetar fortemente a retenção de água e os nutrientes, a infiltração de água, drenagem e aeração, o teor de carbono orgânico, a capacidade de troca de cátions e a porosidade, além de controlar muitas funções e propriedades mecânicas do solo. O carbono orgânico, que tem um papel importante no

ecossistema terrestre, está intimamente associado à fertilidade do solo, por meio de seu controle sobre as propriedades físico-químicas e, além da relação direta e com as mudanças climáticas globais (Akpa et al., 2014; Guo et al., 2015).

Essas propriedades apresentam grande variabilidade espacial e, segundo Florinsky (2012), muito dessa variabilidade pode ser explicada por atributos morfométricos derivados dos modelos digitais de elevação (MDE), que condicionam a pedogênese, em razão do material de origem, fluxo de água e regime de encostas. Nesse sentido, a influência da resolução espacial dos atributos morfométricos sobre as propriedades dos solos foi descrita nos estudos de Smith et al. (2006) e Ruiz-Navarro et al. (2012).

Concomitantemente, diversos estudos têm mostrado a eficiência das técnicas de mapeamento digital de solos, para a predição espacial de atributos do solo. Os métodos mais comumente utilizados têm sido os seguintes: a regressão linear múltipla, para a predição da composição granulométrica, do teor de carbono orgânico, da capacidade de troca catiônica e do pH (Nanni & Demattê, 2006; Carvalho Junior et al., 2014b); a regressão por mínimos quadrados parciais, para a predição do teor de carbono orgânico (Gomez et al., 2008; Stevens et al., 2008); e os métodos geoestatísticos, para a predição da composição granulométrica, do pH, do teor de carbono orgânico e da salinidade do solo (Eldeiry & Garcia, 2010; Carvalho Junior et al., 2014b).

Vários estudos tem mostrado a importância dos atributos morfométricos para a predição das frações granulométricas do solo (Ließ et al., 2012; Akpa et al., 2014) e do teor de carbono orgânico (Grimm et al., 2008; Guo et al., 2015; Bonfatti et al., 2016). O modelo RF tem mostrado algumas vantagens em relação à maioria dos métodos estatísticos de modelagem, conforme destacado por Breiman (2001) e Liaw & Wiener (2002), que são: a habilidade para a modelagem de relações dimensionais altamente não lineares; a utilização de variáveis categóricas e contínuas; a resistência ao overfitting; a relativa robustez ante a presença de "ruídos" nos dados; o fornecimento de uma medida imparcial da taxa de erro; a determinação da importância das variáveis utilizadas; e a exigência de poucos parâmetros para ser implementado. No entanto, uma desvantagem do RF é a limitada interpretação dos resultados, já que as relações entre os preditores e as respostas não podem ser examinadas individualmente, para cada árvore na floresta, e são, por essa razão, frequentemente chamadas de "abordagem caixa-preta" (Grimm et al., 2008).

A utilização de modelos RF, com diferentes combinações de covariáveis preditoras, obteve sucesso na predição espacial do carbono orgânico do solo (Grimm et al., 2008; Viscarra Rossel & Behrens, 2010; Wiesmeier et al., 2011; Guo et al., 2015; Vaysse & Lagacherie, 2015; Bonfatti et al., 2016), e na predição da composição granulométrica (Viscarra Rossel & Behrens, 2010; Ließ et al., 2012; Akpa et al., 2014; Behrens et al., 2014; Vaysse & Lagacherie, 2015).

O objetivo deste trabalho foi avaliar a influência da resolução espacial do modelo digital de elevação e da eficiência de modelos Random Forest sobre a predição dos teores de areia, argila e carbono orgânico, com uso de número reduzido de amostras.

Material e Métodos

Os estudos foram realizados no estado do Mato Grosso do Sul, MS, em área com aproximadamente 19.911 km^2, entre 20º45' e 22º15'S e 55º45' e 57º00'W (Figura 1). A área de estudo engloba parte dos municípios de Antônio João, Bela Vista, Bonito, Caracol, Guia Lopes da Laguna, Jardim, Nioaque, Ponta Porã e Porto Murtinho e, segundo a classificação climática de Köppen-Geiger, apresenta clima do tipo Aw – tropical seco e megatérmico, com estação seca definida –, com temperatura média de 23,1ºC e precipitação anual média próxima de 1.500 mm. A área é constituída predominantemente por rochas pelíticas, calcários, rochas eruptivas ácidas (dacitos), arenitos e biotita gnaisse/granito, mármores e, em menor proporção, por quartzitos, mármores, turfas, sedimentos do Quaternário e anfibólio xisto/metabasito, conforme estudo de Lacerda Filho et al. (2006) elaborado na escala 1:1.000.000.

Para a análise e predição dos teores de areia, argila e carbono orgânico dos solos, utilizaram-se dados da camada superficial (0,0–0,2 m) de 175 perfis de solos (0,009 perfis km^{-2}), coletados durante o Zoneamento Agroecológico do Estado do Mato Grosso do Sul.

Para a avaliação da influência da resolução espacial sobre a eficiência da predição dos atributos do solo, utilizaram-se dados do SRTM (Shutlle Radar Topoghaphy Mission), com resolução espacial de 3

arco-segundos (90 m) e 1 arco-segundo (30 m), para derivar diferentes atributos morfométricos, conforme a seguir: obtidos com a utilização da função Spatial Analyst do ArcGIS Desktop 10.1 – elevação (Elev), declividade (Decl) e aspecto em graus (Aspecto); obtidos no Saga GIS – nível de base do canal de drenagem (NBCD), índice de convergência (IC), plano de curvatura (PC), curvatura longitudinal (CL), fator LS (comprimento x declividade), posição relativa da declividade (PRD), índice de balanço de massa (IBM), comprimento da pendente (CP), índice topográfico de umidade (ITU), profundidade do vale (PV), distância vertical do canal de drenagem (DVCD) e vetor de medida da rugosidade (VMR). A descrição e significância desses atributos pode ser encontrada em Bonfatti et al. (2016).

Foram também utilizados como covariáveis preditoras os dados do sensor TM do Landsat 5 (números digitais), com resolução espacial de 30 m, obtidos de três imagens órbitas/pontos 225/75, 226/74 e 226/75 do ano de 2008, conforme a seguir: banda 2, 0,525-0,605 µm; banda 3, 0,630–0,690 µm; banda 4, 0,755–0,900 µm; banda 5, 1,550–1,750 µm; banda 7, 2,090–2,350 µm; índice de vegetação por diferença normalizada (NDVI), (banda 4 - banda 3/banda 4 + banda 3); e as relações entre a banda 3 e a banda 2 (b3/b2), entre a banda 3 e a banda 7 (b3/b7) e entre a banda 5 e a banda 7 (b5/b7), e entre a banda 5 e a banda

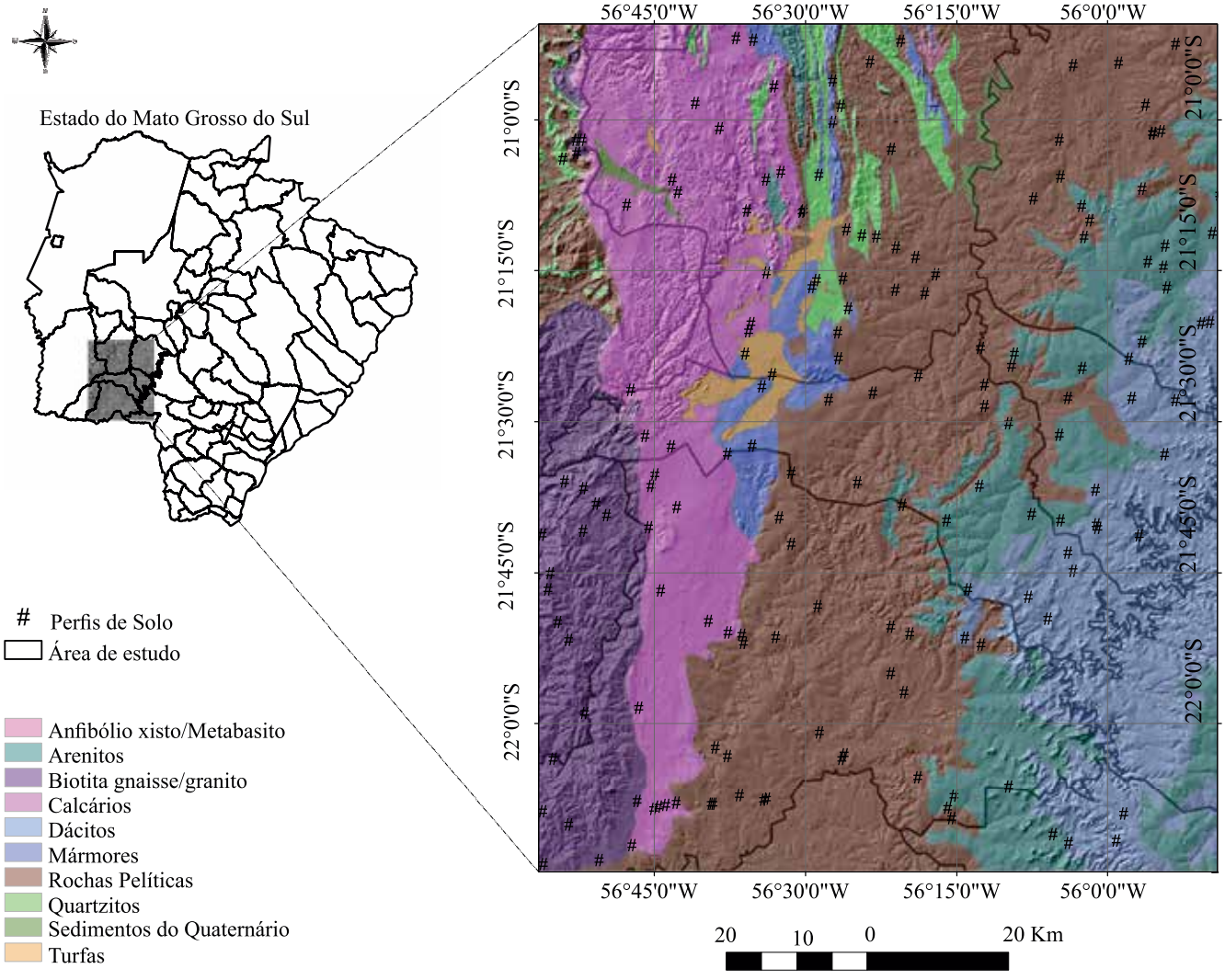

Figura 1. Localização da área de estudo, no Estado do Mato Grosso do Sul, com a distribuição espacial dos perfis de solos utilizados.

2 [(b5-b2)/(b5+b2)]. Utilizou-se também um mapa de litologia da área (Lacerda Filho et al., 2006), para representar o fator de formação material de origem dos solos. Ao final, foram avaliados dois conjuntos, um com resolução espacial de 30 m e outro com 90 m, cada um com 26 covariáveis preditoras. Para compor o segundo conjunto de covariáveis, todas as bandas do Landsat 5 foram reamostradas para 90 m, tendo-se usado a opção nearest no ArcGIS Desktop 10.1.

Na predição dos atributos do solo considerados, utilizou-se o modelo RF, que é uma técnica não paramétrica, desenvolvida por Breiman (2001) como uma extensão do programa CART (Classification and Regression Trees), para melhorar o desempenho de predição do modelo, que consiste 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 variáveis preditoras, escolhido aleatoriamente a partir do total de preditores existentes. No caso da aplicação da RF para regressão, o resultado final consiste da média dos resultados de todas as árvores (Breiman, 2001).

As RFs foram implementadas no pacote randomForest do R (The R Foundation, 2012). Para a utilização de uma RF, três parâmetros precisam ser definidos: o número de árvores na floresta (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, no entanto, foram testados valores para o ntree que variaram de 500 a 1000. Como valor de nodesize, utilizou-se o padrão para os estudos de regressão, que é de cinco para cada nó terminal. Com relação ao mtry, para problemas de regressão, o valor padrão estipulado é de um terço do número total de variáveis preditoras (Liaw & Wiener, 2002); assim, utilizou-se um mtry de valor nove, para vinte e sete variáveis preditoras.

A RF fornece estimativas confiáveis dos erros, por meio de dados conhecidos como out-of-bag (OOB), que é um subconjunto aleatório dos dados não utilizados pelo algoritmo para a construção das árvores. A partir das predições OOB de todas as árvores na floresta, é calculado o erro quadrado médio (MSE_{OOB}), conforme Liaw & Wiener (2002):

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

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

O desempenho dos modelos de predição é idealmente avaliado por meio de um conjunto de dados de validação independente que não tenha sido utilizado no processo de calibração. Deste modo, as amostras da camada superficial (0–20 cm) dos 175 perfis foram divididas em dois conjuntos independentes, um para a calibração (135 amostras) e outro para a validação (40 amostras), obtidos aleatoriamente pelo pacote estatístico R (The R Foundation, 2012). Assim, o desempenho de cada modelo foi calculado a partir das amostras de validação, pelo cálculo da correlação entre os valores observados e os valores estimados, por meio do coeficiente de determinação (R^2) e do RMSE, realizados com a utilização do R (The R Foundation, 2012).

O coeficiente de determinação (R^2) é dado pela relação entre a soma dos quadrados dos resíduos da regressão (SQR) e a soma total dos quadrados (SQT), conforme a seguinte equação:

$$R^2 = \frac{SQR}{SQT} = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

em que, R^2 é o coeficiente de determinação ($0 \leq R^2 \leq 1$); y_i é o valor observado da variável dependente; $\hat{y}_i$é o valor estimado da variável dependente; e $\bar{y}$ é a média da variável dependente. O coeficiente de determinação é sempre positivo e deve ser interpretado como a proporção da variância total da variável dependente y, que é explicada pelo modelo de regressão.

Por sua vez, o RMSE (raiz quadrada do erro médio quadrático) é calculado conforme a seguinte equação:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n} d_i^2}{n}}$$

em que: d é a diferença entre os valores observados e os valores preditos; e n é o número total de amostras consideradas. Assim, quanto maiores são os valores da RMSE, maiores são as discrepâncias entre os conjuntos de dados comparados.

Resultados e Discussão

A estatística descritiva dos atributos da camada superficial (0,0–0,2 m) dos solos, para as amostras de calibração e validação, é apresentada na Tabela 1, enquanto a estatística descritiva das covariáveis preditoras utilizadas é mostrada na Tabela 2.

As amostras de calibração e validação apresentam grande similaridade, quanto aos atributos físicos areia e argila, e não diferem significativamente. No entanto, o teor de carbono orgânico apresentou diferença significativa entre essas amostras. Uma similaridade maior é indicativa de que as amostras de validação representam adequadamente as amostras de

Tabela 1. Estatística descritiva das amostras da camada superficial do solo (0,0–0,2 m), utilizadas para a predição da areia, argila e carbono orgânico do solo (COS)[(1)].

Atributo	Calibração					Validação				
	Mínimo	Máximo	Média	DP	CV (%)	Mínimo	Máximo	Média	DP	CV (%)
Areia (g kg^{-1})	58	917	588,10a	243,13	41	50	880	627,40a	243,25	39
Argila (g kg^{-1})	40	739	234,64a	177,64	76	60	800	222,78a	190,34	85
COS (g kg^{-1})	1,6	64,1	13,24a	10,76	81	2,8	27,4	9,22b	6,38	69

[(1)]Médias com letras iguais, nas linhas, não diferem pelo teste de Tukey, a 5% de probabilidade. DP, desvio-padrão; CV, coeficiente de variação.

Tabela 2. Estatística descritiva das covariáveis preditoras dos conjuntos avaliados[(1)].

Covariável preditora	Conjunto 1 (30 m)				Conjunto 2 (90 m)			
	Mínimo	Máximo	Média	Desvio-padrão	Mínimo	Máximo	Média	Desvio-padrão
Elevação	148	708	324,1a	106,5	149	709	324a	105,2
Declividade	0	38	4,4a	5,4	0,3	34,7	3,9b	4,3
Aspecto em graus	-1	355	174,3a	101,4	4,4	352,6	187,1a	89,9
Curvatura longitudinal	-0,001	0,0006	0,0a	0,0	-0,002	0,001	0a	0,0
NBCD	148,3	696,3	309,9a	101,2	148,9	643,0	299,3a	98,9
Índice de convergência	-57,7	41,7	0,1a	11,4	-46,8	36,0	357,6a	10,9
Plano de curvatura	-0,002	0,0004	0,0a	0,0	-0,001	0,001	0a	0,0
Fator LS	0	5,2	0,5a	0,8	0,0	19,8	0,57a	1,6
PRD	-1,8	7,0	0,6a	1,0	-21,6	536,4	3,3a	40,5
IBM	-0,3	0,5	0,1a	0,1	-0,6	0,9	0,03a	0,1
CP	0	752,1	85,4a	128,2	0,0	1469,8	127,1b	188,3
ITU	4,7	20,0	8,8a	2,7	6	19,9	9,6b	2,5
Profundidade do vale	-24,7	167,8	16,5a	25,9	-83,1	295,5	47,4b	53,6
DVCD	-14,2	82,9	14,2a	17,9	-67,3	144,7	24,9b	27,2
VMR	0	0,03	0,001a	0,004	0,0	0,05	0,0a	0,0
Banda 2 (b2)	21	46	32,8a	4,8	21	45	32,8a	4,6
Banda 3 (b3)	17	66	32,3a	8,7	17	61	32,2a	8,1
Banda 4 (b4)	48	105	72,0a	8,9	49	104	72,2a	8,6
Banda 5 (b5)	42	133	80,7a	19,0	42	133	80,8a	18,5
Banda 7 (b7)	14	69	31,9a	11,2	14	69	32,0a	11,0
Índice NDVI	-0,1	0,7	0,4a	0,1	-0,01	0,6	0,4a	0,1
b3/b2	0,7	1,5	1,0a	0,2	0,7	1,5	1,0a	0,1
b3/b7	0,7	2,5	1,1a	0,2	0,8	2,1	1,0a	0,2
b5/b7	1,5	3,6	2,7a	0,4	1,5	3,7	2,6a	0,4
b5-b2/b5+b2	0,1	0,6	0,4a	0,1	0,2	0,5	0,4a	0,1

[(1)]O conjunto 1 (30 m) corresponde ao SRTM de 1 arco-segundo, e o conjunto 2 (90 m) corresponde ao SRTM de 3 arco-segundos. NBCD, nível de base do canal de drenagem; PRD, posição relativa da declividade; IBM, índice de balanço de massa; CP, comprimento da pendente; ITU, índice topográfico de umidade; DVCD, distância vertical do canal de drenagem; VMR, vetor de medida da rugosidade; e NDVI, índice de vegetação por diferença normalizada.

calibração. Os valores do coeficiente de variação (CV) elevados (>39%), em todos os casos, caracterizam a heterogeneidade dos conjuntos de amostras. O carbono orgânico do solo apresentou o maior CV (81%) nas amostras de calibração, enquanto a argila foi a que mostrou maior CV nas amostras de validação.

A análise de comparação de médias entre o conjunto 1 (30 m) e o conjunto 2 (90 m) mostrou que apenas cinco covariáveis preditoras apresentaram diferenças significativas: Decl, CP, ITU, PV e DVCD. Para as demais covariáveis, não houve diferença significativa entre estes conjuntos.

Alguns estudos já mostraram a influência da resolução dos MDEs sobre o padrão espacial dos atributos morfométricos (Hengl, 2006; Samuel-Rosa et al., 2015). Normalmente, à medida que o tamanho de célula aumenta, diminuem os valores e a acurácia dos atributos morfométricos derivados desses MDEs. Sørensen & Seibert (2007) encontraram consideráveis diferenças nos índices topográficos computados de MDEs com diferentes resoluções espaciais (10, 25 e 50 m). No entanto, Smith et al. (2006) concluíram, em seu estudo, que embora a resolução do MDE tenha um papel importante sobre os atributos morfométricos e, consequentemente, sobre o mapeamento digital de solos, a acurácia das predições ainda é muito dependente das características da paisagem analisada.

Vaze et al. (2010) ressaltaram que, em algumas áreas planas e para alguns processos, uma resolução mais grosseira (25 m ou maior) pode ser adequada para capturar a escala dos processos superficiais, enquanto em outras áreas podem ser necessárias resoluções maiores (~1 m). Em outras palavras, a escala dos processos da paisagem é que determina a resolução espacial do grid ou tamanho da célula adequada para representar esses processos. Dessa maneira, a uniformidade das características do relevo da área, que é predominantemente plano e suave-ondulado, pode ser a responsável pela similaridade entre os conjuntos de dados testados. Cabe ressaltar que a base de dados SRTM foi originalmente produzida com a resolução de 30 m e, posteriormente, foi degradada para 90 m, para a disponibilização em regiões fora dos Estados Unidos.

Uma das vantagens dos modelos RF é a possibilidade de estimativa da importância relativa das covariáveis preditoras, avaliada com base no decréscimo da acurácia da predição quando uma determinada covariável é retirada aleatoriamente do modelo (Breiman, 2001). Em geral, as covariáveis preditoras tiveram comportamento variável quanto à importância para a predição da areia, argila e carbono orgânico do solo (Figura 2). Em todos os atributos do solo, nota-se a grande influência dos atributos morfométricos NBCD e Elev, e da covariável categórica litologia sobre a acurácia da predição, independentemente da resolução do grid utilizada (30 ou 90 m). A importância da litologia e da elevação na predição destes atributos foi destacada nos estudos de Wiesmeier et al. (2011) e Akpa et al. (2014).

A análise da importância das covariáveis preditoras mostrou diferentes combinações de covariáveis preditoras, em razão dos atributos analisados e da resolução espacial do grid. No presente estudo, definiu-se um limiar de importância de aproximadamente 5% para as covariáveis preditoras, abaixo do qual estas foram consideradas sem importância e, consequentemente, foram retiradas dos modelos finais. As covariáveis consideradas nestes modelos estão apresentadas na Tabela 3.

A retirada dessas covariáveis (NBCD, Elev e litologia) produziu decréscimos entre 22 e 25% na predição da areia, 18 a 24% na predição da argila e, entre 7 e 13% para o carbono orgânico do solo, no conjunto 1 (30 m). Com relação ao conjunto 2 (90 m), o decréscimo da acurácia da predição ficou entre 20 e 27% para a areia, 18 e 27% para a argila e 8 e 13% para o carbono orgânico do solo (Figura 2).

O NBCD é definido como a distância vertical até o nível do canal de base da rede hidrográfica local (Hansen et al., 2009). Além de este índice indicar a proximidade dos canais de drenagem e, por consequência, as condições de drenagem, pode também reportar a energia potencial da água (Romão, 2006), a qual ainda atuará na ação dos processos erosivos; o NBCD está, portanto, associado ao relevo e, em contrapartida, atende ao pressupostos dos fatores de formação dos solos. Assim, o maior poder preditivo desta covariável está relacionado à influência que os processos erosivos têm sobre as diferentes litologias encontradas na área e atua diretamente sobre o grau de desenvolvimento dos solos e a variabilidade espacial de seus atributos, conforme destacado por Prates et al. (2012). Além disso, corrobora a importância da covariável litologia para a predição dos atributos avaliados.

Com exceção da relação entre as bandas b5/b7 e [(b5-b2)/(b5+b2)], para o conjunto 1, e [(b5-b2)/

(b5+b2)], para o conjunto 2, todos as demais covariáveis derivadas dos dados do sensor TM do Landsat 5 tiveram importância reduzida na predição. O índice clay minerals (b5/b7) pode ser utilizado para auxiliar a identificação de áreas com diferentes tipos de minerais de argila. Assim, a diversidade litológica da área, que condiciona solos com diferentes tipos de minerais de argila, pode explicar a importância desta

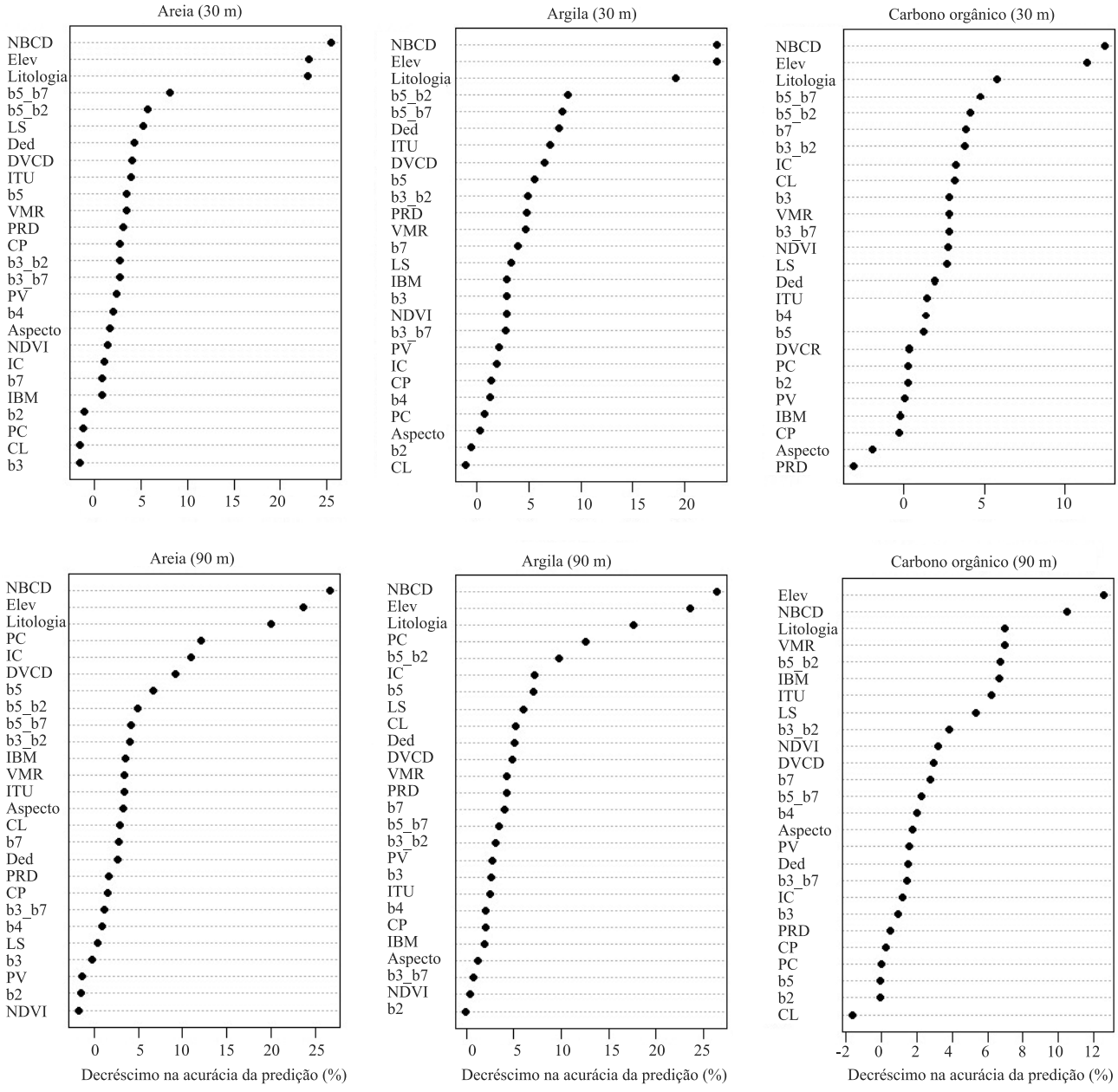

Figura 2. Importância das covariáveis preditoras para os atributos avaliados. O conjunto 1 (30 m) corresponde ao SRTM de 1 arco-segundo, e o conjunto 2 (90 m) corresponde ao SRTM de 3 arco-segundos. Elev, elevação; Decl, declividade; Aspecto, aspecto em graus; NBCD, nível de base do canal de drenagem; IC, índice de convergência; PC, plano de curvatura; LS, fator LS; PRD, posição relativa da declividade; IBM, índice de balanço de massa; CP, comprimento da pendente; CL, curvatura longitudinal; ITU, índice topográfico de umidade; PV, profundidade do vale; DVCD, distância vertical do canal de drenagem; VMR, vetor de medida da rugosidade; b2, banda 2; b3, banda 3; b4, banda 4; b5, banda 5; b7, banda 7.

covariável (b5/b7) para a predição destes atributos com a utilização do conjunto 1.

A presença de cobertura vegetal, na maior parte da área (menos de 10% de solo exposto), pode explicar a reduzida importância da maioria dos dados do Landsat utilizados (Figura 2). Neste sentido, Bartholomeus et al. (2007) ressaltam que estimativas acuradas de atributos do solo, a partir de dados de sensores remotos orbitais, são dificultadas pela presença de cobertura vegetal em percentagens superiores a 20%.

A importância da litologia para a predição das frações granulométricas e do carbono orgânico do solo pode ser observada na Figura 2, que mostra que a retirada desta covariável produziu decréscimos da predição da areia de 23 e 20%, da predição da argila de 18 e 17% e da predição do carbono orgânico do solo de 5 e 7%, respectivamente para os conjuntos 1 e 2. Esta importância, também relatada no estudo de Apka et al. (2014) e Guo et al. (2015), já era esperada, pois a área apresenta uma diversidade litológica considerável, que varia de rochas eruptivas ácidas, a leste, a arenitos, rochas pelíticas, calcários e gnaisses/granitos, a oeste (Figura 1).

Os resultados obtidos na predição espacial dos atributos do solo, pelos modelos RF, por meio de um conjunto de dados de validação independente, estão apresentados na Tabela 4. Em geral, os modelos RF avaliados tiveram comportamentos muito similares entre os conjuntos testados, quanto aos atributos do solo considerados, com exceção da argila, em que a utilização do conjunto 2 (90 m) produziu resultados ligeiramente superiores (R^2 de 0,46 e RMSE de 141,96 g kg^{-1}) aos obtidos pelo conjunto 1 (R^2 de 0,40 e RMSE de 148,11 g kg^{-1}).

A semelhança verificada entre os resultados dos conjuntos 1 e 2 está relacionada às características da área de estudo, que apresenta relevo dominantemente plano e suave-ondulado (0–8%), o que condicionou diferenças não significativas entre os conjuntos, para a maioria das covariáveis preditoras utilizadas (Tabela 2), principalmente as dos atributos morfométricos. Isto corrobora a afirmativa de Smith et al. (2006) de que, apesar da resolução do MDE interferir diretamente sobre estes atributos, a acurácia das predições no mapeamento digital de solos ainda é muito dependente das características da paisagem analisada, o que efetivamente se verificou no presente estudo, em que a resolução espacial dos conjuntos não teve influência sobre os resultados. Além disso, a litologia (covariável categórica), selecionada em todos os modelos como uma das covariáveis mais importantes, apresenta a mesma distribuição de ocorrência das unidades entre os conjuntos 1 e 2 (30 e 90 m, respectivamente).

Os modelos RF tiveram uma capacidade preditiva moderadamente satisfatória, para areia e argila, e pouco satisfatória para o carbono orgânico do solo, de acordo com as medidas de ajustamento utilizadas (R^2 e RMSE). A combinação das covariáveis preditoras nos conjuntos avaliados explicou 44, 40 e 33% da variação da areia, argila e carbono orgânico do solo, respectivamente, para o conjunto 1, e 45, 46 e 33%, respectivamente, para o conjunto 2. Com relação à areia e à argila, os resultados obtidos são similares aos encontrados em outros estudos que utilizaram modelos Random Forest, como os de Ließ et al. (2012) e Akpa et al. (2014), superiores aos de Viscarra Rossel & Behrens (2010) e Vaysse & Lagacherie (2015), e inferiores aos obtidos por Behrens et al. (2014).

Tabela 3. Covariáveis utilizadas na predição dos atributos do solo para cada conjunto[(1)].

Atributo	Covariáveis preditoras	
	Conjunto 1 (30 m)	Conjunto 2 (90 m)
Areia	NBCD, Elev, Litologia, b5/b7, (b5-b2)/(b5+b2), LS, Decl, DVCD e ITU	NBCD, Elev, Litologia, PC, IC, DVCD, b5, (b5-b2)/(b5+b2) e b5/b7
Argila	NBCD, Elev, Litologia, (b5-b2)/(b5+b2), b5/b7, Decl, ITU, DVCD, b5, b3/b2 e PRD	NBCD, Elev, Litologia, PC, (b5-b2)/(b5+b2), IC, b5, LS, CL e Decl
COS	NBCD, Elev, Litologia, b5/b7, (b5-b2)/(b5+b2), b7, b3/b2	Elev, NBCD, Litologia, VMR, (b5-b2)/(b5+b2), IBM, ITU, LS e b3/b2

[(1)]O conjunto 1 (30 m) corresponde ao SRTM de 1 arco-segundo, e o conjunto 2 (90 m) corresponde ao SRTM de 3 arco-segundos. Elev, elevação; Decl, declividade; NBCD, nível de base do canal de drenagem; IC, índice de convergência; PC, plano de curvatura; LS, fator LS; PRD, posição relativa da declividade; IBM, índice de balanço de massa; CL, curvatura longitudinal; ITU, índice topográfico de umidade; DVCD, distância vertical do canal de drenagem; VMR, vetor de medida da rugosidade; b2, banda 2; b3, banda 3; b5, banda 5; b7, banda 7.

Tabela 4. Resultados obtidos pelos modelos Random Forest (RF), na predição dos atributos do solo.

Atributo	Covariáveis preditoras[(1)]			
	Conjunto 1 (30 m)		Conjunto 2 (90 m)	
	R^2	RMSE (g kg^{-1})	R^2	RMSE (g kg^{-1})
Areia	0,44	188,61	0,45	188,82
Argila	0,40	148,11	0,46	141,96
COS	0,33	7,36	0,33	7,01

[(1)]O conjunto 1 (30 m) corresponde ao SRTM de 1 arco-segundo, e o conjunto 2 (90 m) corresponde ao SRTM de 3 arco-segundos. RMSE, raiz quadrada do erro médio quadrático; COS, carbono orgânico do solo.

Ließ et al. (2012) conseguiram explicar somente 30% da variação da areia e 43% da argila, na camada superficial dos solos, por meio do uso de atributos morfométricos e uma densidade de amostragem muito superior à do presente estudo (1,87 perfis km^{-2}). Akpa et al. (2014) relataram percentagens de variação, explicadas pelos modelos RF na camada superficial do solo (0,0–0,15 m), de 48–49% para areia e 53–56% para argila, em um estudo realizado na Nigéria, cuja densidade de amostragem (0,001 perfis km^{-2}) foi inferior à do presente estudo (0,009 perfis km^{-2}).

Os resultados obtidos por Vaysse & Lagacherie (2015) para a areia (33 a 35%) e argila (31 a 35%) foram obtidos com a utilização de atributos morfométricos, dados geológicos, climáticos e do Landsat 7, com uma densidade de amostragem de 0,07 perfis km^{-2}. O baixo desempenho foi atribuído à pequena variação de escala, determinada pelo material de origem, e à relação erosão/deposição ao longo da pendente, que não pode ser capturada pela resolução espacial das covariáveis utilizadas (100 m). Em todos os estudos citados, o baixo desempenho dos modelos RF foi atribuído ao tamanho reduzido do conjunto de dados utilizado.

Quanto ao carbono orgânico do solo, os resultados obtidos para o coeficiente de determinação (R^2) de 0,33, para ambos os conjuntos, foram inferiores aos dos seguintes autores: Viscarra Rossel & Behrens (2010), que obtiveram R^2 de 0,71; aos de Wiesmeier et al. (2011), com R^2 de 0,74; Guo et al. (2015), que alcançaram R^2 de 0,50; e de Vaysse & Lagacherie (2015), cujo R^2 foi de 0,59. Os valores da RMSE – 7,36 e 7,01 g kg^{-1} (conjuntos 1 e 2, respectivamente) – foram mais satisfatórios do que os dos seguintes autores: Grimm et al. (2008), que obtiveram RMSE de 17,20 g kg^{-1}, para a camada superficial; de Viscarra Rossel & Behrens (2010), com RMSE de 12,23 g kg^{-1}; e de Vaysse & Lagacherie (2015), que obtiveram RMSE de 18,31 e 18,47 g kg^{-1}, para as camadas de 0,0–0,05 e 0,05–0,15 m, respectivamente. No entanto, os resultados do presente estudo foram inferiores aos de Wiesmeier et al. (2011), que tiveram RMSE de 5,46 g kg^{-1}, e de Guo et al. (2015), com 2,08 g kg^{-1}.

Conforme destacado por Grimm et al. (2008), resultados pouco satisfatórios para o carbono orgânico do solo, como os obtidos no presente trabalho, podem ser atribuídos ao fato de que o padrão de distribuição espacial desse atributo é altamente variável, em razão das variações em pequena escala dos processos de deposição, redistribuição e estabilização, aliadas à alta variabilidade aleatória intrínseca do carbono orgânico do solo e à baixa densidade de amostragem utilizada no presente estudo (0,009 perfis km^{-2}).

No estudo de Viscarra Rossel & Behrens (2010), foram utilizados dados de reflectância difusa como covariáveis preditoras. Wiesmeier et al. (2011) utilizaram uma densidade de amostragem de 0,03 perfis km^{-2} e dados de uso de terra, geologia, unidades de solo e atributos morfométricos, derivados do SRTM com 90 m de resolução espacial. Guo et al. (2015) usaram como covariáveis preditoras atributos morfométricos, derivados de um MDE com 90 m de resolução, geologia, dados climáticos e dados do sensor Modis, com uma densidade de amostragem superior (2,52 perfis km^{-2}).

Os modelos RF, gerados pelos conjuntos, foram utilizados na modelagem espacial dos atributos de uma parte da área total considerada, em razão do tamanho muito grande dos arquivos que englobam a área toda (Figura 3). Estes modelos produziram mapas de predição muito semelhantes entre os conjuntos, para todos os atributos. O teor de areia variou de 146,96 a 858,69 g kg^{-1}, no conjunto 1, e de 162,57 a 840,19 g kg^{-1}, no conjunto 2. Para a argila, a variação foi de 66,36 a 636,72 g kg^{-1}, no conjunto 1, e de 61,94 a 645,25 g kg^{-1}, no conjunto 2. O carbono orgânico do solo variou de 4,48 a 48,55 g kg^{-1}, no conjunto 1, e de 4,30 a 47,30 g kg^{-1}, no conjunto 2.

A distribuição espacial dos atributos apresentou padrões bastante semelhantes e coerentes entre os dois conjuntos (Figura 3). Assim, verifica-se maior concentração de areia e menor de argila, na porção central da área, correspondente a uma litologia dominada por rochas pelíticas, arenitos e quartzitos. No entanto, menores teores de areia e maiores de argila são verificados nas porções sudeste e noroeste, onde são encontradas rochas eruptivas ácidas (dacitos) e calcários. Valores intermediários de areia e argila estão associados à presença de biotita gnaisse/granito, na porção leste.

O carbono orgânico do solo apresenta uma variação bastante coerente com a distribuição das frações areia e argila preditas pelos dois conjuntos. Os maiores teores de carbono orgânico do solo estão presentes nas porções sudeste e noroeste da área (Figura 3), associadas aos maiores teores de argila nos solos, enquanto os menores teores estão associados a solos mais arenosos. A relação

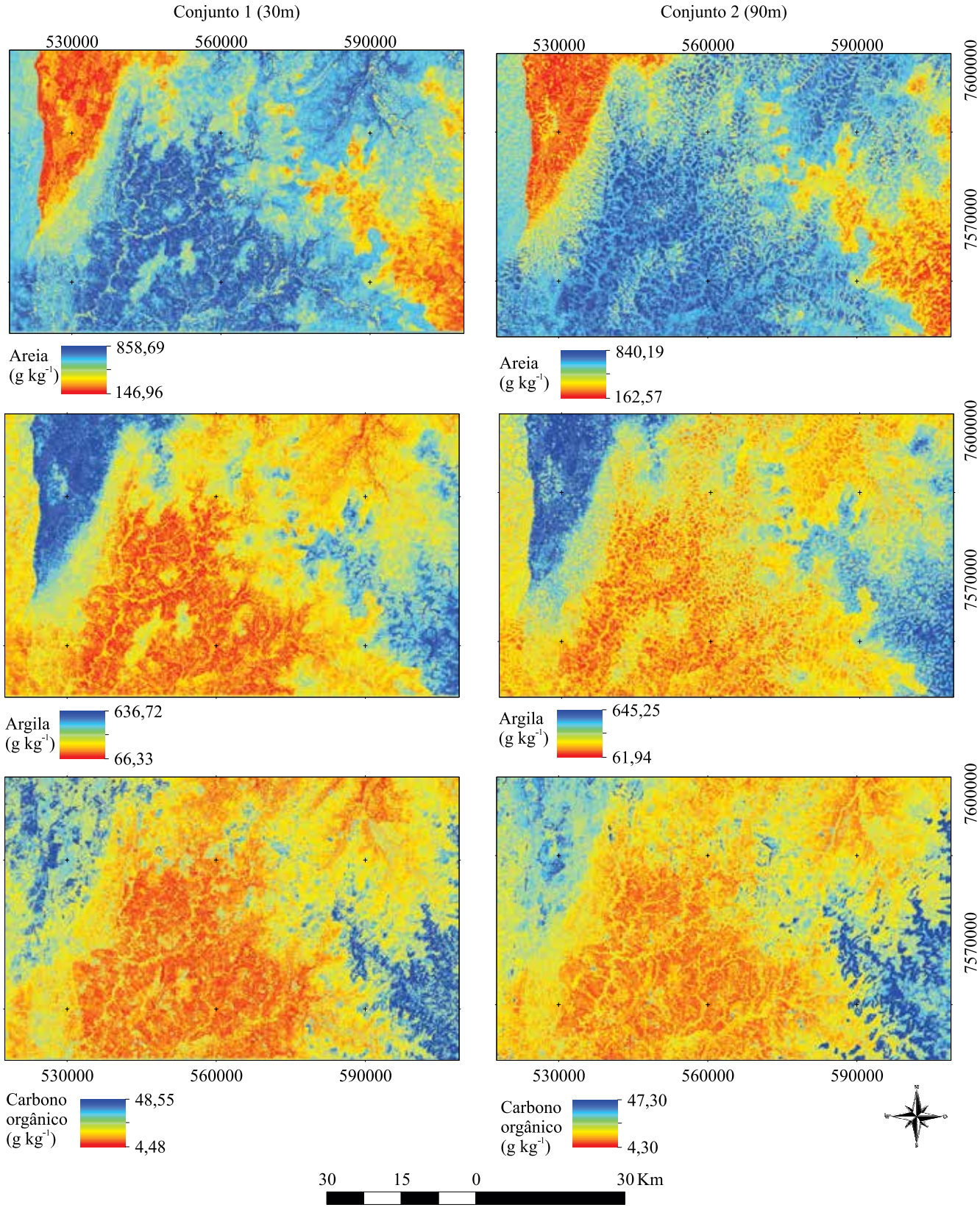

Figura 3. Distribuição espacial dos atributos do solo, estimados pelos modelos Random Forest (RF).

entre a textura do solo e o carbono orgânico do solo foi ressaltada por Akpa et al. (2014).

Finalmente, em razão das características da área de estudo – relevo predominantemente plano e suave-ondulado – e dos dados de solos limitados utilizados, pode-se considerar que os resultados alcançados na predição dos atributos são promissores para uma primeira aproximação. A melhoria destes resultados pode ser alcançada, à medida que novos dados de solos sejam incluídos e novas covariáveis sejam testadas.

Conclusões

1. Os conjuntos de dados testados (30 e 90 m) foram similares para a maioria das covariáveis preditoras utilizadas.

2. O nível de base do canal de drenagem, a elevação e a litologia foram as covariáveis mais importantes testadas pelo modelo Random Forest, independentemente da resolução do grid utilizada.

3. A validação dos modelos apresentou similaridade entre os conjuntos quanto à predição da areia, da argila e do carbono orgânico, o que explica, respectivamente, 44, 40 e 33%, da variabilidade espacial, para a resolução de 30 m, e de 45, 46 e 33%, para a resolução de 90 m.

4. O moderado desempenho dos modelos Random Forest, determinado pelas medidas de ajustamento (R^2 e RMSE), pode ser atribuído ao tamanho reduzido do conjunto de dados utilizado.

5. O uso de covariáveis preditoras, como atributos morfométricos derivados do SRTM, dados do sensor TM do Landsat 5 e a litologia da área, aliado à abordagem Random Forest, mostrou potencial de utilização para estimar valores de areia, argila e carbono orgânico do solo com uso de uma reduzida base de dados de solos.

Referências

AKPA, S.I.C.; ODEH, I.O.A.; BISHOP, T.F.A.; HARTEMINK, A.E. Digital mapping of soil particle-size fractions for Nigeria. **Soil Science Society of America Journal**, v.78, p.1953-1966, 2014.

BARTHOLOMEUS, H.M.; EPEMA, G.F.; SCHAEPMAN, M.E. Determining iron content in Mediterranean soils in partly vegetated areas, using spectral reflectance and imaging spectroscopy. **International Journal of Applied Earth Observation and Geoinformation**, v.9, p.194-203, 2007. DOI: 10.1016/J.JAG.2006.09.001.

BEHRENS, T.; SCHMIDT, K.; RAMIREZ-LOPEZ, L.; GALLANT, J.; ZHU, A.X.; SCHOLTEN, T. Hyper-scale digital soil mapping and soil formation analysis. **Geoderma**, v.213, p.578-588, 2014. DOI: 10.1016/j.geoderma.2013.07.031.

BONFATTI, B.R.; HARTEMINK, A.E.; GIASSON, E.; TORNQUIST, C.G.; ADHIKARI., K. Digital mapping of soil carbon in a viticultural region of Southern Brazil. **Geoderma**, v.261, p.204-221, 2016. DOI: 10.1016/J.GEODERMA.2015.07.016.

BREIMAN, L. Random forests. **Machine Learning**, v.45, p.5-32, 2001.

CARVALHO JUNIOR, W. de C.; CHAGAS, C. da S.; LAGACHERIE, P.; CALDERANO FILHO, B.; BHERING, S.B. Evaluation of statistical and geostatistical models of digital soil properties mapping in tropical mountain regions. **Revista Brasileira de Ciência do Solo**, v.38, p.706-717, 2014a.

CARVALHO JUNIOR, W. de; LAGACHERIE, P.; CHAGAS, C. da S.; CALDERANO FILHO, B.; BHERING, S.B. A regional-scale assessment of digital mapping of soil attributes in a tropical hillslope environment. **Geoderma**, v.232/234, p.479-486, 2014b.

ELDEIRY, A.A.; GARCIA, L.A. Comparison of ordinary kriging, regression kriging, and cokriging techniques to estimate soil salinity using LANDSAT images. **Journal of Irrigation and Drainage Engineering**, v.136, p.355-364, 2010. DOI: 10.1061/(ASCE)IR.1943-4774.0000208.

FLORINSKY, I.V. Influence of topography on soil properties. In: FLORINSKY, I.V. **Digital terrain analysis in soil science and geology**. Amsterdam: Academic Press, 2012. p.145-149. DOI: 10.1016/B978-0-12-385036-2.00008-0.

GOMEZ, C.; VISCARRA ROSSEL, R.A.; MCBRATNEY, A.B. Soil organic carbon prediction by hyperspectral remote sensing and field vis-NIR spectroscopy: an Australian case study. **Geoderma**, v.146, p.403-411, 2008.

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.

GUO, P.-T.; LI, M.-F.; LUO, W.; TANG, Q.-F.; LIU, Z.-W.; LIN, Z.-M. Digital mapping of soil organic matter for rubber plantation at regional scale: An application of random forest plus residuals kriging approach. **Geoderma**, v.237-238, p.49-59, 2015. DOI: 10.1016/J.GEODERMA.2006.07.002.

HANSEN, M.K.; BROWN, D.J.; DENNISON, P.E.; GRAVES, S.A.; BRICKLEMYER, R.S. Inductively mapping expert-derived soil landscape units within dambo wetland catenae using multispectral and topographic data. **Geoderma**, v.150, p.72-84, 2009. DOI: 10.1016/j.geoderma.2009.01.013.

HENGL, T. Finding the right pixel size. **Computers and Geosciences**, v.32, p.1283-1298, 2006. DOI: 10.1016/J.CAGEO.2005.11.008.

LACERDA FILHO, J.W. de; SILVA, M. da G. da; JOST, H. (Org.). **Geologia e recursos minerais do Estado de Mato Grosso do Sul**: texto explicativo dos mapas geológico e de recursos minerais do Estado de Mato Grosso do Sul: escala 1:1.000.000. Campo Grande: Serviço Geológico do Brasil-CRPM, 2006. 121p.

LIAW, A.; WIENER, M. Classification and regression by randomForest. **R News**, v.2/3, p.18-22, 2002.

LIEß, M.; GLASER, B.; HUWE, B. Uncertainty in the spatial prediction of soil texture: comparison of regression tree and Random Forest models. **Geoderma**, v.170, p.70-79, 2012. DOI: 10.1016/J.GEODERMA.2011.10.010.

NANNI, M.R.; DEMATTÊ, J.A.M. Spectral reflectance methodology in comparison to traditional soil analysis. **Soil Science Society of America Journal**, v.70, p.393-407, 2006. DOI: 10.2136/sssaj2003.0285.

PRATES, V.; SOUZA, L.C. de P.; OLIVEIRA JUNIOR, J.C. de. Índices para a representação da paisagem como apoio para levantamento pedológico em ambiente de geoprocessamento. **Revista Brasileira de Engenharia Agrícola e Ambiental**, v.16, p.408-414, 2012. DOI: 10.1590/S1415-43662012000400011.

ROMÃO, P.A. **Modelagem de terreno com base na morfometria e em sondagens geotécnicas - Região de Goiânia, GO**. 2006. 192p. Tese (Doutorado) – Universidade de Brasília, Brasília.

RUIZ-NAVARRO, A.; BARBERÁ, G.G.; GARCÍA-HARO, J.; ALBALADEJO, J. Effect of the spatial resolution on landscape control of soil fertility in a semiarid area. **Journal of Soils and Sediments**, v.12, p.471-485, 2012. DOI: 10.1007/s11368-012-0470-8.

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.

SMITH, M.P.; ZHU, A.-X.; BURT, J.E.; STILES, C. The effects of DEM resolution and neighborhood size on digital soil survey. **Geoderma**, v.137, p.58-69, 2006. DOI: 10.1016/J.GEODERMA.2006.07.002.

SØRENSEN, R.; SEIBERT, J. Effects of DEM resolution on the calculation of topographical indices: TWI and its components. **Journal of Hydrology**, v.347, p.79-89, 2007. DOI: 10.1016/J.JHYDROL.2007.09.001.

STEVENS, A.; WESEMAEL, B. van; BARTHOLOMEUS, H.; ROSILLON, D.; TYCHON, B.; BEN-DOR, E. Laboratory, field and airborne spectroscopy for monitoring organic carbon content in agricultural soils. **Geoderma**, v.144, p.395-404, 2008. DOI: 10.1016/j.geoderma.2007.12.009.

THE R FOUNDATION. **R**: the R project for statistical computing. Vienna: The R Foundation, 2012.

VAYSSE, K.; LAGACHERIE, P. Evaluating digital soil mapping approaches for mapping GlobalSoilMap soil properties from legacy data in Languedoc-Roussillon (France). **Geoderma Regional**, v.4, p.20-30, 2015. DOI: 10.1016/J.GEODRS.2014.11.003.

VAZE, J.; TENG, J.; SPENCER, G. Impact of DEM accuracy and resolution on topographic indices. **Environmental Modelling and Software**, v.25, p.1086-1098, 2010. DOI: 10.1016/J.ENVSOFT.2010.03.014.

VISCARRA ROSSEL, R.A.V.; BEHRENS, T. Using data mining to model and interpret soil diffuse reflectance spectra. **Geoderma**, v.158, p.46-54, 2010. DOI: 10.1016/J.GEODERMA.2009.12.025.

WIESMEIER, M.; BARTHOLD, F.; BLANK, B.; KÖGEL-KNABNER, I. Digital mapping of soil organic matter stocks using Random Forest modeling in a semi-arid steppe ecosystem. **Plant and Soil**, v.340, p.7-24, 2011. DOI: 10.1007/s11104-010-0425-z.

Recebido em 31 de agosto de 2015 e aprovado em 5 de fevereiro de 2016