

MARIANA FARIA VELOSO

**DESENVOLVIMENTO DE FUNÇÕES DE PEDOTRANSFERÊNCIA PARA
ESTIMATIVA DE PROPRIEDADES FÍSICO-HÍDRICAS DO SOLO DO BIOMA
CERRADO**

Dissertação apresentada à Universidade Federal de Viçosa, como parte das exigências do Programa de Pós-Graduação em Engenharia Agrícola, para obtenção do título de *Magister Scientiae*.

Orientador: Lineu Neiva Rodrigues
Coorientador: Elpídio Inácio Fernandes Filho

**VIÇOSA - MINAS GERAIS
2021**

**Ficha catalográfica elaborada pela Biblioteca Central da Universidade
Federal de Viçosa - Campus Viçosa**

T

Veloso, Mariana Faria, 1995-

V443d
2021

Desenvolvimento de funções de pedotransferência para
estimativa de propriedades físico-hídricas do solo do bioma
cerrado / Mariana Faria Veloso. – Viçosa, MG, 2021.
1 dissertação eletrônica (76 f.): il. (algumas color.).

Orientador: Lineu Neiva Rodrigues.

Dissertação (mestrado) - Universidade Federal de Viçosa.
Inclui bibliografia.

DOI: <https://doi.org/10.47328/ufvbbt.2021.086>

Modo de acesso: World Wide Web.

1. Desenvolvimento de recursos hídricos. 2. Irrigação.
3. Hidrologia. 4. Aprendizado do computador. I. Universidade
Federal de Viçosa. Departamento de Engenharia Agrícola.
Programa de Pós-Graduação em Engenharia Agrícola. II. Título.

CDD 22. ed. 333.91

Bibliotecário(a) responsável: Renata de Fátima Alves CRB6/2578

MARIANA FARIA VELOSO

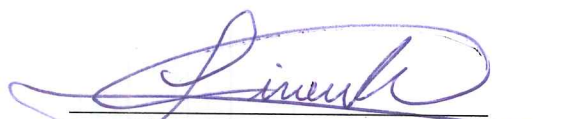
**DESENVOLVIMENTO DE FUNÇÕES DE PEDOTRANSFERÊNCIA PARA
ESTIMATIVA DE PROPRIEDADES FÍSICO-HÍDRICAS DO SOLO DO BIOMA
CERRADO**

Dissertação apresentada à Universidade Federal de Viçosa, como parte das exigências do Programa de Pós-Graduação em Engenharia Agrícola, para obtenção do título de *Magister Scientiae*.

APROVADA: 16 de julho de 2021.

Assentimento:


Mariana Faria Veloso


Lineu Neiva Rodrigues

A Deus, à minha família e à ciência.

AGRADECIMENTOS

Primeiramente agradeço a Deus por ter me dado forças e sabedoria durante toda a minha caminhada, me protegendo e me fazendo ver que a felicidade não é um ponto de chegada e sim, a caminhada até ela.

Aos meus pais Sandra e Welington, os amores da minha vida, que nunca deixaram de faltar nada em nossas casas, por todo amor e carinho incondicional. Às minhas irmãs Camilla e Carol, pelo amor, apoio e ajuda na realização dessa pesquisa. À meu tio Preto, pelo amor, carinho e sabedoria.

À Heverton, pelo amor, companheirismo, amizade, apoio nos projetos de vida e por ter me dado forças nos momentos mais difíceis. .

Aos meus cachorrinhos, Bolota, Barry e Ralf, que tornaram essa caminhada mais leve, alegre e cheia de passeios.

À Universidade Federal de Viçosa e ao Programa de Pós Graduação em Engenharia Agrícola, pela oportunidade de realização do curso. Aos professores e servidores que tornaram tudo isso possível.

Ao Lineu Neiva Rodrigues, pela orientação, confiança, conhecimento e exemplo de profissionalismo repassados a mim.

Ao Elpídio Inácio Fernandes Filho, pela coorientação e paciência nos ensinamentos do R, um exemplo do que é ser um professor.

À Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) pela concessão da bolsa de estudos.

À EMBRAPA Cerrados pela concessão de dados e informações fundamentais utilizadas na realização da pesquisa.

Aos meus “migos e migas” do coração, pela amizade e companheirismo.

À Ciência, não existe desenvolvimento econômico e social sem a ciência, tecnologia e inovação.

Enfim, o meu muito obrigada a todos que contribuíram de alguma forma na execução desse projeto.

“A educação é a arma mais poderosa que você pode usar para mudar o mundo”.

(Nelson Mandela)

RESUMO

VELOSO, Mariana Faria, M.Sc., Universidade Federal de Viçosa, julho de 2021. **Desenvolvimento de funções de pedotransferência para estimativa de propriedades físico-hídricas do solo do Bioma Cerrado.** Orientador: Lineu Neiva Rodrigues.

O Cerrado é a principal região agrícola do Brasil, sendo responsável por uma grande parte da produção de alimentos no país. A falta de dados na escala apropriada sobre os solos da região tem trazido incertezas aos processos de gestão dos recursos hídricos. A obtenção de determinadas propriedades físico-hídricas do solo, como, por exemplo, a condutividade hidráulica do solo saturado (K_s) e a curva de retenção de água, são trabalhosas e custosas, abrindo oportunidade para o uso das Funções de Pedotransferência (FPTs). O objetivo desta dissertação foi desenvolver para o Bioma Cerrado, FPTs para estimar a K_s , as umidades do solo nas tensões de 0, 6, 10, 33, 100 e 1500 kPa e os parâmetros de ajuste da curva de retenção. A partir de diferentes combinações de dados pedológicos, teores de areia, silte e argila, densidade do solo, densidade de partícula, porosidade, microporosidade, macroporosidade, umidade do solo na capacidade de campo (CC) e umidade do solo no ponto de murcha permanente (PMP), as FPTs foram desenvolvidas incorporando tais conjuntos em cinco modelos de aprendizado de máquina: Regressão Linear Múltipla (RLM), Multivariate Adaptive Regression Splines (MARS), Random Forest (RF), Support Vector Regression (SVR) e K Nearest Neighbors (KNN). As FPTs desenvolvidas para K_s apresentaram capacidade preditiva mediana, já para as umidades do solo, os modelos apresentaram valores de RMSE e ME próximos de zero e valores de R^2 superiores a 0,8. No ajuste da curva de retenção, o modelo de van Genuchten (1980) apresentou o melhor desempenho e as FPTs desenvolvidas para os parâmetros, umidade de saturação e umidade residual, apresentaram R^2 iguais a 0,76 e 0,42, respectivamente, e baixos valores de RMSE e ME. Já para os parâmetros, α e n , observou-se baixa capacidade preditiva das FPTs. Os algoritmos RF e SVR apresentaram os melhores desempenhos dentre os modelos avaliados e as variáveis predictoras CC e PMP demonstraram importância no desenvolvimento das FPTs.

Palavras-chave: Recursos Hídricos. Hidrologia. Irrigação. Aprendizado de máquina.

ABSTRACT

VELOSO, Mariana Faria, M.Sc., Universidade Federal de Viçosa, July 2021. **Development of pedotransfer functions to estimate soil hydraulic parameters to the Brazilian Savannah.** Adviser: Lineu Neiva Rodrigues.

The Cerrado Biome is the main agricultural region in Brazil, being responsible for a large part of the country's food production. The lack of soil data on the appropriate scale has brought uncertainty to the water resources process in the region. Obtaining certain soil hydraulic parameters, such as the saturated hydraulic conductivity (Ks) and the water retention curve are laborious and costly, opening up opportunities for the use of Pedotransfer Functions (PTFs). The objective of this dissertation was to develop, for the Brazilian Savannah, PTFs to estimate Ks, soil moistures at tensions of 0, 6, 10, 33, 100 and 1500 kPa and the adjustment parameters of the retention curve. From different combinations of pedological data, sand, silt and clay contents, soil density, particle density, porosity, microporosity, macroporosity, soil moisture at field capacity (FC) and soil moisture at permanent wilting point (PWP), PTFs were developed by incorporating such sets into five machine learning models: Multiple Linear Regression (MLR), Multivariate Adaptive Regression Splines (MARS), Random Forest (RF), Support Vector Regression (SVR) and K Nearest Neighbors (KNN). The PTFs developed for Ks showed median predictive capacity, whereas for soil moisture, the models presented RMSE and ME values close to zero and R² values above 0.8. In the adjustment of the retention curve, the model by van Genuchten (1980) showed the best performance and the PTFs developed for the parameters, saturation moisture and residual moisture, presented R² equal to 0.76 and 0.42, respectively, and low RMSE and ME values. As for the parameters, α and n , a low predictive capacity was observed. The RF and SVR algorithms showed the best performance among the evaluated models and the predictor variables FC and PWP showed importance in the development of PTFs.

Keywords: Water resources. Hydrology. Irrigation. Machine learning.

LISTA DE ILUSTRAÇÕES

Figura 2.1.2. Percentual de dados faltantes para cada variável da base de dados inicial.	26
Figura 2.1.2. Triângulos texturais das bases de dados de treinamento (a) e teste (b) para condutividade hidráulica do solo saturado e umidades do solo nas tensões de 0, 6,10, 33, 100 e 1500 kPa.	27
Figura 2.1.3. Histogramas dos percentuais de (a) argila e (b) areia das amostras de solos.....	28
Figura 2.1.4. Condutividade hidráulica do solo saturado estimada obtida pelos modelos de melhor desempenho nos conjuntos preditores A1 (a), A2 (b), A3 (c) e A4 (d) em relação a condutividade hidráulica do solo saturado observada. (b, c, d) RF; (a) RLM.....	33
Figura 2.1.5. Classificação da importância das variáveis na estimativa de K_s sem o preenchimento de dados faltantes.....	36
Figura 2.1.6. Umidade do solo estimada na tensão de 0 kPa obtidos pelos modelos de melhor desempenho nos conjuntos preditores A1 (a), A2 (b), A3 (c) e A4 (d) em relação a umidade do solo observada na tensão de 0 kPa. (a, c, d) SVR; (b) KNN.	39
Figura 2.1.7. Umidade do solo estimada na tensão de 6 kPa obtidos pelos modelos de melhor desempenho nos conjuntos preditores A1 (a), A2 (b), A3 (c) e A4 (d) em relação a umidade do solo observada na tensão de 6 kPa. (b, c) RF; (a, d) MARS.	39
Figura 2.1.8. Umidade do solo estimada na tensão de 10 kPa obtidos pelos modelos de melhor desempenho nos conjuntos preditores A1 (a), A2 (b), A3 (c) e A4 (d) em relação a umidade do solo observada na tensão de 10 kPa. (a, c, d) RF; (b) RLM.	40
Figura 2.1.9. Umidade do solo estimada na tensão de 33 kPa obtidos pelos modelos de melhor desempenho nos conjuntos preditores A1 (a), A2 (b), A3 (c) e A4 (d) em relação a umidade do solo observada na tensão de 33 kPa. (a) SVR; (b, c, d) RF.....	40
Figura 2.1.10. Umidade do solo estimada na tensão de 100 kPa obtidos pelos modelos de melhor desempenho nos conjuntos preditores A1 (a), A2 (b), A3 (c) e A4 (d) em relação a umidade do solo observada na tensão de 100 kPa. (a, d) RF; (b, c) SVR.	41
Figura 2.1.11. Umidade do solo estimada na tensão de 1500 kPa obtidos pelos modelos de melhor desempenho nos conjuntos preditores A1 (a), A2 (b), A3 (c) e A4 (d) em relação a umidade do solo observada na tensão de 1500 kPa. (a, b, c, d) RF.....	41
Figura 2.1.12. Classificação das importâncias das variáveis predictoras nas estimativas das umidades do solo nas tensões de 0, 6, 10, 33, 100 e 1500 kPa.....	44

Figura 2.2.1. Triângulo textural dos solos do Bioma Cerrado utilizados para o ajuste das curvas de retenção de água no solo.	61
Figura 2.2.2. Umidades do solo em potenciais matriciais específicos utilizados no ajuste dos parâmetros da curva de retenção e desenvolvimento das FPTs.....	62
Figura 2.2.3. Curva retenção de água no solo.	63
Figura 2.2.4. Parâmetros estimados da equação de van Genuchten (1980) obtidos pelos modelos de melhor desempenho no conjunto preditor A1 em relação aos parâmetros observados da equação de van Genuchten (1980). (a, b, c, d) RF.....	66
Figura 2.2.5. Parâmetros estimados da equação de van Genuchten (1980) obtidos pelos modelos de melhor desempenho no conjunto preditor A2 em relação Parâmetros estimados da equação de van Genuchten (1980) observados. A1: (a) RF, A2: (b) RF, A3: (c) RF e A4: (d)RF.	67
Figura 2.2.6. Classificação das importâncias das variáveis preditoras nas estimativas dos parâmetros θ_s e θ_r utilizando o conjunto preditor A1.	70

LISTA DE TABELAS

Tabela 2.1.1. Número de amostras para cada variável a ser estimada após os pré-processamentos.....	27
Tabela 2.1.2. Estatísticas descritivas das variáveis preditoras utilizadas para o desenvolvimento (subconjunto A) e teste (subconjunto B) das funções de pedotransferência da condutividade hidráulica do solo saturado (K_s) ($n = 140$).....	29
Tabela 2.1.3. Estatísticas descritivas das variáveis preditoras utilizadas para o desenvolvimento (subconjunto A) e teste (subconjunto B) das funções de pedotransferência da umidade do solo nas tensões de 6 e 33 kPa ($n = 158$).....	30
Tabela 2.1.4. Estatísticas descritivas das variáveis preditoras utilizadas para o desenvolvimento (subconjunto A) e teste (subconjunto B) das funções de pedotransferência da umidade do solo nas tensões de 0, 10, 100 e 1500 kPa ($n = 268$).....	31
Tabela 2.1.5. Desempenho estatísticos para cada modelo e conjunto preditor na estimativa K_s sem preenchimento dos dados faltantes.....	34
Tabela 2.1.6. Desempenho estatísticos para cada modelo e conjunto preditor na estimativa K_s com preenchimento dos dados faltantes.	37
Tabela 2.1.7. Coeficiente de determinação (R^2) para cada modelo e conjunto preditor na estimativa das umidades do solo nas tensões de 0, 6, 10, 33, 100 e 1500 kPa.....	42
Tabela 2.1.8. Desempenho estatístico dos modelos nulos para cada umidade do solo estimada.	43
Tabela 2.2.1. Desempenho estatístico médio para cada modelo de ajuste	63
Tabela 2.2.2. Estatísticas descritivas das variáveis preditoras utilizadas para o desenvolvimento (subconjunto A) e teste (subconjunto B) das funções de pedotransferência dos parâmetros da equação de van Genuchten	64
Tabela 2.2.3. Estatísticas descritivas dos parâmetros da equação de van Genuchten (1980)..	65
Tabela 2.2.4. Desempenho estatístico das FPTs para estimativa dos parâmetros de van Genuchten (1980) utilizando o conjunto preditor A1.....	67
Tabela 2.2.5. Desempenho estatístico das FPTs para estimativa dos parâmetros de van Genuchten (1980) utilizando o conjunto preditor A2.....	68
Tabela 2.2.6. Desempenho estatístico dos modelos nulos para cada umidade do solo estimada.	69

SUMÁRIO

1.	INTRODUÇÃO GERAL	12
2.	ARTIGOS CIENTÍFICOS	18
2.1	Artigo 1 - Funções de pedotransferência para a estimativa da condutividade hidráulica e da umidade do solo para o Cerrado brasileiro	18
2.1.1	Introdução	19
2.1.2	Material e Métodos	21
2.1.3	Resultados e discussão	26
2.1.4	Conclusões	46
2.1.5	Referências bibliográficas	47
2.2	Artigo 2 - Funções de pedotransferência para a estimativa dos parâmetros de ajustes da curva de retenção de água no solo	52
2.2.1	Introdução	53
2.2.2	Material e métodos	55
2.2.3	Resultados e discussão	60
2.2.4	Conclusões	71
2.2.5	Referências bibliográficas	71
3.	CONCLUSÕES GERAIS	76

1. INTRODUÇÃO GERAL

Considerada a principal fronteira agrícola do Brasil, o Bioma Cerrado é responsável por quase 46% de toda a produção de cereais, leguminosas e oleaginosas no país (IBGE, 2021). No entanto, a abertura de novas áreas agrícolas na região está cada vez mais proibitiva, tornando-se essencial a intensificação da agricultura a fim de suprir as demandas atuais e futuras por alimento. Além disso, o aumento da área irrigada, tecnologia essencial para viabilizar a intensificação da agricultura na região, impacta ainda mais o atual cenário de aumento nas disputas pelo uso de recursos hídricos, sobretudo, em bacias hidrográficas onde já se observam o comprometimento da disponibilidade hídrica.

Reduzir a quantidade de água que é retirada dos mananciais pelos diversos usos é uma das formas mais factíveis de se minimizar os conflitos pelo uso de água. Isto pode ser alcançado por meio de um planejamento integrado de bacias hidrográficas que estabeleça estratégias para aumentar a eficiência de uso de água dos diversos usuários, sobretudo a irrigação. Esse planejamento, entretanto, principalmente nas áreas agrícolas, tem sido dificultado pela inexistência, em grande parte das regiões do país, de dados na escala apropriada sobre clima, solo e planta.

A condutividade hidráulica do solo saturado, que expressa a capacidade do solo em transmitir água, e a curva de retenção, que representa a relação entre a umidade do solo e a energia com a qual a água está retida, são informações fundamentais em qualquer estudo relacionado à dinâmica de água no solo, influenciando diretamente nos processos de escoamento superficial, infiltração e armazenamento.

A obtenção dessas propriedades físico-hídricas do solo, entretanto, apresenta dependência de rotinas trabalhosas para sua aquisição, que muitas vezes, inviabilizam a sua obtenção, principalmente em grandes áreas, como é o caso do Bioma Cerrado. Além disso, à medida que a análise passa do nível macrorregional para o local, aumentando a escala de trabalho, há necessidade de maior detalhamento das amostragens, aumentando o esforço e o custo do trabalho.

A quase inexistência de parâmetros físico-hídricos representativos dos solos do Cerrado abre uma oportunidade para a utilização de Funções de Pedotransferência (FPTs), que são funções que possibilitam estimar propriedades físico-hídricas do solo de difícil obtenção a partir de dados pedológicos facilmente mensurados e de custo acessível, como os teores de areia, silte, argila, teor de matéria orgânica, densidade do solo e entre outros (PACKEPSKY;

RAWLS, 2004; PACHESPKY; PARK, 2015).

Várias FPTs foram desenvolvidas nos últimos anos para estimativa de parâmetros físico-hídricos, tais como, a densidade do solo (CHEN et al., 2018), condutividade hidráulica do solo saturado (OTTONI et al., 2019) e parâmetros da curva de retenção de água no solo (CASTELLINI; LOVINO, 2019; AULER, PIRES; PINEDA, 2017; CONTRERAS; BONILLA, 2018).

A precisão e confiabilidade das FPTs dependem das características do conjunto de dados (escala, variáveis preditoras, tamanho da amostra, heterogeneidade etc.) e das técnicas utilizadas. Vários métodos foram utilizados para a obtenção das FPTs nos últimos anos. Pachepsky e Rawls (2004) categorizam as FPTs em dois tipos, quando as FPTs são desenvolvidas por modelos lineares e não lineares e àquelas desenvolvidas por modelos com técnicas de mineração e exploração dos dados. A regressão linear múltipla (VERECKEN et al., 1989; BERG et al., 1997; TOMASELLA; HODNETT; ROSSATTO, 2000) foi um dos métodos mais utilizados, mas atualmente tem ganhado força as técnicas de aprendizado de máquina, como os algoritmos Random Forest ou Árvores de Decisão (TÓTH et al., 2015; SHIRI et al., 2017; ARAYA; GHEZZEHEI, 2019), K-Nearest Neighbors (GUNARATHNA et al., 2019; KOTLAR; IVERSEN; LIER, 2019), Support Vector Machine (KAINGO et al., 2018; MADY; SHEIN, 2018) e Redes Neurais Artificiais (D'EMILLIO et al., 2018; KALUMBA et al., 2020).

Trabalhos utilizando o aprendizado de máquina para a estimativa da condutividade hidráulica do solo, curva de retenção de água e seus respectivos parâmetros de ajuste apresentaram resultados satisfatórios quando comparados com métodos mais simples (TWARAKAVI et al., 2009; ARAYA; GHEZZEHEI, 2019; GUNARATHNA et al., 2019; KOTLAR; IVERSEN; LIER, 2019). Essa potencialidade de aplicação do aprendizado de máquina está atrelada ao fato de que as relações entre as propriedades físico-hídricas e as variáveis de obtenção simples são complexas, e tendem à um comportamento não-linear, exigindo assim métodos mais robustos capazes de modelar melhor esses dados (SHEN et al., 2018; ARAYA; GHEZZEHEI, 2019).

As FPTs desenvolvidas para estimar a condutividade hidráulica, a curva de retenção e seus parâmetros de ajuste em solos brasileiros (TOMASELLA; HODNETT; ROSSATO, 2000; TOMASELLA; HODNETT, 1997; OTTONI et al., 2019) foram realizadas em diferentes regiões do país, como a região amazônica (TOMASELLA; HODNETT, 1996; TOMASELLA; HODNETT, 1998; KOTLAR; LIER; BRUTO, 2020), nordeste (OLIVEIRA et al., 2002; BARROS et al., 2013), sudeste (SILVA et al., 2008), sul (MICHELON et al.,

2010), assim como em algumas regiões do Cerrado (RODRIGUES; MAIA; DA SILVA, 2011; RODRIGUES; MAIA, 2011, MEDRADO; LIMA, 2014).

No entanto, pode-se dizer que ainda são poucas as FPTs desenvolvidas especificamente para o Bioma Cerrado. Além disso, existem ainda uma incerteza sobre o desempenho dos diferentes algoritmos de aprendizagem de máquina na estimativa das FPTs, sendo importante desenvolver estudos a fim de compreender melhor as dinâmicas de água nos solos do Cerrado.

Adicionalmente a isso, existem diversos modelos que podem ser utilizados para representar a curva de retenção de água no solo, tais como os modelos de Brooks-Corey (1964), Campbell (1974), van Genuchten (1980), Hutson e Cass (1987), Durner (1994), Fredlund-Xing (1994), Kosugi (1994), Seki (2007) e entre outros. Essas equações, entretanto, para serem aplicadas precisam ser parametrizadas e existem dúvidas quando ao desempenho dessas equações para os solos do Cerrado.

A dissertação, cujo objetivo, foi desenvolver FPTs para estimativa de propriedades físico-hídricas do solo do Bioma Cerrado, foi estruturado na forma de dois artigos. No primeiro, são desenvolvidos FPTs para a estimativa da condutividade hidráulica saturada e umidades do solo utilizando algoritmos de aprendizado de máquina. E no segundo artigo, foram avaliados o desempenho de modelos de ajuste da curva de retenção de água e desenvolvidas FPTs para a estimativa dos parâmetros de ajuste da curva de retenção de água utilizando algoritmos de aprendizado de máquina.

1.1 Referências Bibliográficas

ARAYA, S. N.; GHEZZEHEI, T. A. Using machine learning for prediction of saturated hydraulic conductivity and its sensitivity to soil structural perturbations. **Water Resources Research**, v. 55, n. 7, p. 5715-5737, 2019.

AULER, A. C.; PIRES, L. F.; PINEDA, M. C. Influence of physical attributes and pedotransfer function for predicting water retention in management systems. **Revista Brasileira de Engenharia Agrícola e Ambiental**, v.21, n.11, p.746-751, 2017.

BARROS, A. H. C. B.; LIER, Q. J.; MAIA, A. H. N.; SCARPARE, F. V. Pedotransfer functions to estimate water retention parameters of soils in northeastern Brazil. **Revista Brasileira de Ciência do Solo**, v. 37, p. 379-391, 2013.

BERG, M. V. D.; KLAMT, E.; REEUWIJK, L. P. V.; SOMBROEK, W. G. Pedotransfer functions for the estimation of moisture characteristics of Ferralsols and related soils. **Geoderma**, v. 78, p.161-180, 1997.

BROOKS, R. H.; COREY, A. T. Hydraulic properties of porous media, Hydrol. Paper 3,

Colorado State Univ., Fort Collins, CO, USA, 1964.

CAMPBELL, G. S. A simple method for determining unsaturated conductivity from moisture retention data. **Soil Science**, v. 117, p. 311-314, 1974.

CASTELLINI, M.; LOVINO, M. Pedotransfer functions for estimating soil water retention curve of Sicilian soils. **Archives of Agronomy and Soil Science**, v. 65, p. 1401-1416, 2019.

CHEN, S.; RICHER-DE-FORGES, A. C.; SABY, N. P. A.; MARTIN, M. P.; WALTER, C.; ARROUAYS, D. Building a pedotransfer function for soil bulk density on regional dataset and testing its validity over a larger area. **Geoderma**, v. 312, p. 52–63, 2018.

CONTRERAS, C. P.; BONILLA, C. A. A comprehensive evaluation of pedotransfer functions for predicting soil water content in environmental modeling and ecosystem management. **Science of The Total Environment**, v. 644, p. 1580-1590, 2018.

D'EMILIO, A.; AIELLO, R.; CONSOLI, S.; VANELLA, D.; IOVINO, M. Artificial Neural Networks for Predicting the Water Retention Curve of Sicilian Agricultural Soils. **Water**, v. 10, p. 1431, 2018.

DURNER, W., Hydraulic conductivity estimation for soils with heterogeneous pore structure. **Water Resources Research**, v. 32, n. 9, p. 211-223, 1994.

FREDLUND, D. G; XING, A. Equations for the soil water characteristic curve. **Canadian Geotechnical Journal**, v. 31, p. 521-532, 1994.

GHANBARIAN, B.; TASLIMITEHRANI, V.; PACHEPSKY, Y. A. Scale-Dependent Pedotransfer Functions Reliability for Estimating Saturated Hydraulic Conductivity. **Catena**, vol. 149, p. 374-380, 2017.

HUTSON, J. L.; CASS, A. A retentivity function for use in soil-water simulation models. **J. Soil Science**, v. 38, p. 105-113, 1987.

IBGE (Instituto Brasileiro de Geografia e Estatística), Diretoria de Pesquisas, Coordenação de Agropecuária, Levantamento Sistemático da Produção Agrícola – jan. 2021.

KAINGO, J.; TUMBO, S. D.; KIHUPI, N. I.; MBILINYI, B. P. Prediction of soil moisture-holding capacity with support vector machines in dry subhumid tropics. **Applied and Environmental Soil Science**, v. 2018, 2018.

KALUMBA, M.; BAMPS, B. NYAMBE, I.; DONDEYNE, S.; ORSHOVEN, J. V. Development and functional evaluation of pedotransfer functions for soil hydraulic properties for the Zambezi River Basin. **European Journal of Soil Science**, v. 72, n. 4, p. 1559-1574, 2020.

KOSUGI, K. Three-parameter lognormal distribution model for soil water retention, **Water Resources Research**, v. 30, n. 4, p. 891–901, 1994.

KOTLAR, A. M.; IVERSEN, B. V; LIER, Q. J. Evaluation of Parametric and Nonparametric

Machine-Learning Techniques for Prediction of Saturated and Near-Saturated Hydraulic Conductivity. **Vadose Zone Journal**, v. 18, n. 1, 2019.

KOTLAR, A. M.; LIER, Q. J. BRITO, E. S. Pedotransfer functions for water contents at specific pressure heads of silty soils from Amazon rainforest. **Geoderma**, v. 361, 2020.

MADY, A. Y.; SHEIN, E. V. Support vector machine and nonlinear regression methods for estimating saturated hydraulic conductivity. Moscow **University Soil Science Bulletin**, v. 73, n. 3, p. 129–133, 2018.

MEDRADO, E. LIMA, J. E. F.W. Development of pedotransfer functions for estimating water retention curve for tropical soils of the Brazilian savanna. **Geoderma Regional**, v. 1, p. 59-66, 2014.

MICHELON, C. J.; CARLESSO, R.; OLIVEIRA, Z. B.; KNIES, A. E.; PETRY, M. T.; MARTINS, J. D. Funções de pedotransferência para estimativa da retenção de água em alguns solos do Rio Grande do Sul. **Ciência Rural**, v. 40, n. 4, p. 848-853, 2010.

OTTONI, M. V.; OTTONI FILHO, T. B.; LOPEZ-ASSAD, M. L. R. C.; ROTUNNO, O. C. Pedotransfer functions for saturated hydraulic conductivity using a database with temperate and tropical climate soils, **Journal of Hydrology**, v. 575, p. 1345-1358, 2019.

OLIVEIRA, L. B.; RIBEIRO, M. R.; JACOMINE, P. K. T.; RODRIGUES, J. J. V; MARQUES, F. A. Funções de pedotransferência para predição da umidade retida a potenciais específicos em solos do estado de Pernambuco. **Revista Brasileira de Ciência do Solo**, v. 26, n. 2, p. 315-323, 2002.

PACHEPSKY, Y.; RAWLS, W.J. Development of pedotransfer functions in soil hydrology. Elsevier, Amsterdam, Netherlands, 2004.

PACHEPKY, Y.; PARK, Y. Saturated Hydraulic Conductivity of US Soils Grouped According to Textural Class and Bulk Density. **Soil Science Society of America Journal**, vol. 79, n. 4, p. 1094-1100, 2015.

RODRIGUES, L. N.; MAIA, A. H. N. **Funções de pedotransferência para estimar a condutividade hidráulica saturada e as umidades de saturação e residual do solo em uma bacia hidrográfica do Cerrado**. XIX Simpósio Brasileiro de Recursos Hídricos. **Anais...** Macéio - AL: Associação Brasileira de Recursos Hídricos, 2011.

RODRIGUES, L. N; MAIA, A. H. N; SILVA, R. N; **Funções de pedotransferência para estimar capacidade de campo, ponto de murcha permanente e densidade global em solos de uma bacia hidrográfica do Bioma Cerrado**. XL Congresso Brasileiro de Engenharia Agrícola. **Anais...** Cuiabá – MT: Associação Brasileira de Engenharia Agrícola, 2011.

SEKI, K. SWRC Fit - A nonlinear fitting program with a water retention curve for soils having unimodal and bimodal pore structure. **Hydrology and Earth System Sciences**, v.4, p. 407-437, 2007.

SILVA, A. P.; TORMENA, C. A.; FIDALSKI, J.; IMHOFF, S. Funções de pedotransferência para as curvas de retenção de água e de resistência do solo à penetração. **Revista Brasileira de**

Ciência do Solo, v. 32, p. 1-10, 2008.

SHEN, C., LALOY, E., ELSHORBAGY, A., ALBERT, A., BALES, J., CHANG, F. J. HESS
Opinions: Incubating deep-learning-powered hydrologic science advances as a community. **Hydrology and Earth System Sciences**, v. 22, n. 11, p. 5639-5656, 2018.

SHIRI, J.; KESHAVARZI, A.; KISI, O.; KARIMI, S. Using soil easily measured parameters for estimating soil water capacity: Soft computing approaches. **Computers and Electronics in Agriculture**, v. 141, p. 327-339, 2017.

TOMASELLA, J. HODNETT, M. G. Soil hydraulic properties and van Genuchten parameters for an oxisol under pasture in central Amazonia. In: Gash, J. H. C.; Nobre, C. A.; Roberts, J. M.; Victoria, R. L. **Amazonian Deforestation and Climate**. Chichester: John Wiley, p. 101-124, 1996.

TOMASELLA, J., HODNETT, M. G. Estimating unsaturated hydraulic conductivity of Brazilian soils using soil-water retention data. **Soil Science**, v. 162, n. 10, 703–712, 1997.

TOMASELLA, J., HODNETT, M. G. Estimating soil water retention characteristics from limited data in Brazilian Amazonia. **Soil Science**, v. 163, n. 3, 1998.

TOMASELLA, J., HODNETT, M.G., ROSSATO, L. Pedotransfer functions for the estimation of soil water retention in Brazilian soils. **Soil Science Society of America Journal**, v. 64, p. 327–338, 2000.

TÓTH, B., WEYNANTS, M., NEMES, A., MAKÓ, A., BILAS, G., TÓTH, G. New generation of hydraulic pedotransfer functions for Europe. **European Journal of Soil Science**, v. 66, n. 1, p. 226–238, 2015.

TWARAKAVI, N. K. C., SIMUNEK, J., SCHAAP, M. G. Development of pedotransfer functions for estimation of soil hydraulic parameters using support vector machines. **Soil Science Society of America Journal**, v. 73, n. 5, 1443-1452, 2009.

VAN GENUCHTEN, M. T. A closed-form equation for predicting the hydraulic conductivity of unsaturated soils. **Soil Science Society of America Journal**, v. 44, p. 892- 898, 1980.

VERECKEN, H.; MAES, J.; FEYEN, J.; DARIUS, P. Estimating the soil moisture retention characteristic from texture, bulk density, and carbon content. **Soil Science**, v. 148, p. 389-403, 1989.

2. ARTIGOS CIENTÍFICOS

2.1 Artigo 1 - Funções de pedotransferência para a estimativa da condutividade hidráulica e da umidade do solo para o Cerrado brasileiro

Resumo

Nos últimos anos, a principal região agrícola do Brasil, o Bioma Cerrado, tem apresentado uma crescente intensificação da agricultura e conflitos pelo uso da água. A necessidade de estratégias capazes de diminuir a retirada de água dos mananciais, principalmente da irrigação, tem sido prejudicada pela carência de dados de propriedades físico-hídricas do solo representativas da região. Nesse contexto, torna-se necessário a utilização das Funções de Pedotransferência (FPTs). O objetivo do presente trabalho foi desenvolver FPTs por meio de algoritmos de aprendizagem máquina para estimar a condutividade hidráulica do solo saturado (K_s) e as umidades do solo nas tensões de 0, 6, 10, 33, 100 e 1500 kPa para o Bioma Cerrado, sendo que para a estimativa de K_s foram utilizadas duas abordagens, uma sem o preenchimento de dados faltantes e outra, com o preenchimento de dados faltantes. Cinco modelos foram testados: Regressão Linear Múltipla (RLM), Multiple Adaptive Regression Splines (MARS), Random Forest (RF), Support Vector Regression (SVR) e K Nearest Neighbors (KNN). Quatro combinações de dados do solo foram avaliadas, sendo que as variáveis preditoras utilizadas em cada conjunto foram diferentes. No conjunto A1 utilizou-se: teores de areia (Ar), silte (Si) e argila (Ag); no conjunto A2: Ar, Si, Ag e densidade do solo (Ds); no conjunto A3: Ar, Si, Ag, Ds, densidade de partículas (Dp), porosidade total (Pt), microporosidade (Micro) e macroporosidade (Macro); no conjunto A4: Ar, Si, Ag, Ds, Dp, Pt, Micro, Macro, umidade do solo na capacidade de campo (θ_{10}) e umidade do solo no ponto de murcha permanente (θ_{1500}). O conjunto A4 juntamente com os modelos RF e SVR apresentaram os melhores desempenhos na estimativa da K_s , contudo, a capacidade preditiva para tal propriedade foi mediana. O preenchimento de dados faltantes para K_s melhorou o desempenho dos conjuntos A1 e A2. Já para as umidades do solo, os modelos RF, SVR e MARS apresentaram os melhores desempenhos com baixos valores de RMSE e ME, e R^2 superiores a 0,8 utilizando os conjuntos preditores A3 e A4.

Palavras-chave: aprendizado de máquina; dados faltantes; irrigação.

2.1.1 Introdução

Considerada a principal fronteira agrícola do Brasil, o Bioma Cerrado é responsável por quase 46% de toda a produção de cereais, leguminosas e oleaginosas produzidas no país (IBGE, 2021). No entanto, a abertura de novas áreas agrícolas na região está cada vez mais proibitiva, tornando-se essencial a intensificação da agricultura. Além disso, o aumento da área irrigada impacta ainda mais o atual cenário, sobretudo, em bacias hidrográficas onde já se observam o comprometimento da disponibilidade hídrica.

Reduzir a quantidade de água que é retirada dos mananciais pelos diversos usos é uma das formas mais factíveis de se minimizar os conflitos. Isto pode ser alcançado por meio de um planejamento integrado de bacias hidrográficas que estabeleça estratégias para aumentar a eficiência de uso de água dos diversos usuários, sobretudo a irrigação. Esse planejamento, entretanto, principalmente nas áreas agrícolas, tem sido dificultado pela inexistência, em grande parte das regiões do país, de dados na escala apropriada sobre clima, solo e planta.

A curva de retenção de água, que representa a relação entre a umidade do solo e a energia com a qual a água está retida, e a condutividade hidráulica do solo saturado, que expressa a capacidade do solo em transmitir água, são informações fundamentais em qualquer estudo relacionado à dinâmica de água no solo, influenciando diretamente nos processos de escoamento superficial, infiltração e armazenamento.

Contudo, esses parâmetros apresentam grande variabilidade espacial, além de elevado custo amostral e laboratorial pelos métodos diretos, e dependência de rotinas trabalhosas para sua aquisição, que muitas vezes, inviabilizam a sua obtenção, principalmente em grandes áreas, como é o caso do Bioma Cerrado. Além disso, à medida que a análise passa do nível macrorregional para o local, aumentando a escala de trabalho, há necessidade de maior detalhamento das amostragens, aumentando o esforço e o custo do trabalho.

A quase inexistência de parâmetros físico-hídricos representativos dos solos do Cerrado abre uma oportunidade para a utilização das Funções de Pedotransferência (FPTs), que são funções que permitem estimar propriedades físico-hídricas do solo de difícil obtenção a partir de dados pedológicos facilmente mensurados e de custo acessível, como os teores de areia, silte, argila, teor de matéria orgânica, densidade do solo e entre outros (PACKEPSKY; RAWLS, 2004; PACHESPKY; PARK, 2015).

Várias FPTs foram desenvolvidas nos últimos anos para estimativa de parâmetros físico-hídricos, tais como, a densidade do solo (CHEN et al., 2018), condutividade hidráulica do solo saturado (OTTONI et al., 2019) e parâmetros da curva de retenção de água no solo

(CASTELLINI; LOVINO, 2019; AULER, PIRES; PINEDA, 2017; CONTRERAS; BONILLA, 2018).

A precisão e confiabilidade das FPTs dependem das características do conjunto de dados (escala, variáveis preditoras, tamanho da amostra, heterogeneidade etc.) e das técnicas utilizadas. Vários métodos foram utilizados para a obtenção das FPTs nos últimos anos. Pachepsky e Rawls (2004) categorizam as FPTs em dois tipos, quando as FPTs são desenvolvidas por modelos lineares e não lineares e àquelas desenvolvidas por modelos com técnicas de mineração e exploração dos dados. A regressão linear múltipla (VERECKEN et al., 1989; BERG et al., 1997; TOMASELLA; HODNETT; ROSSATTO, 2000) foi um dos métodos mais utilizados, mas atualmente tem ganhado força as técnicas de aprendizado de máquina, como os algoritmos Random Forest ou Árvores de Decisão (TÓTH et al., 2015; SHIRI et al., 2017; ARAYA; GHEZZEHEI, 2019), K-Nearest Neighbors (GUNARATHNA et al., 2019; KOTLAR et al., 2019), Support Vector Machine (KAINGO et al., 2018; MADY; SHEIN, 2018;) e Redes Neurais Artificiais (D'EMILLIO et al., 2018; KALUMBA et al., 2020).

Trabalhos utilizando o aprendizado de máquina para a estimativa de pontos específicos da curva de retenção de água e condutividade hidráulica do solo apresentaram resultados satisfatórios quando comparados com métodos mais simples (ARAYA; GHEZZEHEI, 2019; GUNARATHNA et al., 2019; KOTLAR; IVERSEN; LIER, 2019; AMANABADI et al., 2020). Essa potencialidade de aplicação do aprendizado de máquina está atrelada ao fato de que as relações entre as propriedades físico-hídricas e as variáveis de obtenção simples são complexas, e tendem à um comportamento não-linear, exigindo assim métodos mais robustos capazes de modelar melhor esses dados (SHEN et al., 2018; ARAYA; GHEZZEHEI, 2019).

As FPTs desenvolvidas para estimar a curva de retenção de água e a condutividade hidráulica em solos brasileiros (TOMASELLA; HODNETT; ROSSATO, 2000; TOMASELLA; HODNETT, 1997; OTTONI et al., 2019) foram realizadas em diferentes regiões do país, como a região amazônica (TOMASELLA; HODNETT, 1998; KOTLAR et al., 2020), nordeste (OLIVEIRA et al., 2002; BARROS et al., 2013), sudeste (SILVA et al., 2008), sul (REICHERT et al., 2009; MICHELON et al., 2010), assim como em algumas regiões do Cerrado (RODRIGUES; MAIA; DA SILVA, 2011; RODRIGUES; MAIA, 2011, MEDRADO; LIMA, 2014). No entanto, pode-se dizer que ainda são poucas FPTs desenvolvidas para o Bioma Cerrado, bem como uma avaliação de diferentes algoritmos e parâmetros de entrada, sendo importante desenvolver estudos a fim de compreender melhor as dinâmicas de água no solo, principalmente em modelos de simulações, e suas interações na

agricultura.

Nesse contexto, o objetivo deste trabalho foi desenvolver funções de pedotransferência para estimar a condutividade hidráulica do solo saturado e a umidade do solo para o Bioma Cerrado a partir de algoritmos de aprendizado de máquina.

2.1.2 Material e Métodos

2.1.2.1 Obtenção e pré-processamento da base de dados

Os dados que foram utilizados no desenvolvimento deste trabalho foram obtidos pelo Grupo de Pesquisa em Recursos Hídricos da Embrapa Cerrados e pelo conjunto de dados Hybras (OTTONI et al., 2018).

Inicialmente, foi realizada a compilação dessas bases de dados que apresentam valores de propriedades físico-hídricas dos solos do Brasil, e aquelas amostragens pertencentes ao limite territorial do Bioma Cerrado com um *buffer* de até 100 km foram selecionadas, totalizando 1708 amostras.

Além da localização, o critério de seleção das amostras incluiu a disponibilidade das variáveis a serem estimadas, condutividade hidráulica do solo saturado (K_s) e conteúdo de água em diversos potenciais matriciais, juntamente com os teores de areia, silte, argila, densidade do solo, densidade de partícula, porosidade total, macroporosidade e microporosidade.

Em relação a existência de dados incompletos na base de dados, alguns atributos do solo foram capazes de serem estimados, como, a porosidade total que, quando inexistente, foi estimada com base no conteúdo de água na saturação; a microporosidade, estimada a partir do valor correspondente ao conteúdo de água na tensão de 6 kPa, e a macroporosidade pela diferença entre a porosidade total e a microporosidade (DIAS JÚNIOR et al., 2000).

No pré-processamento dos dados, as seguintes premissas foram avaliadas: (i) somatório dos teores de areia, silte e argila igual a 100%; (ii) valor da densidade de partícula maior que o valor da densidade do solo; e (iii) os valores de porosidade total não devem exceder 60%. As amostras que não atenderam essas premissas foram descartadas.

Posteriormente, um novo conjunto de dados para cada variável a ser estimada foi gerado, descartando-se as variáveis que apresentaram altas porcentagens de valores faltantes. Para tal, foi desenvolvida uma rotina, no software R (R CORE TEAM, 2019) utilizando o pacote *misscompare* (VARGA, 2020), para selecionar uma combinação ótima de parâmetros, conforme a quantidade de amostras e variáveis.

Valores faltantes em bases de dados são comuns, entretanto, pouco é abordado na

literatura as formas de solucionar esse tipo de problema. Em alguns casos, uma opção para lidar com os valores ausentes é realizar o preenchimento dos mesmos. Sendo assim, foi utilizado duas abordagens para a estimativa da K_s , a fim de avaliar se o preenchimento de falhas melhora ou não o desempenho das FPTs. Na primeira abordagem, a K_s foi estimada com base na série sem o preenchimento de dados faltantes, e na segunda abordagem, a estimativa foi feita após o preenchimento de dados faltantes.

O preenchimento de falhas foi realizado utilizando o pacote *misscompare*, que testa dezesseis modelos de preenchimento, sendo assim, foi selecionado o melhor modelo conforme os índices estatísticos: erro médio absoluto (MAE), raiz do erro médio quadrático (RMSE) e teste de Kolmogorov-Smirnov (KS). Para esse preenchimento foi adotado o mecanismo MAR (*Missing at Random*), ou seja, assumiu-se uma probabilidade de que a falta dos dados depende em certa medida dos dados observados (BUUREN, 2012). Na primeira abordagem, a série foi composta por 140 amostras de K_s e na segunda, por 431 amostras de K_s .

2.1.2.2 Desenvolvimento das funções de pedotransferência

Uma vez que a disponibilidade de dados de atributos de solos para a região do Cerrado brasileiro é muito variável, o subconjunto de treinamento foi organizado considerando quatro diferentes conjuntos de preditores, sendo eles, A1: areia, silte e argila; A2: areia, silte, argila e densidade do solo; A3: areia, silte, argila, densidade do solo, densidade de partícula, porosidade total, macroporosidade e microporosidade e A4: areia, silte, argila, densidade do solo, densidade de partícula, porosidade total, macroporosidade, microporosidade, umidade de saturação, umidade na capacidade de campo (10 kPa) e umidade no ponto de murcha permanente (1500 kPa).

No caso da estimativa da umidade do solo nas tensões de 0 e 6 kPa, não foram utilizadas as variáveis porosidade total e microporosidade no conjunto A1, respectivamente, devido ao efeito *dataleakage* que ocorre quando os modelos utilizam informações da própria variável que se pretende estimar, podendo resultar em um sobreajuste (HAREL et al., 2012).

Por fim, para cada variável a ser estimada, utilizou-se os diferentes conjuntos de preditores em cinco modelos de aprendizado de máquina: Regressão Linear Múltipla (RLM), Multivariate Adaptive Regression Splines (MARS), Random Forest (RF), Support Vector Regression (SVR) e K-Nearest Neighbors (KNN).

Além disso, foi obtido para cada variável estimada o modelo nulo que é o modelo mais simples que pode ser definido ou ajustado, e para isso foi utilizada a média da variável estimada

como parâmetro para a obtenção do índices estatísticos ME e RMSE, permitindo verificar se os modelos desenvolvidos para as FPTs apresentam um desempenho melhor ou não que o modelo nulo.

2.1.2.3 Modelos para o desenvolvimento das funções de pedotransferência

Regressão Linear Múltipla

A Regressão Linear Múltipla (RLM) consiste em estimar a relação da variável dependente/resposta (Y_i) a partir de duas ou mais variáveis independentes/preditoras (X_n). Portanto, foram ajustadas equações de RLM, correlacionando as variáveis a serem estimadas para cada conjunto de preditores (Equação 1).

$$Y_i = \beta_{i,0} + \beta_{i,1} \cdot X_1 + \dots + \beta_{i,n} \cdot X_n \quad (1)$$

Em que: Y_i = variável a ser estimada (condutividade hidráulica do solo saturado e umidade do solo nas tensões específicas); $\beta_{i,0}$ = intercepto da regressão linear múltipla; $\beta_{i,1} \dots \beta_{i,n}$ = coeficientes angulares vinculados às variáveis preditoras do solo; $X_1 \dots X_n$ = variáveis preditoras do solo.

Multivariate Adaptive Regression Splines

O Multivariate Adaptive Regression Splines (MARS) é uma técnica de regressão não paramétrica que modela automaticamente a não linearidade e as interações entre variáveis (FRIEDMAN, 1991). Os conjuntos de treinamento foram divididos em segmentos lineares e para cada conjunto foram ajustadas à curvas polinomiais (*splines*), e posteriormente, unidas por meio de nós. Para isto, foi utilizado o pacote *earth* (MILBORROW, 2019), na qual foram construídos modelos com diferentes números de interações e nós, e selecionado o modelo que apresentasse o menor RMSE.

Random Forest

O Random Forest (RF) é um modelo que combina árvores de regressão (BREIMAN, 1993). Para cada árvore gerada foi realizada uma amostragem *bootstrap* e a quantidade de variáveis amostradas é controlada pelo hiperparâmetro *mtry*. A estimativa final foi baseada na média dos valores estimados em cada árvore (RAHMAN et al., 2016).

Para determinar o número ideal de variáveis selecionadas aleatoriamente para construção de cada árvore, foram construídos modelos utilizando números diferentes de variáveis e selecionado o modelo que apresentasse o menor RMSE.

Support Vector Regression

O Support Vector Regression (SVR) tem como princípio o ajuste do hiperplano que separa os pontos em um espaço n-dimensional, sendo n o número de variáveis preditoras (VAPNIK, 1995). Para isso, foi utilizado a função kernel radial, que é um dos kernels mais comumente usados, otimizando os valores dos hiperparâmetros C (custo) e γ (gama), responsáveis pela tolerância de ajuste dos modelos criados, e selecionado o modelo que apresentasse o menor RMSE.

K-Nearest Neighbors

O K-Nearest Neighbors (KNN) é um modelo não paramétrico que estima a variável independente com base na média da distância dos seus vizinhos mais próximos do conjunto de dados, e o número de vizinhos é definido pelo hiperparâmetro k. Para isso, foram construídos modelos com diferentes valores de k, e selecionado o modelo que apresentasse o menor RMSE.

2.1.2.4 Validação e teste dos modelos

Para validação dos modelos gerados, foi utilizado o método *repeated holdout* no qual a base de dados foi dividida em dois subconjuntos considerados independentes e o processo foi repetido 100 vezes, sendo o primeiro subconjunto, o treinamento composto por 70% dos dados, e o segundo, o subconjunto de teste com 30% dos dados.

Para o ajuste dos hiperparâmetros de cada modelo foi aplicado o método de validação cruzada *k-folds* com repetições. Dessa forma, o conjunto de treinamento foi dividido aleatoriamente em k partes ($k = 10$), sendo que uma parte desse conjunto foi retirada para validação do modelo, gerando um novo conjunto de treinamento composto por k-1 partes. Logo após, foram realizados o ajuste do modelo e avaliação de desempenho da estimativa para a parte retirada, e consequentemente seu resultado armazenado. O processo de validação cruzada foi repetido n vezes ($n = 3$), de modo, que cada um das k partes fossem utilizados como teste para validação dos modelos. Por fim, o desempenho final da validação cruzada foi calculado pela média k vezes n resultados obtidos no processo.

Após a otimização dos hiperparâmetros o desempenho dos modelos foram avaliados

usando o conjunto de dados de teste, ou seja, os modelos realizaram uma estimativa em conjunto de dados não utilizado no treinamento, permitindo avaliar sua capacidade de generalização.

2.1.2.5 Desempenho e análise estatística

Para avaliar o desempenho das FPT desenvolvidas para a condutividade hidráulica do solo saturado e umidade do solo foram utilizados os seguintes índices estatísticos: erro médio (ME), raiz do erro médio quadrático (RMSE) e o coeficiente de determinação (R^2), sendo as duas primeiras comumente usadas na avaliação de FPT (SCHAAP et al., 2001, JULIÀ et al., 2004).

O R^2 expressa o grau de concordância entre os valores observados e estimados pelas FPT (Equação 2), assumindo valores entre 0 e 1. O ME expressa se o modelo superestima ($ME > 0$) ou subestima ($ME < 0$) (Equação 3) e o RMSE indica a magnitude do erro (Equação 4).

$$R^2 = \frac{\sum (\hat{y}_j - \bar{y}_j)^2}{\sum (y_j - \bar{y}_j)^2} \quad (2)$$

$$ME = \frac{1}{N} \sum_{j=1}^N y_j - \hat{y}_j \quad (3)$$

$$RMSE = \sqrt{\frac{1}{N} \sum_{j=1}^N (y_j - \hat{y}_j)^2} \quad (4)$$

Em que: y_j e \hat{y}_j são os valores estimados e observados, respectivamente; N é o número de amostras; $\sum (\hat{y}_j - \bar{y}_j)^2$ é a variância explicada pelo modelo e $\sum (y_j - \bar{y}_j)^2$ é a variância total.

Para analisar os resultados de forma mais precisa, foi verificado se há diferença entre os desempenhos dos modelos utilizando o teste não paramétrico de Friedman (DEMNSAR, 2006). O teste baseia-se na comparação de desempenhos (*rank*) e, portanto, para cada um dos modelos avaliados foi determinado a posição, ordenando do melhor para o pior, e o teste retomando à valores de 0 ou 1, sendo 0 há diferença e 1 não há diferença entre os modelos.

No entanto, o teste de Friedman não permite discernir quais modelos apresentam

diferença estatística, sendo assim utilizou-se também o teste de Nemenyi (NEMENYI, 1963). De acordo com esse teste, os modelos são significativamente diferentes entre si quando a subtração dos *ranks* médios dos modelos for igual ou maior que o valor da distância crítica (CD).

2.1.3 Resultados e discussão

2.1.3.1 Base de dados

O conjunto de amostras totais ($n = 1708$) reduziu após as atividades de pré-processamento. Na primeira etapa, foram eliminados 358 amostras inconsistentes, sendo a maioria referentes à valores de porosidade total superior a 60%. A segunda etapa consistiu em remover as variáveis e amostras com altas porcentagens de dados faltantes (Figura 2.1.1).

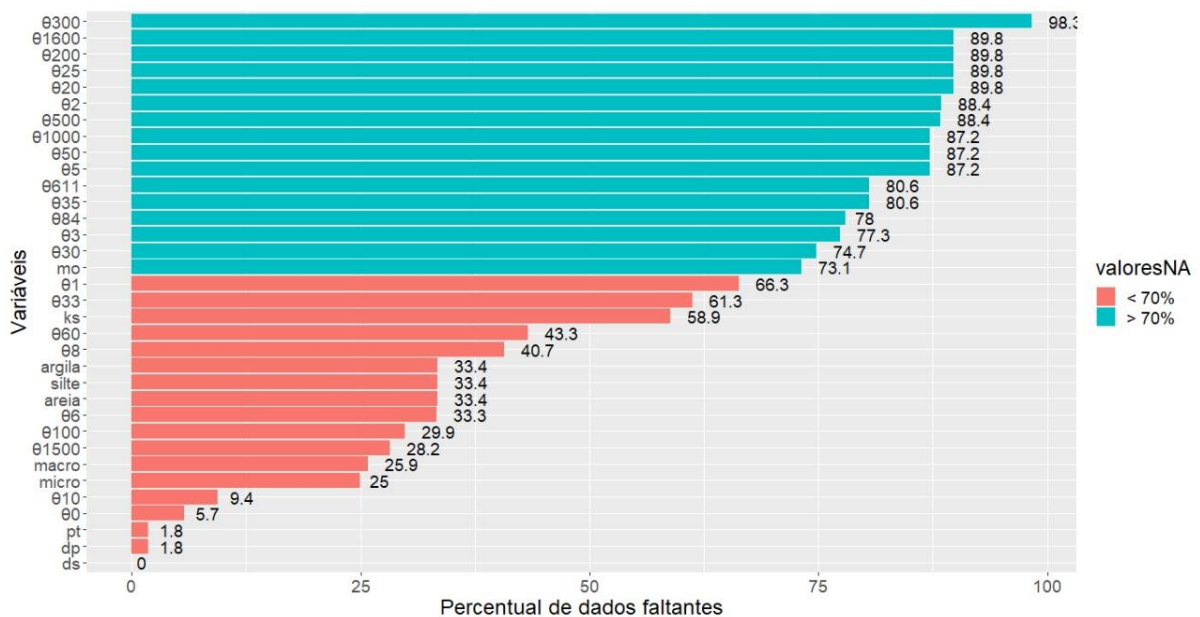


Figura 2.1.1. Percentual de dados faltantes para cada variável da base de dados inicial.

Das 34 variáveis que compõem a base de dados inicial, 33 apresentaram dados faltantes. A proporção de valores faltantes variou de 1,8 a 98,3%. A umidades do solo, de maneira geral, apresentou mais de 25% de dados faltantes, com exceção dos θ_0 e θ_{10} , com apenas 5,7 e 9,4%, respectivamente. Por fim, as variáveis com mais de 70% das amostras como dados faltantes foram excluídas.

Em função dos dados de umidade do solo disponíveis, foram estimados os θ_0 (umidade de saturação), θ_6 , θ_{10} (capacidade de campo), θ_{33} , θ_{100} e θ_{1500} (ponto de murcha permanente).

Desse modo, a Tabela 2.1.1 apresenta o número de amostras utilizadas para o desenvolvimento das FPTs para cada uma das variáveis estimadas.

Tabela 2.1.1. Número de amostras para cada variável a ser estimada após os pré-processamentos

Variável	Nº de amostras
Ks	140
Ks com preenchimento	431
θ_6 e θ_{33}	158
θ_s , θ_{10} , θ_{100} e θ_{1500}	268

Na Figura 2.1.2 apresentam-se os triângulos texturais referentes aos dados de condutividade hidráulica do solo saturado e umidades do solo nas tensões de 0, 6, 10, 33, 100 e 1500 kPa.

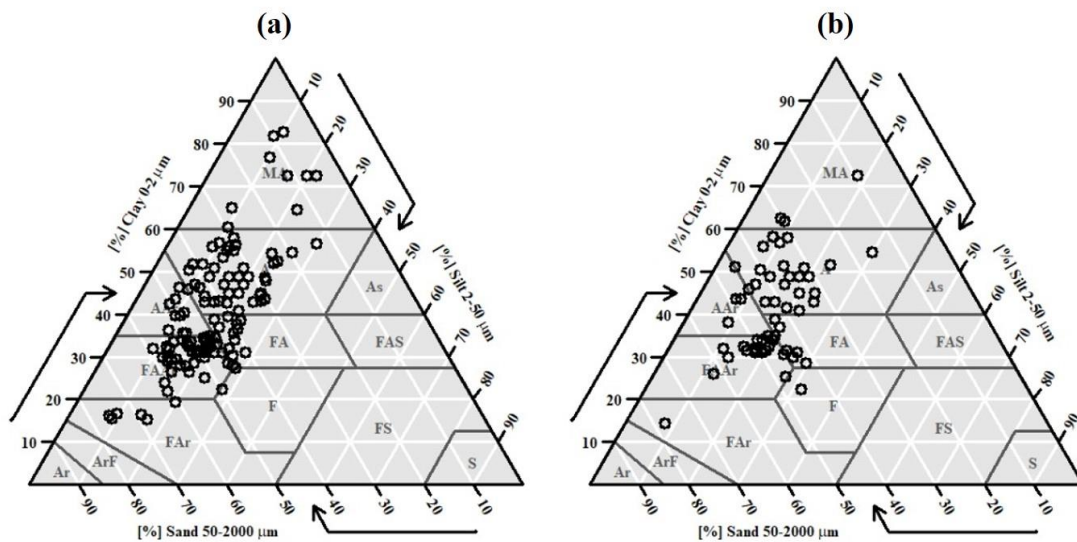


Figura 2.1.2. Triângulos texturais das bases de dados de treinamento (a) e teste (b) para condutividade hidráulica do solo saturado e umidades do solo nas tensões de 0, 6, 10, 33, 100 e 1500 kPa.

Verifica-se que maioria das amostras utilizadas para o desenvolvimento das FPTs no Bioma Cerrado são classificadas como argilosa e franco argilo-arenosa, ou seja, possuem porcentagens consideráveis de argila e areia em sua composição.

A fim de visualizar o comportamento dos teores de argila e areia, na Figura 2.1.3, são apresentados os histogramas dessas variáveis, no qual pode-se verificar que os percentuais de argila variam mais que os percentuais de areia. E a maioria das amostras possuem os percentuais de argila e areia entre 30 e 50%, sendo que para a areia têm-se um pico nos percentuais entre

30 e 40%.

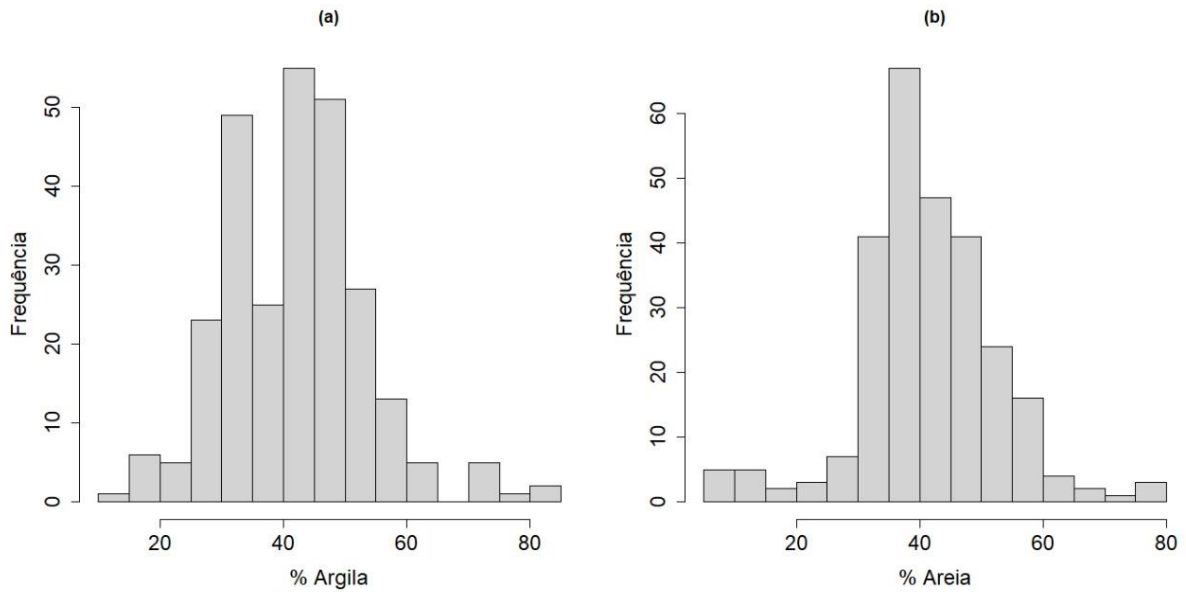


Figura 2.1.3. Histogramas dos percentuais de (a) argila e (b) areia das amostras de solos.

Nas Tabelas 2.1.2, 2.1.3 e 2.1.4 apresentam-se as estatísticas descritivas das variáveis preditoras utilizadas para o desenvolvimento das funções de pedotransferência da condutividade hidráulica do solo saturado, das umidades do solo nas tensões de 6 e 33 e das umidades nas tensões de 0, 10, 100 e 1500 kPa, respectivamente.

Tabela 2.1.2. Estatísticas descritivas das variáveis preditoras utilizadas para o desenvolvimento (subconjunto A) e teste (subconjunto B) das funções de pedotransferência da condutividade hidráulica do solo saturado (Ks) (n = 140)

Subconjunto	Estatística	Ks	Areia	Silte	Argila	Ds	Dp	Pt	Micro	Macro	θ_{10}	θ_{1500}
		mm/h	%	%	%	g/cm ³	g/cm ³	%	%	%	m ³ /m ³	m ³ /m ³
A (n = 100)	Média	34,23	40,6	16,2	43,4	1,39	2,62	46,9	39,59	7,42	0,379	0,299
	Máximo	246,25	62,0	24,0	55,8	1,585	2,70	59,63	50,43	22,92	0,505	0,435
	Mínimo	0,19	23,8	7,3	22,0	1,086	2,5	38,0	29,64	0,29	0,243	0,175
	Desvio Padrão	48,49	7,28	4,41	6,63	0,11	0,06	4,12	3,85	4,83	0,05	0,04
	CV	141,72	17,96	27,34	15,32	7,48	2,38	8,79	9,74	65,15	13,24	16,26
B (n = 40)	Média	33,49	39,3	17,5	43,4	1,38	2,61	47,13	39,93	7,54	0,384	0,311
	Máximo	176,85	60,5	24,0	54,9	1,52	2,70	53,73	49,65	19,86	0,505	0,431
	Mínimo	0,26	25,5	9,28	24,0	1,157	2,5	40,40	27,84	1,08	0,215	0,184
	Desvio Padrão	44,02	8,0	3,70	7,24	0,08	0,06	3,25	3,90	3,97	0,04	0,04
	CV	131,45	20,41	21,24	16,69	5,88	2,30	6,89	9,77	52,66	12,68	13,67

Ds = densidade do solo; Dp = densidade de partícula; Pt = porosidade total; Micro = microporosidade; Macro = macroporosidade; θ_{10} = umidade do solo na tensão de 10 kPa; θ_{1500} = umidade do solo na tensão de 1500 kPa; CV = coeficiente de variação.

Tabela 2.1.3. Estatísticas descritivas das variáveis preditoras utilizadas para o desenvolvimento (subconjunto A) e teste (subconjunto B) das funções de pedotransferência da umidade do solo nas tensões de 6 e 33 kPa (n = 158)

Subconjunto	Estatística	θ_6	θ_{33}	Areia	Silte	Argila	Ds	Dp	Pt	Micro	Macro	θ_{10}	θ_{1500}
		m ³ /m ³	m ³ /m ³	%	%	%	g/cm ³	g/cm ³	%	%	%	m ³ /m ³	m ³ /m ³
A (n = 112)	Média	0,413	0,372	38,8	14,9	46,4	1,39	2,63	46,64	38,71	7,93	0,383	0,305
	Máximo	0,525	0,496	75,7	24,0	82,8	1,59	2,83	56,41	50,43	24,35	0,505	0,435
	Mínimo	0,193	0,157	7,2	7,3	16,4	1,17	2,19	38,71	18,10	0,29	0,155	0,103
	Desvio Padrão	0,06	0,06	7,8	4,7	7,81	0,1	0,08	3,85	5,21	5,47	0,05	0,05
	CV	13,59	15,18	19,94	31,31	16,9	7,15	3,29	8,26	13,48	68,99	14,36	18,6
B (n = 46)	Média	0,394	0,366	38,7	14,32	47,1	1,34	2,57	47,99	36,58	11,40	0,374	0,294
	Máximo	0,496	0,462	77,7	24,0	81,8	1,497	2,70	58,46	46,97	30,54	0,487	0,398
	Mínimo	0,177	0,148	9,7	7,28	14,4	1,039	2,1	40,49	16,26	2,52	0,148	0,098
	Desvio Padrão	0,06	0,06	10,75	4,41	10,87	0,1236	0,136	3,82	5,87	6,52	0,055	0,05
	CV	14,19	15,89	27,76	30,74	22,97	9,23	5,0	7,96	16,05	57,17	14,70	19,16

Ds = densidade do solo; Dp = densidade de partícula; Pt = porosidade total; θ_{10} = umidade do solo na tensão de 10 kPa; θ_{1500} = umidade do solo na tensão de 1500 kPa; θ_6 = umidade do solo na tensão de 6 kPa; θ_{33} = umidade do solo na tensão de 33 kPa; θ_{1500} = umidade do solo na tensão de 1500 kPa; CV = coeficiente de variação.

Tabela 2.1.4. Estatísticas descritivas das variáveis preditoras utilizadas para o desenvolvimento (subconjunto A) e teste (subconjunto B) das funções de pedotransferência da umidade do solo nas tensões de 0, 10, 100 e 1500 kPa (n = 268)

Subconjunto	Estatística	θ_0	θ_{100}	Areia	Silte	Argila	Ds	Dp	Pt	Micro	Macro	θ_{10}	θ_{1500}
		m ³ /m ³	m ³ /m ³	%	%	%	g/cm ³	g/cm ³	%	%	%	m ³ /m ³	m ³ /m ³
A (n = 188)	Média	0,495	0,326	41,00	16,77	42,24	1,36	2,61	47,64	38,28	9,55	0,365	0,286
	Máximo	0,597	0,483	75,68	30,0	82,76	1,65	2,86	59,63	50,43	29,23	0,506	0,436
	Mínimo	0,388	0,138	5,52	7,28	15,25	1,06	2,19	37,64	18,10	0,29	0,156	0,104
	Desvio Padrão	0,05	0,06	11,97	5,17	11,82	0,12	0,08	4,69	4,65	6,16	0,05	0,05
	CV (%)	10,37	17,36	29,2	30,87	28,0	9,08	3,45	9,86	12,15	64,55	15,19	18,44
B (n = 80)	Média	0,493	0,327	41,45	16,58	41,94	1,35	2,609	48,02	38,10	9,71	0,363	0,283
	Máximo	0,521	0,455	77,68	31,5	72,48	1,58	2,735	59,62	48,91	25,54	0,488	0,409
	Mínimo	0,396	0,127	9,52	3,83	14,32	1,03	2,140	39,23	16,26	0,837	0,149	0,099
	Desvio Padrão	0,05	0,07	9,68	5,93	10,39	0,11	0,105	3,90	5,41	5,81	0,06	0,06
	CV (%)	8,75	20,52	23,36	35,77	24,78	8,18	4,052	8,12	14,19	59,88	17,61	22,06

Ds = densidade do solo; Pt = porosidade total; θ_0 = umidade do solo na tensão de 0 kPa; θ_6 = umidade do solo na tensão de 6 kPa; θ_{10} = umidade do solo na tensão de 10 kPa; θ_{33} = umidade do solo na tensão de 33 kPa; θ_{100} = umidade do solo na tensão de 100 kPa; θ_{1500} = umidade do solo na tensão de 1500 kPa; CV (%) = coeficiente de variação.

Analisando as tabelas, observa-se que a maioria das amostras são apresentam consideráveis percentuais de areia e argila, tendo os teores de argila variando de 5 a 88%, silte de 0,04 a 31,5 e areia de 4 a 94,9 %, com valores médios iguais a 44, 16 e 40%, respectivamente. Valores semelhantes de teores de argila foram observados por Medrado e Lima (2014), que desenvolveram FPTs para estimativa de parâmetros da curva de retenção no Bioma Cerrado, contudo, os teores de areia apresentados pelos autores, média de 32%, foram um pouco inferiores aos resultados observados neste trabalho.

Solos de textura argilosa geralmente tem uma maior capacidade de armazenamento de água. Nesse sentido, o teor de argila médio igual a 44% observado nos solos utilizados para o desenvolvimento das FPTs indica que, no geral, os solos do Bioma Cerrado tem uma boa capacidade de armazenamento de água. Além disso, solos com textura argilosa, principalmente, solos bem estruturados como os Latossolos, classe textural considerada predominante no Cerrado (REATTO, 2008), apresentam um comportamento contrário às generalizações aos processos hidráulicos do solo, sendo uma das razões da baixa precisão das FPTs desenvolvidas para regiões temperadas quando aplicadas em regiões tropicais.

Verifica-se, ainda, que solos com textura muito siltosa e siltosa são inexistentes nos conjuntos de dados. Tomasella, Hodnett e Rossato (2000) desenvolveram FPTs para diversas regiões do Brasil e encontraram teores de silte, em sua maioria, menores que 17%.

A densidade média dos solos foi igual a $1,39 \text{ g cm}^{-3}$, um pouco superior aos valores médios de $1,0$ e $1,14 \text{ g cm}^{-3}$ encontrados nos trabalhos de Medrado e Lima (2014) e Rodrigues, Maia e da Silva (2011) obtidos para o Cerrado, respectivamente. Como era de se esperar, a densidade de partícula foi a variável que apresentou menor coeficiente de variação, com mínimo de 2,3% e máxima de 5%. Em contrapartida, K_s apresentou o maior coeficiente de variação, com 141,7%, e isso está atrelado a alta variabilidade inerente dessa propriedade (MENEZES et al., 2006; BAIAMONTE et al., 2017).

Nota-se, ainda, que os valores de desvio padrão para as umidades do solo foram baixos, e os θ_{100} e θ_{1500} , pontos mais distantes da curva, apresentaram os maiores CV dentre as umidades, com valores na faixa de 18%. Os valores médios dos teores de água na capacidade de campo (θ_{10}) e ponto de murcha permanente (θ_{1500}) foram da ordem de $0,37$ e $0,292 \text{ m}^3 \text{ m}^{-3}$, respectivamente.

O modelo de preenchimento de dados faltantes que apresentou o melhor resultado foi o `pcaMethods` PPCA, e após o preenchimento para a estimativa da K_s , verificou-se que os valores médios de todas as variáveis se mantiveram semelhantes aos conjuntos de dados sem o preenchimento, exceto K_s , que apresentou média de $86,64 \text{ mm h}^{-1}$, ou seja, teve-se um aumento

em cerca de duas vezes. Além disso, os valores do CV de todas as variáveis aumentaram, indicando que com a ampliação do tamanho amostral houve um aumento da variabilidade dos dados.

2.1.3.2 Avaliação das funções de pedotransferência para a estimativa da condutividade hidráulica do solo saturado

A capacidade de generalização dos modelos para estimativa da K_s sem o preenchimento de dados faltantes foi avaliada a partir do valor médio dos critérios de desempenho do conjunto de teste. Na Figura 2.1.4 apresentam-se as estimativas de K_s pelos modelos de melhor desempenho em cada conjunto preditor e na Tabela 2.1.5, o desempenho dos demais modelos em cada conjunto preditor.

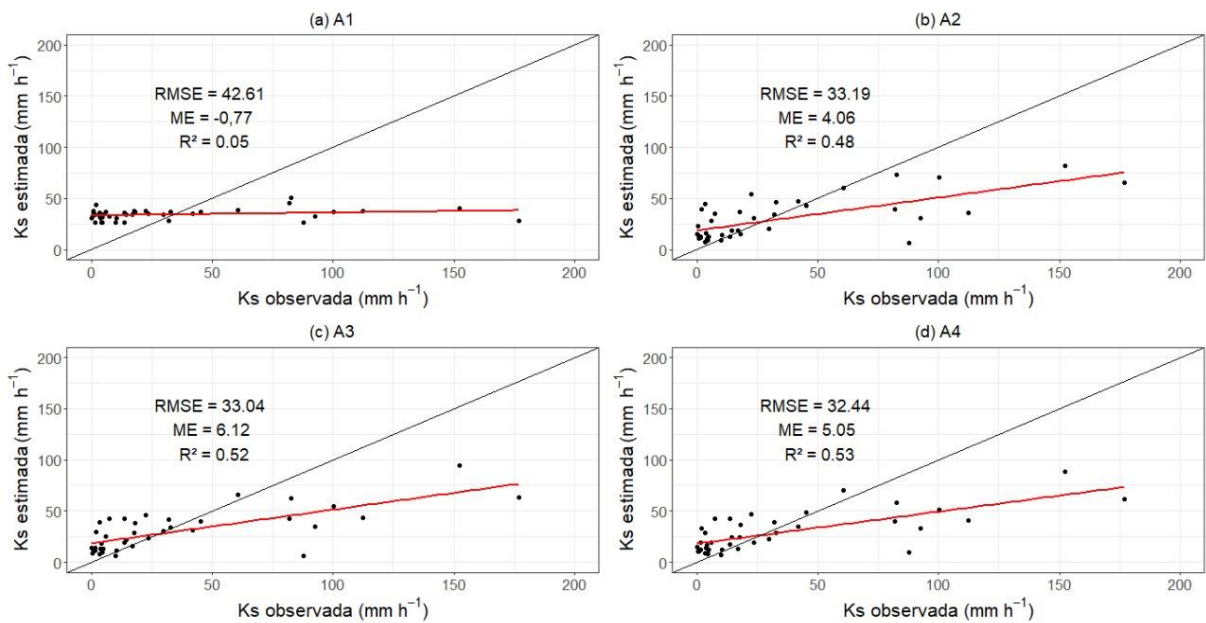


Figura 2.1.4. Condutividade hidráulica do solo saturado estimada obtida pelos modelos de melhor desempenho nos conjuntos preditores A1 (a), A2 (b), A3 (c) e A4 (d) em relação a condutividade hidráulica do solo saturado observada. (b, c, d) RF; (a) RLM.

Tabela 2.1.5. Desempenho estatísticos para cada modelo e conjunto preditor na estimativa Ks sem preenchimento dos dados faltantes.

Ks	Estatística	RLM	MARS	RF	SVR	KNN
A1	R ²	0,05	0,01	0,09	0,16	0,04
	RMSE	42,61	43,48	42,82	43,75	44,53
	ME	-0,77	-0,74	10,01	16,61	9,59
A2	R ²	0,29	0,38	0,48	0,34	0,19
	RMSE	37,01	37,27	33,19	35,92	39,52
	ME	-1,75	2,92	4,06	5,55	5,85
A3	R ²	0,27	0,41	0,52	0,36	0,22
	RMSE	37,74	38,96	32,44	35,51	39,38
	ME	0,79	3,36	4,09	6,12	7,03
A4	R ²	0,34	0,41	0,53	0,38	0,23
	RMSE	36,64	38,96	33,04	35,16	38,84
	ME	3,99	3,36	5,05	6,84	6,31

Os resultados dos testes indicaram que o modelo RF apresentou o melhor R² dentre os modelos avaliados, com valores de 0,48, 0,52 e 0,53 para os conjuntos de preditores A2, A3 e A4, respectivamente, e 0,16 para o conjunto A1 utilizando o modelo SVR. Para os demais modelos dos conjuntos A4, obteve-se valores de R² iguais a 0,34, 0,41, 0,38 e 0,23 para os modelos RLM, MARS, SVR e KNN, respectivamente. Para esses mesmos modelos, obteve-se R² de 0,27, 0,41, 0,36, 0,22 para o conjunto A3 e 0,29, 0,39, 0,34 e 0,19 para o conjunto A2. Já o conjunto A1 apresentou R² iguais a 0,05, 0,01, 0,09 e 0,03 para os modelos RLM, MARS, RF e KNN, respectivamente.

Analisando os valores de ME, observou-se que as FPTs obtidas superestimaram, em média, os valores de Ks, em 5,74 mm h⁻¹, com exceção dos modelos RLM (conjunto preditor A1) e MARS (conjunto preditor A1). No que se refere ao RMSE, os valores variaram de 32,44 a 38,96 mm h⁻¹ para o conjunto A4, 33,04 a 39,38 mm h⁻¹ para o conjunto A3 e 33,2 a 39, 53 mm h⁻¹ para o conjunto A2. Já para o conjunto A1, teve-se um aumento da variação dos valores de RMSE de 42,61 a 44,53 mm h⁻¹. Para todos os conjuntos avaliados, o modelo RF apresentou os menores valores de RMSE, exceto para o conjunto A1, que teve o modelo SVR como o de menor valor RMSE.

Em relação ao modelo nulo, os valores obtidos de ME e RMSE foram 43,47 e -5,34 mm h⁻¹, respectivamente, demonstrando que os modelos obtidos pelos algoritmos de aprendizado de máquina foram superiores ao modelo mais simples, exceto os modelos SVR e KNN quando utilizado o conjunto preditor A1.

No geral, os valores de ME e RMSE para todas FPTs desenvolvidas foram altos, demonstrando uma certa incapacidade de estimativa dos modelos. Esses altos valores de erro

podem ser atribuídos ao número restrito de amostras utilizado neste estudo. Ghanbarian, Taslimitehrani e Pachepsky (2017) utilizaram um conjunto de mais de 19.000 amostras de solo nos Estados Unidos para estimar K_s através de FPTs publicadas, e obtiveram valores de RMSE variando entre 0,56 e 1,27 cm dia⁻¹ e ME variando entre -0,001 e 0,96 cm dia⁻¹, ou seja, valores bem inferiores aos apresentados neste estudo para o Cerrado.

Utilizando o mesmo conjunto de dados nos Estados Unidos, Araya e Ghezzehei (2019) desenvolveram FPTs por meio de modelos de aprendizagem de máquina na estimativa de K_s , e obtiveram valores de RMSE variando de 0,25 a 0,6 cm h⁻¹ para diferentes conjuntos de propriedades físico-hídricas do solo. Os algoritmos que apresentaram os melhores desempenho nesse estudo foram o Boosted Regression Trees (BRT) e RF quando comparado aos modelos SVM e KNN, demonstrando a potencialidade dos modelos baseados em árvores de regressão na estimativa de K_s .

Quando avaliada a performance somente do conjunto A1, observou-se que todos os modelos apresentaram um baixo desempenho, demonstrando uma dificuldade da variável K_s ser explicada apenas por propriedades texturais do solo. Já para os conjuntos A3 e A4, observou-se que a utilização das variáveis estruturais do solo e os pontos de umidades da curva de retenção melhoram o desempenho dos modelos quando combinada aos preditores texturais e densidade do solo.

Na literatura, a acurácia e as incertezas das FPTs para K_s em função somente das propriedades granulométricas e/ou densidade do solo é questionada, apesar da maioria das FPTs publicadas terem sido preditas por essas variáveis (COSBY et al., 1984; SCHAAP et al., 2001; JULIÀ et al., 2004; TÓTH et al., 2015).

Alguns estudos afirmam que a textura do solo não tem impacto predominante sobre K_s (BECKER et al., 2018), apesar da grande relação das distribuições granulométricas com essa propriedade (PACHEPSKY; RAWLS, 2004). A incorporação de variáveis estruturais para a estimativa de K_s é recomendada (O'NEAL, 1949; WÖSTEN et al., 2001; LILLY et al., 2008; WEYNANTS et al., 2009), no entanto, a determinação dessas variáveis em laboratórios, como a porosidade e a umidade do solo, podem ser trabalhosas, mas se apresentam como propriedades capazes de caracterizar melhor a estrutura do solo e conseqüentemente, melhorar a estimativa de K_s .

Otoni et al. (2019) desenvolveram FPTs utilizando a RLM para estimativa de K_s no Brasil, e os modelos baseados em apenas dados texturais apresentaram o menor desempenho, RMSE de 0,9 cm dia⁻¹, quando comparado aos modelos que utilizaram a variável estrutural, macroporosidade, juntamente aos dados texturais e densidade do solo, com RMSE de 0,86 cm

dia⁻¹. Os mesmos autores ainda estimaram Ks somente com a variável macroporosidade e obtiveram um melhor desempenho, com RMSE igual a 0,71 cm dia⁻¹.

Na Figura 2.1.5 apresentam-se a classificação da importância das variáveis para a estimativa de Ks utilizando o conjunto A4. Pode-se verificar que as variáveis estruturais, capacidade de campo (θ_{10}), ponto de murcha permanente (θ_{1500}), porosidade total e macroporosidade foram propriedades importantes na estimativa dos modelos.

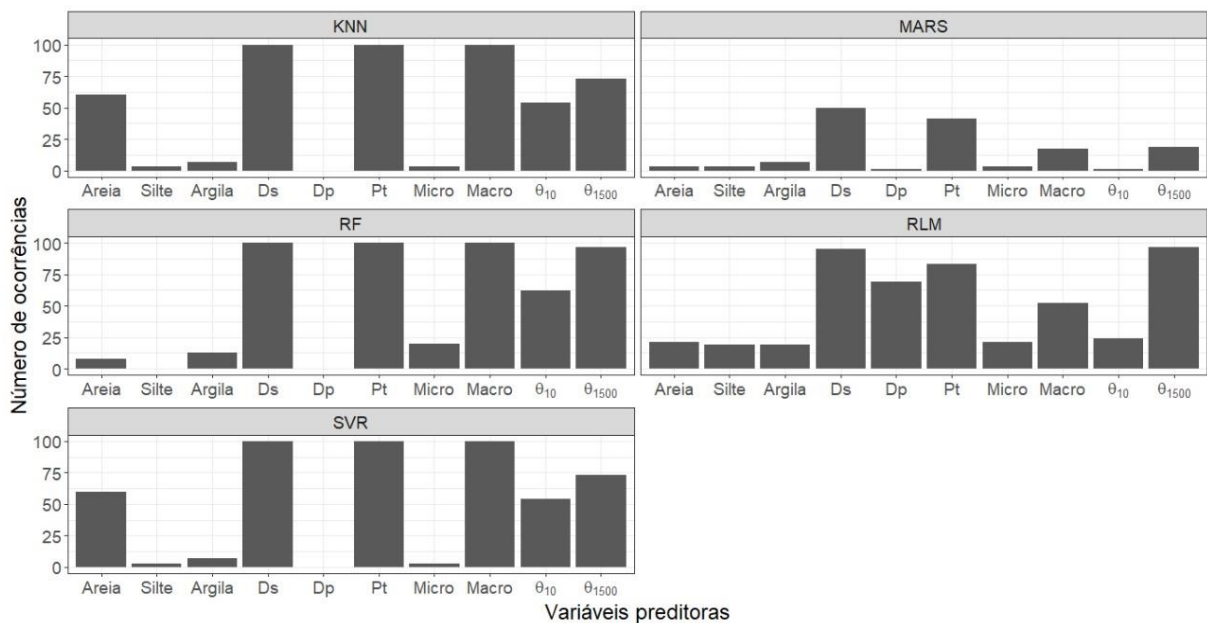


Figura 2.1.5. Classificação da importância das variáveis na estimativa de Ks sem o preenchimento de dados faltantes.

Nota-se ainda, que a densidade do solo foi uma variável eficiente na estimativa de Ks em todos os modelos. E a densidade de partícula foi mais importante que o θ_{10} na RLM, sendo que essa variável não foi considerada importante na estimativa por parte dos modelos mais robustos, exceto o MARS.

No que se refere a performance dos modelos, os resultados dos testes de Friedman indicaram o valor 0 para todas as análises, ou seja, houve diferença entre os modelos. Para os testes de Nemenyi, com intervalo de confiança de 95%, a distância crítica obtida foi de 0,61, o que indica que as distâncias entre os *ranks* médios dos modelos devem ser maiores do que este valor para apresentar diferença significativa. Com base nesse critério e analisando os resultados obtidos do teste de Nemenyi para as FPTs de Ks do conjunto A4, observou-se que os modelos SVR e RF (menores valores *ranks* médios), apresentam uma distância crítica igual a 0,72 (2,13 – 1,41), podendo afirmar que existe diferença estatística significativa entre eles, ou seja, apesar

do RF ter apresentado a maior média de R^2 , o SVR foi considerado melhor.

A RLM, apesar de sido o pior modelo (*rank* 4,17), segundo o teste de Nemenyi, possibilita a FTP ser descrita por uma equação, sendo assim, são apresentadas as equações 5, 6 e 7 obtidas para os conjuntos A2, A3 e A4, respectivamente.

$$K_s = 654,30443 - 143,80209 D_s \quad R^2 = 0,29 \quad (7)$$

$$K_s = 8,766 \times 10^4 - 6,12 \times 10^2 D_s \quad R^2 = 0,27 \quad (6)$$

$$K_s = 8,497 \times 10^4 - 6,108 \times 10^2 D_s - 3,795 \times 10^2 \theta_{1500} \quad R^2 = 0,34 \quad (5)$$

Já para a estimativa da condutividade hidráulica do solo saturado com preenchimento dos dados faltantes, observou-se que os valores de R^2 melhoraram para os modelos SVR e o KNN, mas, no geral, o desempenho dos modelos foi inferior aos resultados sem o preenchimento de dados, exceto os conjuntos A1 e A2, sendo este para os modelos RLM, SVR e KNN. Na Tabela 2.1.6 são apresentados os resultados obtidos para todos os modelos e conjunto preditores avaliados.

Tabela 2.1.6. Desempenho estatísticos para cada modelo e conjunto preditor na estimativa K_s com preenchimento dos dados faltantes.

Ks	Estatística	RLM	MARS	RF	SVR	KNN
A1	R^2	0,18	0,38	0,33	0,46	0,35
	RMSE	169,74	151,12	151,31	164,16	156,16
	ME	19,21	20,53	17,66	48,44	27,44
A2	R^2	0,29	0,36	0,39	0,39	0,35
	RMSE	160,29	152,21	143,13	161,19	156,01
	ME	25,33	28,89	15,45	43,28	31,09
A3	R^2	0,25	0,36	0,41	0,42	0,38
	RMSE	161,71	152,32	145,82	162,57	155,92
	ME	23,34	30,89	22,52	46,48	33,94
A4	R^2	0,27	0,29	0,41	0,42	0,42
	RMSE	159,71	159,63	146,32	161,92	153,09
	ME	25,52	29,15	22,87	45,79	38,28

No conjunto A4, o SVR e KNN foram os modelos que apresentaram os melhores R^2 , valores médios iguais a 0,42. O RF apresentou R^2 igual a 0,41 para os conjuntos A3 e A4, e 0,33 e 0,39 para A1 e A2, respectivamente. Já a RLM apresentou os menores valores de R^2 em todos os conjuntos.

Observou-se uma melhora no desempenho R^2 do modelos SVR e KNN para todos os

conjuntos preditores. No conjuntos A1 e A2, os valores de R^2 variaram de 0,18 a 0,46 e 0,29 a 0,39, respectivamente, ou seja, o aumento do conjunto amostral favoreceu a explicação de K_s em função das variáveis texturais e estruturais. Schaap e Leij (1998) afirmam que a performance das FPTs não dependem somente dos conjuntos preditores mas como também da base de dados utilizada no treinamento e teste das FPTs.

Contudo, ao avaliar os valores de ME e RMSE percebe-se um aumento da magnitude dos erros em relação aos resultados sem o preenchimento dos dados, indicando que com um aumento da base de dados em 291 amostras, os modelos não conseguiram diminuir os erros no ajuste.

Em relação ao modelo nulo, os valores obtidos de ME e RMSE foram 182,99 e 2,077 mm h^{-1} , respectivamente, demonstrando que os modelos obtidos pelos algoritmos de aprendizado de máquina foram superiores ao modelo mais simples.

2.1.3.3 Avaliação das funções de pedotransferência para estimativa da umidade do solo em tensões específicas

A capacidade de generalização dos modelos para estimativa das umidades do solo nas tensões de 0, 6, 10, 33, 100 e 1500 kPa foram avaliadas a partir do valor médio dos critérios de desempenho dos conjuntos de testes. Nas Figuras 2.1.6, 2.1.7, 2.1.8, 2.1.9, 2.1.10 e 2.1.11 são apresentados os gráficos das umidades observadas versus estimadas juntamente com os valores dos índices estatísticos obtidos para o melhor modelo de cada conjunto preditor e na Tabela 2.1.8 são apresentados os valores de R^2 obtidos para todos os modelos de cada conjunto preditor.

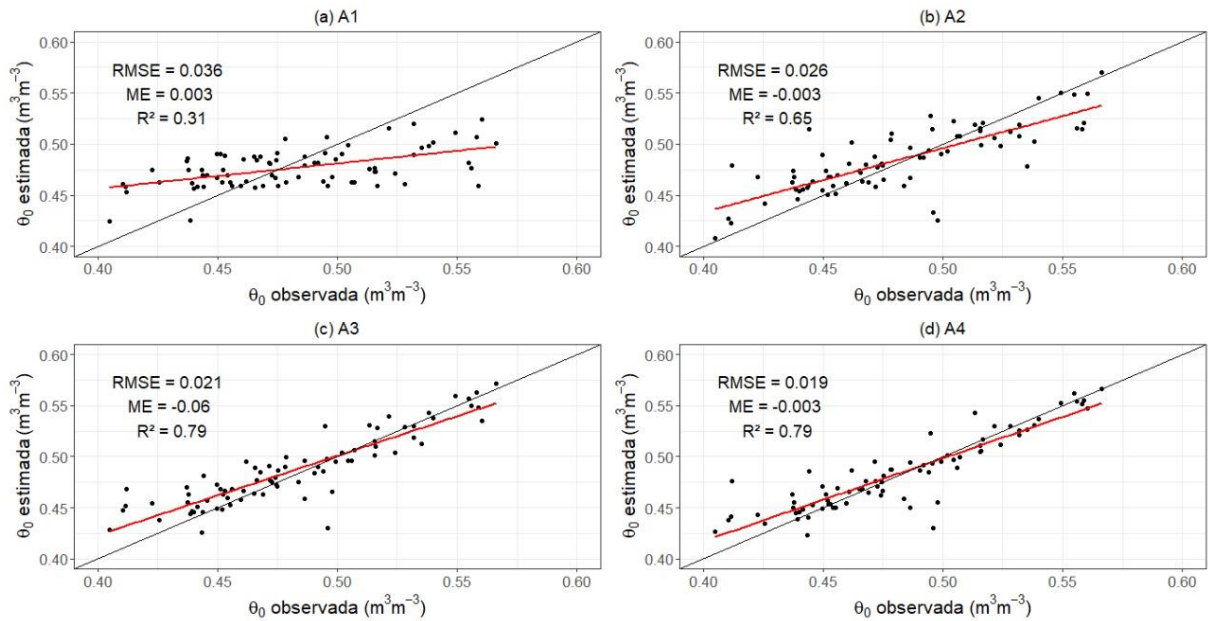


Figura 2.1.6. Umidade do solo estimada na tensão de 0 kPa obtidos pelos modelos de melhor desempenho nos conjuntos preditores A1 (a), A2 (b), A3 (c) e A4 (d) em relação a umidade do solo observada na tensão de 0 kPa. (a, c, d) SVR; (b) KNN.

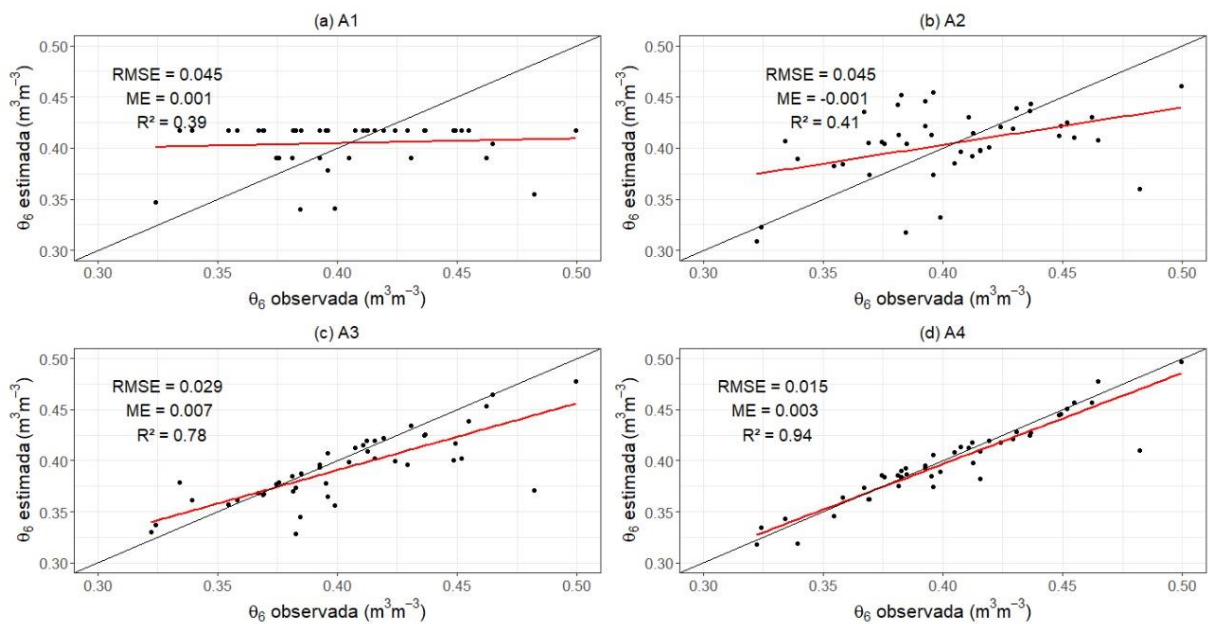


Figura 2.1.7. Umidade do solo estimada na tensão de 6 kPa obtidos pelos modelos de melhor desempenho nos conjuntos preditores A1 (a), A2 (b), A3 (c) e A4 (d) em relação a umidade do solo observada na tensão de 6 kPa. (b, c) RF; (a, d) MARS.

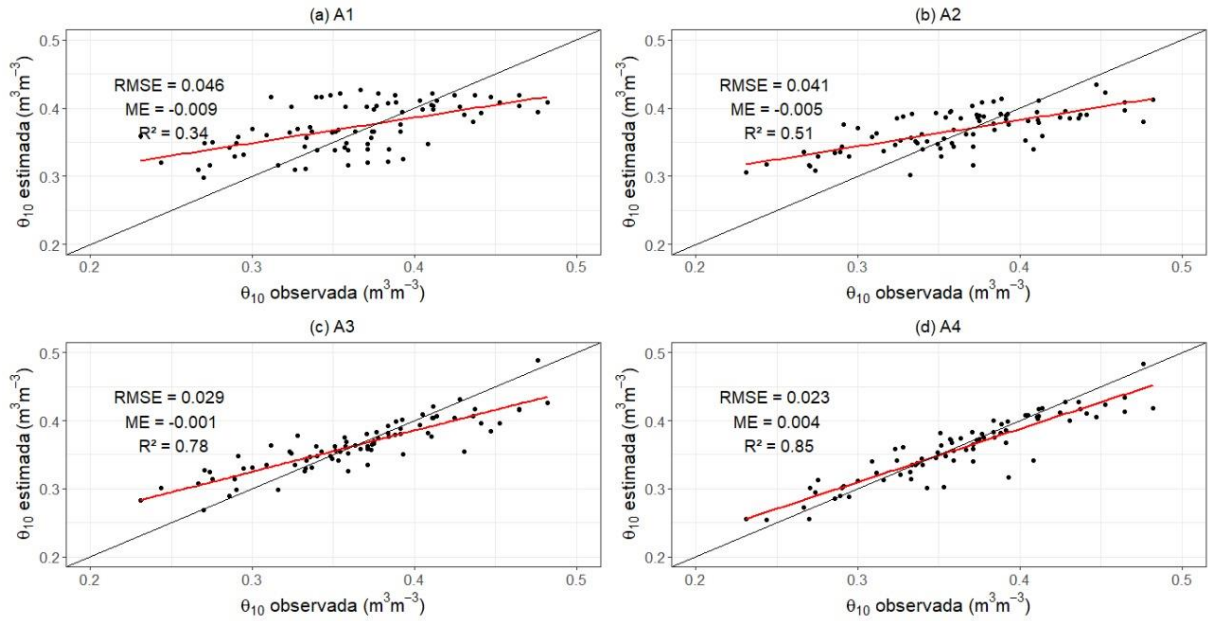


Figura 2.1.8. Umidade do solo estimada na tensão de 10 kPa obtidos pelos modelos de melhor desempenho nos conjuntos preditores A1 (a), A2 (b), A3 (c) e A4 (d) em relação a umidade do solo observada na tensão de 10 kPa. (a, c, d) RF; (b) RLM.

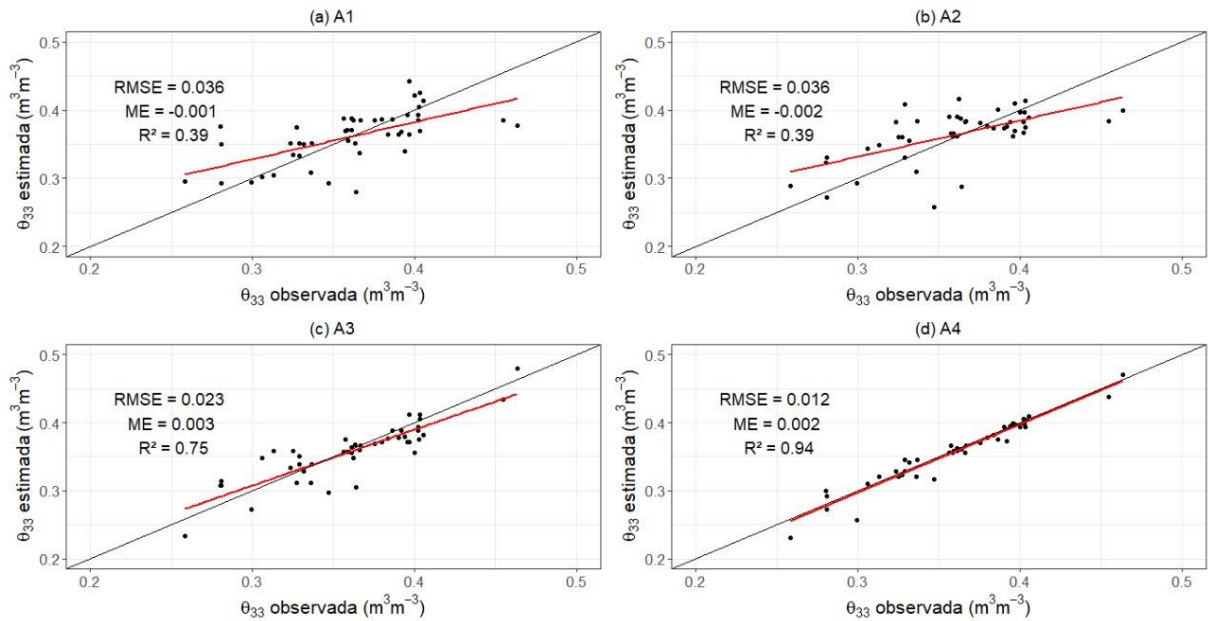


Figura 2.1.9. Umidade do solo estimada na tensão de 33 kPa obtidos pelos modelos de melhor desempenho nos conjuntos preditores A1 (a), A2 (b), A3 (c) e A4 (d) em relação a umidade do solo observada na tensão de 33 kPa. (a) SVR; (b, c, d) RF.

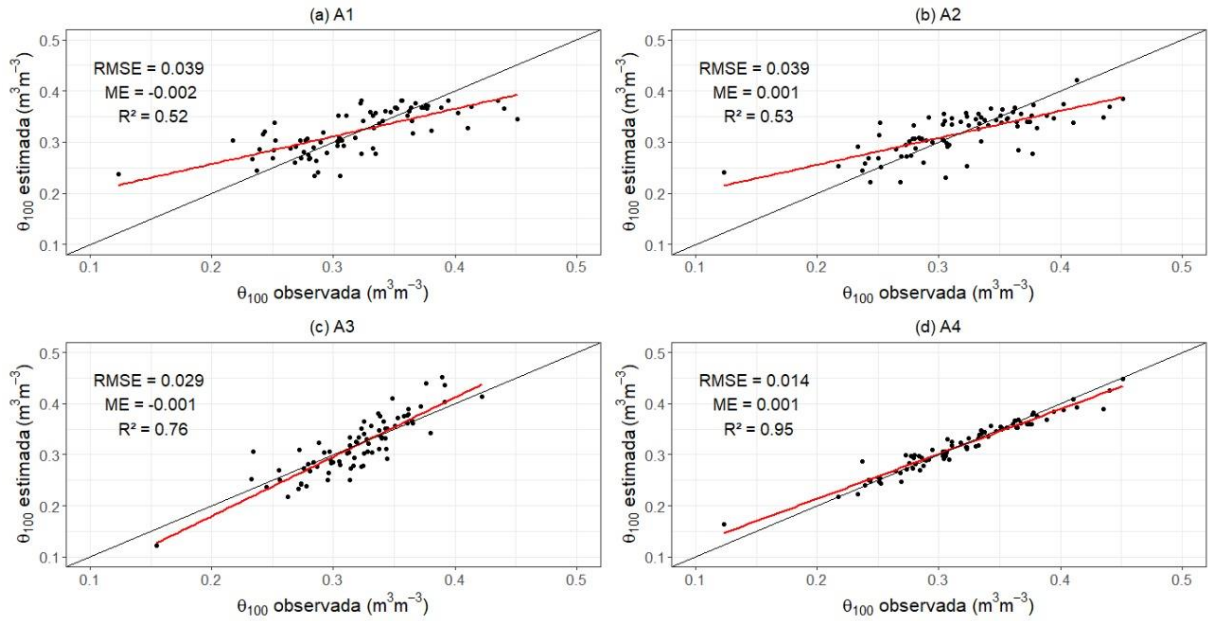


Figura 2.1.10. Umidade do solo estimada na tensão de 100 kPa obtidos pelos modelos de melhor desempenho nos conjuntos preditores A1 (a), A2 (b), A3 (c) e A4 (d) em relação a umidade do solo observada na tensão de 100 kPa. (a, d) RF; (b, c) SVR.

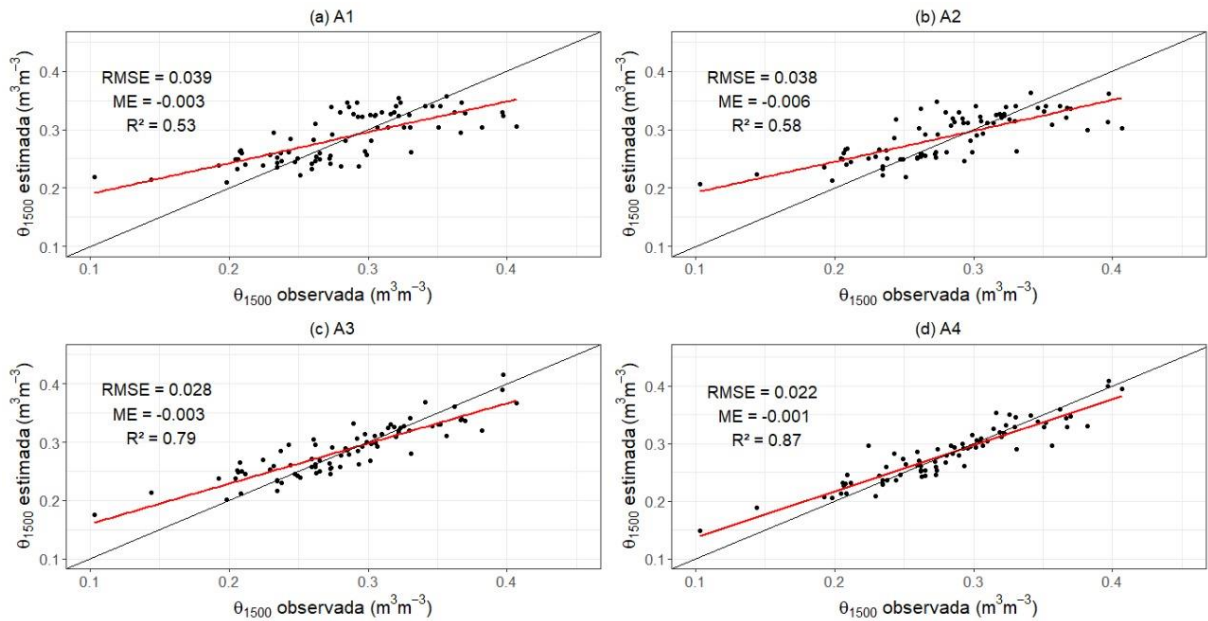


Figura 2.1.11. Umidade do solo estimada na tensão de 1500 kPa obtidos pelos modelos de melhor desempenho nos conjuntos preditores A1 (a), A2 (b), A3 (c) e A4 (d) em relação a umidade do solo observada na tensão de 1500 kPa. (a, b, c, d) RF.

Tabela 2.1.7. Coeficiente de determinação (R^2) para cada modelo e conjunto preditor na estimativa das umidades do solo nas tensões de 0, 6, 10, 33, 100 e 1500 kPa.

Conjunto Preditor	Variável	RLM	MARS	RF	SVR	KNN
A1	θ_0	0,12	0,16	0,31	0,31	0,17
	θ_6	0,27	0,39	0,29	0,16	0,25
	θ_{10}	0,29	0,23	0,34	0,44	0,05
	θ_{33}	0,16	0,13	0,36	0,39	0,19
	θ_{100}	0,39	0,41	0,52	0,47	0,24
	θ_{1500}	0,29	0,44	0,53	0,48	0,33
A2	θ_0	0,64	0,59	0,62	0,61	0,65
	θ_6	0,27	0,39	0,41	0,22	0,26
	θ_{10}	0,51	0,36	0,42	0,44	0,39
	θ_{33}	0,21	0,13	0,39	0,19	0,25
	θ_{100}	0,48	0,39	0,47	0,53	0,39
	θ_{1500}	0,39	0,49	0,58	0,54	0,54
A3	θ_0	0,59	0,72	0,68	0,79	0,72
	θ_6	0,68	0,66	0,78	0,64	0,67
	θ_{10}	0,66	0,75	0,78	0,74	0,72
	θ_{33}	0,52	0,49	0,75	0,67	0,69
	θ_{100}	0,55	0,64	0,74	0,76	0,68
	θ_{1500}	0,64	0,59	0,79	0,76	0,68
A4	θ_0	0,66	0,72	0,75	0,79	0,69
	θ_6	0,93	0,94	0,91	0,78	0,64
	θ_{10}	0,83	0,83	0,85	0,84	0,79
	θ_{33}	0,93	0,94	0,94	0,86	0,79
	θ_{100}	0,95	0,94	0,95	0,95	0,89
	θ_{1500}	0,82	0,82	0,87	0,82	0,78

Os resultados dos testes indicaram que o conjunto A4 foi o que proporcionou os maiores valores R^2 para todas FPTs quando comparada aos demais conjuntos, com destaque para as estimativas dos θ_6 , θ_{33} e θ_{100} , que apresentaram valores superiores a 0,9 nos modelos RLM, MARS, RF e SVR. Na sequência, em relação aos valores de R^2 , vieram os conjuntos A3 e A2, sendo, que o conjunto A1 o que apresentou os menores valores de R^2 , indicando uma dificuldade das umidades dos solos do Cerrado serem explicadas apenas por variáveis granulométricas.

A variação dos valores de ME ficaram entre -0,012 e 0,005, com exceção dos modelos SVR e KNN nas estimativas de θ_{10} (A4), θ_{33} (A1) e θ_{100} (A3) que apresentaram maiores valores, 9,65, 7,07, -6,22 e -8,49 $m^3 m^{-3}$, respectivamente. No que se refere aos valores de RMSE, verificou-se uma variação de 0,01 a 0,04 $m^3 m^{-3}$, indicando uma baixa magnitude dos erros em todas FPTs desenvolvidas.

Tomasella, Hodnett e Rossatto (2000) encontraram valores de RMSE na ordem de 0,032

a $0,427 \text{ m}^3 \text{ m}^{-3}$ para solos tropicais utilizando a RLM. No Rio Grande do Sul, Michelin et al. (2010) obtiveram valores semelhantes de ME e RMSE para as tensões de 0, 33, 100 e 1500 kPa utilizando a RLM e as variáveis do conjunto A3, contudo os valores de R^2 apresentados pelos autores foram um pouco superiores, na faixa de 0,77 a 0,93. Trabalhando em uma bacia hidrográfica localizada no Cerrado, Rodrigues, Maia e da Silva (2011) também obtiveram resultados superiores de R^2 na estimativa da CC e PMP utilizando as variáveis do conjunto A2, com valores de 0,58 e 0,59. Essas diferenças da variação da explicação das umidades em função das variáveis preditoras, se deve a grande extensão territorial trabalhada neste estudo para o Bioma Cerrado e conseqüentemente, a alta variabilidade das características pedológicas quando comparada ao estado do Rio Grande do Sul e uma bacia hidrográfica.

Já os valores de ME e RMSE obtidos pelo modelo nulo para cada variável estimada são apresentados na Tabela 2.1.8.

Tabela 2.1.8. Desempenho estatístico dos modelos nulos para cada umidade do solo estimada.

Estatística/Variável	θ_0	θ_6	θ_{10}	θ_{33}	θ_{100}	θ_{1500}
ME	0,479	4,829	-1,561	2,655	2,256	1,006
RMSE	0,042	0,056	0,055	0,044	0,057	0,058

Observa-se que os algoritmos de aprendizado de máquina foram superiores aos modelos nulos em todas as umidades do solo, exceto para os modelos SVR e KNN nas estimativas de θ_{10} (A4), θ_{33} (A1) e θ_{100} (A3) que apresentaram altos valores de ME como citado anteriormente, ou seja, os modelos desenvolvidos para a estimativa das umidades do solo nas tensões de 0, 6, 10, 33, 100 e 1500 kPa foram capazes, em sua maioria, de serem melhores que o modelo mais simples que pode ser definido.

Em relação as variáveis preditoras incorporadas nos modelos, as umidades do solo θ_{10} e θ_{1500} melhoraram o desempenho das FPTs. Na literatura, poucos estudos foram desenvolvidos utilizando as variáveis de umidade do solo como parâmetros de entrada nas funções, e quando realizada, a maioria foram verificadas em estimativas de parâmetros da curva de retenção e condutividade hidráulica do solo saturado (PAYDAR; CRESSWELL, 1996; SHAAP et al., 2001; ZHANG; SHAAP, 2017).

Em um desses estudos, Gunarathna et al. (2019) desenvolveram FPTs por meio do aprendizado de máquina para estimativa da CC e PMP utilizando umidades como preditoras nos solos do Sri Lanka, e destacaram a importância dessas variáveis nos modelos utilizados RF,

KNN e Redes Neurais Artificiais.

Na Figura 2.1.12 é apresentada a classificação da importância das variáveis preditoras nas estimativas das umidades do solo nas tensões estudadas para cada modelo no conjunto A4.

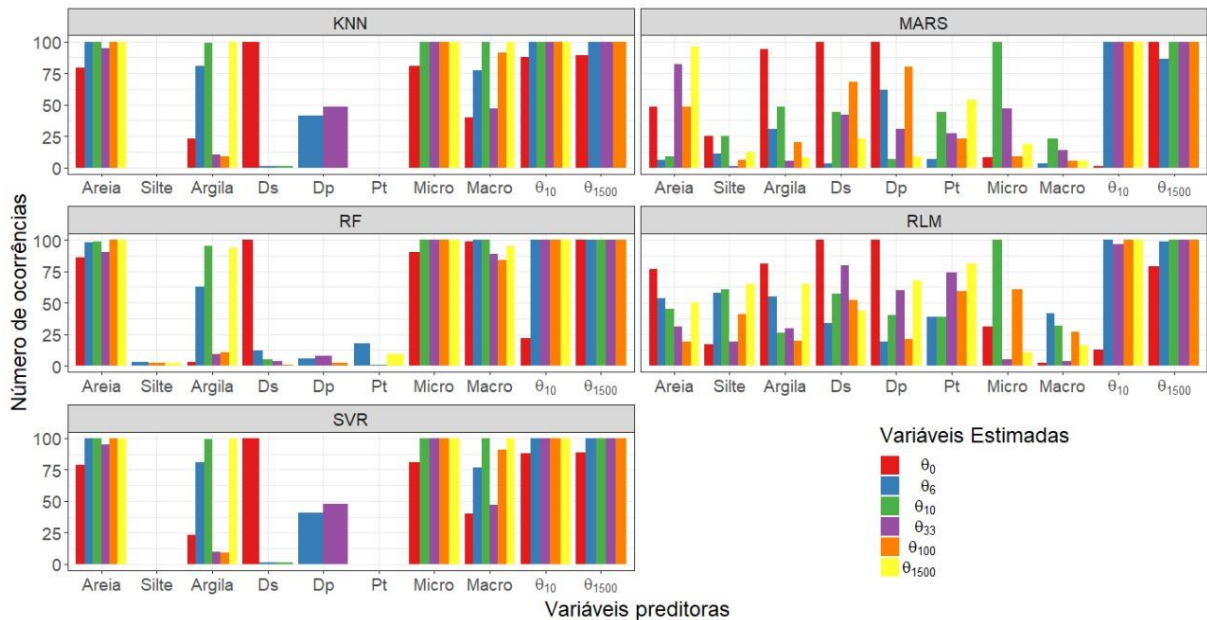


Figura 2.1.12. Classificação das importâncias das variáveis preditoras nas estimativas das umidades do solo nas tensões de 0, 6, 10, 33, 100 e 1500 kPa.

Observa-se que todas as variáveis θ_{10} e θ_{1500} foram importantes em todos os modelos avaliados, exceto o θ_0 , que teve uma menor ocorrência dessas variáveis em sua estimativa. Isso pode ter ocorrido devido a variável Ds mostrar-se mais correlacionada com a umidade de saturação.

A areia e a argila apareceram de forma moderada nos modelos RLM e MARS, porém mais recorrentes no RF, SVR e KNN, principalmente para a CC e PMP. Segundo Hillel (1998) e Assad et al. (2001), nos potenciais mais elevados (0 e 6 kPa), as variáveis estruturais tendem ser mais presentes nas FPTs. Para o Bioma Cerrado, entretanto, observa-se que as variáveis texturais, principalmente a areia e a argila, se apresentaram como variáveis importantes nos modelos para as umidades nesses potenciais. Já nos potenciais mais baixos (menores que 10 kPa), as variáveis texturais e de superfície específica das partículas são mais atuantes (KLEIN et al., 2010), e isso pode ser observado no número de ocorrência da argila nas estimativas do θ_{1500} .

No que se refere ao desempenho dos modelos, os resultados dos testes de Friedman indicaram o valor 0 para todas as análises, ou seja, houve diferença entre os modelos. Para os

testes de Nemenyi, com intervalo de confiança de 95%, a CD obtida foi de 0,61, o que indica que as distâncias entre os *ranks* médios dos modelos devem ser maiores do que este valor para apresentar diferença significativa.

Comparando os dois melhores modelos na estimativa de θ_0 , SVR e MARS (menores valores *ranks* médios), observou-se uma distância crítica igual a 1,05 (2,54 – 1,59), indicando que existe diferença estatística significativa entre eles, ou seja, o modelo SVR foi melhor que o MARS e conseqüentemente, melhor que os demais modelos. Já para o θ_6 e θ_{33} , o MARS foi considerado o melhor modelo. E para os θ_{10} e θ_{1500} , pode-se afirmar que o RF foi o melhor modelo na estimativa dessas umidades. Contudo, ao comparar os modelos RLM e SVR e o MARS e SVR para o θ_{100} , observou-se distâncias críticas iguais a 0,01 (2,27 – 2,26) a 0,29 (2,55 – 2,26), respectivamente, indicando que não existe diferença estatística significativa entre os modelos.

Nota-se ainda, que apesar da simplicidade da RLM, o modelo apresentou desempenhos próximos aos modelos de aprendizagem máquina, e isso se deve ao comportamento quase linear entre as variáveis preditoras e a umidade do solo capaz de ser visualizada na dispersão homogênea dos pontos em relação à linha 1:1 das Figuras 6 a 11.

Nas Equações 8, 9, 10, 11, 12 e 13 são apresentadas as FPTs obtidas pela RLM para os θ_0 , θ_{10} , θ_{13} , θ_{100} e θ_{1500} utilizando os conjuntos A3.

$$\theta_0 = 1,971 - 0,0006902 \text{ Areia} - 0,01312 \text{ Silte} - 0,05946 \text{ Argila} - 0,2804 \text{ Ds} - 0,07274 \text{ Dp} \quad R^2 = 0,59 \quad (8)$$

$$\theta_6 = 2,221 - 0,002782 \text{ Areia} - 0,00171 \text{ Silte} \quad R^2 = 0,68 \quad (9)$$

$$\theta_{10} = 0,105 + 0,0000000197 \text{ Micro} \quad R^2 = 0,66 \quad (10)$$

$$\theta_{33} = 0,915118 + 0,002108 \text{ Areia} - 0,000805 \text{ Silte} + 0,000453 \text{ Argila} \quad R^2 = 0,52 \quad (11)$$

$$\theta_{100} = 0,0489 + 0,0143 \text{ Dp} + 0,0369 \text{ Pt} + 0,0000128 \text{ Micro} \quad R^2 = 0,55 \quad (12)$$

$$\theta_{1500} = 0,1231 - 0,008281 \text{ Areia} - 0,01594 \text{ Silte} - 0,04473 \text{ Argila} + 0,1272 \text{ Dp} - 7,066 \text{ Pt} + 0,00009909 \text{ Micro} \quad R^2 = 0,64 \quad (13)$$

Nas Equações 14, 15, 16, 17, 18 e 19 são apresentadas as FPTs obtidas pela RLM para os θ_0 , θ_6 , θ_{10} , θ_{33} , θ_{100} e θ_{1500} utilizando os conjuntos A4.

$$\theta_0 = 1,715 - 0,0053 \text{ Areia} - 0,010483 \text{ Silte} - 0,05057 \text{ Argila} - 0,303 \text{ Ds} - 0,06769 \text{ Dp} - 0,00006082 \text{ micro} - 0,2927 \theta_{1500} \quad R^2 = 0,66 \text{ (14)}$$

$$\theta_6 = 2,6661428 - 0,0008198 \text{ Areia} - 0,0007967 \text{ Silte} + 2,2930306 \theta_{10} - 0,3849694 \theta_{1500} \quad R^2 = 0,93 \text{ (15)}$$

$$\theta_{10} = 0,7760832 + 0,0001109 \text{ Micro} + 0,6716302 \theta_{1500} \quad R^2 = 0,83 \text{ (16)}$$

$$\theta_{33} = -1,186 + 0,3986 \text{ Ds} - 0,09071 \text{ Dp} + 0,6010 \text{ Pt} + 0,4485 \theta_{10} + 1,937 \theta_{1500} \quad R^2 = 0,93 \text{ (17)}$$

$$\theta_{100} = -1,719 + 0,5317 \theta_{10} + 0,6365 \theta_{1500} \quad R^2 = 0,95 \text{ (18)}$$

$$\theta_{1500} = 7,04 + 1,046 \theta_{10} \quad R^2 = 0,82 \text{ (19)}$$

2.1.4 Conclusões

As atividades de pré-processamento reduziu consideravelmente a base de dados para o desenvolvimento das FPTs.

As FPTs desenvolvidas para a condutividade hidráulica do solo saturado apresentaram capacidade preditiva mediana utilizando as variáveis granulométricas e estruturais, com tendência de superestimação da propriedade físico-hídrica. Além disso, a condutividade hidráulica apresentou dificuldades em ser explicada somente por dados granulométricos e/ou densidade do solo.

O preenchimento dos dados faltantes para a estimativa da condutividade hidráulica melhorou o desempenho das FPTs utilizando as variáveis granulométricas e densidade do solo ou somente as variáveis granulométricas como preditoras.

Os altos valores de R^2 ($> 0,8$) juntamente aos baixos valores de ME e RMSE obtidos nas estimativas das umidades do solo nas tensões de 6, 10, 33, 100 e 1500 kPa utilizando as variáveis granulométricas e estruturais, indicaram uma boa precisão das funções de pedotransferência.

As variáveis preditoras capacidade de campo e ponto de murcha permanente demonstraram importantes nas estimativas da condutividade hidráulica do solo saturado e das

umidades do solo.

Os modelos de aprendizado de máquina foram superiores aos modelos nulos para todas as variáveis estimadas.

Os algoritmos de aprendizado de máquina, Support Vector Regression e Random Forest, foram os melhores modelos para a estimativa da condutividade hidráulica do solo saturado, conforme os índices estatísticos e o teste de Nemenyi. Já para as umidades do solo, os modelos Support Vector Regression, Random Forest e Multiple Adaptive Regression Splines obtiveram os melhores resultados.

2.1.5 Referências bibliográficas

AMANABADI, S.; VAZIRINIA, M.; VEREECKEN, K.; VAKILIAN, K. S.; MOHAMMADI, H. Comparative Study of Statistical, Numerical and machine learning-based pedotransfer functions of water retention curve with particle size distribution data. **Eurasian Soil Science**, v. 52, n. 12, p. 1555-1571, 2019.

ARAYA, S. N.; GHEZZEHEI, T. A. Using machine learning for prediction of saturated hydraulic conductivity and its sensitivity to soil structural perturbations. **Water Resources Research**, v. 55, n. 7, p. 5715-5737, 2019.

ASSAD, M. L. L.; SANS, L. M. A.; ASSAD, E. D.; ZULLO, J. Relação entre água retida e conteúdo de areia total em solos brasileiros. **Revista Brasileira de Agrometeorologia**, v.9, n.3, p.588-596, 2001.

AULER, A. C.; PIRES, L. F.; PINEDA, M. C. Influence of physical attributes and pedotransfer function for predicting water retention in management systems. **Revista Brasileira de Engenharia Agrícola e Ambiental**, v.21, n.11, p.746-751, 2017.

BAIAMONTE, G.; BAGARELLO, V.; D'ASARO, F.; PALMERI, V. Factors influencing point measurement of near-surface saturated soil hydraulic conductivity in a small sicilian basin. **Land Degradation and Development**, v. 28, n. 3, p. 970-982, 2017.

BARROS, A. H. C. B.; LIER, Q. J.; MAIA, A. H. N.; SCARPARE, F. V. Pedotransfer functions to estimate water retention parameters of soils in northeastern Brazil. **Revista Brasileira de Ciência do Solo**, v. 37, p. 379-391, 2013.

BECKER, R.; GEBREMICHAEL, M.; MÄRKER, M. Impact of soil surface and subsurface properties on soil saturated hydraulic conductivity in the semi-arid Walnut Gulch Experimental Watershed, Arizona, USA. **Geoderma**, v. 322, p. 112-120, 2018.

BERG, M. V. D.; KLAMT, E.; REEUWIJK, L. P. V.; SOMBROEK, W. G. Pedotransfer functions for the estimation of moisture characteristics of Ferralsols and related soils. **Geoderma**, v. 78, p.161-180, 1997.

BREIMAN, L. Random Forests. **Machine Learning**, v. 45, n. 1, p. 5–32, 2001.

BUUREN, S. Flexible imputation of missing data. Boca Raton, 2. Ed. FL: Chapman and Hall/CRC Press. 2012.

CASTELLINI, M.; LOVINO, M. Pedotransfer functions for estimating soil water retention curve of Sicilian soils. **Archives of Agronomy and Soil Science**, v. 65, p. 1401-1416, 2019.

CHEN, S.; RICHER-DE-FORGES, A. C.; SABY, N. P. A.; MARTIN, M. P.; WALTER, C.; ARROUAYS, D. Building a pedotransfer function for soil bulk density on regional dataset and testing its validity over a larger area. **Geoderma**, v. 312, p. 52–63, 2018.

CONTRERAS, C. P.; BONILLA, C. A. A comprehensive evaluation of pedotransfer functions for predicting soil water content in environmental modeling and ecosystem management. **Science of The Total Environment**, v. 644, p. 1580-1590, 2018.

COSBY, B.J., HORNBERGER, G.M., CLAPP, R.B., GINN, T.R. A statistical exploration of soil moisture characteristics to the physical properties of soils. **Water Resources Research**, v. 20, p. 682–690, 1984.

DEMSAR, J. Statistical comparisons of classifiers over multiple data sets. **Journal of Machine Learning Research**, v. 7, p. 1-30, 2006.

D'EMILIO, A.; AIELLO, R.; CONSOLI, S.; VANELLA, D.; IOVINO, M. Artificial Neural Networks for Predicting the Water Retention Curve of Sicilian Agricultural Soils. **Water**, v. 10, p. 1431, 2018.

DIAS JUNIOR, M.S.; BERTONI, J.C; BASTOS, A.R.R. Física do solo. Lavras, UFLA, 2000. 147p

FRIEDMAN, J. H. Multivariate adaptive regression splines. **The Annals of Statistics**, v. 19, n. 1, 1991.

GHANBARIAN, B.; TASLIMITEHRANI, V.; PACHEPSKY, Y. A. Scale-Dependent Pedotransfer Functions Reliability for Estimating Saturated Hydraulic Conductivity. **Catena**, vol. 149, p. 374-380, 2017.

HAREL A.; SHABTAI A.; ROKACH, L.; ELOVIC, Y. M-score: Estimating the potential damage of data leakage incident by assigning misuseability weight. Proceedings of the ACM workshop on insider threats, p. 13-20, 2010.

HILLEL, D. Environmental soil physics. Massachusetts: Academic, 1998. 771p.

HODNETT, M. G.; TOMASELLA, J. Marked differences between van Genuchten soil water-retention parameters for temperate and tropical soils: a new water-retention pedo-transfer functions developed for tropical soils. **Geoderma**, v. 108, p. 155-180, 2002.

IBGE (Instituto Brasileiro de Geografia e Estatística), Diretoria de Pesquisas, Coordenação de Agropecuária, Levantamento Sistemático da Produção Agrícola – jan. 2021.

JULIÀ, M. F.; MONTREAL, E.; JIMÉNEZ, A. S. C.; MELÉNDEZ, E. G. Constructing a saturated hydraulic conductivity map of Spain using pedotransfer functions and spatial

prediction. **Geoderma**, v. 123, n. 3-4, p. 257-277, 2004.

KAINGO, J.; TUMBO, S. D.; KIHUPI, N. I.; MBILINYI, B. P. Prediction of soil moisture-holding capacity with support vector machines in dry subhumid tropics. **Applied and Environmental Soil Science**, v. 2018, 2018.

KALUMBA, M.; BAMPS, B. NYAMBE, I.; DONDEYNE, S.; ORSHOVEN, J. V. Development and functional evaluation of pedotransfer functions for soil hydraulic properties for the Zambezi River Basin. **European Journal of Soil Science**, v. 72, n. 4, p. 1559-1574, 2020.

KLEIN, V. A.; BASEGGIO, M.; MADALOSSO, T.; MARCOLIN, C. D. Textura do solo e a estimativa do teor de água no ponto de murcha permanente com psicrômetro. **Ciência Rural**, v.40, n.7, p.1550-1556, 2010.

KOTLAR, A. M.; IVERSEN, B. V.; LIER, Q. J. Evaluation of Parametric and Nonparametric Machine-Learning Techniques for Prediction of Saturated and Near-Saturated Hydraulic Conductivity. *Vadose Zone Journal*, v. 18, n. 1, 2019.

KOTLAR, A. M.; LIER, Q. J. BRITO, E. S. Pedotransfer functions for water contents at specific pressure heads of silty soils from Amazon rainforest. **Geoderma**, v. 361, 2020.

LILLY, A., NEMES, A., RAWLS, W.J., PACHEPSKY, Y. Probabilistic approach to the identification of input variables to estimate hydraulic conductivity. **Soil Science Society of American Journal**, v. 72, p. 16–24, 2008.

MADY, A. Y.; SHEIN, E. V. Support vector machine and nonlinear regression methods for estimating saturated hydraulic conductivity. *Moscow University Soil Science Bulletin*, v. 73, n. 3, p. 129–133, 2018.

MILBORROW, S. Earth: Multivariate adaptive regression splines. R package version 5.1.2, 2019.

MEDRADO, E. LIMA, J. E. F.W. Development of pedotransfer functions for estimating water retention curve for tropical soils of the Brazilian savanna. **Geoderma Regional**, v. 1, p. 59-66, 2014.

MENEZES, S. M.; SAMPAIO, F. M. T.; RIBEIRO, K. D. **Estudo da condutividade hidráulica relacionada com alguns parâmetros físicos do solo**. In: XIII Congresso Brasileiro de Mecânica dos Solos e Engenharia Geotécnica/ IV Simpósio Brasileiro de Mecânica das Rochas/ III Congresso Luso-Brasileiro de Geotecnia, Curitiba, p. 149-153, 2006.

MICHELON, C. J.; CARLESSO, R.; OLIVEIRA, Z. B.; KNIES, A. E.; PETRY, M. T.; MARTINS, J. D. Funções de pedotransferência para estimativa da retenção de água em alguns solos do Rio Grande do Sul. **Ciência Rural**, v. 40, n. 4, p. 848-853, 2010.

NEMENYI, P. B. Distribution-free Multiple Comparisons. PhD thesis, Princeton University. 1963.

O'NEAL, A. M. Some characteristics significant in evaluating permeability. **Soil Science**, v.

67, p. 403-409, 1949.

OTTONI, M. V.; OTTONI FILHO, T. B.; SHAAP, M. G.; LOPES-ASSAD, M. L. R. C.; ROTUNNO FILHO. Hydrophysical database for brazilian soils (HYBRAS) and pedotransfer functions for water retention. **Vadose Zone Journal**, 2018.

OTTONI, M. V.; OTTONI FILHO, T. B.; LOPES-ASSAD, M. L. R. C.; ROTUNNO, O. C. Pedotransfer functions for saturated hydraulic conductivity using a database with temperate and tropical climate soils, **Journal of Hydrology**, v. 575, p. 1345-1358, 2019.

OLIVEIRA, L. B.; RIBEIRO, M. R.; JACOMINE, P. K. T.; RODRIGUES, J. J. V.; MARQUES, F. A. Funções de pedotransferência para predição da umidade retida a potenciais específicos em solos do estado de Pernambuco. **Revista Brasileira de Ciência do Solo**, v. 26, n. 2, p. 315-323, 2002.

PACHEPSKY, Y.; RAWLS, W. J. Development of pedotransfer functions in soil hydrology. Elsevier, Amsterdam, Netherlands, 2004.

PACHEPKY, Y.; PARK, Y. Saturated Hydraulic Conductivity of US Soils Grouped According to Textural Class and Bulk Density. **Soil Science Society of America Journal**, vol. 79, n. 4, p. 1094-1100, 2015.

PAYDAR, Z.; CRESSWELL, H. P. Water retention in Australian soils. II. Prediction using particle size, bulk density, and other properties. **Australian Journal Soil Research**, v. 34, p. 679–693, 1996.

R CORE TEAM (2019). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

RAHMAN, R.; HAIDER, S.; GHOSH, S.; PAL, R. Design of probabilistic random forests with applications to anticancer drug sensitivity prediction. **Cancer Informatics**, v. 14, n. Suppl 5, p. 57–73, 31 mar. 2016.

REATTO, A.; CORREIA, J. R.; SPERA, S. T.; MARTINS, E. S. Solos do bioma Cerrado: aspectos pedológicos. In: Sano, S.M., Almeida, S.P., Ribeiro, J.F. (Eds.), Cerrado: ecologia e flora. Embrapa, Brasília, p. 107–150, 2008.

RODRIGUES, L. N.; MAIA, A. H. N. **Funções de pedotransferência para estimar a condutividade hidráulica saturada e as umidades de saturação e residual do solo em uma bacia hidrográfica do Cerrado**. XIX Simpósio Brasileiro de Recursos Hídricos. **Anais...** Macéio - AL: Associação Brasileira de Recursos Hídricos, 2011.

RODRIGUES, L. N.; MAIA, A. H. N.; SILVA, R. N; **Funções de pedotransferência para estimar capacidade de campo, ponto de murcha permanente e densidade global em solos de uma bacia hidrográfica do Bioma Cerrado**. XL Congresso Brasileiro de Engenharia Agrícola. **Anais...** Cuiabá – MT: Associação Brasileira de Engenharia Agrícola, 2011.

SCHAAP, M. G., LEIJ, F. J. Database-related accuracy and uncertainty of pedotransfer functions. **Soil Science**, v. 163, n. 10, p. 765–779, 1998.

SCHAAP, M.G., LEIJ, F.J., VAN GENUCHTEN, M.TH. ROSSETA: computer program for estimating soil hydraulic parameters with hierarchical pedotransfer functions. **Journal Hydrology**, v. 251, p. 163–176, 2001.

SEKI, K. SWRC Fit - A nonlinear fitting program with a water retention curve for soils having unimodal and bimodal pore structure. **Hydrology and Earth System Sciences**, v.4, p.407-437, 2007.

SILVA, A. P.; TORMENA, C. A.; FIDALSKI, J.; IMHOFF, S. Funções de pedotransferência para as curvas de retenção de água e resistência do solo à penetração. **Revista Brasileira de Ciência do Solo**, v. 32, p. 1-10, 2008.

TOMASELLA, J., HODNETT, M. G. Estimating unsaturated hydraulic conductivity of Brazilian soils using soil-water retention data. **Soil Science**, v. 162, n. 10, 703–712, 1997.

TOMASELLA, J., HODNETT, M. G. Estimating soil water retention characteristics from limited data in Brazilian Amazonia. **Soil Science**, v. 163, n. 3, 1998.

TOMASELLA, J., HODNETT, M.G., ROSSATO, L. Pedotransfer functions for the estimation of soil water retention in Brazilian soils. **Soil Science Society of America Journal**, v. 64, p. 327–338, 2000.

TÓTH, B., WEYNANTS, M., NEMES, A., MAKÓ, A., BILAS, G., TÓTH, G. New generation of hydraulic pedotransfer functions for Europe. **European Journal of Soil Science**, v. 66, n. 1, p. 226–238, 2015.

VAPNIK, V.N. The nature of statistical learning theory. Springer, Berlin. 1995.

VARGA, T. V. Misscompare: Intuitive Missing Data Imputation Framework. 2020.

VERECKEN, H.; MAES, J.; FEYEN, J.; DARIUS, P. Estimating the soil moisture retention characteristic from texture, bulk density, and carbon content. **Soil Science**, v. 148, p. 389-403, 1989.

WEYNANTS, M., VERECKEN, H., JAVAUX, M. Revisiting Vereecken pedotransfer functions: introducing a closed-form hydraulic model. **Vadose Zone Journal**, v. 8, p. 86–95, 2009.

WÖSTEN, J. H. M.; PACHEPSKY, Y. A.; RAWLS, W.J. Pedotransfer functions: bridging gap between available basic soil data and missing soil hydraulic characteristics. **Journal Hydrology**, v. 251, p. 123–150, 2001.

ZHANG, Y.; SCHAAP, M. G. Weighted recalibration of the Rosetta pedotransfer model with improved estimates of hydraulic parameter distributions and summary statistics (ROSETTA3). **Journal Hydrology**, v. 547, 39–53, 2017.

2.2 Artigo 2 - Funções de pedotransferência para a estimativa dos parâmetros de ajustes da curva de retenção de água no solo

Resumo

O Bioma Cerrado é uma região estratégica para a agricultura brasileira, sendo responsável por grande parte da produção de alimentos do país. Visando suprir as demandas atuais e futuras por alimentos de forma sustentável, observa-se uma necessidade de um planejamento integrado dos recursos hídricos, a fim de reduzir a quantidade de água retirada dos mananciais, sobretudo a irrigação. Nesse contexto, a curva de retenção de água é fundamental para a compreensão das dinâmicas de água no solo, contudo sua obtenção é trabalhosa, abrindo oportunidade para a utilização das Funções de Pedotransferência (FPTs). Os objetivos do presente trabalho foram (i) avaliar diferentes modelos de ajuste da curva de retenção de água e (ii) desenvolver FPTs para a estimativa dos parâmetros de ajuste da curva de retenção. Para isso, quatro modelos de ajuste foram utilizados: Brooks e Corey (1964), van Genuchten (1980), Fredlund e Xing (1994) e Durner (1996). No desenvolvimento das FPTs, cinco modelos foram testados: Regressão Linear Múltipla (RLM), Multivariate Adaptive Regression Splines (MARS), Random Forest (RF), Support Vector Regression (SVR) e K Nearest Neighbors (KNN). Duas combinações de dados do solo foram avaliadas, sendo que as variáveis preditoras utilizadas em cada conjunto foram diferentes. No conjunto A1 foram utilizados: teores de areia (Ar), silte (Si), argila (Ag), densidade do solo (Ds), densidade de partículas (Dp), porosidade total (Pt), microporosidade (Micro) e macroporosidade (Macro); e no conjunto A2: Ar, Si, Ag, Ds e Dp. Foi verificado que o modelo de van Genuchten (1980) apresentou o melhor ajuste. O parâmetro θ_s (umidade de saturação) apresentou o melhor desempenho dentre os parâmetros estimados, com destaque para os modelos RF e SVR. Já a θ_r (umidade residual) foi obtido uma média capacidade preditiva por meio do RF e conjunto A1. E os demais parâmetros, α e n , apresentaram baixos desempenhos em todos os modelos e conjunto preditores utilizados.

Palavras-chave: aprendizado de máquina; irrigação; van Genuchten.

2.2.1 Introdução

Considerado o segundo maior bioma em extensão e a principal fronteira agrícola do Brasil, o Cerrado é responsável por uma grande parte dos alimentos produzidos no país, sendo que quase 45% de toda a produção de cereais, leguminosas e oleaginosas é proveniente dessa região (IBGE, 2021). No entanto, existe a necessidade de intensificar a agricultura para suprir, de forma sustentável, as demandas atuais e futuras por alimento.

Na região do Cerrado, a agricultura irrigada, principal usuária de recursos hídricos, é uma das técnicas mais promissoras para se produzir com sustentabilidade econômica e ambiental, tendo como o maior desafio conciliar a sua expansão com a disponibilidade de água por parte dos mananciais, em especial nas regiões que já se encontram em situações de escassez hídrica (RODRIGUES, 2017).

Nesse contexto, um planejamento integrado de bacias hidrográficas no Cerrado se faz necessário, a fim de estabelecer estratégias para aumentar a eficiência de uso da água pelos diversos usuários. Entretanto, a quase inexistência de dados na escala apropriada sobre clima, solo e planta tem prejudicado esse planejamento, bem como a compreensão das dinâmicas de água no solo e seu impacto na produtividade agrícola.

Informações sobre as características dos solos são fundamentais em qualquer estratégia de planejamento agrícola. Diversos trabalhos destacaram a carência de dados de solos nessa região (RODRIGUES; MAIA, 2011; SOUSA NETO et al., 2020), o que tem prejudicado os trabalhos de gerenciamento dos recursos hídricos. Dentre as propriedades do solo, a curva de retenção de água é fundamental para o entendimento da dinâmica de água, sendo utilizada na determinação da água disponível às plantas, em cálculos de balanços hídricos e manejos de irrigação.

A representação da curva de retenção de água é dada pela relação entre o conteúdo de água e o potencial de energia com a qual a água está retida. Essa curva assemelha-se a um S invertido suavizado em que os limites superior e inferior correspondem a umidade de saturação e a umidade residual, respectivamente. Nesse sentido, diversos modelos de ajuste foram desenvolvidos na tentativa de representar melhor a forma geral da curva, tais como os modelos de Brooks-Corey (1964), Campbell (1974), van Genuchten (1980), Hutson e Cass (1987), Durner (1994), Fredlund-Xing (1994), Kosugi (1994), Seki (2007) e entre outros.

O modelo de van Genuchten (1980) é um dos modelos mais comumente utilizados na estimativa da curva de retenção, apresentando resultados satisfatórios em diversos lugares do mundo (TOMASELLA; HODNETT; ROSSATO, 2000; SHARMA et al., 2006; PAN et al.,

2019). A grande aplicação desse modelo ocorre pela flexibilidade de uso em vários tipos de solo e formação de uma curva suave de inclinação contínua, evitando assim problemas de convergência entre as umidades de saturação e residual (D'EMILIO et al., 2018).

No geral, os modelos de ajuste da curva de retenção, possuem vários parâmetros que precisam ser identificados, e para isso, são utilizados métodos de otimização, tais como os mínimos quadrados e algoritmos heurísticos (VAN GENUCHTEN, 1980; GUANGZHOU CHEN; LI, 2016). Contudo, quando se trata de grandes áreas, como o Bioma Cerrado, a obtenção direta da curva de retenção e seus parâmetros é inviável, pois tal determinação demanda tempo e rotinas trabalhosas, sendo interessante a utilização de Funções de Pedotransferência (FPTs).

As FPTs são funções que possibilitam estimar propriedades do solo a partir de atributos físicos de fácil mensuração e baixo custo, tais como teores de areia, silte, argila, matéria orgânica, densidade do solo entre outros (PACHEPSKY; RAWLS, 2004; PACHESPKY; PARK, 2015). Vereecken et al. (2010) classifica as FPTs em dois tipos: pontuais, quando se estima o conteúdo de água em diferentes potenciais matriciais e as paramétricas, quando se estima os parâmetros empíricos da curva de retenção de água. Alguns estudos publicados destacam as utilidades das FPTs paramétricas, principalmente por fornecer diretamente os parâmetros necessários dos modelos de ajuste que descrevem a dinâmica de água no solo e suas interações planta-atmosfera (VEREECKEN et al., 2010; MEDEIROS et al., 2014).

Pachepsky e Rawls (2004) categorizam as FPTs em dois tipos, quando as FPTs são desenvolvidas por modelos lineares e não lineares e àquelas desenvolvidas por modelos com técnicas de mineração e exploração dos dados. No Brasil, o desenvolvimento de FPTs para estimar os parâmetros de ajuste da curva de retenção foram obtidas, em sua maioria, por regressão linear múltipla. Tomasella e Hodnett (1996) e Medeiros et al. (2014) desenvolveram FPTs para os solos da Amazônia utilizando os modelos de ajuste de Brooks e Corey e de van Genuchten, respectivamente. Utilizando um banco de dados de solos de diversos locais do Brasil, Tomasella, Hodnett e Rossato (2000) produziram FPTs para estimar os parâmetros de Brooks e Corey. Barros et al. (2013) desenvolveram FPTs para estimar os parâmetros da equação de van Genuchten para as curvas de retenção no nordeste do Brasil. Medrado e Lima (2014) estimaram parâmetros do modelo de van Genuchten através de modelos não lineares e apresentaram resultados significativos para o Cerrado.

Trabalhos utilizando técnicas mais avançadas como os algoritmos de aprendizado de máquina e redes neurais artificiais são menos comuns para a estimativa dos parâmetros da curva de retenção, contudo quando realizadas, apresentaram resultados satisfatórios (TWARAKAVI

et al., 2009; EBRAHIMI et al., 2014; D'EMILIO et al., 2018). Essa potencialidade de aplicação de métodos mais avançados está atrelada ao fato de que as relações entre as propriedades físico-hídricas e as variáveis de obtenção simples são complexas, e tendem à um comportamento não-linear, exigindo assim métodos mais robustos capazes de modelar melhor esses dados (SHEN et al., 2018; ARAYA; GHEZZEHEI, 2019).

Além de poucas FPTs desenvolvidas para o Bioma Cerrado, é importante avaliar o desempenho de diferentes modelos de ajuste da curva de retenção e modelos matemáticos de obtenção das FPTs, a fim de compreender melhor as dinâmicas de água no solo e suas intervenções na agricultura.

Nesse sentido, os objetivos deste trabalho foram: (i) avaliar o desempenho de modelos no ajuste das curvas de retenção de água; e (ii) desenvolver funções de pedotransferência para estimar os parâmetros de ajuste da curva de retenção de água utilizando algoritmos de aprendizagem máquina.

2.2.2 Material e métodos

2.2.2.1 Obtenção da base de dados

Os dados utilizados no desenvolvimento deste trabalho foram obtidos das bases de dados do Grupo de Pesquisa em Recursos Hídricos da Embrapa Cerrados e do Hybras (OTTONI et al., 2018).

Inicialmente, foi realizada a compilação desses bancos de dados que apresentam valores de propriedades físico-hídricas dos solos do Brasil e aquelas amostragens pertencentes ao limite territorial do Bioma Cerrado com um *buffer* de até 100 km foram selecionadas. Além da localização, utilizou-se como critério de seleção das amostras a disponibilidade de curvas de retenção de água (umidades do solo em diversas tensões matriciais) juntamente com os teores de areia, silte, argila, densidade do solo, densidade de partícula, porosidade total, macroporosidade e microporosidade, totalizando 606 amostras.

No que se refere ao pré-processamento dos dados, as seguintes premissas foram avaliadas: (i) somatório dos teores de areia, silte e argila igual a 100%; (ii) valor da densidade de partícula maior que o valor da densidade do solo; e (iii) os valores de porosidade total não devem exceder 60%. As amostras que não atenderam essas premissas foram descartadas.

2.2.2.2 Parâmetros de ajuste das curvas de retenção

Para a obtenção dos parâmetros de ajuste das curvas de retenção, foi utilizado o software SWRC Fit (SEKI, 2007) e selecionado os modelos: Brooks e Corey (1964) (Equação 1), van Genuchten (1980) (Equação 3), Fredlund-Xing (1994) (Equação 4) e Durner (1994) (Equação 5).

$$Se = \begin{cases} \left(\frac{\psi_b}{\psi}\right)^\lambda & (\psi > \psi_b) \\ 1 & (\psi \leq \psi_b) \end{cases} \quad (1)$$

Sendo,

$$Se = \frac{\theta - \theta_r}{\theta_s - \theta_r} \quad (2)$$

Em que: Se = grau de saturação efetiva, adimensional; ψ = potencial matricial, kPa; ψ_b = potencial matricial de entrada de ar, L; λ = parâmetro característico do solo, que indica a distribuição do tamanho dos poros, adimensional; θ = umidade do solo, m^{-3} ; θ_r = umidade residual, m^{-3} ; θ_s = umidade de saturação, m^{-3} .

$$Se = \left[\frac{1}{1 + (\alpha\psi)^n} \right]^m \quad (m = 1 - 1/n) \quad (3)$$

Em que: α = fator de escala, m; n = fator de forma, adimensional; $m = 1 - (1/n)$ (MUALEM, 1976).

$$Se = \left[\frac{1}{\ln[e + (\psi/a)^n]} \right]^m \quad (4)$$

Em que: a = parâmetro de ajuste do modelo, adimensional.

$$Se = w_1 \left[\frac{1}{1 + (\alpha_1\psi)^{n_1}} \right]^{m_1} + (1-w_1) \left[\frac{1}{1 + (\alpha_2\psi)^{n_2}} \right]^{m_2} \quad (5)$$

Em que: w_i = fator de peso da subcurva ($0 < w_i < 1$ e $\sum w_i = 1$); α_1, n_1, m_1 = parâmetros de ajuste para baixos potenciais; α_2, n_2, m_2 = parâmetros de ajuste para altos potenciais.

Os parâmetros ajustados para cada equação foram: $\theta_s, \theta_r, \psi_b, \lambda$ (Brooks e Corey); $\theta_s, \theta_r, \alpha, n$ (van Genuchten); $\theta_s, \theta_r, a, m, n$ (Fredlund-Xing) e $\theta_s, \theta_r, w_1, \alpha_1, n_1, w_2, \alpha_2, n_2$ (Durner).

Vale ressaltar que para o ajuste dos modelos, são necessários uma quantidade mínima de pontos da curva, que deverá ser igual ou maior que o número de parâmetros de ajuste do modelo, ou seja, para aquelas curvas que apresentaram no mínimo seis pontos, esta será capaz de se ajustar para todos os modelos.

Após os ajustes das curvas, foi selecionado o modelo que apresentou o melhor desempenho com base no coeficiente de determinação (R^2), erro médio quadrático (RMSE) e erro médio (ME), e assim, desenvolvidas as FPTs para estimativa dos seus respectivos parâmetros de ajuste.

2.2.2.3 Desenvolvimento das funções de pedotransferência

Para estimar os parâmetros do melhor modelo de ajuste das curvas de retenção do Bioma Cerrado, o subconjunto de treinamento foi organizado considerando dois diferentes conjuntos de preditores, sendo eles, A1: areia, silte, argila, densidade do solo, densidade de partícula, porosidade total, macroporosidade e microporosidade; A2: areia, silte, argila, densidade do solo e densidade de partícula. Em seguida, esses conjuntos foram incorporados em cinco modelos de aprendizado de máquina: Regressão Linear Múltipla (RLM), Multivariate Adaptive Regression Splines (MARS), Random Forest (RF), Support Vector Machine (SVR) e K-Nearest Neighbors (KNN).

Além disso, foi obtido para cada variável estimada o modelo nulo que é o modelo mais simples que pode ser definido ou ajustado, e para isso foi utilizada a média da variável estimada como parâmetro para a obtenção do índices estatísticos ME e RMSE, permitindo verificar se os modelos desenvolvidos para as FPTs apresentam um desempenho melhor ou não que o modelo nulo.

2.2.2.4 Modelos para o desenvolvimento das funções de pedotransferência

Regressão Linear Múltipla

A Regressão Linear Múltipla (RLM) consiste em estimar a relação da variável dependente/resposta (Y_i) a partir de duas ou mais variáveis independentes/preditoras (X_n). Portanto, foram ajustadas equações de RLM, correlacionando as variáveis estimadas para cada

conjunto de preditores (Equação 6).

$$Y_i = \beta_{i,0} + \beta_{i,1} \cdot X_1 + \dots + \beta_{i,n} \cdot X_n \quad (6)$$

Em que: Y_i = variável a ser estimada (parâmetros da curva de retenção de água); $\beta_{i,0}$ = intercepto da regressão linear múltipla; $\beta_{i,1} \dots \beta_{i,n}$ = coeficientes angulares vinculados às variáveis preditoras do solo; $X_1 \dots X_n$ = variáveis preditoras do solo.

Multivariate Adaptive Regression Splines

O Multivariate Adaptive Regression Splines (MARS) é uma técnica de regressão não paramétrica que modela automaticamente a não linearidade e as interações entre variáveis (FRIEDMAN, 1991). Para isso, os conjuntos de treinamento foram divididos em segmentos lineares e para cada conjunto, foram ajustadas à curvas polinomiais (*splines*), e posteriormente, unidas por meio de nós.

Dessa forma, foi utilizado o pacote *earth* (MILBORROW, 2019), na qual foram construídos modelos com diferentes números de interações e nós, e selecionado o modelo que apresentasse o menor RMSE.

Random Forest

O Random Forest (RF) é um modelo que combina árvores de regressão (BREIMAN, 1993), e para cada árvore, foi realizada uma amostragem *bootstrap*, e a quantidade de variáveis amostradas é controlada pelo hiperparâmetro *mtry*. A estimativa final foi baseada na média dos valores estimados em cada árvore (RAHMAN et al., 2016).

Para determinar o número ideal de variáveis selecionadas aleatoriamente para construção de cada árvore, foram construídos modelos utilizando números diferentes de variáveis e selecionado o modelo que apresentasse o menor RMSE.

Support Vector Regression

O Support Vector Machine (SVR) tem como princípio o ajuste do hiperplano que separa os pontos em um espaço n-dimensional, sendo n o número de variáveis preditoras (VAPNIK, 1995). Para isso, foi utilizado a função kernel radial, que é um dos kernels mais comumente usados, otimizando os valores dos hiperparâmetros C (custo) e γ (gama), responsáveis pela tolerância de ajuste dos modelos criados, e selecionado o modelo que apresentasse o menor

RMSE.

K-Nearest Neighbors

O K-Nearest Neighbors (KNN) é um modelo não paramétrico que estima a variável independente com base na média da distância dos seus vizinhos mais próximos do conjunto de dados, e o número de vizinhos é definido pelo hiperparâmetro k . Para isso, foram construídos modelos com diferentes valores de k , e selecionado o modelo que apresentasse o menor RMSE.

2.2.2.5 Validação e teste dos modelos

Para validação dos modelos gerados, foi utilizado o método *repeated holdout* no qual a base de dados foi dividida em dois subconjuntos considerados independentes e o processo foi repetido 100 vezes, sendo o primeiro subconjunto, o treinamento composto por 70% dos dados, e o segundo, o subconjunto de teste com 30% dos dados.

Para o ajuste dos hiperparâmetros de cada modelo foi aplicado o método de validação cruzada *k-folds* com repetições. Dessa forma, o conjunto de treinamento foi dividido aleatoriamente em k partes ($k = 10$), sendo que uma parte desse conjunto foi retirada para validação do modelo, gerando um novo conjunto de treinamento composto por $k-1$ partes. Em seguida, foram realizados o ajuste do modelo e avaliação de desempenho da estimativa para a parte retirada, e conseqüentemente seu resultado armazenado. O processo de validação cruzada foi repetido n vezes ($n = 3$), de modo, que cada um das k partes fossem utilizados como teste para validação dos modelos. Por fim, o desempenho final da validação cruzada foi calculado pela média k vezes n resultados obtidos no processo.

Após a otimização dos hiperparâmetros o desempenho dos modelos foram avaliados utilizando o conjunto de dados teste, ou seja, os modelos realizaram a estimativa em um conjunto de dados não utilizado no treinamento, permitindo avaliar a sua capacidade de generalização.

2.2.2.6 Desempenho e análise estatística

Para avaliar o desempenho das FPTs desenvolvidas para os parâmetros da curva de retenção foram utilizados os seguintes índices estatísticos: erro médio (ME), raiz do erro médio quadrático (RMSE) e o coeficiente de determinação (R^2), sendo as duas primeiras comumente usadas na avaliação de FPT (SCHAAP et al., 2001, JULIÀ et al., 2004).

O R^2 expressa o grau de concordância entre os valores observados e estimados pelas FPTs (Equação 7), assumindo valores entre 0 e 1. O ME expressa se o modelo superestima (ME > 0) ou subestima (ME < 0) (Equação 8) e o RMSE indica a magnitude do erro (Equação 9).

$$R^2 = \frac{\sum (\hat{y}_j - \bar{y}_j)^2}{\sum (y_j - \bar{y}_j)^2} \quad (7)$$

$$ME = \frac{1}{N} \sum_{j=1}^N y_j - \hat{y}_j \quad (8)$$

$$RMSE = \sqrt{\frac{1}{N} \sum_{j=1}^N (y_j - \hat{y}_j)^2} \quad (9)$$

Em que y_j e \hat{y}_j são os valores estimados e observados, respectivamente; N é o número de amostras; $\sum (\hat{y}_j - \bar{y}_j)^2$ é a variância explicada pelo modelo e $\sum (y_j - \bar{y}_j)^2$ é a variância total.

Para analisar os resultados de forma mais precisa, foi verificado se há diferença entre os desempenhos dos modelos utilizando o teste não paramétrico de Friedman (DEMNSAR, 2006). O teste baseia-se na comparação de desempenhos (*rank*) e, portanto, para cada um dos modelos avaliados foi determinado a posição, ordenando do melhor para o pior, e o teste retomando à valores de 0 ou 1, sendo 0 há diferença e 1 não há diferença entre os modelos.

No entanto, o teste de Friedman não permite discernir quais modelos apresentam diferença estatística, sendo assim utilizou-se também o teste de Nemenyi (NEMENYI, 1963). De acordo com esse teste, os modelos são significativamente diferentes entre si quando a subtração dos *ranks* médios dos modelos for igual ou maior que o valor da distância crítica (CD).

2.2.3 Resultados e discussão

2.2.3.1 Base de dados

A base de dados utilizada nesse estudo apresentou solos com textura com altos teores de areia e argila, sendo a maioria dos solos classificados como argiloso e franco argilo-arenoso

(Figura 2.2.1).

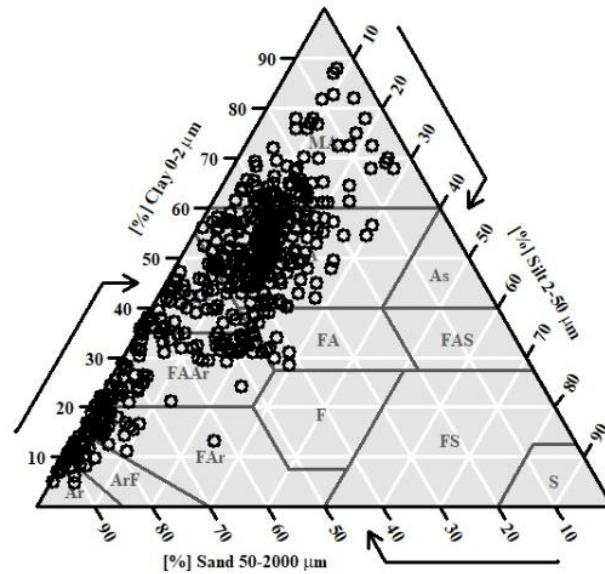


Figura 2.2.1. Triângulo textural dos solos do Bioma Cerrado utilizados para o ajuste das curvas de retenção de água no solo.

Destes dados, os valores de umidade do solo foram verificados em diversos potenciais matriciais, sendo formados por diferentes quantidades de pontos (umidade do solo versus potencial matricial). Os potenciais matriciais presentes na base de dados foram 0, 1, 3, 5, 6, 8, 10, 20, 25, 30, 33, 60, 50, 100, 200, 300, 500, 1000 e 1500 kPa, e a variação das umidades em alguns desses potenciais são apresentadas em formato Box-Plot na Figura 2.2.2.

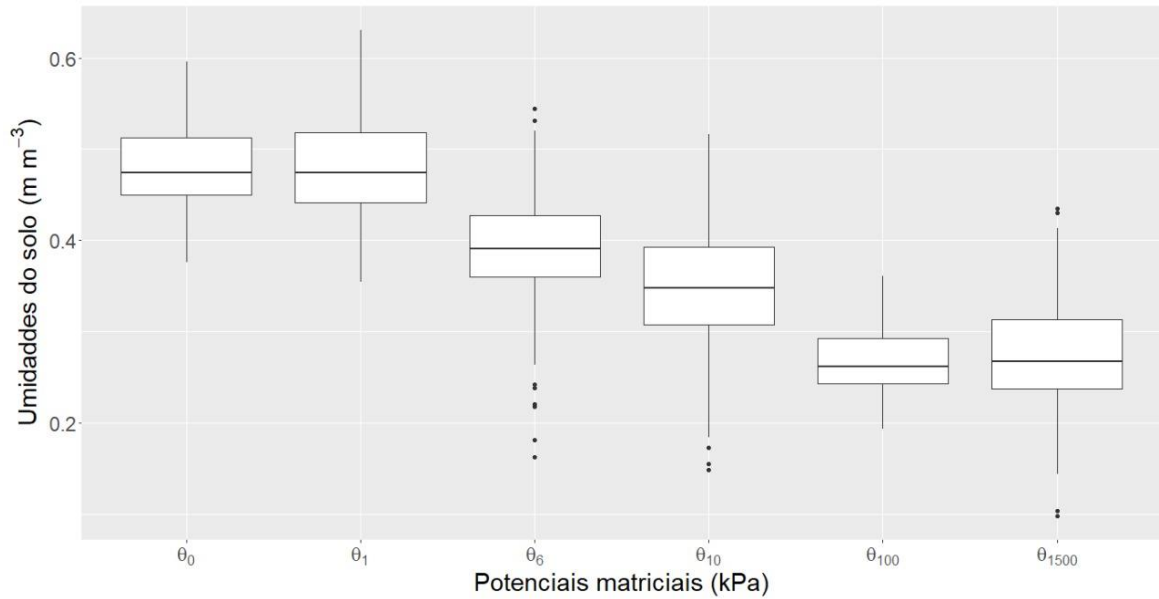


Figura 2.2.2. Umidades do solo em potenciais matriciais específicos utilizados no ajuste dos parâmetros da curva de retenção e desenvolvimento das FPTs.

Os teores umidade na capacidade de campo e no ponto de murcha permanente, correspondentes aos potenciais matriciais de 10 e 1500 kPa, apresentaram valores médios na faixa de 0,314 e 0,276 $\text{m}^3 \cdot \text{m}^{-3}$, respectivamente. Já as θ_0 (umidade de saturação) e θ_1 apresentaram médias 0,49 e 0,478 $\text{m}^3 \cdot \text{m}^{-3}$, respectivamente. De forma geral, as umidades apresentaram variações semelhantes.

A quantidade de pontos verificada para cada uma dessas umidades foram: 0 (556 pontos), 1 (268 pontos), 6 (473 pontos), 10 (218 pontos), 100 (373 pontos) e 1500 (374 pontos). O valor mínimo de pontos observados nas curvas de retenção foram cinco e máximo de nove pontos.

2.2.2.3 Ajuste das curvas de retenção de água e seleção do modelo

Na Figura 2.2.3 são apresentados os comportamentos médios das retenções de água obtidos para cada modelo de ajuste, e na Tabela 2.2.1, os desempenhos estatísticos obtidos para cada modelo de ajuste.

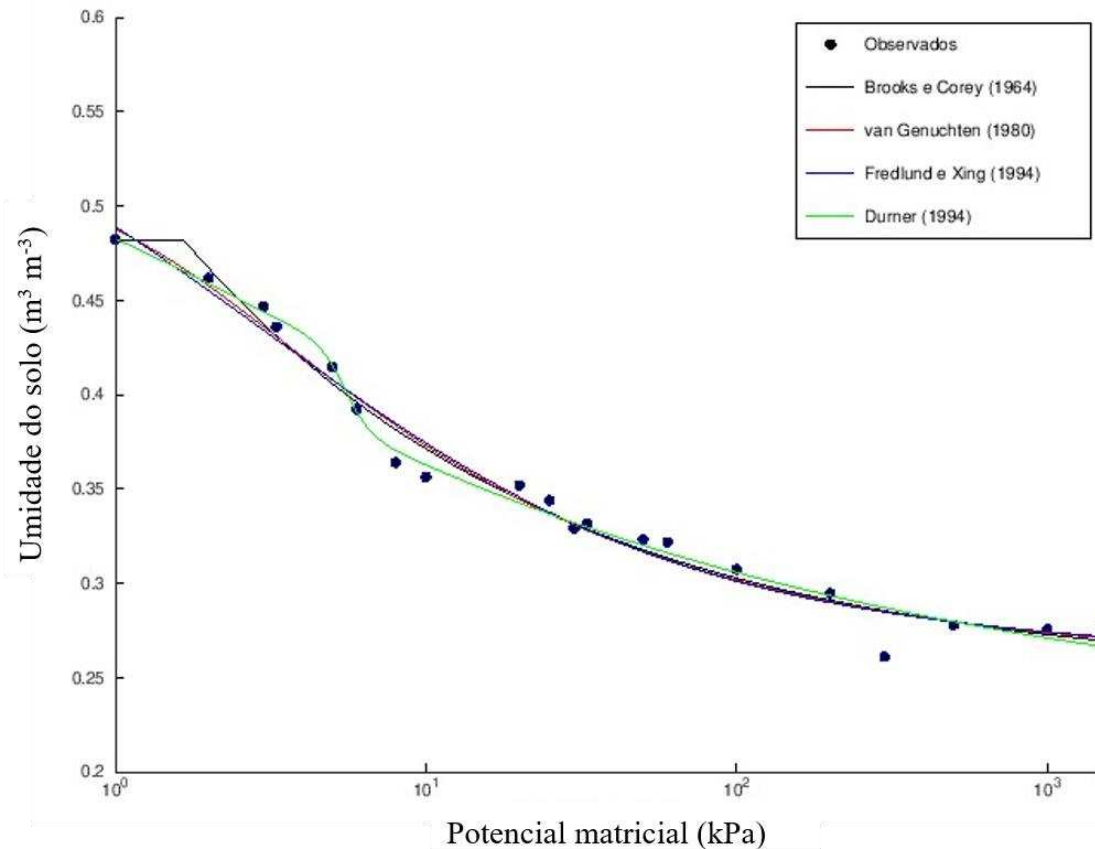


Figura 2.2.3. Curva retenção de água no solo.

Tabela 2.2.1. Desempenho estatístico médio para cada modelo de ajuste

Modelo	R²	RMSE	ME
Brooks e Corey (1964)	0,975	0,014	-0,001
van Genuchten (1980)	0,984	0,001	-0,001
Fredlund e Xing (1994)	0,971	0,125	-0,123
Durner (1994)	0,974	0,016	-0,002

A equação de van Genuchten (1980) foi o modelo que melhor se ajustou aos dados com valores estatísticos médios de: $R^2 = 0,984$; $RMSE = 0,009 \text{ m}^3 \text{ m}^{-3}$; $ME = -0,002 \text{ m}^3 \text{ m}^{-3}$. Para os demais modelos, Brooks e Corey (1964), Fredlund e Xing (1994) e Durner (1996), os resultados estatísticos também foram satisfatórios para o Bioma Cerrado.

Costa, Oliveira e Kato (2008) estimaram a curva de retenção de água para Latossolos Vermelho-Amarelo da região do Cerrado e o modelo de van Genuchten apresentou um R^2 igual a 0,91, sendo superior ao outro modelo (Hutson e Cass) avaliado pelos autores. Pan et al. (2019) utilizaram três modelos de ajuste para as curvas de retenção na região do Tibete, Ásia, e

verificaram que os modelos de van Genuchten e Brooks e Corey obtiveram os melhores ajustes com R^2 superiores a 0,9, diferentemente do outro modelo (Gardner) utilizado.

Por apresentar o melhor desempenho para o Bioma Cerrado, a equação de van Genuchten foi selecionada para o desenvolvimento das FPTs dos seus parâmetros de ajuste. Nas Tabelas 2.2.2 e 2.2.3 são apresentadas as estatísticas descritivas das variáveis preditoras utilizadas para o desenvolvimento das FPTs e dos parâmetros da equação de van Genuchten, respectivamente.

Tabela 2.2.2. Estatísticas descritivas das variáveis preditoras utilizadas para o desenvolvimento (subconjunto A) e teste (subconjunto B) das funções de pedotransferência dos parâmetros da equação de van Genuchten

Subconjunto	Estatística	Areia	Argila	Silte	Ds	Dp	Pt	Macro	Micro
		%	%	%	G cm ⁻³	g cm ⁻³	%	%	%
A (n = 420)	Média	43,68	44,43	11,89	1,34	2,67	47,71	9,44	38,33
	Máximo	94,68	88,00	20,01	1,76	2,98	59,56	30,54	50,43
	Mínimo	4,01	4,79	0,021	0,82	2,31	37,64	0,29	16,26
	Desvio Padrão	0,102	16,79	6,21	0,15	0,13	4,47	5,86	4,85
	CV (%)	53,61	39,58	59,29	10,93	4,51	9,37	62,06	12,65
B (n = 178)	Média	46,10	43,11	10,78	1,35	2,67	47,85	9,97	37,98
	Máximo	93,02	82,76	27,0	1,61	2,98	59,63	30,54	50,43
	Mínimo	4,00	4,79	0,021	0,82	2,925	38,58	0,29	16,26
	Desvio Padrão	20,18	16,79	6,21	0,15	0,13	4,45	6,47	4,97
	CV (%)	42,84	39,58	59,29	10,93	4,51	9,30	64,91	13,01

Ds = densidade do solo; Dp = densidade de partícula; θ_s = umidade de saturação; Pt = porosidade total; macro = macroporosidade; micro = microporosidade.

Tabela 2.2.3. Estatísticas descritivas dos parâmetros da equação de van Genuchten (1980)

Subconjunto	Estatística a	θ_s	θ_r	α	n
		$m^3 m^{-3}$	$m^3 m^{-3}$	m^{-1}	-
A (n = 420)	Média	0,490	0,191	12,289	1,653
	Máximo	0,571	0,365	109,426	15,527
	Mínimo	0,385	0,113	0,065	0,116
	Desvio Padrão	0,06	0,11	34,14	0,74
	CV (%)	1,09	53,66	278,09	76,23
B (n = 178)	Média	0,505	0,193	3,084	1,659
	Máximo	0,576	0,381	73,235	4,535
	Mínimo	0,381	0,105	0,081	0,115
	Desvio Padrão	0,06	0,09	9,01	0,74
	CV (%)	11,86	50,99	292,06	45,15

θ_s = umidade de saturação; θ_r = umidade residual; α , n = parâmetros de ajuste; CV = coeficiente de variação.

Observa-se que os teores de argila variaram de 4,79 a 88%, silte de 0,02 a 20,01% e a areia de 4,01 a 94,68%, com valores médios iguais a 44,43, 11,89 e 43,68%, respectivamente. Valores semelhantes de teores de argila foram observados por Medrado e Lima (2014), que desenvolveram FPTs para estimativa de parâmetros da curva de retenção no Bioma Cerrado, indicando uma boa capacidade de armazenamento de água da região. Contudo, os teores de areia apresentados pelos mesmos autores, média de 32%, foram um pouco inferiores aos resultados observados neste trabalho.

A densidade média dos solos foi igual a 1,34 g cm⁻³, um pouco superior aos valores médios de 1,0 e 1,14 g cm⁻³ encontrados nos trabalhos de Medrado e Lima (2014) e Rodrigues, Maia e da Silva (2014) obtidos para o Cerrado, respectivamente. Como era de se esperar, a densidade de partícula foi a variável que apresentou menor coeficiente de variação, com mínima de 2,31 e máxima de 2,98 g cm⁻³.

Os parâmetros θ_s e θ_r apresentaram uma variação com médias de 0,49 e 0,192 m m⁻³, respectivamente, e CV superior a 50% para a umidade residual. O parâmetro n também apresentou um alto CV com valores de 76,23% para o conjunto de treinamento e 45,15% para o conjunto de teste. Além disso, foi verificado uma variabilidade acentuada para o parâmetro α , que pode ser visualizada pela diferença entre os valores médios dos conjuntos de treinamento e teste, bem como os altos valores de CV, superiores a 278%. Vereecken et al. (1989) ressalta a alta variabilidade do parâmetro α e afirma que curvas de retenção com altos valores de α indicam que os solos apresentam consideráveis teores de areia, como foi verificado para o Bioma Cerrado.

Além disso, os parâmetros α e n determinam a forma da curva de retenção, então uma alta variação desses parâmetros causam alterações na forma das curvas, que conseqüentemente pode prejudicar a estimativa da umidade do solo e o uso racional da água. Contudo, a combinação de todos os parâmetros da equação de van Genuchten que vai determinar o bom desempenho ou não das FPTs para a estimativa da umidade do solo.

2.2.2.4 Avaliação das funções de pedotransferência para a estimativa dos parâmetros de ajuste

A capacidade de generalização dos modelos para estimativa dos parâmetros de van Genuchten (1980) foi avaliada a partir do valor médio dos critérios de desempenho do conjunto de teste. Nas Figuras 2.2.4 e 2.2.5 são apresentadas as estimativas obtidas pelo melhor modelo para cada parâmetro (θ_s , θ_r , α , n) e conjunto predictor, e nas Tabelas 2.2.4 e 2.2.5 são indicados os índices estatísticos obtidos para os demais modelos.

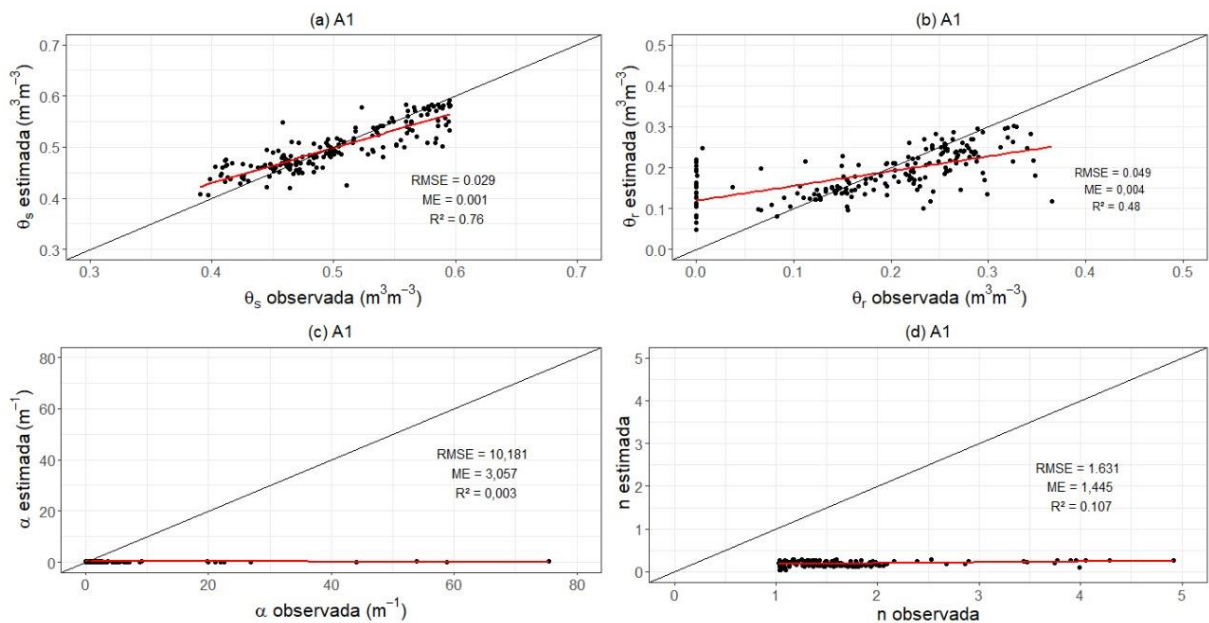


Figura 2.2.4. Parâmetros estimados da equação de van Genuchten (1980) obtidos pelos modelos de melhor desempenho no conjunto predictor A1 em relação aos parâmetros observados da equação de van Genuchten (1980). (a, b, d) RF, (c) SVR.

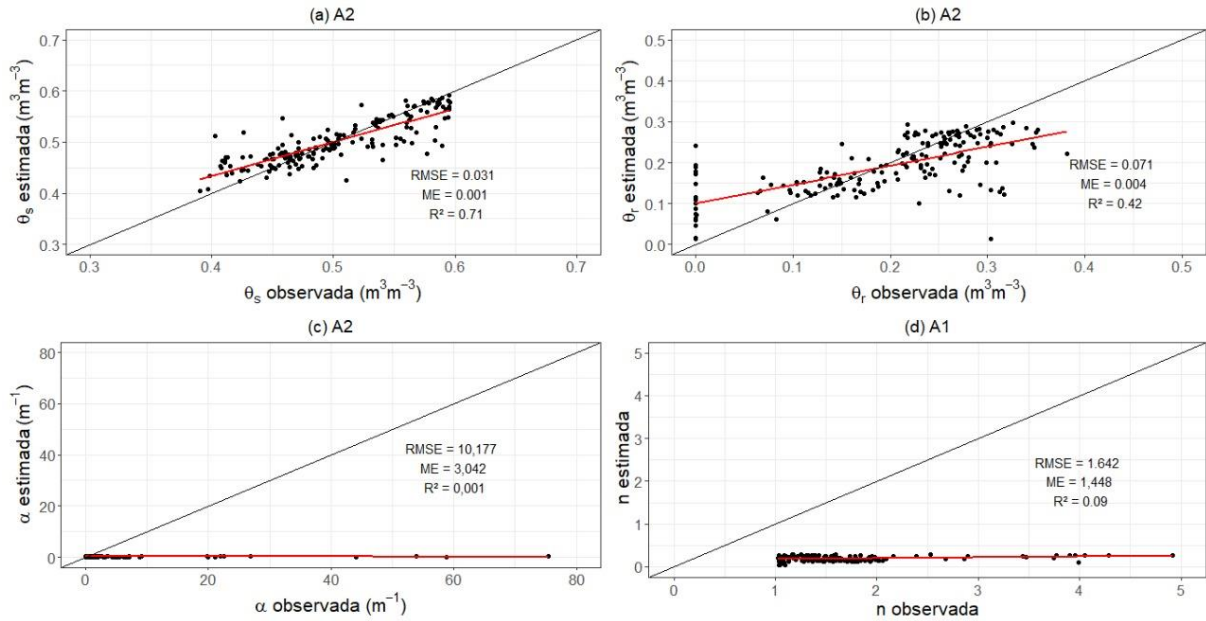


Figura 2.2.5. Parâmetros estimados da equação de van Genuchten (1980) obtidos pelos modelos de melhor desempenho no conjunto preditor A2 em relação Parâmetros estimados da equação de van Genuchten (1980) observados. (a, b, d) RF, (c) SVR.

Tabela 2.2.4. Desempenho estatístico das FPTs para estimativa dos parâmetros de van Genuchten (1980) utilizando o conjunto preditor A1

Parâmetro	Estatística	RLM	MARS	RF	SVR	KNN
θ_s	R^2	0,69	0,65	0,76	0,75	0,74
	RMSE	0,031	0,034	0,029	0,029	0,029
	ME	0,001	0,002	0,001	0,002	0,001
θ_r	R^2	0,14	0,21	0,48	0,29	0,38
	RMSE	0,097	0,075	0,049	0,869	0,078
	ME	0,002	-0,001	0,004	-0,001	-0,003
α	R^2	0,001	0,001	0,001	0,003	0,001
	RMSE	10,179	10,179	10,184	10,181	10,179
	ME	3,056	3,053	3,059	3,047	3,051
n	R^2	0,019	0,029	0,107	0,043	0,078
	RMSE	1,631	1,632	1,631	1,631	1,632
	ME	1,443	1,441	1,445	1,428	1,439

Tabela 2.2.5. Desempenho estatístico das FPTs para estimativa dos parâmetros de van Genuchten (1980) utilizando o conjunto preditor A2

Parâmetro	Estatística	RLM	MARS	RF	SVR	KNN
θ_s	R ²	0,69	0,69	0,71	0,72	0,64
	RMSE	0,032	0,032	0,031	0,029	0,034
	ME	0,001	6,291	0,001	5,061	-0,001
θ_r	R ²	0,21	0,23	0,42	0,38	0,36
	RMSE	0,088	0,081	0,071	0,079	0,079
	ME	0,002	-0,001	0,004	-0,009	-0,001
α	R ²	0,001	0,001	0,001	0,001	0,001
	RMSE	10,179	10,179	10,184	10,177	10,179
	ME	3,054	3,052	3,063	3,042	3,056
n	R ²	0,015	0,023	0,09	0,034	0,032
	RMSE	1,592	1,645	1,642	1,641	1,639
	ME	1,445	1,443	1,448	1,431	1,442

Os modelos RF apresentaram os melhores desempenhos nas estimativas dos parâmetros θ_s e θ_r , com destaque para a θ_s com R² iguais a 0,91 e 0,71 para os conjuntos preditores A1 e A2, respectivamente. Apesar do SVR ter apresentado o melhor valor de R² e RMSE para o conjunto A2, o modelo superestimou a θ_s em 5,06 m³ m⁻³, fazendo o modelo RF mais adequado para a estimativa do parâmetro. Para a umidade residual, os valores de R² foram iguais a 0,48 e 0,42 para os conjuntos preditores A1 e A2, respectivamente.

Já para os parâmetros, α e n, os valores de R² foram aproximadamente zero para todos os modelos e conjuntos preditores, tendo o SVR como melhor modelo dentre os demais. Nota-se que as retas de ajuste (linha vermelha) estão distantes da linha 1:1, indicando que os ajustes foram ruins.

Os valores de ME para os θ_s e θ_r variaram entre -0,009 e 0,004 m³ m⁻³ para os conjuntos A1 e A2, com exceção dos modelos MARS e SVR no conjunto A2. Já o valor de ME para o parâmetro α , observa-se uma superestimação em média de 3,047 e 3,042 m⁻¹ para os conjuntos A1 e A2, respectivamente. Já para o parâmetro n, os valores de ME indicaram uma superestimação entorno de 1,445.

No que se refere aos valores de RMSE, o conjunto A1, apresentou uma variação próxima de zero para θ_s e θ_r , e valores mais altos para α e n, com médias no entorno de 10,18 m⁻¹ para os dois conjuntos preditores. Esses altos valores de RMSE e ME, bem como os baixos valores de R² para os parâmetros α e n podem ser atribuídos a própria característica inerente dos mesmos, causada pela alta variabilidade e conseqüentemente, a dificuldade no ajuste, como citado por Vereecken et al. (1989).

Já os valores de ME e RMSE obtidos pelo modelo nulo para cada variável estimada são

apresentados na Tabela 2.2.6.

Tabela 2.2.6. Desempenho estatístico dos modelos nulos para cada umidade do solo estimada.

Estatística/Variável	θ_s	θ_r	α	n
ME	1,581	9,001	-0,001	-0,001
RMSE	0,056	0,099	9,711	0,798

Observa-se que os algoritmos de aprendizado de máquina foram superiores aos modelos nulos para as umidades de saturação e residual, contudo, os modelos nulos para os parâmetros α e n apresentaram valores de ME e RMSE inferiores aos erros encontrados para os modelos de aprendizado máquina, indicando que o modelo mais simples foi capaz de ajustar melhor aos dados.

Barros et al. (2013) desenvolveram FPTs para estimativa dos parâmetros de van Genuchten na região nordeste do Brasil e encontraram valores de R^2 iguais a 0,12 e 0,21 para os parâmetros α e n, respectivamente, utilizando dados granulométricos, densidade do solo e matéria orgânica. Outros autores como Scheinost et al. (1997), Pachespky e Rawls (2004) e Wosten et al. (2001) encontraram também dificuldades na estimativa dos parâmetros α e n considerando diversos tipos de solos e regiões diferentes do mundo.

Ao analisar os resultados do conjunto A1, percebe-se uma melhora no desempenho das FPTs para a estimativa de θ_s e θ_r . Tomassela e Hodnett (1998) destacam a importância das propriedades estruturais do solo para a estimativa das umidades do solo próximos à saturação e Pachespky e Rawls (2004) acrescentam que esse fato está atrelado às dificuldades dos modelos em representar a distribuição de água nos poros em função do tamanho das partículas do solo nessas faixas de sucções.

Na Figura 2.2.6 são apresentadas as variáveis que mais contribuíram para a estimativa de θ_s e θ_r , pelos modelos obtidos pelo conjunto preditor A1. Observa-se que as variáveis estruturais, macroporosidade e microporosidade, foram importantes em todos os modelos avaliados, com exceção da RLM.

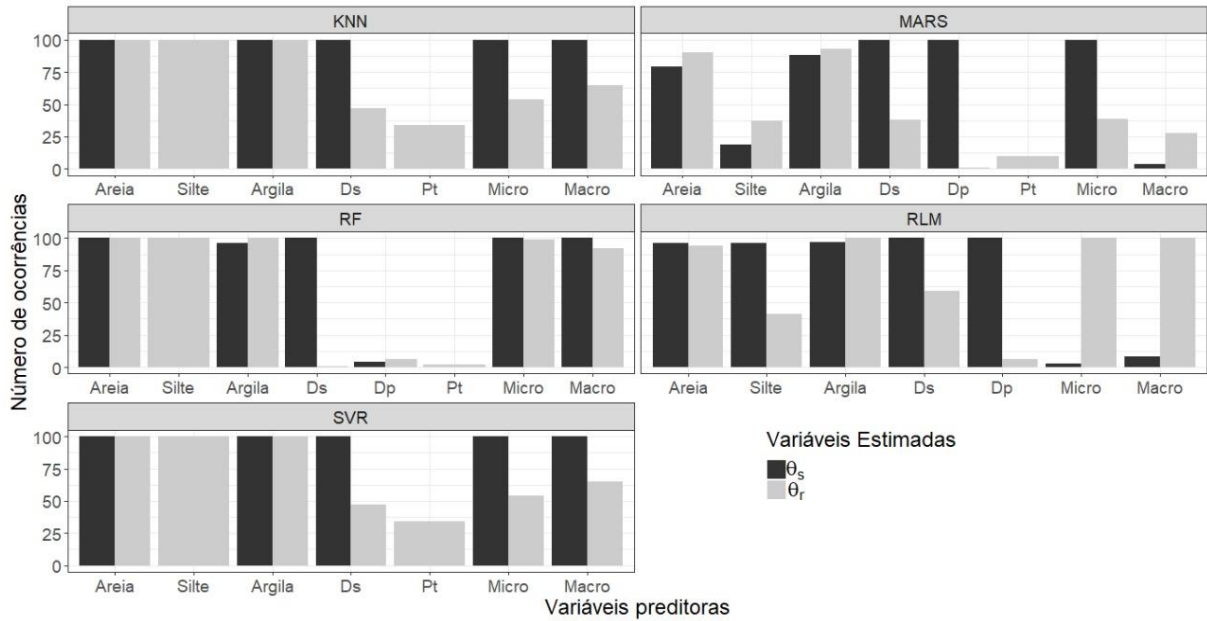


Figura 2.2.6. Classificação das importâncias das variáveis predictoras nas estimativas dos parâmetros θ_s e θ_r utilizando o conjunto predictor A1.

Observa-se que os teores granulométricos, areia e argila, também se destacaram no desenvolvimento dos modelos. Já o silte não contribuiu muito para os modelos mais robustos, KNN, MARS, RF e SVR, na estimativa do θ_s . E as variáveis Ds e Dp alcançaram uma maior ocorrência na estimativa da umidade de saturação do que na residual.

No que se refere ao desempenho dos modelos para os parâmetros θ_s e θ_r , os resultados dos testes de Friedman indicaram o valor 0 para todas as análises, ou seja, houve diferença entre os modelos. Para os testes de Nemenyi, com intervalo de confiança de 95%, a CD obtida foi de 0,61, o que indica que as distâncias entre os *ranks* médios dos modelos devem ser maiores do que este valor para apresentar diferença significativa.

Então, comparando os dois melhores modelos na estimativa de θ_s , RF e SVR (menores valores *ranks* médios), observou-se uma distância crítica igual a 0,12 (1,57 – 1,45), indicando que não existe diferença estatística significativa entre os modelos. Já para o θ_r , RF e KNN (menores valores *ranks* médios), observou-se uma distância crítica igual a 1,26 (2,32 – 1,06), indicando que existe diferença estatística significativa entre eles, ou seja, o modelo RF foi melhor que o KNN e consequentemente, melhor que os demais modelos.

Apesar da RLM não ter apresentado o melhor desempenho dentre os modelos avaliados, esta possibilita a FTP ser representada na forma de equação, sendo assim, as FPTs obtidas para os parâmetros θ_s e θ_r utilizando o conjunto predictor A1 e seus referentes R^2 são apresentados pelas Equações 10 e 11, respectivamente.

$$\theta_s = 0,7475 - 0,02008 \text{ Areia} - 0,003137 \text{ Silte} - 0,00122 \text{ Argila} - 0,2118 \text{ Ds} + 0,1487 \text{ Dp} \quad R^2 = 0,69 \quad (10)$$

$$\theta_r = -0,8568502 + 0,041881 \text{ Areia} - 0,0076599 \text{ Silte} + 0,0062293 \text{ Argila} + 0,0062293 \text{ Micro} - 0,0351842 \text{ Macro} \quad R^2 = 0,21 \quad (11)$$

Para o conjunto A2, as FPTs obtidas para os parâmetros θ_s e θ_r são apresentadas nas equações 12 e 13, respectivamente.

$$\theta_s = 0,720269 - 0,0200064 \text{ Areia} - 0,0030131 \text{ Silte} - 0,0012113 \text{ Argila} + -0,1402536 \text{ Dp} \quad R^2 = 0,69 \quad (12)$$

$$\theta_r = -0,5194270 + 0,02254812 \text{ Areia} + 0,0061362 \text{ Silte} + 0,0029232 \text{ Argila} + 0,0878161 \text{ Dp} \quad R^2 = 0,14 \quad (13)$$

2.2.4 Conclusões

O modelo de van Genuchten (1980) apresentou o melhor ajuste aos dados da curva de retenção de água para o Bioma Cerrado. Já os demais modelos, Brooks e Corey (1964), Fredlund e Xing (1994) e Durner (1996), também apresentaram resultados satisfatórios, com R^2 superiores a 0,9 e valores de RMSE e ME próximos de zero.

As FPTs desenvolvidas para os parâmetros α e n da equação de van Genuchten (1980) apresentaram baixos desempenhos em todos os modelos e conjuntos de preditores, tendo os modelos nulos como superiores aos modelos de aprendizado de máquina. Para o parâmetro θ_r , o melhor modelo foi o RF utilizando as variáveis preditoras granulométricas, densidade do solo, densidade de partícula, porosidade, microporosidade e macroporosidade, contudo, a capacidade preditiva foi mediana.

Já para as FPTs obtidas para o parâmetros θ_s , foi obtido o melhor desempenho dentre os parâmetros, com R^2 igual a 0,76 e valores RMSE e ME próximos de zero utilizando o algoritmo RF. E segundo o teste de Nemenyi, os modelos RF e SVR não se diferenciaram estatisticamente na estimativa do mesmo.

2.2.5 Referências bibliográficas

ARAYA, S. N.; GHEZZEHEI, T. A. Using machine learning for prediction of saturated hydraulic conductivity and its sensitivity to soil structural perturbations. **Water Resources Research**, v. 55, n. 7, p. 5715-5737, 2019.

BARROS, A. H. C. B.; LIER, Q. J.; MAIA, A. H. N.; SCARPARE, F. V. Pedotransfer functions to estimate water retention parameters of soils in northeastern Brazil. **Revista Brasileira de Ciência do Solo**, v. 37, p. 379-391, 2013.

BREIMAN, L. Random Forests. **Machine Learning**, v. 45, n. 1, p. 5–32, 1 out. 2001.

BROOKS, R. H.; COREY, A. T. Hydraulic properties of porous media, Hydrol. Paper 3, Colorado State Univ., Fort Collins, CO, USA, 1964.

CAMPBELL, G. S. A simple method for determining unsaturated conductivity from moisture retention data. **Soil Science.**, v. 117, p. 311-314, 1974.

COSTA, W. A.; OLIVEIRA, C. A. S.; KATO, E. Modelos de ajuste e métodos para a determinação da curva de retenção de água de um latossolo vermelho-amarelo. **Revista Brasileira de Ciência do Solo**, v. 32, n.2, 2008.

D'EMILIO, A.; AIELLO, R.; CONSOLI, S.; VANELLA, D.; IOVINO, M. Artificial Neural Networks for Predicting the Water Retention Curve of Sicilian Agricultural Soils. **Water**, v. 10, p. 1431, 2018.

DEMSAR, J. Statistical comparisons of classifiers over multiple data sets. **Journal of Machine Learning Research**, v. 7, p. 1-30, 2006.

DURNER, W., Hydraulic conductivity estimation for soils with heterogeneous pore structure. **Water Resources Research**, v. 32, n. 9, p. 211-223, 1994.

EBRAHIMI, E.; BAYAT, H.; NEYSHABURI, M. R.; ABYANEH, H. Z. Prediction capability of different soil water retention curve models using artificial neural networks. **Archives of Agronomy and Soil Science**, v. 60, n. 6, 2014.

FREDLUND, D. G; XING, A. Equations for the soilwater characteristic curve. **Canadian Geotechnical Journal**, v. 31, p. 521-532, 1994.

FRIEDMAN, J. H. Multivariate adaptive regression splines. **The Annals of Statistics**, v. 19, n. 1, 1991.

GUANGZHOU CHEN, L.J.; LI, X. Sensitivity Analysis and Identification of Parameters to the van Genuchten Equation. **Journal Chemistry**, v. 2016, 2016.

HUTSON, J. L.; CASS, A. A retentivity function for use in soil-water simulation models. **J. Soil Science**, v. 38, p. 105-113, 1987.

IBGE (Instituto Brasileiro de Geografia e Estatística), Diretoria de Pesquisas, Coordenação de Agropecuária, Levantamento Sistemático da Produção Agrícola – jan. 2021.

JULIÀ, M. F.; MONTREAL, E.; JIMÉNEZ, A. S. C.; MELÉNDEZ, E. G. Constructing a

saturated hydraulic conductivity map of Spain using pedotransfer functions and spatial prediction. **Geoderma**, v. 123, n. 3-4, p. 257-277, 2004.

KOSUGI, K. Three-parameter lognormal distribution model for soil water retention, **Water Resources Research**, v. 30, n. 4, p. 891–901, 1994.

MEDEIROS, J. C.; COOPER, M.; ROSA, J. D.; GRIMALDI, M.; COQUET, Y. Assessment of pedotransfer functions for estimating soil water retention curves for the amazon region. **Revista Brasileira de Ciência do Solo**, v. 38, p. 730-743, 2014.

MEDRADO, E. LIMA, J. E. F.W. Development of pedotransfer functions for estimating water retention curve for tropical soils of the Brazilian savanna. **Geoderma Regional**, v. 1, p. 59-66, 2014.

MICHELON, C. J.; CARLESSO, R.; OLIVEIRA, Z. B.; KNIES, A. E.; PETRY, M. T.; MARTINS, J. D. Funções de pedotransferência para estimativa da retenção de água em alguns solos do Rio Grande do Sul. **Ciência Rural**, v. 40, n. 4, p. 848-853, 2010.

MILBORROW, S. Earth: Multivariate adaptive regression splines. R package version 5.1.2, 2019.

NEMENYI, P. B. Distribution-free Multiple Comparisons. PhD thesis, Princeton University. 1963.

PACHEPSKY, Y.; RAWLS, W.J. Development of pedotransfer functions in soil hydrology. Elsevier, Amsterdam, Netherlands, 2004.

PACHEPKY, Y.; PARK, Y. Saturated Hydraulic Conductivity of US Soils Grouped According to Textural Class and Bulk Density. **Soil Science Society of America Journal**, vol. 79, n. 4, p. 1094-1100, 2015.

PAN, T.; HOU, S.; LIU, Y.; TAN, Q. Comparison of three models fitting the soil water retention curves in a degraded alpine meadow region. **Scientific reports**, v. 9, 2019.

RAHMAN, R.; HAIDER, S.; GHOSH, S.; PAL, R. Design of probabilistic random forests with applications to anticancer drug sensitivity prediction. **Cancer Informatics**, v. 14, n. Suppl 5, p. 57–73, 31 mar. 2016.

RODRIGUES, L. N.; MAIA, A. H. N. **Funções de pedotransferência para estimar a condutividade hidráulica saturada e as umidades de saturação e residual do solo em uma bacia hidrográfica do Cerrado**. XIX Simpósio Brasileiro de Recursos Hídricos. **Anais...** Macéio - AL: Associação Brasileira de Recursos Hídricos, 2011.

RODRIGUES, L. N.; MAIA, A. H. N.; SILVA, R. N; **Funções de pedotransferência para estimar capacidade de campo, ponto de murcha permanente e densidade global em solos de uma bacia hidrográfica do Bioma Cerrado**. XL Congresso Brasileiro de Engenharia Agrícola. **Anais...** Cuiabá – MT: Associação Brasileira de Engenharia Agrícola, 2011.

SEKI, K. SWRC Fit - A nonlinear fitting program with a water retention curve for soils having unimodal and bimodal pore structure. **Hydrology and Earth System Sciences**, v.4, p.407-

437, 2007.

SCHAAP, M.G., LEIJ, F.J., VAN GENUCHTEN, M.TH. ROSSETA: computer program for estimating soil hydraulic parameters with hierarchical pedotransfer functions. **Journal Hydrology**, v. 251, p. 163–176, 2001.

SHARMA, S.K.; MOHANTY, B.P.; ZHU, J. Including topography and vegetation attributes for developing pedotransfer functions. **Soil Science Society American Journal**, v. 70, p. 1430–1440, 2006.

SHEN, C., LALOY, E., ELSHORBAGY, A., ALBERT, A., BALES, J., CHANG, F. J. HESS Opinions: Incubating deep-learning-powered hydrologic science advances as a community. **Hydrology and Earth System Sciences**, v. 22, n. 11, p. 5639-5656, 2018.

SCHENOIST, A. C.; SINOWSKI, W.; AUERSWALD, K. Regionalization of soil water retention curves in a highly variable soilscape, I. Developing a new pedotransfer function. **Geoderma**, v. 48, n. 3-4, p. 129-143, 1997.

SOUSA NETO, E.; SMALLMAN, L.; OMETTO, J.; WILLIAMS, M. Carbon dynamics in the Brazilian Cerrado: stocks and fluxes estimated by a model data fusion framework (CARDAMOM). **EGU General Assembly**, 2020.

TOMASELLA, J. HODNETT, M. G. Soil hydraulic properties and van Genuchten parameters for an oxisol under pasture in central Amazonia. In: Gash, J. H. C.; Nobre, C. A.; Roberts, J. M.; Victoria, R. L. **Amazonian Deforestation and Climate**. Chichester: John Wiley, p. 101-124, 1996.

TOMASELLA, J., HODNETT, M.G., ROSSATO, L. Pedotransfer functions for the estimation of soil water retention in Brazilian soils. **Soil Science Society of America Journal**, v. 64, p. 327–338, 2000.

TWARAKAVI, N. K. C., SIMUNEK, J., SCHAAP, M. G. Development of pedotransfer functions for estimation of soil hydraulic parameters using support vector machines. **Soil Science Society of America Journal**, v. 73, n. 5, 1443-1452, 2009.

VAPNIK, V.N. The nature of statistical learning theory. Springer, Berlin. 1995.

VAN GENUCHTEN, M. T. A closed-form equation for predicting the hydraulic conductivity of unsaturated soils. **Soil Science Society of America Journal**, v. 44, p. 892- 898, 1980.

VEREecken, H.; MAES, J.; FEYEN, J.; DARIUS, P. Estimating the soil moisture retention characteristic from texture, bulk density, and carbon content. **Soil Science**, v. 148, p. 389-403, 1989.

VEREecken, H.; WEYNANTS, M.; JAVAUX, M.; PACHEPSKY, Y.; SCHAAP, M. G; VAN GENUCHTEN, M.T. Using pedotransfer functions to estimate the van Genuchten-Mualem soil hydraulic properties: A review. **Vadose Zone Journal**, v. 9, p. 1-26, 2010.

WÖSTEN, J. H. M.; PACHEPSKY, Y. A.; RAWLS, W.J. Pedotransfer functions: bridging gap between available basic soil data and missing soil hydraulic characteristics. **Journal**

Hydrology, v. 251, p. 123–150, 2001.

3. CONCLUSÕES GERAIS

As FPTs desenvolvidas para o Bioma Cerrado apresentaram desempenhos satisfatórios na estimativa das umidades do solo nas tensões de 0, 6, 10, 33, 100 e 1500 kPa. Já para a estimativa da K_s , as FPTs apresentaram média capacidade preditiva e o preenchimento de dados faltantes para tal variável melhorou o desempenho dos modelos utilizando os conjuntos de preditores constituídos por dados granulométricos e densidade do solo ou somente por dados granulométricos.

As variáveis predictoras, umidades do solo na capacidade de campo e no ponto de murcha permanente, foram considerados importantes quando utilizados no desenvolvimento das FPTs.

As FPTs desenvolvidas utilizando algoritmos mais robustos (MARS, RF, SVR e KNN) apresentaram, na maioria dos casos, os melhores desempenhos quando comparado com o método mais simples da RLM, contudo, a utilização da RLM é útil devido a sua capacidade de ser apresentar os resultados na forma de uma expressão matemática.

A equação de van Genuchten (1980) se mostrou como o melhor modelo para o ajuste da curva de retenção no Bioma Cerrado, o que não exclui a utilização dos outros modelos avaliados, Brooks e Corey (1964), Fredlund e Xing (1994) e Durner (1994), que também apresentaram resultados satisfatórios.

Na estimativa dos parâmetros de van Genuchten (1980), a umidade de saturação apresentou o melhor desempenho, com R^2 igual a 0,76 utilizando o conjunto de preditor composto por dados granulométricos e estruturais. A umidade residual obteve média capacidade preditiva e os parâmetros, α e n , resultados baixos de R^2 e altos de RMSE e ME.

Os modelos que mais se destacaram nas estimativas da K_s , umidades do solo e dos parâmetros de ajuste da equação de van Genuchten (1980) foram o RF e SVR.