

UNIVERSIDADE FEDERAL DE PELOTAS

Faculdade de Agronomia Eliseu Maciel

Programa de Pós-Graduação em Manejo e Conservação do Solo e da Água



Tese

**Modelagem da retenção de água em solos sob clima subtropical usando funções
de pedotransferência e redes neurais artificiais**

Miguel David Fuentes Guevara

Pelotas, 2022

Miguel David Fuentes Guevara

Modelagem da retenção de água em solos sob clima subtropical usando funções de pedotransferência e redes neurais artificiais

Tese apresentada ao Programa de Pós-Graduação em Manejo e Conservação do Solo e da Água da Faculdade de Agronomia Eliseu Maciel da Universidade Federal de Pelotas, como requisito parcial à obtenção do título de Doutor em Ciências (área do conhecimento: Manejo e Conservação do Solo e da Água).

Orientador: Prof. Dr. Luís Carlos Timm

Coorientador (es): Prof. Dr. Ulisses Brisolara Corrêa

Prof^a Dra. Luciana Marini Kopp

Prof. Dr. Tonismar dos Santos Pereira

Pelotas, 2022

G939m Fuentes Guevara, Miguel David

Modelagem da retenção de água em solos sob clima subtropical usando funções de pedotransferência e redes neurais artificiais / Miguel David Fuentes Guevara ; Luís Carlos Timm, orientador ; Ulisses Brisolara Corrêa, coorientador. — Pelotas, 2022.

157 f.

Tese (Doutorado) — Programa de Pós-Graduação em Manejo e Conservação do Solo e da Água, Faculdade de Agronomia Eliseu Maciel, Universidade Federal de Pelotas, 2022.

1. Bacia hidrográfica. 2. Atributos topográficos. 3. Atributos físico-hídricos. 4. Pedofunção pseudo-contínua. 5. Algoritmos de aprendizado de máquina. I. Timm, Luís Carlos, orient. II. Corrêa, Ulisses Brisolara, coorient. III. Título.

CDD : 631.521

Miguel David Fuentes Guevara

MODELAGEM DA RETENÇÃO DE ÁGUA EM SOLOS SOB CLIMA SUBTROPICAL
USANDO FUNÇÕES DE PEDOTRANSFERÊNCIA E REDES NEURAIS ARTIFICIAIS

Tese aprovada, como requisito parcial para obtenção do grau de Doutor em Ciências (área do conhecimento: Manejo e Conservação do Solo e da Água), Programa de Pós-Graduação em Manejo e Conservação do Solo e da Água, Faculdade de Agronomia Eliseu Maciel, Universidade Federal de Pelotas.

Data da defesa: 01/04/2022

Banca examinadora:

Prof. Dr. Luís Carlos Timm (Orientador)

Doutor em Agronomia pela Universidade de São Paulo (Escola Superior de Agricultura Luiz de Queiroz/USP)

Prof. D.Sc. Robson André Armindo

Doutor em Ciências pela Universidade de São Paulo

Prof. Dra. Cláudia Liane Rodrigues de Lima

Doutora em Solos e Nutrição de Plantas pela ESALQ/Universidade de São Paulo

Dra. Virginia Ortiz Andersson

Doutora em Ciência da Computação pela Universidade Federal de Pelotas.

*Dedico a mis padres Miguel y Yallexi,
a mi querida hermana Katy
y a mi amado sobrino Abdiel,
porque todo lo que soy hoy es gracias a ustedes.*

Agradecimientos

Agradezco primeramente a Dios porque me permitió entender el propósito de mi vida, para que a cada paso que yo dé, sea más semejante a Él.

A mi querido padre Miguel que me demuestra todo su grande amor, sacrificando sus propios intereses y sueños por verme feliz triunfando. A ti padre por importarte con los más mínimos detalles, siempre preocupado para que nunca me falte nada, trabajando con todo el sudor para sustentarnos como familia, siempre queriéndonos dar los mejores cuidados y la mejor educación;

A mi querida madre Yallexi por toda la disposición en verme crecer como persona y profesional, porque su grande amor incondicional lo pude ver reflejado en sus cuidados por mi vida y mi futuro. A ti madre por ser una vencedora y luchadora contra el cáncer, que a pesar de no estar a tu lado en estos 6 años muy difíciles en tu vida, siempre me enseñaste a mantener la Fé y firmeza en Dios aceptando su voluntad;

A mi querida hermana Katy por todo el amor y cuidado, porque pude ver cada día tu preocupación por verme feliz y protegerme con cautela de todo lo que el mundo me deparaba, siempre atenciosa, cariñosa y temerosa de Dios. Y que más agradecerte por esa felicidad que trajiste al mundo, nuestro querido Abdiel.

A mis orientadores Luís Carlos Timm y Ulisses Brisolara Côrrea por todo el apoyo que me brindaron para seguir adelante, por los momentos de amistad y alegría. Además, agradecer por toda la dedicación, por soportarme y enseñarme durante el tiempo de mi doctorado, que con seguridad es un grande conocimiento transmitido que me hizo crecer en mi vida profesional. También agradecido por los incentivos y colaboraciones de mis coorientadores Luciana Marini Kopp y Tonismar dos Santos Pereira.

A todos los profesores del PPG MACSA por las enseñanzas brindadas y todo el incentivo en el proceso de aprendizaje;

A los participantes de esta investigación Gabriel Gomes, Thiago Rodrigues da Rosa, Ricardo Matsumura Araújo, Robson Armindo y a los laboratorios colaboradores DATALAB y GPSAPA, los cuales fueron claves para para la realización de este trabajo y todos los momentos de compañerismo;

A mis colegas de estudios durante el doctorado en el PPG MACSA y PPG Recursos hídricos, especialmente a Stefan, Yasmin, Rômulo, Mauricio, Ângelo, Nathan, Rodrigo y Luana por los momentos de apoyo en los estudios, además agradecer por los momentos de unión, compañerismo y alegrías.

A la Universidad Federal de Pelotas, Departamento de Suelos y Programa de Postgraduación en Manejo y Conservación del Suelo y del Agua por permitir realizar mis estudios de doctorado, ofreciéndome todo lo necesario para llevar a cabo la investigación;

A la CAPES por concederme la beca de estudios;

A todos los demás amigos que de una u otra forma aportaron un granito de arena en estos años de doctorado en lo académico y personal, en especial a Thais, Tamara, Eliana, Karol, Nicolas y Jeferson.

Muchas Gracias!

Resumo

FUENTES GUEVARA, Miguel David. **Modelagem da retenção de água em solos sob clima subtropical usando funções de pedotransferência e redes neurais artificiais.** Orientador: Luís Carlos Timm. 2022. 157 f. Tese (Doutorado) – Programa de Pós-graduação em Manejo e Conservação do Solo e da Água). Universidade Federal de Pelotas, Pelotas, 2022.

As funções de pedotransferência (FPTs) surgem como uma alternativa para estimar a retenção de água no solo a partir de atributos do solo mais facilmente disponíveis nos bancos de dados de solo. As redes neurais artificiais (RNAs) são uma ferramenta que possibilita a geração de FPTs que melhoram esta estimativa. Portanto, o objetivo do presente estudo foi explorar as ferramentas que permitem a avaliação e geração de FPTs para estimar a retenção de água no solo em escala de bacia hidrográfica. Para isso, a área de estudo selecionada foi a bacia hidrográfica Arroio Pelotas (BHAP) localizada no sul do estado do Rio Grande do Sul. Na bacia foram estabelecidos 100 pontos amostrais ao longo de uma transecção espacial de 25 km, nos quais se coletaram amostras de solo na profundidade de 0-20 cm para determinação dos atributos do solo, oriundos do trabalho de dissertação da Luciana Oliveira e atributos topográficos. Assim, obtiveram-se dados do teor de argila, silte e areia, densidade do solo, densidade de partículas, macroporosidade, microporosidade, carbono orgânico, retenção de água no solo nos potenciais 0, -1, -6, -10, -33, -100 e -1.500 kPa, elevação, aspecto, declividade, curvatura plana, curvatura do perfil, e curvatura do terreno. O primeiro estudo foi a avaliação da transferibilidade de FPTs desenvolvidas em outras condições climáticas na estimativa da retenção de água no solo nos potenciais de -33 e 1.500 kPa. A escolha das melhores FPTs foi feita de acordo com a proximidade na estimativa da taxa de variação da evapotranspiração real, obtendo-se como melhor FPT a desenvolvida por Veerecken *et al.* (1989). O segundo estudo foi a geração de uma FPT pseudo-contínua (FPT-PC) mediante a técnica de RNAs para estimar a retenção de água no solo nos diversos potenciais matriciais. Para isso, foi gerado um banco de dados de solo pela junção dos dados da BHAP e 105 dados coletados em um grid na sub-bacia hidrográfica de cabeceira Sanga Ellert/Canguçu-RS a qual está inserida na BHAP e no município de canguçu. A técnica de *random search* foi usada no intuito de explorar os efeitos dos hiper-parâmetros da rede, selecionar os atributos de entrada e obter um modelo que melhore as estimativas da retenção de água no solo. O melhor conjunto de atributos de entrada na RNA foi densidade do solo, carbono orgânico, microporosidade, argila, curvatura do terreno e plano de curvatura do terreno. Assim, o modelo com melhor desempenho para estimar a retenção de água no solo para o conjunto de todos os potenciais matriciais teve a configuração de hiper-parâmetros: semente = 98,186; taxa de aprendizado = 0,2; número de épocas = 5.000; e momento = 0,2, obtendo métricas estatísticas de desempenho $r^2 = 0,73$; $ME = 0,052 \text{ cm}^3 \text{ cm}^{-3}$; e $RMSE = 0,065 \text{ cm}^3 \text{ cm}^{-3}$.

Palavras-chave: Bacia hidrográfica. Atributos topográficos. Atributos físico-hídricos. Água no solo. Pedofunção pseudo-contínua. Algoritmos de aprendizado de máquina.

Abstract

FUENTES GUEVARA, Miguel David. **Modeling water retention in soils under subtropical climate using pedotransfer functions and artificial neural networks.**

Advisor: Luís Carlos Timm. 2022. 157 f. Thesis (Doctoral Degree). Graduate Program in Management and Conservation of Soil and Water (PPG-MACSA). The Federal University of Pelotas, Pelotas, 2022.

Pedotransfer functions (PTFs) emerge as an alternative to estimate the soil water content from soil attributes that are easier available in soil databases. The *artificial neural networks* (ANNs) are a tool that enables PTFs generation with better estimates. Therefore, the objective of this study was to explore the tools that allow the assessment and generation of PTFs to estimate the soil water content on a soil watershed. To achieve this, the study area selected was the Pelotas River Watershed (PRW), located in the southern Rio Grande do Sul - Brazil. In the PRW were established 100 sample points along a 25-km-spatial transect in which were collected soil samples from the surface soil (0-20 cm) to determine soil and topographical attributes. Thus, the following attributes were obtained: clay, silt, and sand content, bulk density, particle soil density, macroporosity, microporosity, soil organic content, soil water content in the potentials 0, -1, -6, -10, -33, -100 e -1,500 kPa, elevation, aspect, soil slope, plan curvature, profile curvature, and curvature. The first study evaluated the PTFs transferability developed in other climate conditions for estimating the soil water content in the matric potentials -33 and 1,500 kPa. The bests PTFs were chosen according to the performance to estimate the variation rates of actual evapotranspiration, obtaining that the PTF developed by Vereecken *et al.* (1989) had the better performance. The second study carried out the generation of a pseudo-continuous PTF (PC-PTF) through the ANN technique to estimate the soil water content in several matric potentials. A soil database was generated to achieve this, joining the PRW data and the 105-point grid collected in the Ellert Creek/Canguçu-RS Watershed (ECW), which is a headwater of the PRW. The *random search* technique was used to explore the hyper-parameter effects of the ANN, to find the input attributes selection, and to obtain one model for improving soil water content estimations. The ANN's best input attributes were soil bulk density, soil organic carbon content, microporosity, clay, curvature, and plane curvature. Thus, the model with best performance to estimate the soil water content for the set of all matric potentials had hyper-parameters set: seed = 98.186; learning rate = 0.2; number of epochs = 5,000; and moment = 0.2. This model obtained performance statistical metrics $r^2 = 0.73$; $ME = 0.052 \text{ cm}^3 \text{ cm}^{-3}$; and $RMSE = 0.065 \text{ cm}^3 \text{ cm}^{-3}$.

Keywords: Watershed. Topographical attributes. Physical-hydric attributes. Soil water. Pseudo-continuous pedotransfer function. Machine learning algorithms.

Lista de Figuras

Figura 1 - Exemplo de trade-off entre bias e variância apresentando o ponto de complexidade ótimo do desempenho de um modelo. Adaptado de Aggarwal (2018)....	42
Figura 2 - Ilustração do underfitting e overfitting em relação ao erro de treinamento e teste. Adaptado de Chollet (2020).....	43
Figura 3 - a) arquitetura de um perceptron sem bias b) arquitetura de um perceptron com bias. Apresentando os nós de entrada, nó de saída e neurônio de bias. Adaptado de Aggarwal (2018).....	48
Figura 4 - Exemplo de arquitetura de uma RNA de múltiplas camadas. Apresentando a camada de entrada, camadas ocultas e camada de saída. Adaptado de Aggarwal (2018).	48
Figura 5 - Exemplos de dados linearmente (a) separáveis e (b) inseparáveis em duas classes. Fonte: Aggarwal (2018).	52
Figura 6 - Perceptron modificado: modelo base de um neurônio para projetos de RNAs. Fonte: Soares (2013).	53
Figura 7 - Arquitetura básica de uma rede feed-forward com duas camadas ocultas e uma única camada de saída, apresentando a rede (a) sem neurônios de bias e (b) com neurônios de bias. Adaptado de Aggarwal (2018).....	54
Figura 8 – Ruptura do neurônio e valores de pré-ativação e pós-ativação dentro de um neurônio. Adaptado de Aggarwal (2018).	55
Figura 9 - Representações gráficas das funções de ativação mais usadas. Fonte: Aggarwal (2018).	57
Figura 10 - Separações de conjuntos pelas funções lógicas AND, OR e XOR. Fonte: Araújo (2018).....	58
Figura 11 - Solução de separação de um conjunto pelo operador OR com um único neurônio linear “perceptron” usando uma função de ativação sigmóide. Fonte: Araújo (2018).....	58
Figura 12 - Exemplo do comportamento da função loss MSE em função das épocas de treinamento de uma RNA. Representando o erro quadrado médio dos conjuntos treino e testes. Adaptado de Brownlee (2019).	60
Figura 13 - Exemplificação da loss em função do ajuste dos pesos. Apresentado o ponto aleatório de início dos pesos e da loss e o mínimo global (ponto mais baixo de loss onde se quer chegar). Adaptado de Campbell (2017).....	61
Figura 14 - Comportamento do gradiente descendente durante o treinamento da rede neural, apresentando o ponto de início dos pesos loss de forma aleatória até a chegada no mínimo global (ponto mais baixo de loss).	62
Figura 15 - Topologias típicas de FPTs baseadas em redes neurais artificiais para solucionar problemas de regressão: pontuais, paramétricas e pseudo-contínuas (FPT-	

PC). Nos nodos de entradas se encontram os atributos de solo e na FPT-PC é adicionado o potencial matricial ($-\psi$). Adaptado de Haghverdi, Cornelis e Gharaman (2012).70

Lista de Abreviaturas e Siglas

AM	Aprendizado de Máquina
BHAP	Bacia hidrográfica Arroio Pelotas
CNNs	Rede neurais convolucionais
CRAS	Curva de retenção de água no solo
FPT	Função de pedotransferência
FPT _s	Funções de Pedotransferência
FPT-PC	FPT pseudo-contínua
h	Potencial matricial
IA	Inteligência artificial
MAE	Erro absoluto médio
ME	Erro médio
MSE	Erro quadrado médio
r^2	Coeficiente de determinação
RMSE	Raiz quadrada do erro médio
RNA	Rede neural artificial
RNAs	Redes neurais artificiais
RS	Rio Grande do Sul
SVM	Máquina de vetores de suporte
θ	Umidade do solo

Sumário

1 Introdução.....	14
1.1 Organização do texto	17
2 Hipóteses	17
3 Revisão Bibliográfica.....	18
3.1 Importância das funções e propriedades hidráulicas do solo	18
3.1.1 Curva de retenção de água no solo.....	19
3.2 Aspectos gerais das funções de pedotransferência	22
3.3 Funções de pedotransferência para estimativa da CRAS	26
3.4 Disponibilidade de bases de dados de solo	27
3.5 Aspectos gerais dos métodos mais utilizados no desenvolvimento das funções de pedotransferência.....	28
3.6 Avaliação das funções de pedotransferência	30
3.7 Aplicabilidade e transferibilidade das funções de pedotransferência	31
3.8 Diferenças entre funções de pedotransferência de solos de clima tropical e temperado	34
3.9 Funções de pedotransferência geradas e aplicadas no Rio Grande do Sul.....	37
3.10 Aprendizado de máquina e técnicas para a avaliação de modelos.....	40
3.10.1 Generalização, otimização e avaliação dos modelos de AM.....	41
3.10.1.1 Bias e variância	41
3.10.1.2 Sobreajuste e subajuste	42
3.10.1.3 Conjuntos de treinamento, validação e testes no AM.....	44
3.10.2 Aspectos gerais das redes neurais artificiais.....	47
3.10.2.1 Redes neurais artificias	47
3.10.2.2 Arquitetura básica de uma rede neural.....	47

3.10.2.3 O perceptron: rede de camada única e o seu modelo matemático	49
3.10.2.4 Redes de múltiplas camadas e modelo matemático	53
3.10.2.5 Funções de ativação, nós de saída e função de custo (<i>loss</i>)	55
3.10.2.6 Função de custo (<i>loss</i>)	59
3.10.2.7 Treinamento das redes neurais e modelos de aprendizagem	62
3.10.2.8 Treinamento da rede neural com <i>backpropagation</i>	62
3.10.2.9 Principais problemas práticos no treinamento das redes neurais	63
3.10.2.9.1 Sobreajuste nas RNAs	64
3.10.2.9.2 Regularização.....	65
3.10.2.9.3 Parada antecipada	65
3.10.2.9.4 <i>Random search</i>	67
3.10.3 Aplicações de redes neurais artificiais para estimativa da CRAS	67
4 Artigo 1.....	72
5 Artigo 2.....	110
6 Considerações Finais	145
Referências	146
Anexos	155
Anexo A – Disponibilidade do repositório	156

1 Introdução

A retenção de água no solo é uma propriedade físico-hidráulica importante na ciência do solo, pois o seu entendimento contribui à tomada de decisões e ao estabelecimento de práticas agrícolas adequadas, como o manejo eficiente da água, manejo e conservação do solo e da água, e técnicas de irrigação e drenagem. Além disso, o conhecimento da retenção de água no solo é a chave para o entendimento dos processos do ciclo hidrológico e modelos terrestres, permitindo entender a realidade desses processos em uma bacia hidrográfica (JANA; MOHANTY; SHENG, 2012; PINEDA *et al.*, 2018; ZIMMERMANN; BASILE, 2011).

Normalmente, a determinação em campo e laboratório da retenção de água no solo em relação às diferentes tensões exercidas pelo solo (curva de retenção de água no solo) demanda tempo e necessita de equipamentos caros para a sua obtenção, sobretudo em estudos de grande escala como é o caso das bacias hidrográficas (BARROS *et al.*, 2013). Diante disso, estão sendo desenvolvidos métodos alternativos para estimar esse parâmetro de forma rápida e de baixo custo, dentre esses métodos, as Funções de Pedotransferência (FPTs) merecem destaque. Bouma (1989) redefiniu o conceito de FPTs como sendo “funções preditivas de certas propriedades do solo de obtenção mais difícil a partir de outros atributos que são facilmente disponíveis”. Atualmente, as FPTs são consideradas como ferramentas indiretas para estimar certos parâmetros do solo de difícil obtenção a partir de outros atributos do solo regularmente encontrados em bancos de dados, tais como granulometria do solo, carbono orgânico e densidade do solo (BARROS *et al.*, 2013; JANA; MOHANTY; SHENG, 2012).

As FPTs são amplamente utilizadas pelos cientistas do solo durante décadas, visto que elas permitem minimizar a escassez de informações de certos parâmetros em banco de dados de solo (AULER; PIRES; PINEDA, 2017; BOUMA, 1989; COSTA *et al.*, 2013; PINEDA *et al.*, 2018), sendo a curva de retenção de água no solo, um dos atributos do solo mais comumente estimados pelas FPTs (DA SILVA *et al.*, 2015; NEBEL *et al.*, 2010).

O principal inconveniente encontrado durante a utilização das FPTs a nível global é que, a grande maioria foram desenvolvidas sob condições específicas de outros países com pedogênese e zonificação climática diversa, predominando a geração de FPTs em

zonas de clima temperado em relação aos solos de clima tropical. Os solos de clima tropical apresentam um processo de formação distinto em relação aos de clima temperado, diferindo nos tipos de solos, características e propriedades físico-químicas, pedológicas e mineralógicas, devido às diferenças nos processos de intemperismo induzidos pelo clima (BOTULA *et al.*, 2012).

Frente à falta de bancos de dados de solos grandes e com diversidade de informações nas regiões de clima tropical, a geração de FPTs para este tipo de clima são ainda escassas, o que tem levado à utilização de forma generalizada das FPTs de clima temperado para estimar a retenção de água no solo, tendendo-se a aumentar os erros de estimativa (BARROS *et al.*, 2013; COSTA *et al.*, 2013; DA SILVA *et al.*, 2015; MEDRADO; LIMA, 2014; NEBEL *et al.*, 2010; PINEDA *et al.*, 2018). Em nível do Brasil, onde predominam os climas tropical e subtropical, as FPTs que estimam a retenção de água no solo ainda são poucas, sendo necessária uma maior exploração de técnicas que contribuam na geração de novas FPTs com maior precisão e exatidão para estes tipos de climas, e que manifestem os efeitos reais das características dos solos brasileiros sobre essa propriedade físico-hidráulica.

Dentro das variadas técnicas clássicas de obtenção das FPTs, a técnica de regressão linear múltipla mostra-se como a mais prática e simples de ser aplicada na geração de FPTs. Apesar da complexidade que a retenção de água no solo possui como variável de estudo, esse método pode resultar na maioria das vezes em uma alta exatidão da sua estimativa. Contudo, as FPTs obtidas a partir das técnicas de inteligência artificial como o caso das redes neurais artificiais (RNAs), podem apresentar melhor estimativa e exatidão em relação aos métodos tradicionais, uma vez que as RNAs são descritas como aproximadores universais de funções e possuem a capacidade de aprender e mapear qualquer função não-linear contínua, resolvendo problemas complexos (AGGARWAL, 2018; SOARES *et al.*, 2014). As FPTs e RNAs têm sido principalmente utilizadas para a estimativa da água no solo em escala de vertentes e continental, porém poucas têm sido geradas e aplicadas para bacias hidrográficas, que é a unidade de estudo fundamental para o planejamento e gestão dos recursos hídricos (SCHAAP; LEIJ; VAN GENUCHTEN, 2001; VEREECKEN *et al.*, 1989; WÖSTEN *et al.*, 1999).

Dentre as bacias hidrográficas localizadas no sul do estado do Rio Grande do Sul (RS) sob as condições de clima subtropical, a Bacia Hidrográfica do Arroio Pelotas (BHAP) é estratégica para o desenvolvimento econômico e social da região, em particular para os municípios de Pelotas, Morro Redondo, Arroio do Padre, Capão do Leão e Canguçu. A BHAP é responsável pelo fornecimento de cerca de 50% da água usada no abastecimento do município de Pelotas para uma população de 329.000 habitantes. Além do abastecimento de Pelotas, a BHAP possui um valor histórico e cultural incalculável para a região sul do RS, destacando-se que o curso d'água principal da BHAP é um importante afluente do canal São Gonçalo, que abastece o município de Rio Grande – RS e o canal ainda é uma via navegável importante que liga a laguna dos Patos à Lagoa Mirim (DONGLI *et al.*, 2017).

A BHAP sofre impactos das atividades antrópicas provenientes das atividades agrícolas intensivas, os quais a tornam vulnerável à erosão e alagamentos. Assim, a compreensão e quantificação dos processos hidrológicos como a retenção de água no solo nesta bacia hidrográfica, desenvolve um papel importante no instante de serem aplicados modelos hidrológicos que possam suportar as decisões de manejo dos recursos hídricos e contribuir para a caracterização do efeito do uso do solo nas vazões da bacia. Por esse motivo, a geração de FPTs específicas do clima subtropical para a estimativa da água no solo, tem a expectativa de contribuir fortemente ao melhor uso dos modelos hidrológicos na BHAP sob condições de clima subtropical, permitindo entender com mais abrangência os modelos.

A importância econômica e social, e a carência de estudos na bacia supracitada merecem atenção de pesquisadores, em especial para estudos relacionados a um melhor entendimento do meio físico. Frente ao exposto, esta pesquisa visa a exploração de técnicas que permitam a avaliação e geração de FPTs no contexto dos diversos climas, no intuito de identificar modelos que permitam estimar a retenção de água no solo na bacia hidrográfica Arroio Pelotas.

1.1 Organização do texto

A presente tese será apresentada no nível de descrição em artigos de acordo com o manual de normas UFPel para trabalhos acadêmicos. Em primeiro lugar será apresentada a seção de revisão bibliográfica que visa fazer um levantamento do estado-da-arte em relação às funções de pedotransferência (FPT) e o seu potencial para estimativa das propriedades físico-hidráulicas do solo, abordando-se os conceitos fundamentais, metodologias utilizadas na atualidade e resultados relevantes na área científica. Além disso, esta seção aborda os conceitos básicos das técnicas de aprendizado de máquinas (AM), redes neurais artificiais (RNAs) e métodos de avaliação que serão utilizados no desenvolvimento dos artigos produto desta tese.

Em seguida, serão anexados os artigos 1 e 2 estabelecidos para o desenvolvimento da tese, sendo o primeiro artigo referente à aplicabilidade e transferibilidade de FPTs para solos de clima subtropical em escala de bacia hidrográfica na estimativa da retenção de água no solo e o segundo artigo sobre a geração de uma FPT pseudo-contínua baseada em RNAs e aplicação de técnicas de AM para estimativa da retenção de água no solo em escala de bacia hidrográfica sob clima subtropical. Seguindo as instruções do manual de normas, os artigos serão apresentados de acordo às normas das revistas, sendo o artigo 1 nas normas da revista *Soil & Tillage Research* e o artigo 2 nas normas da revista *Computers and Electronics in Agriculture*.

2 Hipóteses

- As funções de pedotransferência desenvolvidas para as condições de clima temperado e tropical estimam satisfatoriamente a retenção de água no solo na bacia Arroio Pelotas.
- Funções de pedotransferência desenvolvidas a partir de dados de solo e de clima subtropical mediante a técnica de redes neurais apresentam maior exatidão e precisão em relação às funções da literatura para a estimativa da retenção de água no solo na bacia Arroio Pelotas.

- A técnica de *random search* aplicada às redes neurais artificiais permite a seleção de atributos topográficos, atributos do solo e hiper-parâmetros que têm importância na estimativa da retenção de água no solo na bacia Arroio Pelotas.

3 Revisão Bibliográfica

3.1 Importância das funções e propriedades hidráulicas do solo

As propriedades hidráulicas são frequentemente utilizadas e têm múltiplas aplicações em diversos campos como a hidrologia, agricultura, ciência do solo, geologia, geografia, botânica, engenharia civil e engenharia ambiental, entre outras. Assim, o conhecimento das propriedades hidráulicas torna-se essencial também nas áreas de manejo e uso da água e do solo, devido a que os sistemas de irrigação e drenagem agrícola são dependentes destas, precisando-se do uso racional e econômico da água, visando uma adequada gestão e manejo desse recurso, relacionado com as alterações temporais das propriedades hidráulicas do solo (BOTULA *et al.*, 2012; PACHEPSKY; RAJKAI; TÓTH, 2015; PACHEPSKY; RAWLS; LIN, 2006).

Além disso, a quantificação das propriedades hidráulicas têm um papel importante na produção de culturas, nas necessidades hídricas das plantas, nos efeitos da erosão hídrica, nos fenômenos relacionados com o movimento e retenção de água no solo, infiltração, evapotranspiração, transporte de gases e solutos (DA SILVA *et al.*, 2017; SOARES *et al.*, 2014), que futuramente servirão como informações para a alimentação dos modelos ambientais e hidrológicos, com o intuito do melhor planejamento e gestão dos recursos hídricos e a dinâmica da água no sistema solo-planta-atmosfera.

A determinação das funções hidráulicas do solo, particularmente a condutividade hidráulica, a curva de retenção de água e a difusividade de água no solo, permitem entender a transferência de energia na interface solo-planta-atmosfera e melhorar a eficiência dos modelos hidrológicos (VAN LOOY *et al.*, 2017). Além disso, elas são amplamente utilizadas em vários modelos referentes à simulação de processos no solo e aos relacionados com o movimento de água e solutos entre a superfície do solo e o lençol freático (ZHANG; SCHAAP, 2017). Do mesmo modo, essas funções tornam-se

essenciais para alimentar a carência de informações existentes nos modelos de gestão ambiental (MINASNY; HARTEMINK, 2011).

3.1.1 Curva de retenção de água no solo

Dentro das funções hidráulicas a curva de retenção de água no solo se destaca como uma das mais importantes, uma vez que o conhecimento adequado desta propriedade pode auxiliar os produtores agrícolas no que diz respeito ao manejo da irrigação (CASTELLINI; IOVINO, 2019; SOARES *et al.*, 2014). Nas condições de saturação do solo e no estado de equilíbrio, a água livre encontra-se unicamente sob pressão atmosférica, e uma vez o solo no ponto de saturação é submetido a uma força de sucção, parte da água será retirada do sistema poroso do solo permitindo a entrada do ar.

Conforme as forças de sucção são aumentadas, o sistema poroso começa a ser preenchido em maior quantidade pelo ar, possibilitando que poros de menor diâmetro percam cada vez mais água. Aumentando-se gradativamente as forças de sucção, o processo continuará repetindo-se, contribuindo com a redução da água entre os agregados (macroporos), seguido da retirada da água entre os intra-agregados e partículas do solo (microporos), até finalmente chegar na redução da espessura do filme de água que rodeia as partículas do solo. Continuando o procedimento de sucção, a água ficará mais fortemente retida nas partículas precisando que forças maiores de sucção sejam aplicadas para a retirada da água presa na matriz do solo (LIBARDI, 2012; SOARES, 2013).

Deste modo, a curva de retenção de água no solo (CRAS) é a relação que existe entre o teor de água e a energia na qual está retida a água na matriz do solo, podendo-se gerar um gráfico entre a umidade do solo (θ ou w) e as forças de sucção ou potenciais matriciais da água (h ou Ψ_m), conhecido como curva de retenção de água no solo (URACH, 2007). A retenção de água na matriz do solo é regida pelas forças capilares e de adsorção, comumente denominadas de forças matriciais (potencial matricial) (LIBARDI, 2012; MCNEILL *et al.*, 2018).

A CRAS apresenta como principal vantagem, a manifestação do aspecto dinâmico da água no solo e é específica de cada tipo de solo, horizonte ou camada do perfil do solo. Por outro lado, a CRAS apresenta relação dependente com a distribuição de partículas e a estrutura do solo, e manifesta uma relação inversa entre o teor de água no solo e o seu potencial matricial, permitindo identificar que o solo se encontra na saturação quando sucções pequenas são suficientes para extrair água da matriz do solo e vice-versa (LIBARDI, 2012).

A CRAS é útil para determinar o teor de água que pode ser retida em certos potenciais matriciais ou dentro de determinados limites, como por exemplo, a estimativa da água disponível para as plantas. O conhecimento de toda a curva é fundamental, porque permite estabelecer a armazenagem de água no solo e a sua relação com as necessidades hídricas de determinado cultivo, já que as respostas das plantas vão além do teor de água necessário para a planta, sendo mais relacionadas às respostas com o potencial energético da água no solo (SOARES, 2013).

Outro aspecto da CRAS é que é possível estabelecer a distribuição de poros por tamanho, já que a cada potencial matricial aplicado é possível conhecer o tamanho dos poros esvaziados no solo (OTTONI, 2017). Em condições de alta umidade no solo predominam os fenômenos de capilaridade, sendo a curva dependente nessa condição do arranjo e dimensões dos poros, refletidos na densidade e a porosidade do solo, os quais são fortemente influenciados pela estrutura e compactação do solo. Enquanto em condições de baixa umidade no solo, predominam os fenômenos de adsorção, influenciados pela granulometria e a superfície específica das partículas (OTTONI, 2017). Outros fatores que afetam a CRAS encontram-se, a granulometria do solo, o tipo de mineral, a profundidade do solo e a matéria orgânica (BAYAT *et al.*, 2013; BRUAND, 2004; FAMIGLIETTI; RUDNICKI; RODELL, 1998; LI *et al.*, 2018; REYNOLDS *et al.*, 2009; VAN LOOY *et al.*, 2017).

A dinâmica da água no solo manifestada pela CRAS pode ser afetada pela compactação do solo, já que o aumento da compactação ocasiona a redução da porosidade total do solo, por conta do decréscimo dos poros maiores intra-agregados para poros de tamanhos médios, provocando que a quantidade de água na saturação seja reduzida e também a água retida em baixos potenciais. Contudo, o formato da curva

na faixa de altas sucções não é afetado pela compactação, assim podendo a curva ser idêntica ao dos solos não compactados nestas altas suções (ASCHONITIS; KOSTOPOULOU; ANTONOPOULOS, 2012; URACH, 2007; VAN LOOY *et al.*, 2017).

A granulometria do solo como característica do solo pode afetar a CRAS, devido a que um maior teor de argila no solo tem a capacidade de reter maior umidade sob um determinado potencial matricial. Os solos argilosos têm uma maior capacidade de adsorção de água, devido a maior área superficial específica das partículas de argila em relação às demais partículas do solo, o que lhes confere maior potencial de adsorção de água. Consequentemente, a inclinação da curva nos solos argilosos será mais suave, atribuída à distribuição mais uniforme do tamanho de poros na matriz do solo (REYNOLDS *et al.*, 2009; VEREECKEN *et al.*, 1989).

Pelo contrário, em solos com maior teor de areia, a distribuição de poros é menos uniforme e os poros são maiores, possibilitando que a água seja retirada mais rapidamente em baixas tensões, sobrando assim menor água a ser retida nos potenciais menores, por isso, a curva em solos arenosos têm uma inclinação mais acentuada (BAYAT *et al.*, 2013; REYNOLDS *et al.*, 2009).

Em casos menos prováveis, o comportamento da CRAS é afetado pelo tipo de mineral presente no solo, o qual pode influenciar o formato da curva. O caso mais conhecido onde pode ocorrer esta variação na umidade do solo é nos Vertissolos, nos quais predominam muitos minerais expansivos ou minerais 2:1 que atribuem a propriedade da contração e expansão nas argilas, permitindo também a concentração da solução do solo em torno das mesmas (BRUAND, 2004; URACH, 2007).

Em relação à profundidade vários estudos têm relatado o comportamento da CRAS em relação às diferentes profundidades do solo, observando que a retenção de água em potenciais matriciais maiores é dependente da variação da composição granulométrica das diferentes camadas, porém quando o solo é analisado para potenciais menores, observa-se que a retenção de água não é dependente da granulometria, e o seu comportamento é contrário à variação da composição granulométrica (FAMIGLIETTI; RUDNICKI; RODELL, 1998; LI *et al.*, 2018).

A matéria orgânica pode apresentar diversas relações com a CRAS porque esta pode sofrer alterações frequentes, seja pelas condições naturais ou pelos manejos

diferenciados do solo (RAWLS *et al.*, 2003; URACH, 2007), pelo qual vários estudos têm mostrado tanto efeitos positivos como efeitos limitados da matéria orgânica sobre a retenção da água. De forma geral, a matéria orgânica por possuir uma alta capacidade adsortiva e a influência que a mesma tem sobre a estrutura do solo, essa à sua vez tem efeito sobre a retenção de água no solo, porém variável de acordo à granulometria do solo, ao tipo de material orgânico, ao grau de decomposição do material orgânico (grau de estabilidade), e ao tamanho das partículas orgânicas (MINASNY; MCBRATNEY, 2018; SOARES, 2013). Exemplificando esses diversos efeitos, Minasny e McBratney (2018) indicaram que aumentos do carbono orgânico no solo ocasionaram uma mudança em toda a forma da CRAS ou em todos os potenciais matriciais.

A medição da CRAS completa no laboratório ou a medição da CRAS parcial no campo pode ser custosa, difícil e demorada devido à forte não linearidade da função de retenção de água e parcialmente porque esta função hidráulica está sujeita a uma variabilidade espacial substancial, sendo requerido o uso de funções de pedotransferência (FPTs) quando se precisa a determinação em grande escala (ZHANG; SCHAAP, 2017). Uma vez obtida a CRAS para cada tipo de solo, ela permite conhecer os limites onde a água pode diminuir sem afetar o desenvolvimento das plantas e, efetivamente a sua produtividade, além de fornecer outros elementos necessários para os projetos de irrigação racionais.

3.2 Aspectos gerais das funções de pedotransferência

A definição de funções de pedotransferência (FPTs) foi estabelecida por Bouma (1989) como “*traduzindo dados que temos para o que precisamos*”, em outras palavras, são aquelas funções preditivas de certos atributos do solo mais dificilmente de serem medidas a partir de outros atributos que são facilmente mensuráveis. Com o passar do tempo, surgiu um conceito mais atualizado das FPTs definindo-se como a estimativa indireta das propriedades hidráulicas complexas do solo mediante o uso de modelos que utilizam atributos de fácil obtenção e acesso em rotina de laboratório, levantamentos de solo e bancos de dados, tais como, textura do solo (areia, silte e argila), densidade,

conteúdo de matéria orgânica do solo (ou carbono orgânico) (MINASNY; HARTEMINK, 2011; PINEDA *et al.*, 2018).

As mensurações dessas propriedades hidráulicas no laboratório ou no campo tornam-se difíceis, devido ao tempo necessário para realizar as análises, aos equipamentos caros e às modificações intrínsecas que podem ocorrer nas amostras durante as determinações (SOARES *et al.*, 2014). Diante das dificuldades encontradas, surgem as FPTs como uma solução na determinação dessas propriedades de difícil mensuração, podendo ser estimadas mediante diferentes técnicas estatísticas, algoritmos, métodos de lógica e inteligência artificial. As FPTs podem ser normalmente expressas na forma de equações matemáticas ou modelos que contêm propriedades e características básicas do solo como parâmetros de entrada correlacionados entre si, as quais são relevantes na prática agrícola e no manejo dos recursos naturais (PEREIRA, 2017; PEREIRA *et al.*, 2018; SOARES *et al.*, 2014). Contudo, deve-se sempre seguir o principal princípio das FPTs destacado por Minasny e McBratney (2002), que estabelece como princípio básico das FPTs a eficiência, o qual baseia-se em não estimar uma variável que seja fácil ou de baixo custo de ser determinada, com o fim de evitar o mau uso do conceito.

Apesar da maioria de FPTs terem sido desenvolvidas para estimar funções e propriedades hidráulicas, elas não só são restringidas às propriedades hidráulicas, mas também, para a estimativa de atributos físicos, mecânicos, químicos e biológicos (MINASNY; HARTEMINK, 2011). McBratney *et al.* (2002) e Minasny e Hartemink (2011) relataram alguns exemplos de algumas funções e propriedades estimadas a partir das FPTs, tais como, taxa de infiltração de água no solo, condutividade térmica do solo, repelência à água, resistência mecânica do solo, fator de erodibilidade, taxa da carga estrutural, índice de plasticidade, capacidade de intercâmbio catiônico, sorção de fósforo, cádmio e arsênico, saturação de alumínio, conteúdo de hematita, entre outros.

Assim as propriedades e funções estimadas pelas FPTs posteriormente podem ser utilizadas para:

- avaliação do movimento da água e lixiviação de poluentes;

- estimativa da lixiviação de poluentes para lençóis freáticos em uma bacia hidrográfica;
- desenvolvimento de mapas regionais de vulnerabilidade às águas subterrâneas que indiquem o impacto da lixiviação;
- avaliação da magnitude de acumulação de cádmio na escala regional;
- estimativa de produção de grãos de soja e produção regional de culturas;
- avaliação da produção de cultura e lixiviação de nitrato como resultado de diferentes práticas de manejo;
- o maior entendimento dos processos físicos que ocorrem no solo;
- melhora na estimativa das funções e propriedades do solo a partir de atributos que não são usualmente disponíveis, como por exemplo, características mineralógicas, tipo de argila, conectividade de poros, entre outros.

Os respectivos modelos têm sido desenvolvidos para melhorar o entendimento de importantes processos do solo e também para atuar como ferramentas nas avaliações agrícolas, problemas ambientais, e conseqüentemente os modelos de simulação são utilizados em pesquisas e sistemas de gestão (MCBRATNEY *et al.*, 2002).

Um dos princípios fundamentais das FPTs é a sua facilidade de uso e a parcimônia em termos de variáveis de entrada, podendo ser implementadas em vários modelos bem como em estruturas de softwares para simular o comportamento de sistemas agrícolas complexos. Embora várias FPTs tenham sido propostas na literatura, relativamente há pouca informação disponível acerca do seu desempenho e comparações entre os dados estimados e medidos, os quais são de crucial importância para determinar sua aplicabilidade, por exemplo, a confiabilidade e precisão nos dados estimados e medidos de retenção de água em função de considerado intervalo de potencial matricial para uma

ampla faixa de texturas do solo (CASTELLINI; IOVINO, 2019). As FPTs disponíveis na literatura não são sempre aplicáveis com exatidão e precisão aceitáveis para diferentes regiões, por tanto tentativas têm sido feitas para recalibrar as FPTs publicadas pela adaptação dos parâmetros para às condições do solo diferentes daquelas utilizados no seu desenvolvimento (PICCIAFUOCO *et al.*, 2019).

O solo é um sistema com certa complexidade, devido às interações entre os processos químicos, físicos e biológicos que nele acontecem, dificultando assim a modelagem desses processos (MULAZZANI, 2016). O uso das FPTs tem permitido o entendimento físico dos processos que ocorrem no solo, através da busca de relações existentes entre as variáveis de entrada e estimativas. Frente a isso, as FPTs surgem como uma alternativa para o entendimento desses processos, permitindo mediante vários métodos explorar novas e diversas relações entre as propriedades do solo. Contudo, a exatidão da estimativa das FPTs não é somente pelas boas relações que existem entre as variáveis de entrada e as propriedades hidráulicas a ser estimadas, mas também é afetada pela qualidade da base de dados utilizadas nos processos de treinamento da FPT e do teste (OTTONI, 2017).

Entre um dos princípios das FPTs encontra-se a eficiência, o qual estabelece que não se deve estimar algo que é mais fácil e mais barato medir ou determinar, ou seja, o custo e esforço gerado para obter o valor da variável preditora deve ser menor ao da variável a ser estimada. Não obstante, existem estimativas de variáveis que podem ser melhoradas quando se tem acesso a dados de variáveis preditoras mais caras. Um exemplo típico é a estimativa da densidade do solo mediante uma FPT, que mesmo sendo utilizada para estimar outros atributos do solo mais complexos, pode ser considerada uma FPT que cumpre o princípio da eficiência, uma vez que a densidade do solo raramente é encontrada nos bancos de dados de solos que são oriundos dos levantamentos para classificação de solos. Embora se utilizem dados de argila, areia, silte e matéria orgânica para a estimativa da densidade do solo, essa FPT continua cumprindo o princípio da eficiência porque usa dados facilmente disponíveis nos relatórios de levantamentos para estimar uma variável que não foi determinada (BUDIMAN *et al.*, 2003).

3.3 Funções de pedotransferência para estimativa da CRAS

A dificuldade que envolve a medição das funções ou propriedades hidráulicas e a falta da disponibilidade de grandes e diversas bases de dados é a principal razão da necessidade de pesquisas que desenvolvam FPTs para a estimativa da CRAS (MCBRATNEY *et al.*, 2002). Além disso, na determinação CRAS requerem-se esforços, tempo e custos de medição para obter informação em estudos de grande escala (BOTULA; RANST; CORNELIS, 2014), porém a determinação da CRAS mediante as FPTs não é objetivo principal em si, mas visa fornecer dados de entrada essenciais para serem utilizados em modelos de simulação em estudos agrícolas, hidrológicos e ambientais (MINASNY; HARTEMINK, 2011; NGUYEN *et al.*, 2015). Consequentemente, a estimativa imprecisa dessa função hidráulica do solo pode influenciar a qualidade geral dos resultados de todo o processo de modelagem (BOTULA *et al.*, 2012).

As FPTs têm sido amplamente utilizadas para estimar a CRAS, destacando-se três tipos de métodos de estimativa: métodos de estimativa pontual, métodos de estimativa de parâmetros e métodos semi-físicos (DOBARCO *et al.*, 2019). FPTs pontuais estimam a retenção de água em um determinado potencial de água no solo (BOTULA; RANST; CORNELIS, 2014; RAWLS; BRAKENSIEK; SAXTON, 1982; VEREECKEN *et al.*, 2010). Os métodos de estimativa de parâmetros estabelecem relações empíricas entre os atributos do solo e os parâmetros da CRAS, descrevendo a evolução contínua de água em diferentes potenciais do solo com equações matemáticas (BOTULA; RANST; CORNELIS, 2014; VAN GENUCHTEN, 1980; VEREECKEN *et al.*, 2010). As abordagens semi-físicas baseiam-se em suposições mecanicistas sobre as relações entre a retenção da água, estrutura do solo e distribuição do tamanho de poros (ARYA; PARIS, 1981; BOTULA; RANST; CORNELIS, 2014).

Outra categorização das FPTs para estimar a CRAS é a classificação em FPTs de classes e FPTs contínuas. Na primeira classificação, se dá a atribuição de valores dos parâmetros da CRAS ou da retenção de água no solo a classes definidas pelas variáveis de entrada (por exemplo, classes de textura do solo, intervalos de conteúdo de matéria orgânica ou faixas de densidade do solo) e a segunda classificação usa atributos físicos ou químicos do solo como variáveis de entradas em modelos de regressão multivariada

ou aprendizado de máquina para estimar os parâmetros do modelo da CRAS ou a água do solo em um específico potencial da água (DOBARCO *et al.*, 2019).

Adicionalmente, outro tipo de FPTs foi introduzido por Haghverdi, Cornelis e Ghahraman (2012) referido à função de pedotransferência (FPT) pseudo-contínua, a qual é capaz de determinar CRAS quase contínuas sem usar nenhuma equação hidráulica do solo, ou seja, esse tipo de FPT permite prever o teor de água em qualquer potencial matricial desejável, sem a necessidade de usar uma equação específica como a de van Genuchten.

Como solução ao empirismo inerente de alguns métodos estatísticos utilizados para desenvolver FPTs, surge o desenvolvimento de FPTs baseadas em considerações físicas, as quais procuram utilizar equações físicas que ligam a granulometria do solo às propriedades físico-hidráulicas. A título de exemplo, Silva *et al.* (2017) mencionam que o Splintex PTF é uma modificação do modelo físico-empírico Arya and Paris (1981), o qual usa dados limitados de textura que são ajustados à função *spline* cúbica e fornece suas saídas na forma dos parâmetros de van Genuchten (1980). Por outro lado, o Splintex PTF permite incluir como parâmetro de entrada a umidade na saturação e um ponto de retenção de água medido em uma determinada pressão arbitrária. Os autores ressaltam que a inclusão de um ou mais pontos de retenção medidos pode melhorar significativamente a estimativa de parâmetros hidráulicos, permitindo que a FPT se ajuste às variações específicas do solo que podem ser muito difíceis de identificar ou quantificar.

3.4 Disponibilidade de bases de dados de solo

No ano de 2001 foi gerado pelos Estados Unidos o maior banco de dados de solos com informações de propriedades físico-hidráulicas, denominado de UNSODA, (NEMES *et al.*, 2001), porém todos os solos utilizados na geração do banco pertenciam ao ambiente temperado. Esse banco possuía inicialmente 902 dados referentes à retenção de água de solo e 4174 dados de granulometria. Depois na Europa foi desenvolvido o banco hidropedológico EU-HYDI (TÓTH *et al.*, 2015), também de solos de clima temperado, com 18637 amostras de solos extraídas de 6460 perfis de solo distribuídos pelo continente europeu.

No Brasil durante a época de 2014 surgiu a necessidade do desenvolvimento de um banco hidrofísico que representasse os solos brasileiros com predominância de solos de clima tropical. Esse banco hidrofísico foi denominado como HYBRAS, consolidando dados de retenção de água e de condutividade hidráulica saturada em conjunto com os dados básicos de solo, totalizando 1075 dados. Todos os dados foram extraídos de publicações inventariadas e 60 trabalhos científicos desenvolvidos no Brasil. Apesar de ser uma base com dados de solos de clima tropical, esta apresenta limitações quanto à distribuição nos tipos de solos, nos locais de amostragem, biomas e climas, concentrando-se uma maior densidade de dados nos estados de Rio Grande do Sul, São Paulo e Rio de Janeiro, devido a que nesses estados se concentram importantes centros de pesquisas pedológicas no País (OTTONI, 2017). Frente a isso, existe a necessidade de aumento da quantidade e diversidade de dados para obter uma base de dados mais representativa dos solos brasileiros, o qual precisa da ajuda e compromisso dos pesquisadores da área de solo que disponibilizem os dados obtidos nos trabalhos científicos para melhorar a qualidade do HYBRAS.

3.5 Aspectos gerais dos métodos mais utilizados no desenvolvimento das funções de pedotransferência

Uma extensa revisão realizada por Vereecken et al. (2010) e Botula, Ranst e Cornelis (2014) resumindo sobre os diversos métodos de abordagem empírica que podem ser utilizados na geração das FPTs, distinguiram duas categorias, as técnicas de regressão estatística (modelos lineares e não lineares) e técnicas de mineração e exploração de dados (redes neurais artificiais e métodos de grupos de tratamento de dados). Vereecken et al. (2010) afirmaram que de forma geral, os métodos baseados em redes neurais artificiais levaram a um melhor desempenho das FPTs em termos de indicadores básicos, como a raiz quadrada do erro médio (RMSE), devido ao fato que esta abordagem não requer uma forma funcional a priori com a qual se possa relacionar a FPT de entrada com a FPT de saída.

O desenvolvimento das FPTs para muitas aplicações na modelagem do solo e da água tem sido bem documentado, utilizando-se recentemente novos métodos (BAYAT; MAZAHARI; MOHANTY, 2019; GUNARATHNA *et al.*, 2019; RASTGOU *et al.*, 2020), sendo a maioria dos modelos existentes desenvolvidos com base em alguns dos seguintes enfoques: pesquisa em tabelas simples por classes texturais específicas (WÖSTEN; FINKE; JANSEN, 1995), modelos baseados em regressão linear e não linear (RAJKAI; KABOS; VAN GENUCHTEN, 2004), redes neurais artificiais (RNA) (BAYAT; MAZAHARI; MOHANTY, 2019), máquinas de vetores de suporte (NGUYEN *et al.*, 2017), métodos de grupo de manipulação de dados (BOTULA; RANST; CORNELIS, 2014) técnicas não paramétricas de vizinho mais próximo (NEMES *et al.*, 2006), redes neurais artificiais multiescala Bayesiana (JANA; MOHANTY; SHENG, 2012).

Dentro dos métodos de geração das FPTs os modelos de regressão múltiplas lineares e as RNAs são os mais utilizados na literatura. Os modelos de regressão baseiam-se na correlação entre a variável objetivo e as covariáveis, destacando-se o uso das regressões múltiplas lineares devido ao seu elevado potencial na estimativa das propriedades hidráulicas do solo, atribuído a que é uma ferramenta estatística comum para a estimativa de uma variável dependente a partir de um número de covariáveis independentes (HERBST; DIEKKRÜGER; VEREECKEN, 2006). Por outro lado, as RNAs vêm sendo aplicadas em inúmeros estudos com sucesso, como uma técnica de processamento de dados adequada para modelar as relações entre os atributos do solo e posteriormente desenvolver FPTs (ZHAO *et al.*, 2016). As RNAs são uma tentativa de desenvolver um modelo que funciona de maneira semelhante ao cérebro humano, compreendendo uma rede densa de conexões entre os dados de entrada, os neurônios dispostos nas camadas ocultas e os dados de saída (MOREIRA DE MELO; PEDROLLO, 2015).

Exemplo do uso das regressões múltiplas lineares encontram-se no trabalho de Da Silva *et al.* (2015) que desenvolveram FPT para solos das planícies costeiras brasileiras, gerando FPTs específicas para cada classe de solo e nos respectivos horizontes A e B e no *solum*, obtendo-se ajustes com adequada precisão para estimar a retenção de água nos potenciais de -33 kPa e -1500 kPa, partindo de dados de textura e

matéria orgânica como parâmetros de entrada e resultando em coeficientes de determinação entre 0,41 e 0,75.

Soares et al. (2014) utilizaram RNAs para estimar a CRAS nos solos de Rio Grande do Sul a partir de um banco de dados com informações disponíveis na literatura, observando que a arquitetura com melhor desempenho foi a 4-27-7 (camada de entrada, oculta e de saída), obtendo-se o índice de desempenho “ótimo” e concluindo que as RNAs são uma ferramenta com alta capacidade preditiva da CRAS resultando em coeficientes de determinação entre 0,85 e 0.99.

3.6 Avaliação das funções de pedotransferência

O dado de entrada é um dos mais importantes fatores que influencia a eficiência de predição das FPTs (BAYAT; MAZAHARI; MOHANTY, 2019). Recentemente os pesquisadores têm avaliado a aptidão de diferentes variáveis para estimar a CRAS encontrando diferentes resultados (BAYAT *et al.*, 2013). Da Silva et al. (2017) reportaram que há a necessidade de utilizar variáveis que melhorem a estimativa dos parâmetros da CRAS, recomendando-se procurar novas informações e facilmente disponíveis que possam auxiliar na estimativa da CRAS, já que é um desafio duradouro e sem respostas adequadas ainda.

De acordo com Schaap, Nemes e van Genuchten (2004) a resposta de uma FPT pode ser unicamente avaliada utilizando dados que não foram inseridos na calibração (dados independentes) da FPT em questão. As equações preditivas como as FPTs são rotineiramente avaliadas em termos de correspondência entre valores estimados e medidos. Quando os valores medidos são utilizados para o desenvolvimento da equação, então a exatidão e precisão da estimativa da equação é avaliada. No caso de testar uma FPT já desenvolvida a partir de um conjunto de dados de outra região, a confiabilidade do modelo está sendo avaliada, examinando-se a correspondência entre dados medidos e estimados.

Existe uma grande quantidade de estatísticas disponíveis para avaliar a exatidão e precisão, e unicamente um número limitado de técnicas são utilizadas nos projetos de modelagem devido a restrições de tempo e recursos. Entre os índices estatísticos mais

comumente usados para avaliação das FPTs destacam-se, o erro médio (ME), o erro médio absoluto (MAE), o coeficiente de determinação (r^2) e a raiz do erro quadrático médio (RMSE) (DONATELLI; WÖSTEN; BELOCCHI, 2004).

A exatidão e precisão das estimativas pode servir como referência, no entanto, deve ser comparada à variabilidade nos dados de entrada medidos, cuidando-se que em geral os modelos não devem ser mais precisos que os dados usados no desenvolvimento do modelo. Por tanto, as estimativas realizadas pelas FPTs podem ser consideradas suficientemente precisas, se a variabilidade dos erros das FPTs não difere significativamente da variabilidade em outros dados, e se o erro médio não difere significativamente de zero (DONATELLI; WÖSTEN; BELOCCHI, 2004).

3.7 Aplicabilidade e transferibilidade das funções de pedotransferência

O desenvolvimento de novas FPTs é uma tarefa muito árdua, além de recomendável sempre que possível (MINASNY *et al.*, 2008; NGUYEN *et al.*, 2015). Apesar de se tenham realizado diversos estudos universalmente, nos quais tem-se tentado aplicar modelos para estimar a retenção de água no solo, verificou-se que as FPTs possuem maior desempenho na área onde estas foram desenvolvidas, no lugar de serem usadas em regiões de diferentes condições de formação do solo, por essa razão é necessário ter cuidado na aplicação e transferência das FPTs fora da área de desenvolvimento, sendo necessária avaliar-se as suas incertezas de estimativa (PATIL; SINGH, 2016; PINEDA *et al.*, 2018).

Muitas FPTs têm aplicabilidade limitada porque foram desenvolvidas para solos específicos dentro de um domínio geomórfico e geográfico limitado, além de serem projetadas para uma aplicação à escala regional. Assim, FPTs de regiões específicas podem ser precisas unicamente para os dados originais utilizados no treinamento, mas não são confiáveis para solos em outras condições. Cornelis *et al.* (2001) e McBratney *et al.* (2002) alertaram que a extrapolação das FPTs além dos limites estatísticos do conjunto de dados de calibração e das localizações geográficas dos solos a partir das quais foram desenvolvidas devem ser evitadas, ou avaliadas cuidadosamente por sua qualidade estimativa.

McBratney et al. (2002) descrevem que as FPTs possuem outro princípio básico, a incerteza que se refere a não utilizar uma FPT sem ser previamente avaliada a incerteza associada, escolhendo-se aquela que estima o atributo de interesse com menor variância associada às estimativas da variável. Santos et al. (2013) e Medrado e Lima (2014) ressaltam o cuidado que deve ser tomado no momento de extrapolar a estimativa de uma FPT além da região e classe de solo para a qual foi calibrada, recomendando-se a aplicação das FPT em bancos de dados de solos semelhantes, garantindo maior desempenho das estimativas.

Dobarco et al. (2019) desenvolveram um estudo objetivando a geração de FPTs de classes e contínuas para a estimativa da retenção de água no solo na capacidade de campo e no ponto de murcha permanente com as suas incertezas associadas para a avaliação da capacidade de água disponível e discutir sua aplicabilidade em grandes territórios utilizando a zona metropolitana da França, como exemplo. Os autores mencionam que a aplicabilidade das FPTs propostas, deve ser avaliada devido à alta variabilidade nos fatores de formação do solo na escala nacional. Assim, a avaliação da transferibilidade e aplicabilidade das FPTs a partir de dados de atributos diferentes dos originais pode-se obter resultados imprecisos.

Outro estudo proposto para avaliação da aplicabilidade de FPTs calibradas a novas amostras de solos foi realizado por Tranter, McBratney e Minasny (2009) que aplicaram a métrica de distância *Mahalanobis* para determinar o domínio da aplicabilidade de uma FPT. Esse método utiliza uma distância limite ao centróide do conjunto de dados de calibração para definir se uma amostra é considerada similar ao conjunto de dados de calibração e, portanto, pertence ao domínio de aplicabilidade da FPT ou se é diferente, nesse caso é descartada a aplicação da FPT para essa amostra.

Nemes, Schaap e Wösten (2003) utilizaram três bases de dados para desenvolver FPTs a diferentes escalas usando RNA para estimar a retenção de água no solo, a primeira base continha dados de solo da Hungria (escala nacional), a segunda base com dados do HYPRES (escala continental) e a terceira base com dados de USA e Europa (escala intercontinental). Posteriormente a exatidão das FPTs foram testadas utilizando uma base de dados independentes da Hungria mediante o índice estatístico RMSE, encontrando que as FPTs da escala continental e intercontinental proporcionaram erros

muito maiores em relação às FPTs desenvolvidas na escala nacional (solos da Hungria), sugerindo-se que tendo um pequeno conjunto de dados relevantes, quando disponíveis, é melhor que usar um conjunto de dados grande, mas muito geral.

Apenas um número limitado de estudos tem buscado a estimativa da retenção de água com dados de solos de clima tropical, como encontrado até o momento na literatura, sendo mais restrito para solos de clima subtropical (MICHELON *et al.*, 2010; REICHERT *et al.*, 2009; SANTOS *et al.*, 2013; SOARES *et al.*, 2014; ZOU *et al.*, 2016), limitando a transferibilidade e aplicabilidade das FPTs, afetando assim a exatidão das suas estimativas nessas condições. Frente a isso, existe a necessidade de ampliar as pesquisas relacionadas tanto à geração como à avaliação de FPTs em zonas de clima subtropical.

Além disso, Nemes *et al.* (2006) observaram que a maioria dos estudos de avaliação do desempenho das FPTs utilizando conjunto de dados independentes permanecem incertos no que se refere às principais fonte de erros nas estimativas. Devido a que as diferenças entre os conjuntos de dados utilizados para gerar as FPTs, as diferenças nos algoritmos de desenvolvimento das FPTs ou as diferenças entre as variáveis de entrada utilizadas, provavelmente podem resultar no erro geral da estimativa (NGUYEN *et al.*, 2015). Alguns exemplos dos erros em termos de RMSE que ocorrem durante a aplicabilidade das FPTs da literatura em regiões diferentes para as quais foram desenvolvidas são apresentados na Tabela 1.

Tabela 1. Análise da exatidão de FPTs da literatura em outras regiões.

FPTs	Clima - Origem	Clima - Destino	RMSE	
			θ -33 kPa	θ -1500 kPa
Gupta e Larson (1970) - Pontual	Temperado - USA	Tropical – Deltas no Vietnã	0,113	0,118
Vereecken et al. (1989) - Paramétrica	Temperado - Bélgica	Tropical - Congo	0,1168	0,1016
Rosetta – H3 Schaap et al. (2001)	Temperado – Norte America e Europa	Tropical – Brasil/HYBRAS	0,080	0,100
Aina and Periaswamy (1985)	Tropical - Nigéria	Tropical - IGBP/T	0,088	0,084
Minasny and Hartemink (2011)	Tropical - ISRIC	Tropical - Indonésia	0,130	0,100

Na Tabela 1 de forma geral é observado que as FPTs desenvolvidas em clima temperado, quando aplicadas em regiões de clima tropical, mostraram RMSE elevados para a estimativa da retenção de água nos potenciais -33 kPa e -1500 kPa, demonstrando as incertezas que acontecem durante a transferibilidade e aplicabilidade das FPTs. Por outro lado, as FPTs desenvolvidas em clima tropical e aplicadas em regiões similares de clima tropical, também tiveram a tendência de apresentar RMSE elevados, possivelmente atribuído às diferenças que existem entre os tipos de solos, conjuntos de dados utilizados para gerar as FPTs, às diferenças nos algoritmos de desenvolvimento das FPTs ou aos diferentes atributos utilizados como variáveis de entrada que podem resultar no aumento do erro geral de estimativa como afirmado por (NGUYEN *et al.*, 2015).

Conhece-se que infelizmente existem muitas FPTs na literatura e a escolha da melhor se torna laboriosa. Além disso, há muito progresso no desenvolvimento de FPTs e na identificação apropriada de preditores mais eficientes na estimativa de algumas propriedades hidráulicas, contudo, existem ainda variabilidades nas FPTs não resolvidas ou inexplicadas, as quais desempenham um papel importante quando se estão analisando as suas influências sobre os modelos numéricos ou hidrológicos. Exemplo disso é observado no estudo de Weihermüller *et al.* (2021), os quais analisaram o impacto de diferentes FPTs nos resultados das simulações de fluxo de água realizadas no HYDRUS-1D. Com base nisso, eles identificaram que as FPTs Rosetta, Wösten, e Tóth são as mais robustas para a estimativa das funções de propriedades hidráulicas de van Genuchten Mualem e a FPT de Cosby para as funções de Brooks Corey. Os mesmos autores encontraram que a escolha da FPT impacta os resultados da simulação, causando variabilidade substancial nos fluxos de água estimados e consequentemente a água armazenada no perfil do solo, concluindo que os resultados obtidos fornecem diretrizes para a seleção de FPTs a serem usadas em modelos de grande escala.

3.8 Diferenças entre funções de pedotransferência de solos de clima tropical e temperado

Nas últimas décadas a maioria das FPTs existentes e publicadas têm sido desenvolvidas utilizando base de dados originadas de regiões temperadas (BOTULA *et al.*, 2012; NGUYEN *et al.*, 2015; OTTONI *et al.*, 2018; TOMASELLA; HODNETT;

ROSSATO, 2000). Frente a falta de informações nos bancos de dados de solo e da carência de modelos capazes de estimar certos atributos hidráulicos do solo em algumas regiões tropicais e subtropicais, FPTs desenvolvidas em solos de clima temperado têm sido utilizadas para estimar atributos nos solos de clima tropical e subtropical (MEDEIROS *et al.*, 2014; NEBEL *et al.*, 2010; NGUYEN *et al.*, 2015; OTTONI *et al.*, 2018; ZOU *et al.*, 2016).

As FPTs desenvolvidas no clima temperado apresentam a peculiaridade de não representar as funcionalidades hidráulicas típicas de solos de clima tropical, devido às diferenças da composição dos solos. Exemplo disso é observado no comportamento híbrido de Latossolos com predomínio de argilas cauliníticas e oxidativas, que apresentam discrepâncias em termos de características e propriedades químicas, físicas e mineralógicas. Além disso, as FPTs das regiões de clima temperado não conseguem abranger a faixa dos teores de argila que são comuns no ambiente tropical (TOMASELLA; HODNETT, 2004). Apesar dos fatores de formação serem semelhantes nos climas temperado e tropical, a extensão e distribuição espacial desses fatores é diferente (NGUYEN *et al.*, 2015).

Ottoni *et al.* (2018) ressaltam que a maior diferença no comportamento hidráulico entre os solos temperados e solos brasileiros intemperizados é devido ao conteúdo de argila e a sua natureza mineralógica, afetando fortemente a formação da estrutura do solo, sendo necessário o desenvolvimento de FPTs para estimar a retenção da água que incluam informações sobre a estrutura do solo e sugerindo-se que sejam desenvolvidas FPTs específicas para solos intemperizados. Do mesmo modo, Botula *et al.* (2012) argumentam que as diferenças químicas e físicas entre solos tropicais úmidos e temperados poderiam ser a causa do pobre desempenho de FPTs temperadas quando aplicados em solos tropicais altamente intemperizados, sendo essas distinções originadas pelas diferenças marcantes nas características mineralógicas.

Na mesma linha, considera-se que uma baixa qualidade da estimativa das funções e propriedades hidráulicas do solo pode influenciar de forma geral a qualidade das saídas de todo o processamento de um modelo. Exemplo disso, é encontrado no estudo de Espino *et al.* (1995) quem utilizou FPTs temperadas de Vereecken *et al.* (1989) e Vereecken, Maes e Feyen (1990) para determinar os parâmetros das funções hidráulicas

do solo $\theta(h)$ e condutividade hidráulica $k(h)$ que foram usadas como entrada no modelo SWATRER (DIERCKX; BELMANS; PAUWELS, 1986), tendo como objetivo a simulação da resposta hidráulica de um perfil de solo em condições climáticas naturais de contorno por um período de 1 ano na Alemanha, obtendo como resultado nas saídas do modelo que, as simulações baseadas na retenção de água estimadas indiretamente superestimam a retenção de água medida em todo o perfil do solo. Os mesmos autores recomendam que mais estudos devem incluir FPTs em um procedimento de teste, antes de serem aplicados em modelos mais avançados, utilizando-se um número limitado de dados medidos de retenção de água do local de interesse para ser utilizado na avaliação das FPTs.

Minasny e Hartemink (2011) reportaram outras diferenças relevantes entre solos temperados e tropicais, tais como a densidade do solo e a capacidade de troca de cátions. Diante dos argumentos anteriormente expostos, é necessário ampliar o conhecimento da interface entre o comportamento hidrofísico de solos tropicais e temperados e contribuir para um melhor entendimento das interações entre os processos pedogenéticos e hidráulicos, além de contribuir com o esclarecimento da extensão da aplicabilidade das FPTs. Contudo, a necessidade de mais estudos e geração de informações precisas e atualizadas das funções e propriedades hidráulicas do solo no clima tropical é muito mais urgente do que em outras regiões, especialmente quando refere-se à situação atual das mudanças climáticas globais, porque os dados de solo são no geral escassos e desatualizados nessas regiões (MINASNY; HARTEMINK, 2011; NGUYEN *et al.*, 2015).

Além disso, a carência de bancos de dados a escala nacional com informações das funções e propriedades hidráulicas do solo em algumas regiões é destacada como a principal razão que podem dificultar o desenvolvimento de FPTs para solos tropicais em quando comparadas com o número de FPTs de solos temperados no mundo (NGUYEN *et al.*, 2015; OTTONI *et al.*, 2018).

No Brasil tem-se desenvolvido FPTs utilizando dados de solo coletados em todos os estados para solos de clima tropical e subtropical (TOMASELLA *et al.*, 2003) e algumas de forma mais localizadas para as regiões Amazônica e Nordeste, estados de Pernambuco, Santa Catarina, São Paulo e Rio Grande do Sul, e no Bioma cerrado

(ARRUDA; ZULLO; OLIVEIRA, 1987; BARROS *et al.*, 2013; GIAROLA; SILVA; IMHOFF, 2002; MEDRADO; LIMA, 2014; MICHELON *et al.*, 2010; OLIVEIRA *et al.*, 2002; REICHERT *et al.*, 2009; SANTOS *et al.*, 2013; SOARES *et al.*, 2014; TOMASELLA; HODNETT, 1998), com intuito de melhorar as estimativas preditivas dos atributos do solo sob condições mais próximas do real às regiões tropicais e subtropicais no Brasil, além de aumentar o número de parâmetros disponíveis que servem como entrada nos modelos hidrológicos.

Recentemente, Ottoni *et al.* (2018) estabeleceram uma base de dados físico-hidráulicas de diferentes solos brasileiros (HYBRAS), disponíveis para o desenvolvimento de FPTs, o qual contém dados de 445 perfis do solo com 1075 amostras, sendo representativos de uma ampla faixa de solos brasileiros. Outro objetivo deste estudo foi usar a base de dados com o intuito de comparar a precisão e exatidão da estimativa da retenção de água no solo, baseados em FPTs desenvolvidas para as regiões brasileiras e de clima temperado. Os autores concluíram que de forma geral, as FPTs desenvolvidas para as condições do Brasil tiveram melhor desempenho que os modelos desenvolvidos nas regiões de clima temperado, especialmente para solos mais intemperizados e de textura fina, mostrando que a retenção de água nos solos intemperizados difere dos de clima temperado, devido às diferenças na estrutura dos poros, conteúdo de argila e natureza mineralógica.

3.9 Funções de pedotransferência geradas e aplicadas no Rio Grande do Sul

No Rio Grande do Sul têm sido aplicadas e geradas diversas FPTs a escala estadual, regional e de vertente (GIAROLA; SILVA; IMHOFF, 2002; MICHELON *et al.*, 2010; REICHERT *et al.*, 2009; SANTOS *et al.*, 2013; SOARES *et al.*, 2014), procurando melhorar as informações faltantes em relação às funções e propriedades hidráulicas dos solos e objetivando também a melhoria na precisão e exatidão da estimativa dos modelos.

A escala estadual Reichert *et al.* (2009) geraram FPTs pontuais para estimar a retenção de água do solo e a capacidade de água disponível, criando dois bancos de dados, o primeiro banco de dados gerado a partir de informações de dados da literatura

(725 dados) e o segundo banco com dados obtidos de solos de um sistema de monitoramento e manejo de irrigação (239 dados). O primeiro banco com dados da literatura era caracterizado por representar diferentes classes de solos e horizontes em diversas regiões do estado de Rio Grande do Sul (RS), incluindo dados de atributos como curvas de retenção de água, matéria orgânica, argila, silte, areia, densidade do solo e de partícula.

A partir do primeiro banco de dados da literatura organizado por Reichert et al. (2009) foram geradas FPTs mediante a técnica de regressão múltipla e com o segundo banco de dados foi realizado o processo de validação e avaliação das FPTs desenvolvidas. Além disso, os dois bancos foram unidos e utilizados para avaliar nove FPTs disponíveis na literatura, com as quais foi calculada a capacidade de água disponível. As FPTs geradas por Reichert et al. (2009) a partir do conjunto de dados da literatura (primeiro banco de dados) tiveram capacidade na estimativa obtendo-se coeficientes r^2 na faixa de 0,56 a 0,66. Por outro lado, ao avaliar as nove FPTs da literatura desenvolvida para outras regiões, as FPTs não foram apropriadas para estimar a retenção de água nos solos de RS, destacando-se assim o melhor desempenho das FPTs desenvolvidas para o estado. Os autores concluíram que as variações nas estimativas da capacidade de água disponível foram mais dependentes das frações areia e silte do que da argila, demonstrando a importância e influência da distribuição do tamanho das partículas nos processos de retenção de água e a relevância do uso dessas características como parâmetros de entrada nas FPTs.

A escala regional Michelon et al. (2010) desenvolveram e validaram FPT para estimar a retenção de água em alguns solos do RS, utilizando informações das regiões do planalto, missões e fronteira oeste do estado, pertencentes ao banco de dados de análises físicas de áreas irrigadas por aspersão. O banco de dados foi particionado em conjunto de geração (60% do total = 720 amostras) e para validação (40% do total = 480 amostras). As FPTs foram desenvolvidas mediante o critério de seleção das variáveis independentes mais significativas por análise de correlação simples entre todas as variáveis (teor de areia fina, silte, argila, densidade do solo, densidade de partículas, macroporosidade, microporosidade e porosidade total) e seguidamente foi feita a análise

de regressão múltipla, utilizando a opção *stepwise* incluindo as variáveis ao 5% de probabilidade de erro.

De forma geral, as FPTs mostraram-se com resultados de alta precisão ($r^2 > 0,70$) para estimar a retenção de água nos potenciais selecionados. Principalmente, as FPTs geradas para os potenciais matriciais de 0 e -1 kPa apresentaram associações apenas com variáveis relacionadas à estrutura do solo, e nos menores potenciais (-33, -100, -500 e -1500 kPa) estiveram associadas com os atributos texturais e estruturais.

À escala de vertente no município de Capão do Leão, Nebel et al. (2010) avaliaram a aplicabilidade e transferibilidade de 8 FPTs na literatura, desenvolvidas a partir de uma variedade de solos e características. Os autores utilizaram as FPTs para estimar a umidade gravimétrica (teor de água medido em proporções de massa) em dois solos de terras baixas no RS (Organossolos e Planossolos) destinados à lavoura de arroz irrigado com rotação de culturas sob o cultivo convencional. As oito FPTs testadas para avaliar a umidade gravimétrica nas tensões de 33 e 1500 kPa, apresentaram a tendência de superestimar e subestimar respectivamente a umidade gravimétrica do solo, quando avaliadas mediante a métrica estatística ME. A FPT de Oliveira et al. (2002) teve melhor desempenho na tensão de 33 kPa em relação às demais (ME= 0.043 e RMSE= 0.049 kg kg⁻¹), enquanto para a tensão 1500 kPa a FPT de Pidgeon (1972) foi a que apresentou o melhor desempenho (ME= 0.0054 e RMSE= 0.0216 kg kg⁻¹). No entanto, baseando-se no critério de seleção da melhor FPT de Cornelis et al. (2001) por ranqueamento das métricas estatísticas ME, RMSE e r^2 , a FPT de Pidgeon (1972) apresentou o melhor desempenho para ambos potenciais matriciais, seguida das FPTs de Lal (1979) e Oliveira et al. (2002).

Observando os trabalhos realizados referentes à geração de FPTs no estado do Rio Grande do Sul e especificamente no clima subtropical, esses estudos possuem principalmente o viés em avaliações a escala agrícola e de lavouras. Juntamente é observado que os estudos carecem do desenvolvimento de FPTs para estimar a retenção de água a escala de bacia hidrográfica que tenham como enfoque o de contribuir com o aumento da qualidade das informações necessárias para o estudo das bacias como unidade de manejo, planejamento dos recursos hídricos e modelagem hidrológica. Além disso, os estudos realizados nas condições subtropicais são escassos, limitando-se

assim a aplicação e avaliação de outras metodologias para o desenvolvimento das FPTs, sendo necessário explorar outras relações que podem existir entre as variáveis de entradas nos modelos e as estimativas da retenção de água nos diversos potenciais.

3.10 Aprendizado de máquina e técnicas para a avaliação de modelos

Aprendizado de Máquina (AM) é um campo central da inteligência artificial (IA). O objetivo do AM é transmitir à máquina certos conceitos com uma precisão para que ela tenha a capacidade de converter os exemplos em conhecimentos, mediante algoritmos e técnicas de aprendizado automático (ERTEL, 2017). De acordo com Mitchell (1997) o AM é o estudo dos diversos algoritmos de computação que melhoram automaticamente com a experiência (RICH, 1983).

Geralmente, os algoritmos de AM se enquadram em quatro categorias: aprendizado supervisionado, aprendizado não supervisionado, aprendizado semi-supervisionado e aprendizado por reforço, sendo as categorias mais utilizadas na literatura o aprendizado supervisionado e não supervisionado. O aprendizado supervisionado requer a apresentação de exemplos rotulados de classe conhecidas, ou seja, o objetivo é apreender as relações que existem entre os exemplos e seus rótulos. Por outro lado, no aprendizado não supervisionado a prévia estruturação dos exemplos ou dados por parte do usuário não é necessária, podendo os dados de treinamento estarem sem rótulos ou mesmo usar exemplos cujas classes não são conhecidas (AGGARWAL, 2018).

Dentre os mais importantes algoritmos de aprendizado de acordo às categorias mais utilizadas, podemos citar no aprendizado supervisionado, a indução de árvores de decisão (classificação), as redes neurais artificiais (classificação + aproximação) e as máquinas de vetores de suporte (SVM). E no aprendizado não supervisionado, o algoritmo do vizinho mais próximo, o algoritmo do vizinho mais distante e o agrupamento *k-means* (*k-means clustering*) (ERTEL, 2017).

3.10.1 Generalização, otimização e avaliação dos modelos de AM

Conhecendo-se que a meta do AM é obter modelos que tenham bom desempenho com dados nunca antes vistos, então se faz necessário avaliar o potencial de generalização dos modelos. O fundamental no AM é a relação entre a otimização e generalização, onde a otimização se refere ao processo de ajuste de um modelo para conseguir o melhor desempenho possível sobre o conjunto de dados de treinamento (o aprendizado), enquanto a generalização se refere ao quão bem o modelo que foi treinado, desempenha-se sobre dados nunca antes vistos (conjunto de dados de testes) (CHOLLET, 2020).

3.10.1.1 Bias e variância

Muitos estudos utilizam os termos bias e variância ou “*bias-variance trade-off*” para avaliar o desempenho de um modelo, afirmando que uma alta variância é proporcional ao *overfitting* e um alto *bias* é proporcional ao ajuste insuficiente (*underfitting*). O *bias-variance trade-off* estabelece que o erro quadrático de um algoritmo de AM pode ser particionado em três componentes: o *bias*, a variância e o ruído (RASCHKA, 2018).

O *bias* é definido como o erro que demonstra a incapacidade do algoritmo assimilar corretamente relações entre os atributos e os valores dependentes, podendo um modelo causar sempre erros, mesmo se o modelo tiver acesso a uma fonte infinita de dados de treinamento, sendo assim, o *bias* não poderá ser removido (AGGARWAL, 2018). Então, o *bias* é basicamente expressado como a diferença entre o valor estimado e o valor observado (RASCHKA, 2018).

A variância é o erro que indica que o modelo é sensível a ruídos e ajustou seu comportamento a esses resultados. A presença de uma alta variância é manifestada pelo *overfitting* específico aos dados de treinamento em questão. Assim, a variância é simplesmente definida como a diferença entre o valor estimado ao quadrado e o valor observado ao quadrado. Por último, o ruído é causado pelos erros inerentes nos dados (AGGARWAL, 2018; RASCHKA, 2018).

Um algoritmo deve ter um equilíbrio entre *bias* e variância, não tendo um *bias* muito elevado porque pode ter uma maior taxa de erro no conjunto de testes, e não deve ter uma alta variância, uma vez que o algoritmo pode considerar ruídos e dados aleatórios, aprendendo muitas particularidades do conjunto de treino (AGGARWAL, 2018; RASCHKA, 2018).

Frente ao anterior, o objetivo do *bias-variance trade-off* é quantificar o erro esperado do algoritmo em termos de *bias*, variância e ruído, podendo ser o erro do modelo quantificado pelo erro quadrado médio (MSE) entre os valores estimados e observados, porém definido sobre os exemplos do conjunto de testes. O *trade-off* entre *bias* e variância à medida que a complexidade do modelo é aumentada pode ser ilustrada na Figura 1, onde claramente é apresentado que existe um ponto de complexidade ótimo do modelo em que o desempenho é otimizado (AGGARWAL, 2018; RASCHKA, 2018).

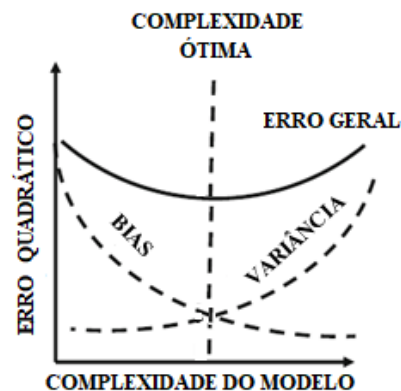


Figura 1 - Exemplo de *trade-off* entre *bias* e variância apresentando o ponto de complexidade ótimo do desempenho de um modelo. Adaptado de Aggarwal (2018).

3.10.1.2 Sobreajuste e subajuste

Durante o aprendizado, precisa-se realizar o controle entre os valores estimados e observados. Esse é o trabalho da função de perda "*loss function*", a qual considera as estimativas do algoritmo e o verdadeiro alvo, calculando uma pontuação das distâncias entre eles e capturando o desempenho do determinado algoritmo (CHOLLET, 2020).

No início do treinamento, a otimização e generalização estão correlacionadas: Quanto menor *loss* nos dados de treinamento, menor *loss* nos dados de testes, indicando que o modelo sofre subajuste (*underfitting*) e ainda há um processo a ser feito sobre os dados de treinamento (Figura 2). Porém, após um certo número de iterações sobre os dados de treinamento, a generalização para de melhorar e as métricas de validação são interrompidas, começando o modelo a sofrer sobreajuste (*overfitting*), ou seja, o modelo está começando a aprender padrões específicos para os dados de treinamento, porém errôneos ou irrelevantes quando são apresentados novos dados (CHOLLET, 2020).

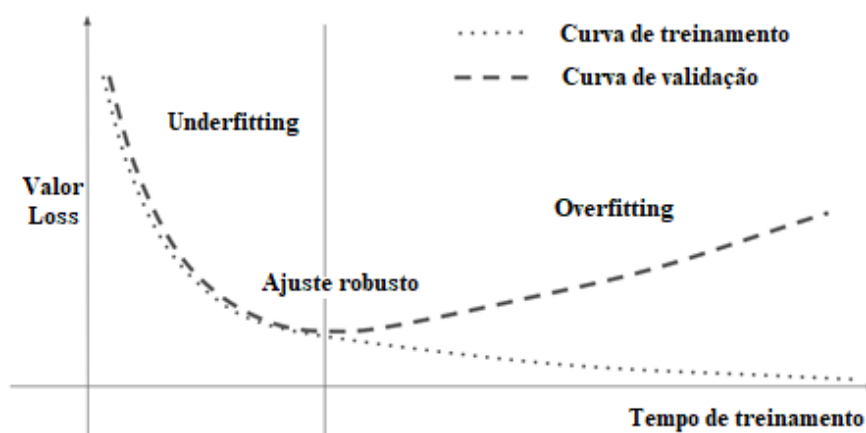


Figura 2 - Ilustração do *underfitting* e *overfitting* em relação ao erro de treinamento e teste. Adaptado de Chollet (2020).

Um modelo treinado com uma grande quantidade de dados naturalmente vai generalizar melhor, no entanto, frente a carência de dados, uma segunda opção é modular a quantidade de informação que o modelo pode armazenar ou adicionar algumas restrições em relação às informações que ele pode armazenar (CHOLLET, 2020).

Do mesmo modo, existem outras estratégias para a mitigação do *overfitting* e maximizar a generalização, além dessas contribuírem para avaliar os modelos de AM (AGGARWAL, 2018). Não obstante, o *overfitting* não é o único obstáculo presente na avaliação dos modelos, destacando-se também o *underfitting*, porém de mais fácil identificação, já que um modelo que apresenta *underfitting* não se desempenha bem, nem no conjunto de treinamento nem no conjunto de testes, sendo as taxas de erro semelhantes entre eles, indicando que o modelo tem um alto *bias* e sendo necessário

melhorar o seu desempenho pelo ajuste dos parâmetros do algoritmo. Por outro lado, um modelo que realiza *overfitting* pode ser usualmente reconhecido pela alta precisão do conjunto de treinamento, mas com baixa precisão no conjunto de testes, indicando que o modelo tem baixo *bias*, sendo necessário deixar livre ao algoritmo para conseguir generalizar melhor (CHOLLET, 2020; RASCHKA, 2018).

A avaliação do desempenho dos modelos pode ser feita mediante diferentes métricas estatísticas sobre um conjunto de testes predeterminado, podendo serem essas métricas específicas para o modelo e o conjunto de dados. Nesse sentido, as métricas são direcionadas a capturar a precisão e exatidão da estimativa, estabelecendo se um modelo tem bom ou ruim desempenho na generalização (RASCHKA, 2018).

Muitos algoritmos de AM têm a capacidade de adaptar a complexidade do modelo aprendido à complexidade dos dados de treinamento, porém na presença de ruídos, o problema de *overfitting* pode ser potencializado (MITCHELL, 1997). De fato, o *overfitting* é possível mesmo quando os dados de treinamento não apresentam ruído, especialmente quando números pequenos de exemplos são associados ao algoritmo. Frente a isso, muitas técnicas procuram otimizar a complexidade do modelo, tal que, os erros de aproximação ou classificação sejam minimizados quando se apresenta um conjunto de dados desconhecidos (conjunto de testes).

3.10.1.3 Conjuntos de treinamento, validação e testes no AM.

Como procedimento imprescindível durante a avaliação de modelos, é requerido que os dados sejam divididos em dois conjuntos: treinamento e teste. A partir dos dados do conjunto de treinamento se realiza o treino, e dependendo do tipo do algoritmo escolhido, a partir do conjunto de treino pode-se gerar um conjunto de validação que permita a avaliação dos modelos gerados durante o processo de treinamento do algoritmo. Uma vez que o modelo esteja pronto, este é testado uma última vez com o conjunto de testes.

Durante o desenvolvimento de um modelo sempre é necessário ajustar a sua configuração, por exemplo, em uma rede neural é possível ajustar os hiper-parâmetros¹ como o número de camadas, neurônios, as funções de ativação, as taxas de aprendizado, métodos de otimização e seus parâmetros, entre outros, realizando-se os ajustes da rede utilizando como sinal *feedback* o desempenho da rede nos dados do conjunto de validação (AGGARWAL, 2018).

Entre as formas de avaliação de modelos encontra-se o método de validação “*hold-out method*” sendo a técnica de avaliação mais simples e resumida da seguinte forma: Primeiro, tomando um conjunto de dados e dividindo-se em duas partes, um conjunto de treinamento e um conjunto de testes, seguidamente o modelo é ajustado ao conjunto de treinamento e as estimativas são realizadas com o conjunto de testes (RASCHKA, 2018). A fração de estimativas corretas que podem ser calculadas pela comparação entre os dados estimados com os dados medidos do conjunto de testes, constitui a precisão da estimativa do modelo. Em si, o ajuste do modelo é uma forma de aprendizado na qual se busca uma boa configuração dentro de um espaço de parâmetros. Contudo, o ajuste da configuração do modelo baseado no seu desempenho no conjunto de validação, pode resultar rapidamente em sobreajuste do conjunto de validação, mesmo que o modelo nunca seja treinado diretamente sobre ele. No final do treinamento, o modelo pode ter um bom desempenho unicamente sobre os dados de validação, porque foi otimizado para essa situação, sendo assim, é necessária a avaliação do desempenho do modelo com dados novos e nunca antes vistos por ele (conjunto de dados de testes), obtendo-se o real desempenho do modelo e o grau de generalização do mesmo (AGGARWAL, 2018).

¹ Os hiper-parâmetros são parâmetros utilizados no aprendizado de máquina cujos valores podem ser escolhidos arbitrariamente antes de iniciar o treinamento de um determinado algoritmo e definem seu comportamento durante esse processo.

Frequentemente, a divisão do conjunto de dados em conjuntos de treinamento e de testes, se faz como um processo simples de subamostragem aleatória, assumindo-se que cada dado foi extraído da mesma distribuição de probabilidade, optando-se por escolher aleatoriamente $2/3$ das amostras para o conjunto de treinamento e $1/3$ das amostras para o conjunto de testes. Contudo, um dos tipos de problemas que ocorrem quando um conjunto de dados é dividido em conjuntos de treinamento e de teste separados é que a técnica unicamente é viável quando tem-se grande conjuntos de dados e as distribuições de probabilidades são semelhantes nos dois conjuntos, reduzindo-se as problemáticas relacionadas ao viés pessimista e às grandes variâncias das estimativas (RASCHKA, 2018).

Apesar da divisão dos dados em conjuntos de treinamento, validação e testes parecer uma tarefa simples, existem outras técnicas avançadas para fazer esta divisão, e sobretudo, podem ser muito úteis quando há pouca quantidade de dados disponíveis. Dentre essas técnicas encontram-se os mais clássicos métodos de avaliação de modelos: validação cruzada (*cross validation* ou *k-fold validation*) e a validação cruzada k vezes (*k-fold cross validation* ou *k-fold validation com embaralhamento*) (RASCHKA, 2018).

A *k-fold cross validation* (validação cruzada k -vezes), é um método que permite a travessia das etapas de treinamento e validação em sucessivas rodadas. Como primeira instância o princípio da *cross validation* é que cada amostra no conjunto de dados tem a oportunidade de ser testado e a *k-fold cross validation* é um caso especial de *cross validation* onde o usuário pode iterar sobre o conjunto de dados definido k vezes. Em cada rodada, o conjunto de dados é dividido em k -partes, sendo uma parte utilizada para a validação e as partes $k-1$ restantes são misturadas (embaralhadas) em um subconjunto de treinamento para avaliação do modelo. Desta forma o método utiliza todos os dados para treinamento e teste, reduzindo-se o viés pessimista ao usar mais dados para o treinamento e evitando-se deixar de lado uma porção relativamente grande do conjunto de dados como dados de testes (RASCHKA, 2018).

Cuidados devem ser tomados durante a avaliação dos modelos, não devendo-se treinar e avaliar um modelo no mesmo conjunto de dados de treinamento, já que pode apresentar um viés ou tendência (*bias*) muito otimista devido ao sobreajuste (*overfitting*),

em outras palavras, é duvidoso dizer se o modelo simplesmente memoriza os dados de treinamento ou se generaliza bem para dados novos e independentes, assim o viés de otimismo (*Optimism bias*) pode ser estimado como a diferença entre a precisão do treinamento e do teste.

3.10.2 Aspectos gerais das redes neurais artificiais

Entre os tipos de algoritmos de aproximação de funções mais populares no aprendizado supervisionado, encontram-se as redes neurais artificiais (RNAs), as quais serão apresentadas com maior abrangência nesta seção.

3.10.2.1 Redes neurais artificiais

As RNAs foram desenvolvidas para simular o comportamento do sistema nervoso de organismos biológicos na realização de tarefas de aprendizado, mediante o uso de unidades computacionais (neurônios) que têm a capacidade de aprender de forma similar aos neurônios humanos (AGGARWAL, 2018). O processo de desenvolvimento de uma RNA inicia com a determinação da arquitetura (construção da rede), onde se escolhe o número de camadas e de neurônios em cada camada, para ajustar depois no processo de treinamento os seus parâmetros livres (hiper-parâmetros). Essa arquitetura pode ser montada antes do treinamento, já que se procura mediante tentativas e erros durante o treinamento, a arquitetura adequada, sendo também dependente da complexidade do problema (SOARES *et al.*, 2014).

3.10.2.2 Arquitetura básica de uma rede neural

Dentre as estruturas básicas das RNAs existem classes de arquiteturas diferentes, sendo a classificação mais comum a realizada quanto ao número de camadas, tendo-se quatro grupos: as redes com uma única camada, as redes de múltiplas camadas, as redes acíclicas (*Feed-forward*) e as cíclicas (*Recurrent Neural Network-RNNs*) (SOARES, 2013)

Nas redes neurais de uma única camada, um conjunto de entradas é mapeado diretamente para uma saída usando uma variação generalizada de uma função linear, a qual é comumente chamada de *perceptron* (Figura 3) e que só consegue resolver problemas linearmente separáveis (AGGARWAL, 2018).

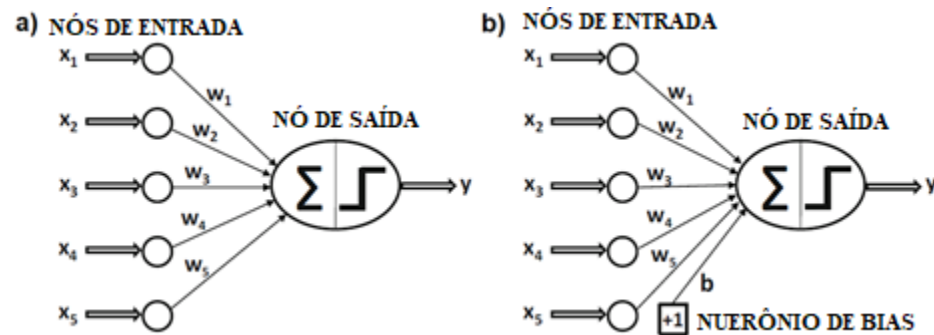


Figura 3 - a) arquitetura de um perceptron sem *bias* b) arquitetura de um perceptron com *bias*. Apresentando os nós de entrada, nó de saída e neurônio de *bias*. Adaptado de Aggarwal (2018).

No caso das redes de múltiplas camadas, os neurônios são organizados na forma de camadas, nas quais as camadas de entrada e saída são separadas por um grupo de camadas ocultas (Figura 4) (AGGARWAL, 2018).

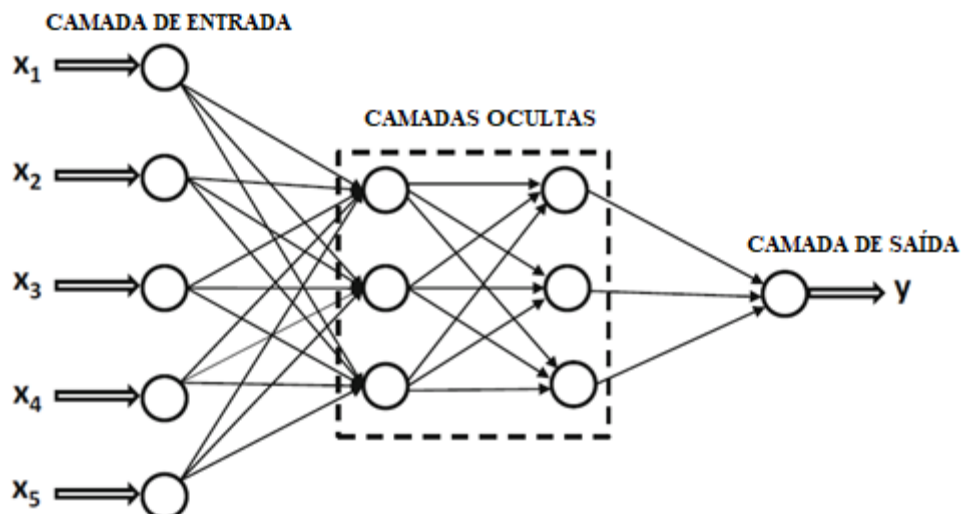


Figura 4 - Exemplo de arquitetura de uma RNA de múltiplas camadas. Apresentando a camada de entrada, camadas ocultas e camada de saída. Adaptado de Aggarwal (2018).

3.10.2.3 O perceptron: rede de camada única e o seu modelo matemático

O *perceptron* é conhecido como a rede neural mais simples, a qual contém uma única camada de entrada e um nó de saída, como o mostrado na sua arquitetura básica (Figura 3a). Considerando uma situação onde cada exemplo de treinamento é da forma $(\bar{x}, y)$, onde cada $\bar{X} = [x_1 \dots x_d]$ contém variáveis de características d (entradas), e $y \in \{-1, +1\}$ contém os valores observados da variável de classe binária (saídas). Além disso, a camada de entrada contém d nodos que transmitem as d características $\bar{X} = [x_1, \dots, x_d]$ com conexões de peso $\bar{w} = [w_1 \dots w_d]$ para um nó de saída. Porém, a camada de entrada não realiza nenhum cálculo por si só, sendo a função linear $\bar{w} \cdot \bar{x} = \sum_{i=1}^d w_i x_i$ calculada no nó de saída. Seguidamente, o sinal desse valor real é usado para estimar a variável $\hat{y}$, sendo calculada pela equação 1 (AGGARWAL, 2018).

$$\hat{y} = \text{sin}al\{\bar{w} \cdot \bar{X}\} = \text{sin}al\left\{\sum_{j=1}^d w_j x_j\right\} \quad (1)$$

Assim, a função sinal atribui um valor real para +1 ou -1, a qual é apropriado para uma classificação binária e serve como função de ativação. Frente a isso, diferentes funções de ativação podem ser usadas para simular diferentes tipos de modelos que são utilizados no AM. Cabe ressaltar que o *perceptron* tem duas camadas, embora a camada de entrada não execute nenhum cálculo, esta camada pode somente transmitir os valores das características. Por tanto, a camada de entrada do *perceptron* não é incluída na contagem do número de camadas, tendo assim uma única camada computacional que lhe atribui a característica de uma rede de camada única (AGGARWAL, 2018).

Em algumas configurações de redes para casos de análises binárias, existe uma parte invariável da estimativa, comumente denominada de *bias*, a qual não pode ser capturada pela abordagem do *perceptron* apresentada na equação 1 e na Figura 3a, sendo necessário incorporar o *bias* como uma variável adicional “ b ”, permitindo capturar a parte invariável da estimativa, como o apresentado na equação 2 e na Figura 3b, em outras palavras o *bias* corresponde a um neurônio ou entrada especial que serve para aumentar os graus de liberdade, permitindo ao *perceptron* conseguir uma melhor

aproximação da função dos exemplos de treinamento apresentados, podendo ser considerado como um ajuste fino da rede (AGGARWAL, 2018).

$$\hat{y} = \text{sinal}\{\bar{w} \cdot \bar{X} + b\} = \text{sinal}\left\{\sum_{j=1}^d w_j x_i + b\right\} \quad (2)$$

O *bias* pode ser incorporado como o peso de uma conexão, utilizando um neurônio de *bias*, o qual é obtido através da adição conceitual de um neurônio que sempre transmite um valor de 1 para o nó de saída (Figura 3b), assim, o peso da conexão que conecta o neurônio do *bias* ao nó de saída fornece a variável *bias*. O algoritmo *perceptron* foi criado com o objetivo de reduzir o número de classificações incorretas, ou minimizar o erro de estimativa. Como resultado, seu objetivo é expressado como mínimos quadrados em relação a todos os exemplos de treinamento de um específico conjunto “D” de dados, que inclui pares de características-rótulos (equação 3) (AGGARWAL, 2018).

$$\text{Minimização}_{\bar{W}} L = \sum_{(\bar{X}, y) \in D} (y - \text{sinal}\{\bar{W} \cdot \bar{X}\})^2 \quad (3)$$

Esse tipo de função objetivo de minimização apresentada na equação 3 é a anteriormente mencionada função de perda “*loss function*”. Embora a função objetivo seja definida sobre todos os dados de treinamento, o algoritmo de treinamento de redes neurais funciona alimentando instâncias de dados de entrada $\bar{X}$ na rede uma por uma (ou pequenos grupos de exemplos) para criar a estimativa $\hat{y}$, onde os pesos são atualizados com base no valor do erro $E(\bar{X}) = (y - \hat{y})$. Especificamente, quando o ponto de dados $\bar{X}$ é alimentado na rede, o vetor de peso $\bar{W}$ é atualizado da forma como o apresentado na equação 5, em que o parâmetro α regula a taxa de aprendizagem da rede neural, método conhecido como gradiente descendente (AGGARWAL, 2018).

$$\bar{W} \leftarrow \bar{W} + \alpha(y - \hat{y})\bar{X} \quad (4)$$

O algoritmo *perceptron* percorre repetidamente (ciclos) por todos os exemplos de treinamento em ordem aleatória e ajusta de forma iterativa os pesos até que a convergência seja atingida, assim um único ponto dos dados de treinamento pode ser percorrido várias vezes, sendo cada ciclo conhecido como uma época. Sabendo que o *perceptron* otimiza uma função desconhecida usando o método gradiente descendente, este pode ser expresso em termos de erro $E(\bar{X}) = (y - \hat{y})$. de acordo com a equação 5 (AGGARWAL, 2018).

$$\bar{W} \leftarrow \bar{W} + \alpha E(\bar{X}) \bar{X} \quad (5)$$

O algoritmo básico do *perceptron* pode ser considerado um método estocástico de gradiente descendente, que implicitamente minimiza o erro quadrático da estimativa, executando atualizações do gradiente descendente com relação aos pontos de treinamento escolhidos aleatoriamente. O pressuposto é que a rede neural percorre os pontos em ordem aleatória durante o treinamento e altera os pesos com o objetivo de reduzir o erro de estimativa nesse ponto. Deste modo, na equação 5 é possível concluir que atualizações diferentes de zero são feitas nos pesos somente quando $y \neq \hat{y}$, o que ocorre apenas quando acontecem erros na estimativa. No mesmo sentido, durante o gradiente descendente estocástico de pequenos grupos “*mini-batches*”, as atualizações referentes à equação 5 são implementadas sobre um subconjunto escolhido aleatoriamente dos pontos de treinamento S (equação 6) (AGGARWAL, 2018):

$$\bar{W} \leftarrow \bar{W} + \alpha \sum_{\bar{X} \in S} E(\bar{X}) \bar{X} \quad (6)$$

O tipo de modelo proposto no *perceptron* é um modelo linear, no qual a equação $\bar{W} \bar{X} = 0$ define um hiperplano linear, onde $\bar{W} = [w_1, \dots, w_d]$ é um vetor d -dimensional que é normal ao hiperplano. Esse tipo de modelo desempenha-se particularmente bom quando os dados são linearmente separáveis, como o apresentado na Figura 5 onde são apresentados exemplos de dados linearmente separáveis e inseparáveis (AGGARWAL, 2018).

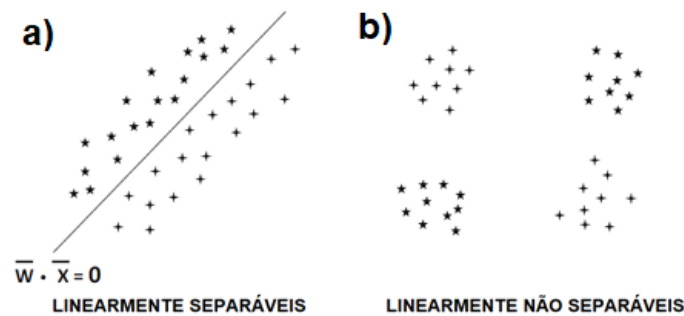


Figura 5 - Exemplos de dados linearmente (a) separáveis e (b) inseparáveis em duas classes. Fonte: Aggarwal (2018).

A Figura 5a mostra que o *perceptron* é um algoritmo bom na classificação dos conjuntos de dados, quando estes são linearmente separáveis, e por outro lado, este algoritmo tende a apresentar um desempenho ruim em conjuntos de dados como o apresentado na Figura 5b, manifestando a limitação inerente à modelagem de um *perceptron*, sendo necessário o uso de arquiteturas neurais mais complexas (AGGARWAL, 2018).

Com o passar do tempo, os cientistas realizaram mais ajustes no modelo original do *perceptron*, sendo a principal modificação a adição de uma função que determina o estado de ativação da saída do neurônio conhecido como função de ativação, a qual recebe o resultado do somatório do produto de cada entrada pelos seus pesos respectivos, gerando um único valor que poder ser a resposta da rede neural ou ser usado como entrada para um próximo neurônio (HAYKIN, 1994; SOARES, 2013). Assim o modelo de perceptron ficou modificado como um modelo de neurônio base para os projetos de RNAs, o qual é composto por três elementos básicos, o conjunto de sinapses ou conexões, a junção aditiva (somatório) e a função de ativação, estas duas últimas localizadas antes do neurônio de saída, como o apresentado na Figura 6.

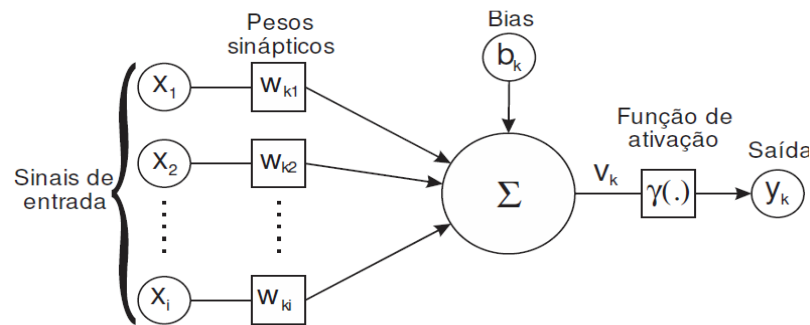


Figura 6 - Perceptron modificado: modelo base de um neurônio para projetos de RNAs. Fonte: Soares (2013).

O conjunto de sinapses é caracterizado pelos produtos entre os sinais de entrada e os pesos sinápticos, referidos ao terminal da entrada da sinapse. A junção aditiva, se refere ao somatório do conjunto de sinapses e constitui um combinador linear. A função de ativação vai receber o resultado da junção aditiva e tem a função de restringir a amplitude de saída do neurônio, limitando o intervalo permissível de amplitude do sinal de saída do neurônio a um valor finito, que normalmente é normalizada no intervalo unitário $[0, 1]$ ou $[-1, -1]$. Assim, a escolha da função de ativação deve ser feita de acordo com o problema a ser solucionado pela rede neural (SOARES, 2013).

3.10.2.4 Redes de múltiplas camadas e modelo matemático

Frente às limitações que têm as redes de camada única (*perceptron*) para a solução de problemas que não são linearmente separáveis, surgiram as redes neurais de múltiplas camadas “*multilayer*” que têm mais de uma camada computacional, contendo camadas intermediárias entre as camadas de entrada e saída que comumente são chamadas de camadas ocultas “*hidden layers*”, devido a que os cálculos desempenhados não são visíveis para o usuário. A arquitetura específica das redes *multilayer* é chamada alimentação para frente “*feedforward*”, porque as camadas sucessivas se alimentam uma na outra na direção *forward* (da entrada para a saída), onde pressupõe-se que todos os nós em uma camada estão conectados aos da camada seguinte (AGGARWAL, 2018).

Similar ao *perceptron*, neurônios de *bias* podem ser utilizados tanto nas camadas ocultas como nas de saída, como o representado na Figura 7, onde se mostram redes

que contém três camadas, tendo em conta que a camada de entrada normalmente não é contada, porque essa simplesmente transmite os dados e não é desempenhado nenhum cálculo nela.

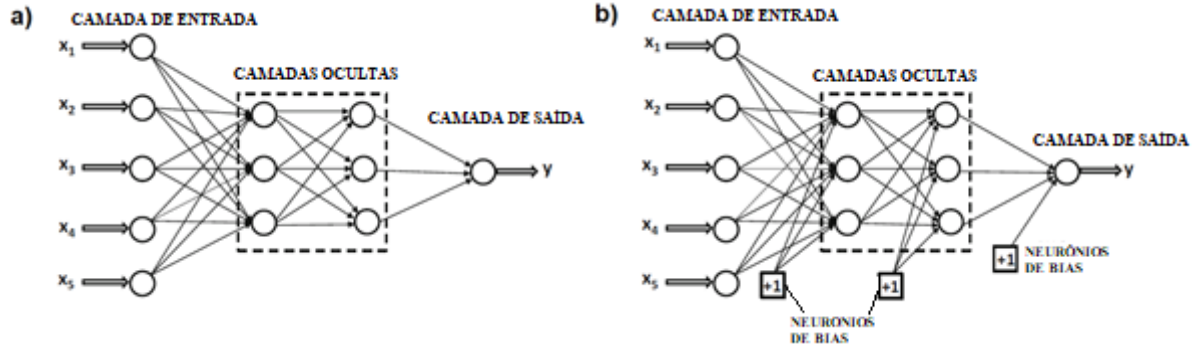


Figura 7 - Arquitetura básica de uma rede *feed-forward* com duas camadas ocultas e uma única camada de saída, apresentando a rede (a) sem neurônios de *bias* e (b) com neurônios de *bias*. Adaptado de Aggarwal (2018).

Se uma rede neural tiver $p_1 \dots p_k$ unidades em cada uma de suas k camadas, então as representações do vetor (coluna) dessas saídas, denotadas por $\bar{h}_1 \dots \bar{h}_k$ tem dimensionalidades $p_1 \dots p_k$. Por tanto, o número de unidades em cada camada é chamado de dimensionalidade dessa camada. Os pesos das conexões entre a camada de entrada e a primeira camada oculta estão contidos na matriz W_1 com tamanho $d \times p_1$, enquanto os pesos entre a r – ésima camada oculta e a $(r + 1)$ – ésima camada oculta são indicados pela matriz $p_r \times p_{r+1}$ denotada por W_r e se a camada de saída contém o nós, então a matriz final W_{k+1} é de tamanho $p_k \times o$. O vetor de entrada d -dimensional $\bar{x}$ é transformado nas saídas usando as seguintes equações recursivas:

$$\bar{h}_1 = \phi(W_1^T \bar{x}) \quad [\text{entrada para camada oculta}] \quad (7)$$

$$\bar{h}_{p+1} = \phi(W_{p+1}^T \bar{h}_p) \forall p \in \{1 \dots k - 1\} \quad [\text{oculta para camada oculta}] \quad (8)$$

$$\bar{h}_{p+1} = \phi(W_{k+1}^T \bar{h}_k) \quad [\text{oculta para camada oculta}] \quad (9)$$

3.10.2.5 Funções de ativação, nós de saída e função de custo (loss)

A escolha da função de ativação é uma parte fundamental do design de uma rede neural. Assim, o uso de funções de ativação não lineares torna-se significativo quando se parte do perceptron de camada única para as arquiteturas de múltiplas camadas. Diante disso, existem diferentes tipos de funções não lineares, como sinal, sigmóide, ou tangente hiperbólica que podem ser usadas em várias camadas, sendo denotada a função de ativação pelo símbolo ϕ :

$$\hat{y} = \phi(\bar{W} \cdot \bar{X}) \quad (10)$$

Por tanto, um neurônio calcula realmente duas funções dentro de si (equação 10). Assim, pode ser observada a ruptura dos cálculos do neurônio em dois valores separados como o mostrado na Figura 8 onde ocorre a incorporação do símbolo Σ , bem como o símbolo de ativação ϕ dentro de um neurônio (AGGARWAL, 2018).

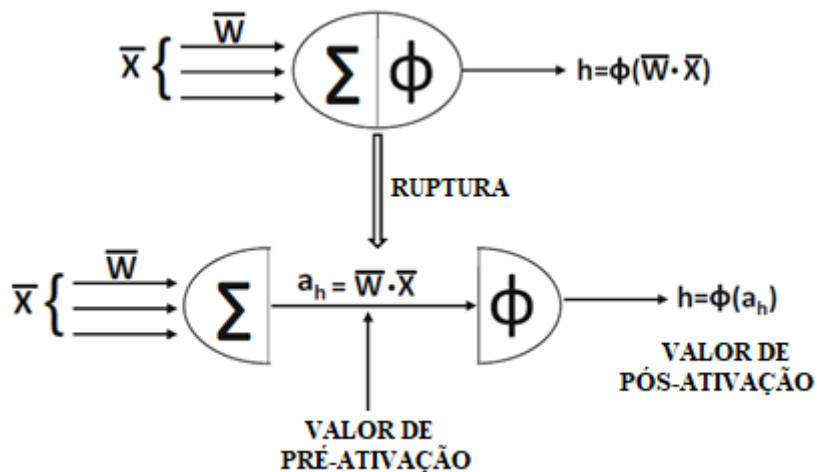


Figura 8 – Ruptura do neurônio e valores de pré-ativação e pós-ativação dentro de um neurônio. Adaptado de Aggarwal (2018).

O valor calculado antes de aplicar da função de ativação $\phi(\cdot)$ será referido como o valor de pré-ativação, enquanto o valor calculado após a aplicação da função de ativação será referido como o valor de pós-ativação. Assim, a saída do neurônio é sempre

o valor de pós-ativação, embora as variáveis de pré-ativação sejam frequentemente usadas em diferentes tipos de análises, como por exemplo nos cálculos do algoritmo de retropropagação “*backpropagation*”. Entre as diversas funções de ativação, a mais básica função de ativação $\phi(\cdot)$ é a identidade ou ativação linear “*identity*”, a qual não fornece nenhuma não-linearidade (equação 10).

$$\phi(v) = v \text{ (função identidade ou linear)} \quad (11)$$

A função de ativação linear é frequentemente utilizada no nó de saída, quando o objetivo é um valor real. No início do desenvolvimento das RNAs as funções de ativação mais clássicas foram as funções sinal “*sign*”, sigmóide “*sigmoid*” e a tangente hiperbólica (*tanh*).

$$\phi(v) = \text{sign}(v) \text{ (função sinal)} \quad (12)$$

$$\phi(v) = \frac{1}{1+e^{-v}} \text{ (função sigmóide)} \quad (13)$$

$$\phi(v) = \frac{e^{2v}-1}{e^{2v}+1} \text{ (função tanh)} \quad (14)$$

Enquanto a função de ativação sinal pode ser usada para aproximar saídas binárias no momento da estimativa, sua não diferenciabilidade impede seu uso para criar a função *loss* durante o treinamento. A função sigmóide gera um valor entre (0, 1), o qual é útil na realização de cálculos que devem ser interpretados como resultados probabilísticos. A função *tanh* tem uma forma similar à da função sigmóide, exceto que é re-escalada horizontalmente e verticalmente convertida/re-escalada para [-1,1], assim as funções sigmóides e *tanh* estão relacionadas como mostra a equação 15.

$$\tanh(h) = 2 \cdot \text{sigmóide}(2v) - 1 \quad (15)$$

A função *tanh* é preferível à sigmóide quando se deseja que as saídas dos cálculos sejam positivas e negativas, uma vez que a sua maior inclinação (devido ao alongamento) e centralização da média com relação ao sigmóide facilita o treinamento. Embora, as funções sigmóides e *tanh* tenham sido ferramentas historicamente escolhidas para a incorporação da não linearidade nas RNAs, nos últimos anos várias funções de ativação linear por segmentos (*piecewise*) tornaram-se mais populares nas redes modernas, como a *ReLU* e *hard tanh*, devido à facilidade no treinamento com essas funções nas redes neurais de várias camadas.

$$\phi(v) = \max\{v, 0\} \quad (\text{Rectified Linear Unit |Relu|}) \quad (16)$$

$$\phi(v) = \max\{\min[v, 1], -1\} \quad (\text{hard tanh}) \quad (17)$$

As representações das funções de ativação mais populares são apresentadas na Figura 9. O uso de uma função de ativação não linear permite aumentar o potencial de modelagem de uma rede, em relação a um *perceptron*.

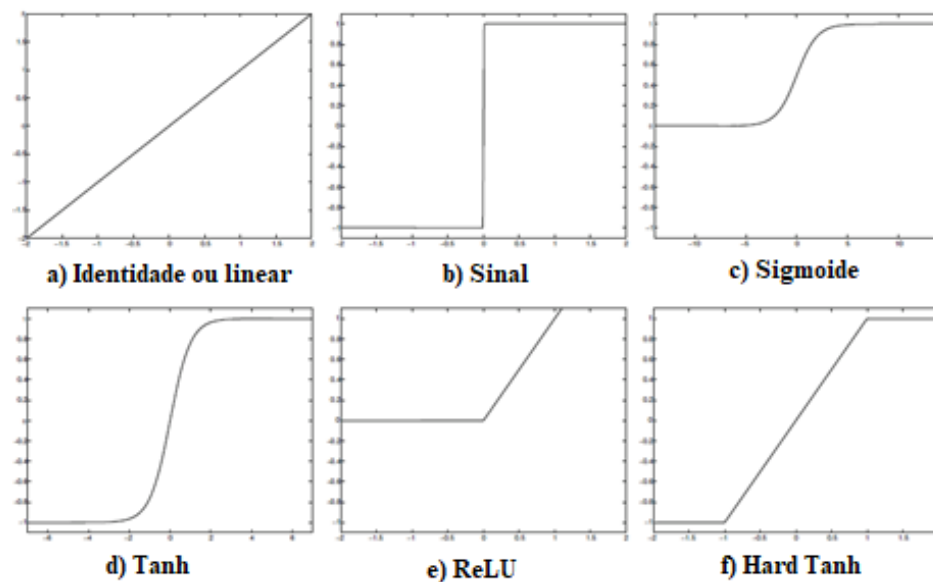


Figura 9 - Representações gráficas das funções de ativação mais usadas. Fonte: Aggarwal (2018).

No caso hipotético de usar as funções lógicas, AND, OR e XOR para separar problemas como o apresentado na Figura 13, observa-se que a função XOR (OU exclusivo) não é linearmente separável e é realmente impossível para um único hiperplano separar os conjuntos.

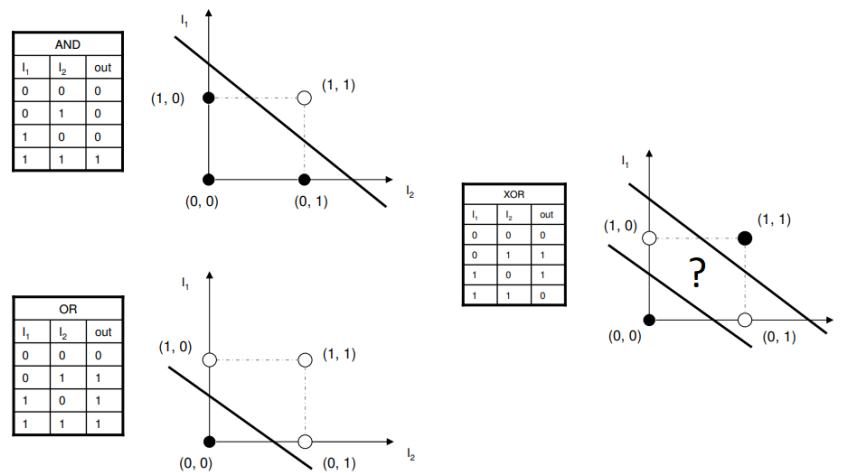


Figura 10 - Separações de conjuntos pelas funções lógicas AND, OR e XOR. Fonte: Araújo (2018).

Utilizando as funções de ativação na separação de um conjunto pelo operador lógico OR com um único neurônio linear, a função de ativação sigmóide, leva o conjunto de dados de entrada para outra dimensão onde se torna linearmente separáveis, como o apresenta a Figura 11 (ARAÚJO, 2018).

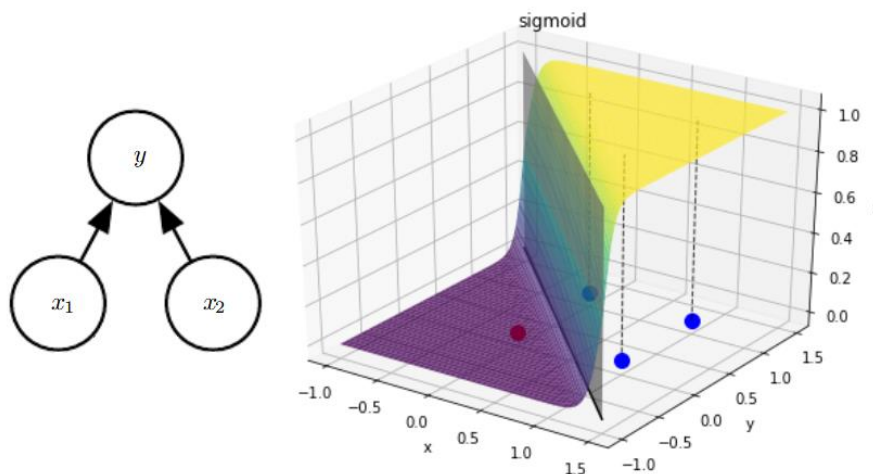


Figura 11 - Solução de separação de um conjunto pelo operador OR com um único neurônio linear “perceptron” usando uma função de ativação sigmóide. Fonte: Araújo (2018).

A Figura 11 mostra que o hiperplano aprendido pelo perceptron é determinado pela equação $y: 15(x_1 + x_2 - 0,5) = 0$, a qual é decomposta em duas partes, um fator constante que afeta diretamente a nitidez da curva sigmoidal, e a equação para um hiperplano que separa o espaço de entrada do neurônio (ARAÚJO, 2018).

3.10.2.6 Função de custo (loss)

A escolha da função *loss* é fundamental para definir as saídas em uma forma que é sensível à aplicação em questão e a sua meta é medir e verificar quão bem o algoritmo utilizado está desempenhando-se no conjunto de treinamento. Entre as funções *loss* mais populares no aprendizado de máquinas usadas atualmente encontram-se, o erro quadrado médio (MSE), *loss* de probabilidade (*likelihood loss*) e *loss* de *cross-entropy* (ou *log loss*) (BROWNLEE, 2019).

A função MSE é a mais fácil de entender, implementar e geralmente funciona bem na maioria dos algoritmos de aprendizado, o qual considera o somatório das diferenças quadráticas entre os valores estimados e observados, calculando a média em todo o conjunto de dados como o apresentado na equação 18. O seu resultado é sempre positivo independente do sinal dos valores observados e estimados, sendo o seu valor perfeito 0,0 (BROWNLEE, 2019).

$$MSE = \frac{\sum_{i=1}^n (y - \hat{y})^2}{n} \quad (18)$$

Tomando como exemplo o treinamento de uma rede neural a partir de um determinado conjunto de dados de treino e testes, é possível que dada a natureza estocástica do algoritmo, seus resultados específicos possam variar. Assim, a *loss* MSE durante os ciclos de treinamento (épocas) pode ser representada como na Figura 12 para os conjuntos de treino e testes (BROWNLEE, 2019).

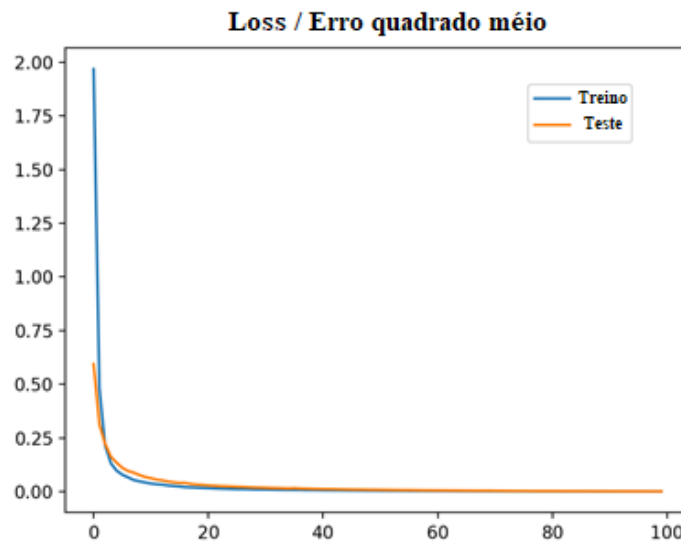


Figura 12 - Exemplo do comportamento da função *loss* MSE em função das épocas de treinamento de uma RNA. Representando o erro quadrado médio dos conjuntos treino e testes. Adaptado de Brownlee (2019).

Conforme mostrado na Figura 12, o MSE do modelo convergiu com rapidez, e os resultados do treinamento e dos testes foram os mesmos. Como resultado, o comportamento dos resultados e a convergência do modelo sugerem que o MSE é uma boa métrica estatística para avaliar se a rede neural aprende um problema em específico (BROWNLEE, 2019).

O objetivo do treinamento das redes é encontrar os pesos e *biases* que minimizem a função de custo, o que pode ser feito plotando a perda em relação ao número de pesos. No entanto, para fazê-lo com precisão, seria necessário ver de forma gráfica múltiplas dimensões que permitam analisar os diversos pesos e *biases* na rede. Como é difícil ver mais de três dimensões, é razoável supor, por exemplo, que a rede requer achar dois valores de pesos. Neste caso, a terceira dimensão pode ser usada para representar a função de custo (*loss*) em função dos pesos, conforme mostrado na Figura 13 (CAMPBELL, 2017).

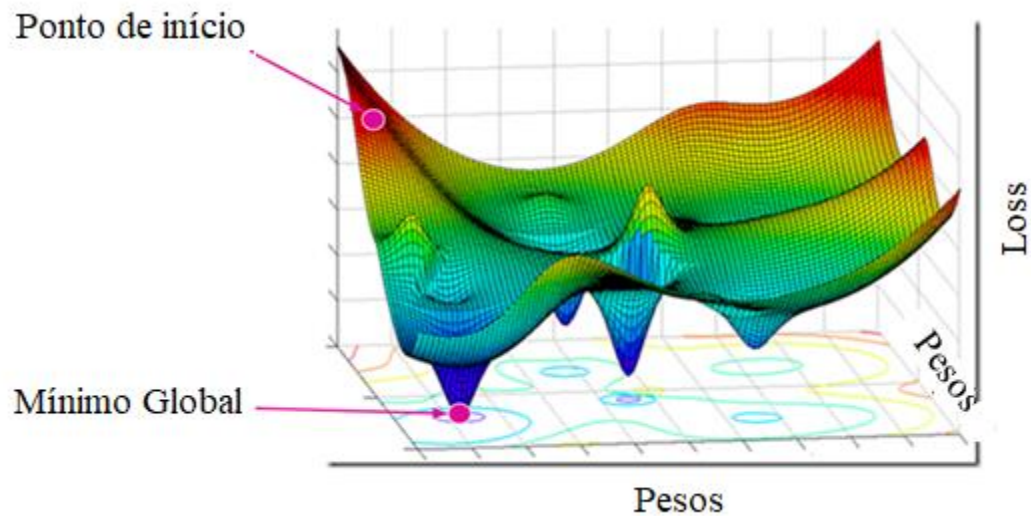


Figura 13 - Exemplificação da *loss* em função do ajuste dos pesos. Apresentado o ponto aleatório de início dos pesos e da *loss* e o mínimo global (ponto mais baixo de *loss* onde se quer chegar). Adaptado de Campbell (2017).

Antes do treinamento da rede, os pesos e *biases* são inicializados aleatoriamente, de modo que a função *loss* será provavelmente alta, provocando que a rede tenha uma quantidade de parâmetros errados. Assim, o objetivo é encontrar o ponto mais baixo da função *loss*, e então ver quais são os valores de pesos correspondentes para essa minimização do erro. Na Figura 13 pode ser visto facilmente onde é o ponto de início dos pesos e a sua respectiva *loss* quando inicializado o treinamento, além de observar o ponto mais baixo da *loss* e os seus correspondentes valores de pesos, porém essa visualização na vida real não é tão fácil porque a rede não tem uma boa visão geral da função *loss*, apenas podendo-se saber qual é a *loss* atual, os pesos e *biases* atuais (CAMPBELL, 2017).

Diante disso, a procura do ponto mais baixo da função *loss* é feita pelo uso da técnica chamada de gradiente descendente, onde se procura a direção que tem a inclinação mais íngreme, dando-se passos graduais nessa direção cuidando em não dar passos muitos grandes ou muitos pequenos. Por tanto, em termos de função *loss* da rede neural pode-se encontrar a direção da inclinação mais acentuada para baixo “gradiente descendente” na qual se deem pequenos passos pelo incentivo dos pesos nessa mesma direção. Inicialmente, a função *loss* terá um valor alto e a rede fará estimativas incorretas (Figura 13 e Figura 14), mas à medida que os pesos são ajustados e a função *loss* diminui, a rede começa a melhorar dando respostas mais corretas, como o apresentado

na Figura 14, onde é visualizada a descida da função *loss* à medida que se desce na paisagem “*landscape*”, melhorando as estimativas da rede, assim como sua confiança (CAMPBELL, 2017).

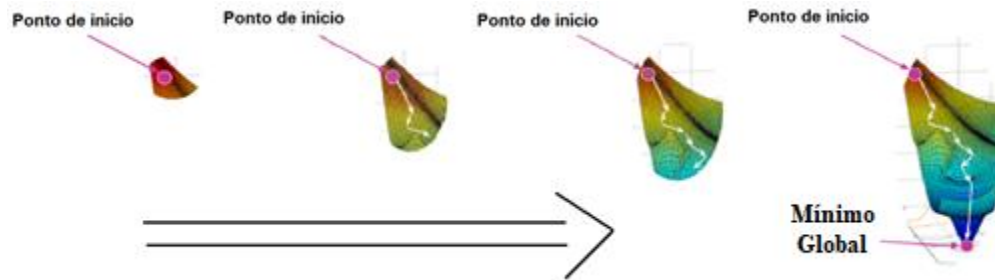


Figura 14 - Comportamento do gradiente descendente durante o treinamento da rede neural, apresentando o ponto de início dos pesos *loss* de forma aleatória até a chegada no mínimo global (ponto mais baixo de *loss*).

3.10.2.7 Treinamento das redes neurais e modelos de aprendizagem

No treinamento das redes de múltiplas camadas o problema à *loss*, a qual é uma função de composição complexa dos pesos nas camadas anteriores e onde o gradiente de uma função de composição é calculado utilizando o algoritmo de retropropagação (*backpropagation*), o qual é o método mais conhecido para o treinamento das redes de múltiplas camadas (AGGARWAL, 2018). Em sínteses, o processo de treinamento nesses tipos de arquiteturas, baseia-se no ajuste dos parâmetros da rede e dos pesos das conexões entre as unidades de processamento, mediante um processo iterativo que armazena no final o conhecimento que a rede obteve sobre o ambiente na qual está funcionando (SOARES, 2013).

3.10.2.8 Treinamento da rede neural com *backpropagation*

A partir do desenvolvimento do algoritmo *backpropagation* o treinamento de redes neurais com camadas intermediárias foi possível, permitindo o uso em grande extensão do *perceptron* de múltiplas camadas para a solução de diversos problemas de classificação e identificação, conhecida hoje em dia como arquitetura do tipo *Multilayer Perceptron* (MLP) (MITCHELL, 1997).

O algoritmo de *backpropagation* utiliza a regra da cadeia do cálculo diferencial para calcular os gradientes de erro em termos de somas de produtos do gradiente local sobre os diversos caminhos desde um nó para a saída. Embora esse somatório tem um número exponencial de componentes (caminhos), este pode ser calculado eficientemente utilizando programação dinâmica. Esse algoritmo possui duas fases principais, conhecidas como as fases de *forward* (para frente) e *backward* (para trás), a primeira é necessária para calcular os valores de saída e as derivadas locais em vários nós, e a segunda fase é necessária para acumular os produtos desses valores locais em todos os caminhos desde o nó para a saída (DEISENROTH; FAISAL; ONG, 2020).

Na fase *forward* as entradas para um exemplo de treinamento são inseridas na rede neural, resultando em uma cascata de cálculos para frente através das camadas, usando o conjunto atual de pesos. A saída final estimada pode ser comparada com o exemplo de treinamento, e a derivada da função *loss* em relação à saída é calculada, assim a derivada dessa *loss* precisa ser calculada com respeito aos pesos em todas as camadas na fase *backward* (DEISENROTH; FAISAL; ONG, 2020).

Durante a fase *backward* o principal objetivo é aprender o gradiente da função *loss* com relação aos diferentes pesos pelo uso da regra da cadeia do cálculo diferencial, sendo os gradientes utilizados para atualizar os pesos, uma vez que os gradientes são aprendidos na direção *backward*, começando desde o nó de saída (DEISENROTH; FAISAL; ONG, 2020).

3.10.2.9 Principais problemas práticos no treinamento das redes neurais

Apesar do grande desempenho das redes neurais para realizar ótimas aproximações de funções universais, ainda existem desafios consideráveis em relação ao treinamento de redes neurais para fornecer esse nível de desempenho, os quais estão relacionados a vários problemas práticos durante o treinamento, destacando-se o *overfitting*; problema de fuga; explosão e dissipação de gradientes.

3.10.2.9.1 Sobreajuste nas RNAs

Como anteriormente mencionado, o problema de *overfitting* refere-se ao fato que o ajuste de um modelo a um conjunto de treinamento específico não garante que ele proporcionará um ótimo desempenho de estimativa sobre dados não vistos, mesmo se o modelo estimar perfeitamente as saídas nos dados de treinamento. Em outras palavras, há sempre uma lacuna entre o desempenho dos dados de treinamento e testes, que sempre é particularmente grande quando os modelos usados são complexos e o conjunto de dados é pequeno. Aumentar o número de exemplos de treinamento melhora o potencial de generalização do modelo, enquanto aumentar a complexidade do modelo reduz seu potencial de generalização. Ao mesmo tempo, quando muitos dados de treinamento estão disponíveis é pouco provável que um modelo muito simples consiga capturar as relações complexas entre as entradas e saídas (RASCHKA, 2018).

No geral, modelos com um grande número de parâmetros têm alta capacidade e requerem uma quantidade maior de dados para obter poder de generalização para dados de testes não vistos. A noção do *overfitting* no AM é entendida no *trade-off* entre *bias* e *variância*, onde a principal vantagem é que nem sempre se ganha com modelos mais poderosos (menos tendenciosos) quando se trabalha com dados de treinamento limitados, devido à maior variância desses modelos (JAMES *et al.*, 2017).

As redes neurais têm sido sempre conhecidas por serem poderosas o suficiente para aproximar qualquer função, no entanto, a falta de dados disponíveis pode resultar em baixo desempenho, sendo um dos motivos pelos quais as redes neurais só ganharam destaque recentemente, devido à maior disponibilidade de dados, os quais têm revelado as vantagens das RNAs sobre o AM tradicional. No geral as RNAs exigem um design cuidadoso para minimizar os efeitos prejudiciais do *overfitting*, mesmo quando se tem acesso a uma grande quantidade de dados. Assim existem algumas técnicas de design que são mais utilizadas para mitigar o impacto do *overfitting* como, a regularização, a parada antecipada “*early stopping*”, a arquitetura neural e o compartilhamento de parâmetros (BORGESSEN; SCHAAP, 2005).

3.10.2.9.2 Regularização

Uma vez que um número grande de parâmetros causa *overfitting*, uma abordagem comum é restringir o modelo para utilizar menos parâmetros diferentes de zero, mediante a adição do valor quadrado de cada parâmetro (multiplicado pelo parâmetro de regularização $\lambda > 0$) à função objetivo. Um exemplo de regularização seria o aplicado para os *mini-batches* do subconjunto de treinamento S na equação 6 e a atualização do $\alpha > 0$ como o apresentado na equação 19 (CHOLLET, 2020).

$$\bar{W} \leftarrow \bar{W}(1 - \alpha\lambda) + \sum_{\bar{x} \in S} E(\bar{X})\bar{X} \quad (19)$$

Onde $E(\bar{X})$ representa o erro atual ($y - \hat{y}$) entre os valores estimados e observados dos exemplos de treinamento $\bar{X}$. A regularização é particularmente importante quando a quantidade de dados é limitada, que pode ser interpretada biologicamente como o esquecimento gradual, como resultado do qual os padrões menos importantes (ou seja, ruidosos) são removidos. Em geral, é aconselhável usar modelos mais complexos com regularização, no lugar de modelos mais simples sem regularização (CHOLLET, 2020). Como o observado na equação 19, esta é uma forma general usada por muitos modelos de AM regularizados como a regressão de mínimos quadrados, em que $E(\bar{X})$ é substituído pela função de erro desse modelo específico.

3.10.2.9.3 Parada antecipada

As redes neurais são treinadas utilizando variações dos métodos de gradiente descendente com o objetivo da sua otimização, nas quais esse método é executado para atingir a convergência e otimizar a *loss* nos dados de treinamento, porém não necessariamente a *loss* nos dados de testes é otimizada, o qual ocorre porque ao final do processo de treinamento as redes podem sobreajustar às sutilezas específicas dos dados de treinamento, não generalizando bem para os dados de testes.

Outra forma comum de regularização é a parada antecipada ou *early stopping*, na qual o gradiente descendente termina após apenas algumas iterações. Uma maneira de decidir o ponto de parada é reter uma parte dos dados de treinamento para depois usá-la como conjunto de dados para testar o erro do modelo (conjunto de validação), sendo a abordagem de gradiente descendente terminada quando o erro com o conjunto de validação começa a aumentar. O treinamento baseado no algoritmo de *backpropagation* é unicamente aplicado à porção de dados de treinamento que não inclui o conjunto de validação e ao mesmo tempo o erro do modelo no conjunto de validação é continuamente monitorado. Em algum momento, o erro começa a aumentar no conjunto de validação, mesmo que continue a diminuir no conjunto de treinamento, indicando que o treinamento nesse exato momento causa *overfitting*, podendo ser esse ponto escolhido para finalizar o treinamento.

O monitoramento com base nos cálculos sobre os dados de validação deve ser feito até alcançar a melhor solução no processo de aprendizado, uma vez que o *early stopping* não é realizado após pequenos aumentos no erro no conjunto de validação que podem ser causados por variações ruidosas, senão que se recomenda continuar com o treino para verificar se o erro continua a aumentar. Em outras palavras, o ponto de finalização do treino é escolhido em retrospectiva, depois que o erro no conjunto de validação continua aumentando, não se tendo esperança de melhorar o desempenho do erro no conjunto de validação.

Uma vantagem do *early stopping* é que este pode ser facilmente adicionado ao treinamento da rede neural sem alterar significativamente o processo de treinamento, enquanto métodos como redução de peso (*weight decay*) exigem que sejam tentados diferentes valores de parâmetros de regularização λ , que pode ser de alto custo computacional. Assim, o *early stopping* tem a facilidade de ser combinado com algoritmos existentes, o que lhe permite ser usado em combinação com outras técnicas de regularização de maneira relativamente direta, sendo quase sempre usado porque não se perde muito ao adicioná-lo ao processo de aprendizagem. Nesse sentido, o *early stopping* pode ser visto como um tipo de restrição no processo de otimização, porque restringe o número de etapas no gradiente descendente e o espaço entre a solução final e o ponto de inicialização é restringido (AGGARWAL, 2018; BORGESSEN; SCHAAP,

2005; MANZAN, 2016). Em conclusão, a adição de restrições ao modelo de um problema de AM geralmente é uma forma de regularização.

3.10.2.9.4 *Random search*

Na seleção dos hiper-parâmetros mais ótimos para melhorar o processo de aprendizado de um determinado algoritmo, existem técnicas como o *grid search* e *random search*. Durante o desenho de uma RNA é mais comum aplicar experimentos com a técnica *grid search* para explorar os melhores hiper-parâmetros. No entanto, esta técnica realiza a busca sobre um grid de tentativas baseadas em seu desempenho sobre um conjunto de validação, o que torna esta técnica mais custosa em tempo computacional. Diante disso, a técnica *random search* surge como uma alternativa mais fácil de otimização dos hiper-parâmetros da RNA. O *random search* gera aleatoriamente um número grande de soluções a um problema de otimização, normalmente geradas mediante uma distribuição uniforme, para escolher os melhores e possíveis valores de hiper-parâmetros dentro de um domínio limitado, tomando aleatoriamente alguns pontos de amostragem desses hiper-parâmetros no domínio assumido e aguardando por resultados subótimos (BERGSTRA; BENGIO, 2012; GALLEGO, 2004). De acordo com Shi e Ólafsson (2000) o *random search* na maioria das vezes atinge uma solução ótima e pode ser aplicado na maior parte dos problemas de otimização, devido aos poucos pressupostos que são necessários para o seu uso. Por outro lado, esta técnica tem como desvantagem que pode ser lento para achar uma solução ótima ou muito próxima ao ótimo, no entanto, este problema cada vez mais está mitigado pela grande evolução que têm os sistemas computacionais na atualidade.

3.10.3 *Aplicações de redes neurais artificiais para estimativa da CRAS*

O uso de RNAs nas ciências do solo aumentaram nos últimos anos com o objetivo de melhorar as estimativas dos atributos do solo, porque estas permitem um melhor desempenho em relação aos algoritmos tradicionais, tais como a regressão múltipla (PEREIRA *et al.*, 2018; PHAM *et al.*, 2019; SOARES *et al.*, 2014), apresentando potencial

para derivar FPTs que estimem os diversos atributos (HAGHVERDI; CORNELIS; GHAMRAN, 2012).

Entre as funções hidráulicas, o conhecimento e entendimento do comportamento da água no solo mediante a curva de retenção de água no solo (CRAS) é necessário para muitas aplicações na agricultura, engenharia e o manejo e conservação dos recursos naturais, assim a sua estimativa torna-se necessária. Frente a isso, diversos autores priorizaram a estimativa da água no solo mediante o uso de técnicas de inteligência artificial como as RNAs. Koekkoek e Booltink (1999) utilizaram RNAs para estimar a retenção de água no solo em vários potenciais matriciais (0, -100 e -15,000 hPa) com base nos bancos de dados de solos da Escócia e Holanda consideradas regiões de clima temperado, utilizando 3 combinações de variáveis de entrada com os atributos texturais, densidade e matéria orgânica do solo para estimar o volume de água. Os autores obtiveram que as redes neurais se desempenharam melhor que as funções de regressão com r^2 entre 0,80 e 0,93 e RMSE entre 0,0264 e 0,0476 $\text{cm}^3 \text{cm}^{-3}$.

Schaap, Leij e van Genuchten (2001) desenvolveram o programa Rosetta que implementa quatro FPT hierárquicas a partir de RNAs para a estimativa dos parâmetros de van Genuchten (1980) a partir de três bancos de dados de solos com a maioria de amostras variando de solos em clima temperado a tropical de Norte América e Europa. Esse programa, mediante a combinação de análise de redes neurais com o método *bootstrap*, permite além de proporcionar estimativas das incertezas dos parâmetros hidráulicos estimados. O ótimo desempenho geral do Rosetta para estimativa da retenção de água no solo foi caracterizado pela redução do RMSE de 0.078 a 0.044 $\text{cm}^3 \text{cm}^{-3}$ quando mais preditores foram usados.

Minasny e McBratney (2002) desenvolveram um pacote neural “*Neuropack*” para derivar FPT utilizando RNAs, o qual consiste de dois programas: Neuropath e Neuroman, ambos com uma interface amigável com o usuário a pesar de usar algoritmos robustos, onde o Neuropath permite desenvolver uma rede neural geral de camada única que pode modelar qualquer relação entre entradas e saídas ou gerar FPT baseadas no método de *bootstrap* que estimam pontos de retenção de água em um ciclo de calibração, validação e predição, enquanto Neuroman é uma rede neural que estima FPT paramétricas mediante a combinação de RNAs e o método *bootstrap*. Os mesmos autores

demonstraram que o *Neuropack* quando comparado ao programa Rosetta teve 30% melhor exatidão ($RMSE = 0,04 \text{ m}^3 \text{ m}^{-3}$) e menor *bias*.

Mais recente Haghverdi, Cornelis e Ghahraman (2012) estabeleceram um novo enfoque de geração de FPT conhecidas como pseudo-contínuas para estimar a retenção de água no solo a partir de uma RNA, a qual tem um desempenho contínuo porque é capaz de estimar o volume de água em qualquer potencial matricial desejável sem a necessidade de usar uma equação específica. Nessa abordagem pseudo-contínua o logaritmo natural do potencial matricial é considerado como um parâmetro de entrada, assim permitindo aumentar o número de amostras no conjunto de dados de treinamento por um fator igual ao número de pares de potenciais matriciais utilizados para determinar a CRAS no conjunto de dados (BOTULA; RANST; CORNELIS, 2014). Contudo, existe apenas um parâmetro de saída, gerando o volume de água (θ) no potencial matricial predefinido (h), quer dizer, diferentes valores do potencial matricial produzem diferentes volumes de água. Ao utilizar uma ampla gama de potenciais matriciais como parâmetros de entrada será estimada uma faixa correspondente do volume de água e obtida uma curva (pseudo) contínua (HAGHVERDI; CORNELIS; GHAHRAMAN, 2012).

As amostras de solo dos bancos de dados utilizados no desenvolvimento da FPT pseudo-contínua eram provenientes de solos no norte e nordeste do Irã (clima subtropical a tropical) e solos da Austrália pertencentes ao banco de dados *Neuropack* com predominância de climas desérticos e semiáridos. Os autores encontraram que as funções pseudo-contínuas desenvolvidas para esse banco de dados tiveram desempenho ligeiramente melhor ($RMSE = 0,027 \text{ m}^3 \text{ m}^{-3}$) que as FPT pontuais e paramétricas desenvolvidas ($RMSE = 0,029 \text{ m}^3 \text{ m}^{-3}$).

O tipo de FPT pseudo-contínua pode ser útil para a melhoria e desenvolvimento de FPTs baseadas em redes neurais artificiais (RNAs) quando se apresentam restrições como o número limitado de amostras de solos. Na maioria dos países existe uma carência de grandes e consistentes bancos de dados de solo, frente a isso, a FPT pseudo-contínua mostra-se como uma ferramenta potencial nas estimativas dos parâmetros das funções e de propriedades hidráulicas do solo. Apesar de ser gerada uma RNA pseudo-contínua a partir de poucas amostras, o analista não deve ignorar as limitações físicas de estimar a CRAS usando dados limitados, enfatizando-se que ao usar uma FPT de RNA de

abordagem pseudo-contínua a complexidade do relacionamento entre entrada-entrada e entrada-saída também aumentou em relação com as FPTs de RNA tradicionalmente geradas, abrangendo mais a não linearidade (HAGHVERDI; ÖZTÜRK; CORNELIS, 2014).

Resumindo as principais tipologias de FPTs que podem ser desenvolvidas a partir de RNAs, a Figura 15 desenvolvida por Haghverdi, Cornelis e Ghahraman (2012) apresenta uma representação esquemática das FPTs pontuais, paramétricas e pseudo-contínuas. As três FPTs baseadas em RNAs são arquiteturas para solucionar problemas de regressão, sendo as duas primeiras utilizadas para estimar múltiplas saídas.

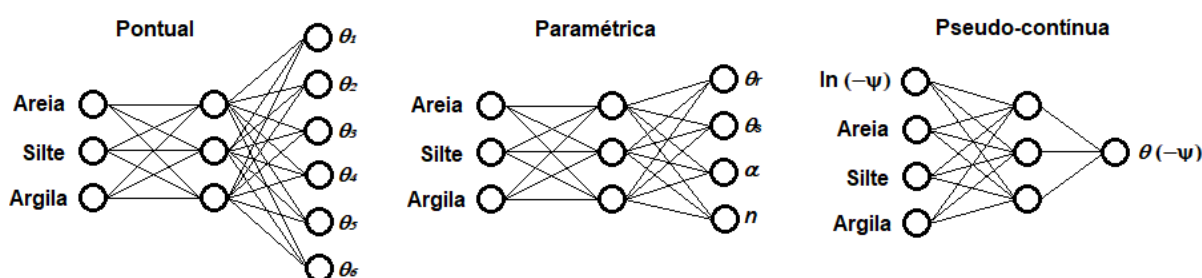


Figura 15 - Topologias típicas de FPTs baseadas em redes neurais artificiais para solucionar problemas de regressão: pontuais, paramétricas e pseudo-contínuas (FPT-PC). Nos nodos de entradas se encontram os atributos de solo e na FPT-PC é adicionado o potencial matricial $(-\psi)$. Adaptado de Haghverdi, Cornelis e Ghahraman (2012).

Bayat, Mazaheri e Mohanty (2019) desenvolveram FPTs para estimar a umidade no solo em diferentes potenciais, baseado no modelo de van Genuchten. As variáveis preditoras foram atributos topográficos, propriedades térmicas, propriedades físicas e características geográficas de solos da parte central dos Estados Unidos. Os autores utilizaram a técnica de RNAs para gerar 21 FPTs com as diversas variáveis de entrada no software Neuroman. Durante o treinamento da RNA foi utilizada a tipologia multilayer perceptron *feedforward* com uma camada oculta (6 neurônios), com a função de ativação sigmoidal ($\tanh$) na camada oculta e uma função linear na camada de saída. Posteriormente ao treinamento da rede, o desempenho da rede em estimar a água do solo em todos os potenciais foi avaliado mediante as métricas estatísticas de critério de informação de Akaike, RMSE e r^2 . Os autores observaram que, mediante as RNAs, as propriedades físicas do solo (textura, densidade, carbono orgânico e condutividade hidráulica) são as mais apropriadas para estimar a retenção de água no solo e os

parâmetros do modelo de van Genuchten. Entre as propriedades geográficas e topográficas, unicamente as características de vegetação melhoraram significativamente a acurácia das estimativas ($0,88 \leq r^2 \leq 0,91$). Utilizando as propriedades térmicas como preditores, os autores notaram que a acurácia das estimativas aumentou significativamente ($r^2 = 0,99$) em comparação com as FPTs que usaram as propriedades geográficas e topográficas, demonstrando que tanto as propriedades físicas do solo como as propriedades térmicas são úteis na estimativa da retenção de água no solo.

Apesar do diverso desenvolvimento de RNAs para estimar a retenção de água no solo, as RNAs têm sido raramente utilizadas para estimar esse atributo em solos de condições de clima subtropical, representativo do sul do Brasil. Por esse motivo, se faz necessário que neste estudo se aprimore o desenvolvimento de FPTs para condições de clima subtropical, as quais podem contribuir com a redução das incertezas geradas ao estimar a retenção de água no solo, incertezas as quais são produto do uso de FPTs desenvolvidas para outras condições climáticas, como as FPTs de clima temperado e tropical. Durante o desenvolvimento desta pesquisa se propõe o uso de uma RNA que utilize como variáveis preditoras atributos físicos e topográficos do solo para a estimativa da retenção de água no solo de clima subtropical, o qual tem sido pouco explorado na literatura. Além disso, a proposta de desenvolver uma RNA mediante a técnica de função de pedotransferência pseudo-contínua para esta região, permitirá o desenvolvimento de modelos com maior precisão em regiões que tenham poucos dados de atributos que são utilizados como variáveis preditoras da retenção de água no solo.

Para o desenvolvimento dos artigos da presente tese foram utilizados os dados coletados na Bacia Hidrográfica Arroio Pelotas, produto da dissertação de Oliveira (2015) e os dados da Bacia Hidrográfica Sanga Ellert, produto da dissertação de Soares (2018).

4 Artigo 1

Examining the performance of pedotransfer functions to estimate the soil water status and evapotranspiration in a Brazilian subtropical watershed

Miguel David Fuentes-Guevara, Robson André Armindo, Luís Carlos Timm, Atilla Nemes, Alessandra Calegari da Silva, Luciana Köpp Marini, Tonismar dos Santos Pereira, Leandro Sanzi Aquino.

Abstract

Soil hydraulic properties and functions are essential for modeling and evaluating soil processes as water storage, solute transport, infiltration, drainage, and water balance. The use of pedotransfer functions (PTFs) for their estimation is a viable alternative when measurements are difficult, expensive, and time-consuming. Nonetheless, the applicability of PTFs developed for soils under temperate and tropical climates has rarely been evaluated for subtropical soils. This study aimed to examine the performance of parametric- and point-PTFs developed with data from temperate and tropical climates to estimate soil water retention under a subtropical climate. In addition, the PTF estimates were applied to simulate the actual evapotranspiration behavior as a function of the soil water content (θ) in different land-uses on a watershed scale. The study was carried out in the Pelotas River Watershed (PRW), Southern Brazil, where one hundred points were sampled in the 0-0.20 m soil layer, equidistantly spaced along a 25-km-spatial transect. For each point, clay, silt, fine sand, total sand, and organic carbon contents, bulk density, particle density, macroporosity, microporosity, total porosity, and the soil water retention curve (SWRC) were measured. The

parametric PTFs were used to estimate the parameters of the van Genuchten-Mualem model that were applied for estimating θ at matric potentials (h) of -33 and -1,500 kPa as well as the available water (AW) content, which is widely used in crop and water management modeling. The point-based PTFs selected were also used for estimating $\theta(-33 \text{ kPa})$, $\theta(-1,500 \text{ kPa})$, and AW. Each PTF was evaluated according to statistical measures of correlation and uncertainty. Results showed that the parametric-PTFs were more accurate than point-PTFs, highlighting that Barros *et al.* (2013)-4V (tropical climate) and Vereecken *et al.* (1989) (temperate climate) models presented the best performance to estimate $\theta(-33 \text{ kPa})$, $\theta(-1,500 \text{ kPa})$, and AW. Measured and estimated ET varied according to the land use and as a function of the θ_{fc} and θ_{pwp} results.

Keywords: parametric PTFs; point PTFs; soil hydraulic parameters; soil water retention; transect sampling.

Introduction

Soil hydraulic properties and soil-water functions (*e.g.*, saturated hydraulic conductivity and the water retention curve) are key characteristics to help understand processes of the hydrological cycle and are critical parameters that help depict the physical reality in a watershed in simulation models. However, these soil properties are not easily available from soil databases since they are difficult, time-consuming, and expensive to obtain (Freitas *et al.*, 2019), mainly at a watershed scale. To overcome this, pedotransfer functions (PTFs) have been applied to the soil and environmental sciences (Pringle *et al.*, 2007) for estimating saturated soil hydraulic conductivity, water content retained at specific matric potentials, and water retention curve at the watershed scale. PTFs have been developed to estimate soil-water retention (Bouma, 1989) from simpler, widely available, and easy-to-obtain soil databases determined *in situ* or laboratory.

Bouma (1989) has divided PTFs into continuous and class, however, a more detailed classification as continuous, point, parametric, and pseudo-continuous PTFs has been used (Wösten, 1995; Minasny *et al.*, 1999; Botula *et al.*, 2014). Botula *et al.* (2014) reported that both texture-class-based- and parametric-PTFs are used to estimate soil hydraulic properties. When estimates of water content (θ) at specific matric potentials (h) are desired, *e.g.*, field capacity or permanent wilting point, parametric and point-based PTFs have been used (Schaap *et al.* 2001; Wösten *et al.* 1999; Gupta and Larson, 1979; Rawls and Brakensiek, 1989, among others). The advantage of parametric over the point-based PTFs is that they can generate any $\theta(h)$ data between $\theta(0)$ and $\theta(1,500 \text{ kPa})$, whereas the point-based PTFs usually provide specific values, *e.g.*, $\theta(-33 \text{ kPa})$ and $\theta(-1,500 \text{ kPa})$. For this reason, parametric PTFs are more commonly applied in hydrological models.

Although the majority of these PTFs were developed for temperate climate conditions (temperate PTFs), they have been widely applied for tropical climate conditions (Botula *et al.*, 2012), which may differ in soil types, mineralogy, and physicochemical properties due to climate conditions.

Botula *et al.* (2012) evaluated the capability of several published parametric- and point-PTFs derived from tropical and temperate soils to predict water retention at -33 and -1500 -kPa-matric potentials of soils under tropical conditions in the Lower Congo. They found that the majority of examined parametric- and point-PTFs derived from tropical soils performed well for estimating $\theta(-33 \text{ kPa})$ and $\theta(-1,500 \text{ kPa})$. They also argued that the estimation of $\theta(h)$ data by those temperate and tropical PTFs may induce output errors when watershed models are applied to agricultural studies due to large discrepancies in the derived soil hydraulic data, which substantially diminish the quality of the modeling results. These discrepancies can be associated

with sparse and old data, soil-forming factors, different climatic and pedological conditions, and the type of soil (Minasny and Hartemink, 2011).

Nebel *et al.* (2010) assessed the applicability of eight point PTFs of tropical climate developed by several authors (Pidgeon, 1972; Lal, 1972; Arruda *et al.*, 1987; Dijkerman, 1988; Bell and van Keulen 1995, 1996; Oliveira, 2002; Peraza, 2003; Urach, 2007) to estimate soil water contents at -33 and -1,500 kPa at the field-scale for two Brazilian lowland soils under subtropical climate. The authors addressed acceptable results for the tested tropical PTFs and applied them to the experimental area. Thus, the use of suitable PTFs can support better watershed management or the simulation of water flow in the heterogeneous vadose zone when measurements of soil water content are required on a greater scale. Besides, the use of specific computer models to the understanding of the water flow and hydrological processes in a watershed requires the determination of soil hydraulic properties in a large number of points within the area of interest (Romano and Santini, 1997). On the other hand, few studies have assessed the performance of temperate or tropical PTFs to estimate soil hydraulic properties and functions for subtropical climate conditions mainly at the watershed scale (Zou *et al.*, 2016).

The Pelotas River Watershed (PRW) is an important subtropical watershed in the south of Rio Grande do Sul (RS) state, Brazil. Its main watercourse, the Pelotas River, supplies drinking water for the population of Pelotas city, which has about 329,000 inhabitants (Beskow *et al.*, 2016). The PRW is susceptible to water erosion and floods, and has suffered the impacts of human activities that have caused soil and nutrient losses resulting in economic and social issues. Therefore, the comprehension and quantification of hydrological processes at the watershed scale (e.g., surface and subsurface runoff, water balance, and sediment transport, among others) play a crucial role in the application of hydrological models in decision-support on water resource

management and to characterize the effect of land use on the watershed streamflow (Alvarenga *et al.*, 2016; Mello *et al.*, 2019). In this sense, PTFs used for estimating soil hydraulic properties and soil-water functions over temperate or tropical watersheds, using attributes more easily determined in the field or laboratory, are expected to strongly contribute to the better use of hydrological models for subtropical watersheds and the understanding of their outputs as well. Hence, the objectives of this study were to analyze the performance of temperate PTFs, tropical PTFs, and a physically-based PTF model for estimating soil water retention along a spatial transect established in the Pelotas River Watershed under subtropical climate and to examine the potential of these PTFs to estimate the actual evapotranspiration behavior as a function of water retention in different land-uses.

Material and Methods

Study area

This study was carried out in the Pelotas River Watershed (PRW), situated in Southern Rio Grande do Sul (RS) state, Brazil, where the drainage area upstream from the outlet is named Ponte Cordeiro de Farias (Figure 1). PRW has an area of about 370 km² and an elevation range between 37 and 517 m. Based on the Köppen classification, the climate of the region is Cfa type, indicating a wet subtropical climate with mean temperatures above 22°C during the hottest month of the year (Kuinchtner and Buriol, 2001). The soils found in the PRW are Hapludalf + Udorthent, Hapludalf + Hapludult, Udorthent, and Paleudult, according to the US Soil Taxonomy classification (Soil Survey Staff, 2014). Further information about the climatic conditions and description of this study area can be found in Dongli *et al.* (2017).

Insert figure 1

Soil data and experimental design

A 25-km-spatial transect was established in the PRW, starting from the control point, “Ponte Cordeiro de Farias”, up to the point furthest away in the north-west direction, transecting the watershed to include the greatest number of soil types. Along the transect, one hundred points were located and spaced at equal distances (250 m). Field coordinates were marked using a Global Positioning Satellite instrument and then plotted on the watershed map using *ArcGIS* software.

Soil samples were taken in the 0-20 cm soil layer along the transect at each point marked, collecting disturbed samples in the middle of this layer to determinate the particle density (Blake and Hartge, 1986), the particle fractions contents of sand, fine sand, silt, and clay using the USDA classification scheme by the sedimentation method (Gee and Bauder, 1986). In this method, sampling particle fractions were taken using a pipette. The textural distribution of the soils along the 25-km-spatial transect was coarse-textured with large sand content with ranges of textures of sand (35.7-91.6%), fine sand (4.6-21.3%), silt (1.5-42.3%), and clay (6.5-32.5%). Additionally, the organic carbon (OC) content of the soil samples was analyzed by the wet combustion (Walkley-Black) method and multiplied by 1.724 to conversing of the organic matter (OM) content. OC and OM were used to feed different PTFs as input parameters.

Undisturbed soil cores were also sampled at each point along the transect in the middle of the layer using sampling rings of 5-cm-diameter $\times$ 3-cm-length to measure the soil bulk density (BD) by the ratio of dried soil mass (at 105°C until reaching a constant dried mass) and the inner ring volume. The soil water retention curves (SWRCs) were built by measuring the volumetric soil water content (θ) at seven matric potentials, 0 kPa determined after sample saturation, -1, -6, and

-10 kPa using tension table and after that in Richards pressure chamber apparatus at -33, -100, and -1,500 kPa. As the majority of the analyzed point-PTFs estimates $\theta(-33 \text{ kPa})$, this value was assumed in this study as field capacity (Turek *et al.*, 2020; Armindo and Wendroth, 2016; Ottoni Filho *et al.*, 2016). The value of -1,500 kPa was considered as an approximation of the permanent wilting point (PWP). Many authors have considered these two points in the development of point-PTFs of many soils (Gupta and Larson, 1979; Rawls *et al.*, 1982; Aina and Periaswamy, 1985; Arruda *et al.*, 1987; Lal, 1987; Dijkerman, 1988; Bell and van Keulen, 1996; Oliveira *et al.*, 2002; Minasny and Hartemink, 2011; Botula, 2013; Santos *et al.*, 2013; Gunarathna *et al.*, 2019).

Land use data

A land use map was generated to the PRW. The map was constructed with various images obtained from the Brazilian Institute for Space Research (INPE) website and Landsat satellite. The utilized images belonged to Scenes 221_81 and 222_82, which were interpreted using the Environment for Visualization Images (ENVI) software (Perrota, 2005). Afterward, the land use map was developed using a maximum likelihood supervised classification method and the data of land use were validated based on the interpretation of the satellite images and through field visits. The validation was realized along with the 100 training samples established in the PRW with a portable GPS. Thus, these samples allowed us to identify six main land-uses in the PRW and the construction of its land use map, which showed the following distributions along the transect: cultivated pasture (6%), native grassland (32%), native forest (24%), silviculture (7%), annual (19%), and permanent (12%) crops.

Exploratory data analysis

All data sets were submitted to classical statistical analysis in which mean, standard deviation, coefficient of variation (CV), minimum and maximum values, and skewness and kurtosis coefficients were calculated. The normality of all data distributions was also assessed using the Kolmogorov-Smirnov (K-S) test at a 5% probability level. The R software, package “agricolae” version 4.3.1 was used to perform all calculations (R Core Team, 2019). The same descriptive statistics procedure was applied to the PTFs estimates together with the Kolmogorov-Smirnov test for the assessment of the normality.

Parameterizing soil water content

The parameters of the van Genuchten (1980)-Mualem (1976) (VGM) model were fitted to experimental SWRCs as:

$$\theta(h) = \theta_r + \frac{\theta_s - \theta_r}{\left[1 + (\alpha|h|)^n\right]^{1-1/n}} \quad (1)$$

in which θ is the measured volumetric soil water content ($\text{cm}^3 \text{ cm}^{-3}$) as a function of the matric potential h (kPa), θ_r is the residual water content ($\text{cm}^3 \text{ cm}^{-3}$), θ_s is the measured saturated water content ($\text{cm}^3 \text{ cm}^{-3}$), and n and α (kPa^{-1}) are fitting parameters. The SWRC software (version 3.0 beta) was used to obtain θ_r , n , and α parameters (Dourado-Neto *et al.*, 2001).

Examined PTFs

Six PTFs were selected from literature to estimate specific $\theta(h)$ data because they had the best performance in the dataset in a previous analysis. Four parametric-PTFs were contrasted against two-point PTFs in the estimation of specific $\theta(h)$ data, *e.g.*, $\theta(-33 \text{ kPa})$ and $\theta(-1,500 \text{ kPa})$

(Table 1). These PTFs were chosen based on the following criteria: according to the localization where their soil data calibrations are originated, representing temperate or tropical climatic conditions and because they were fitted with a large number and variability of soil samples. Furthermore, two options of the PTF1 (texturally-based) of the physical-empirical-model Splintex 2.0 (Silva et al., 2020) was analyzed, since it does not require calibration and can be applied to any porous medium (Table 1).

Insert table 1

The values of $\theta(-33 \text{ kPa})$ (hereafter assumed as θ_{fc}) and $\theta(-1,500 \text{ kPa})$ (assumed as θ_{pwp}) are point-based data widely used in crop and water management modeling. They were calculated through the fitted parameters of equation 1 using measured and estimated data by parametric- and point-based PTFs and the first hierarchical PTF of Splintex model (PTF1), which presents two options of simulation: SimA, where only parameters θ_r , α , and n are estimated and θ_s is equated to total porosity and SimB, where all parameters θ_s , θ_r , α , and n are estimated (Table 1). The second hierarchical PTF of Splintex (PTF2), where an experimental $\theta(h)$ value is inserted as a predictor was not assessed in this study. Furthermore, the available water (AW) was assessed for all points in the transect by $AW = (\theta_{fc} - \theta_{pwp})$ for all PTF scenarios.

Evapotranspiration estimation analysis

Measured and estimated results of θ_{fc} and θ_{pwp} were used to estimate the actual evapotranspiration (ET) behavior in each land use through the system of equations of Bergström (1992):

$$ET(\theta) = \begin{cases} PE, & \theta \geq \theta_{fc} \\ \left(\frac{\theta - \theta_{pwp}}{\theta_{fc} - \theta_{pwp}} \right) PE, & \theta_{pwp} < \theta < \theta_{fc} \\ 0, & \theta \leq \theta_{pwp} \end{cases} \quad (6)$$

in which PE is the potential evapotranspiration (mm month⁻¹) and θ is the soil water content (cm³ cm⁻³). According to Turek *et al.* (2018), this system of equations provides ET as a function of θ , depending on the range of available water, $AW = (\theta_{fc} - \theta_{pwp})$, and PE.

The measured data of PE represent the PWR and were taken from the Pelotas agro-meteorological station in the municipality of Capão do Leão-RS. The climatological normal data set from 1971 to 2000 showed that the maximum monthly PE in this region was 153.7 mm month⁻¹. PE was fixed to the maximum monthly value, considering the same climatological period with normal distribution, to perform a sensitivity analysis of the parameters θ , θ_{fc} , and θ_{pwp} measured and estimated by PTFs, because PE is used for irrigation design and management in the region.

Graphs of $ET(\theta)$ were plotted to compare the measured and predicted ET in each land use aforementioned. Consequently, the ET lines yielded by PTFs that stayed closer to the ET lines calculated with the measured data were used as a selection criterion of the best PTF to estimate the soil water status in each land use of the PRW under subtropical climate conditions along with the previous analysis.

Statistical indices of performance evaluation

The performance of predictive capacity of the examined PTFs was done by means of precision and accuracy analyses. Precision was associated with random errors, quantified by the correlation coefficient (r) (or the determination coefficient, r^2) and the standard deviation of the

estimation error (SDE). The accuracy evaluation was done by means of systematic errors taking into account the mean error (ME) and the root mean square error (RMSE), expressed as follow:

$$ME = \frac{1}{N} \sum_{i=1}^n (e_i - m_i) \quad (2)$$

$$RMSE = \sqrt{\frac{1}{N} \sum_{i=1}^n (e_i - m_i)^2} \quad (3)$$

$$SDE = \sqrt{\frac{1}{N-1} \sum_{i=1}^n [(e_i - m_i) - ME]^2} \quad (4)$$

$$MSE = ME^2 + \left(\frac{N-1}{N} \right) SDE^2 \quad (5)$$

in which N is the number of sampling points, e_i and m_i are, respectively, the estimated and measured data. While RMSE, SDE, and MSE generate positive uncertainties of estimation, ME provides positive or negative results showing whether PTFs tend to overestimate or underestimate the aim-variable. The index MSE of eq. (5) associates random and systematic uncertainties when a big number of data is analyzed. In our study the term $N/(N-1)$ means more than 99% due to the amount of data investigated, thus integrating precision and accuracy bias into MSE. Check De Vos et al. (2005) for more information about theoretical aspects of ME, RMSE, SDE, and MSE.

Results and discussion

Exploratory analysis of soil properties along the transect.

The summary statistics with the exploratory analysis of the soil properties used to examine the PTFs are shown in table 2.

Insert table 2

Following the Wilding and Drees (1983) criterion for classifying the variation of the key-properties, the coefficient of variation (CV) ranged from 9.82% (BD) to 47.97% (macroporosity) (Table 2). BD, particle density, OC, OM, and total porosity presented small variability concerning mean ($CV < 15\%$), whereas clay, silt, sand, fine sand contents, and microporosity exhibited moderate variability with CV ranging from 15 to 35% and the macroporosity showed high variability ($CV > 35\%$).

The skewness coefficient ranged from -1.41 to 0.87 and kurtosis varied from -0.73 to 1.52. The majority of kurtosis results were positive, which indicates a leptokurtic distribution with a greater concentration of data, except for clay content and macroporosity, which presented a negative value describing a platykurtic distribution. Press *et al.* (2007) suggested that the skewness can be useful if its absolute value is out of the range quantified by $\sqrt{6/N}$ (referred to standard error when the mean ($\bar{x}$) is estimated by the sample mean). In this study ($N=100$), the skewness should be out of the range -0.24 to 0.24 to be considered reliable, therefore, silt, OM, and total porosity contents presented positive skewness, whereas BD and fine sand content showed negative skewness, and there was no skewness in the data of other variables.

Skewness and kurtosis of clay, sand, BD, macroporosity, and microporosity were close to zero, which suggests a normal distribution with the mean values indicating the center point of the distribution. This exploratory analysis was confirmed by the results of the Kolmogorov-Smirnov test (Table 2). The pattern frequency of all properties followed the normal distribution, since the found value is below the critical value of 0.136 ($P\text{-value} < 0.05$).

Estimation and evaluation of the $\theta(h)$ points using the selected PTFs

The accuracy analysis of the PTFs used to estimate $\theta(h)$ values of soils along the transect in the PRW was quantified by means of ME, SDE, and RMSE, respectively. The integrated analysis of the precision and accuracy of the PTFs was quantified by means of MSE values.

Insert table 3

The PTFs selected had contrasting estimation accuracy. RMSE was used as the main evaluation criterion because it can show the overall estimation error of the PTFs. Among parametric-PTFs (Table 3), the best PTF to estimate $\theta(-33 \text{ kPa})$ was found with Vereecken *et al.* (1989)-PTF (RMSE= $0.0553 \text{ cm}^3 \text{ cm}^{-3}$), followed by the PTF of Barros *et al.* (2013)-4V (RMSE= $0.0636 \text{ cm}^3 \text{ cm}^{-3}$). On the other hand, both simulations of the Splintex 2.0-PTF1 yielded similarly and had the worst RMSE values (0.1217 and $0.1315 \text{ cm}^3 \text{ cm}^{-3}$, respectively) to estimate $\theta(-33 \text{ kPa})$. Regarding the estimation of $\theta(-1500 \text{ kPa})$, the PTF of Wösten *et al.* (1999) had the best accuracy with RMSE= $0.0510 \text{ cm}^3 \text{ cm}^{-3}$ followed closely by Barros *et al.* (2013)-4V (RMSE= $0.0580 \text{ cm}^3 \text{ cm}^{-3}$), while Splintex 2.0-PTF1-sim A and B had the worst accuracy for estimating $\theta(-1,500 \text{ kPa})$, with RMSE of 0.0994 and $0.1526 \text{ cm}^3 \text{ cm}^{-3}$, respectively. Across all matric potential values, the parametric-PTF of Barros *et al.* (2013)-4V overall resulted in the lowest RMSE amongst all tested parametric-PTFs.

Among point-PTFs (Table 3), the subtropical-PTF of Michelin *et al.* (2010) showed the best accuracy for $\theta(-33 \text{ kPa})$ ($0.0558 \text{ cm}^3 \text{ cm}^{-3}$) estimation and the poorest accuracy was found for

the tropical-PTF of Obalum and Obi (2012) ($0.1665 \text{ cm}^3 \text{ cm}^{-3}$). This behavior was the opposite for $\theta(-1,500 \text{ kPa})$.

Several authors report that PTFs tend to yield lower RMSE values for $\theta(-1,500 \text{ kPa})$ estimation than for $\theta(-33 \text{ kPa})$ (Tomasella and Hodnett, 2004; Vereecken *et al.*, 2010; Botula *et al.*, 2012; Medeiros *et al.* 2014; Nguyen *et al.* 2015; Touil *et al.*, 2016; Rustanto *et al.*, 2017; Contreras and Bonilla, 2018), which is in agreement with our findings.

Regarding the ME values (Table 3), the parametric-temperate-PTF of Vereecken *et al.* (1989) showed a decrease from -0.0095 to $-0.0339 \text{ cm}^3 \text{ cm}^{-3}$ when $\theta(-33 \text{ kPa})$ and $\theta(-1,500 \text{ kPa})$ values were estimated, respectively. The same trend was found using Splintex 2.0-PTF1-Sim B. Differently, the point-based PTF of Michelin *et al.* (2010) overestimated $\theta(-33 \text{ kPa})$ and underestimated $\theta(-1,500 \text{ kPa})$ values.

The overall estimation error expressed by MSE values (equation 5) indicated that the parametric-PTF of Vereecken *et al.* (1989) yielded the lowest MSE ($0.0031 \text{ cm}^6 \text{ cm}^{-6}$) for estimating $\theta(-33 \text{ kPa})$, while the PTF of Wösten *et al.* (1999) yielded the lowest MSE ($0.0026 \text{ cm}^6 \text{ cm}^{-6}$) to estimate $\theta(-1,500 \text{ kPa})$ (Table 3).

The best performance of the temperate-PTF of Vereecken *et al.* (1989) and the tropical-PTF of Barros *et al.* (2013)-4V for subtropical soils in the PRW is due to the similarity of textural class (sandy loam) amongst the soil database (Figure 2), and because the close relationship that the texture soil has in the water retention. More specifically, Vereecken *et al.* (1989) used clay (ranging from 0 to 54.5 %) and sand (ranging from 5.6 to 97.8%) contents as main input predictors to estimate the VGM parameters α and n . The authors concluded that detailed information on the particle-size distribution (PSD) can better describe the SWRC shape employing α and n . On the other hand, the tropical-PTF of Barros *et al.* (2013) was developed to estimate the VGM

parameters using clay contents ranging from 6 to 54.2% and sand contents ranging from 24.9 to 94.5%. These results may contribute to our study with the indirect quantification of key-soil-hydraulic properties, as highlighted by De Jong van Lier *et al.* (2015), who estimated the available water content for several soils worldwide and addressed that PTFs can successfully contribute to water balance simulations emphasizing their importance in hydrological and agricultural models.

Insert figure 2

The performance of tropical PTFs in Brazilian soils was also reported in other studies (Nebel *et al.*, 2010 and Ottoni *et al.*, 2018). For subtropical soils, the study carried out by Zou *et al.* (2016) in central China was one of the few that evaluated the performance of two temperate parametric PTFs (Wösten *et al.*, 1999 and Li *et al.*, 2007). They found that both temperate-PTFs produced overestimates of the VGM parameters of soils in which clay contents ranged from 10.9 to 35.2% and sand contents ranged from 7.6 to 67.3%. Zou *et al.* (2016) still developed a parametric subtropical PTF to estimate the VGM parameters using BD, soil textural fractions (contents of sand, silt, and clay), geometric mean, and geometric standard deviation of the particle diameters as predictors. However, the last two predictors have not been included as covariates to estimate VGM parameters in the most developed PTFs because they are not easily available properties from soil databases.

Exploratory analysis of measured and estimated soil water contents along the transect

The summary statistics of measured and estimated $\theta(-33 \text{ kPa})$ and $\theta(-1,500 \text{ kPa})$ values are shown in Table 4, as well as the Kolmogorov-Smirnov test. The CV values of the two measured

$\theta(h)$ data sets [CV value of 27.90% for $\theta(-33 \text{ kPa})$ and 34.69% for $\theta(-1,500 \text{ kPa})$] were higher than those ones of the two estimated $\theta(h)$ data sets [CV value ranging from 6.23 to 26.64% for $\theta(-33 \text{ kPa})$ and from 19.40 to 34.00% for $\theta(-1,500 \text{ kPa})$] by the majority evaluated PTFs excepting for both $\theta(-33 \text{ kPa})$ and $\theta(-1,500 \text{ kPa})$ simulations through Splintex-PTF1-SimB. These results show that most PTFs smoothed the behavior of the $\theta(h)$ values, as found by Turek *et al.* (2018).

Insert table 4

There was an increase in CV values for $\theta(-1,500 \text{ kPa})$ data sets when compared to those CV values for $\theta(-33 \text{ kPa})$ data sets (Table 4), i.e., as the soil is drying the dispersion of θ values around their mean is increasing, corroborating with results found by Romano and Santini (1997), Timm *et al.* (2006), and Nebel *et al.* (2010). The PTFs developed by Vereecken *et al.* (1989), Barros *et al.* (2013), and Michelin *et al.* (2010) generated high CV values that are closer to the CV of the measured data at the two matric potentials.

The values of the skewness coefficient were -0.10 to 0.02 (Table 4) for the measured $\theta(-33 \text{ kPa})$ and $\theta(-1,500 \text{ kPa})$, respectively, which can be considered unreliable in our study following the criterion suggested by Press *et al.* (2007). The skewness coefficient value found in $\theta(-33 \text{ kPa})$ estimation by the PTF of Michelin *et al.* (2010) was equal to the found in the measured $\theta(-33 \text{ kPa})$, while the skewness coefficient for $\theta(-1500 \text{ kPa})$ estimated according to Wösten *et al.* (1999) was similar to that one found in measured data (Table 4).

The kurtosis coefficients of all measured and estimated $\theta(-33 \text{ kPa})$ and $\theta(-1,500 \text{ kPa})$ data sets had the same tendency. All measured and estimated data sets followed the normal distribution according to the Kolmogorov-Smirnov test at 5% of significance level, except for $\theta(-33 \text{ kPa})$

estimated by Splintex-PTF1-sim A and $\theta(-1,500 \text{ kPa})$ estimated by the PTF of Michelin *et al.* (2010) (Table 4).

Estimation of the AW data using PTFs

The available water (AW) content was estimated with all PTFs because AW has been the most indirectly estimated parameter after first estimating FC and WP values using PTFs (Ottoni *et al.*, 2018). Thus, these AW estimates were compared to the AW calculated from measured $\theta(-33 \text{ kPa})$ and $\theta(-1,500 \text{ kPa})$ data, hereafter attributed as “measured” AW (Figure 3).

Insert figure 3

The tropical- and parametric-PTFs of Barros *et al.* (2013)-4V was more accurate to estimate AW with RMSE ($0.02 \text{ cm}^3 \text{ cm}^{-3}$) and ME ($-0.002 \text{ cm}^3 \text{ cm}^{-3}$). In the second place, the temperate- and parametric-PTFs of Vereecken *et al.* (1989) presented ME= $0.024 \text{ cm}^3 \text{ cm}^{-3}$ and RMSE= $0.037 \text{ cm}^3 \text{ cm}^{-3}$ values higher than Barros’s PTFs (Figure 4), overestimating the measured AW data (ME>0). Our AW results are in agreement with those presented by Tomasella and Hodnett (2004), Wang *et al.* (2012), Patil *et al.* (2010), Costa *et al.* (2013), Silva *et al.* (2017), and Ottoni *et al.* (2018), which have also shown that the PTFs failed in estimating indirectly AW after first estimating FC and WP values. However, Silva *et al.* (2020) comment that the determination of soil properties, for example, AW, is a difficult and onerous task and often unfeasible in large-scale monitoring. The estimation of AW based on PTFs can be an economical alternative, of short term and recognized technical viability, since, used with caution, depending on the reliability required in the specific interest of the study.

Evapotranspiration estimations assessment

Soil water content is one of the most important soil hydraulic attributes in modeling hydrological processes at the watershed scale (Dongli et al., 2017; Mallet et al., 2020). At the soil surface, soil water content influences the partitioning of precipitation into infiltration and runoff and controls evapotranspiration by controlling evaporation and water availability to plants, therefore affecting also the partitioning of latent and sensible heat (Grayson et al., 1997). Soil water content measurements might also be used to assess whether or not the calibration of different rainfall-runoff models provides results that are compatible with the physical reality of the watershed (e.g., Hu et al. 2015; Christanto et al., 2019).

The monitoring of soil water content is recognized as a complex phenomenon at a watershed scale, due mainly to the natural and anthropogenic processes (for instance, land-use systems) that influence soil water retention characteristics predominantly at the surface layer (Becker et al., 2018). Therefore, the comprehension and quantification of the soil water balance components (e.g., infiltration, evapotranspiration, direct surface runoff) are also affected due to changes in land-use systems at the watershed scale.

The determination of soil water retention characteristics can be costly and much effort is required to obtain its representative at the watershed scale. In this sense, pedotransfer functions used for estimating the soil water content over watersheds, using attributes more easily determined in the field or laboratory, are expected to strongly contribute to the better use of hydrological models and the understanding of their outputs. Fig. 4 shows the values of $\theta(-33 \text{ kPa})$ (Fig. 4a) and $\theta(-1,500 \text{ kPa})$ (Fig. 4b) measured in areas under native land-use systems (native grassland and native forest) were higher than those other land-use systems. Generally, cultivated soils under native vegetation when compared to soils impacted by agricultural land-use (e.g., cultivated

pasture, annual cropping) present higher soil water retention values as a result of high organic inputs and abundant burrowing fauna (Saha and Kukal, 2013; Dionizio and Costa, 2019; among others).

Insert figure 4

Furthermore, we assessed the influences of observed $\theta(-33 \text{ kPa})$ and $\theta(-1,500 \text{ kPa})$ as well as those estimated values from all punctual and parametric-PTFs on estimating the actual evapotranspiration (ET) under cultivated pasture, native grassland, native forest, silviculture, annual and permanent crop land-use systems, using Bergström (1992)'s equation (Eq. 6) (Fig. 5). The results showed that estimated ET values using observed and estimated $\theta(-33 \text{ kPa})$ and $\theta(-1,500 \text{ kPa})$ varied with each type of land use (Figure 5). Therefore, the uses of native forest, silviculture, and cultivated pasture presented the lines of ET with higher slopes as a function of the θ_{fc} and θ_{pwp} variations.

Insert figure 5

Conclusions

In this study, parametric- and point-PTFs calibrated with temperate, tropical, and subtropical soil data were applied to estimate $\theta(h)$ values in a 25 km-spatial transect in the Pelotas River Watershed (PRW), which is located under subtropical climate. Furthermore, we also investigated the performance of those PTFs to estimate the available water content (AW) and actual evapotranspiration (ET) behavior, which are parameters used as input for running some crop

models to support water management decisions. Overall, the examined parametric-PTFs had better accuracy tending to slightly underestimate $\theta(-33 \text{ kPa})$ and $\theta(-1,500 \text{ kPa})$ in comparison with the point-PTFs for the subtropical condition. Within the parametric-PTFs, Barros *et al.* (2013)-4V (tropical climate) and Vereecken *et al.* (1989) (calibrated with data from temperate climate) presented the best performance to estimate $\theta(-33 \text{ kPa})$, $\theta(-1,500 \text{ kPa})$, and AW, whereas the point-PTF of Obalum and Obi (2012) presented the worst performance to estimate $\theta(-33 \text{ kPa})$ and AW. The measured ET varied according to the land use, where the uses of native forest, silviculture, and cultivated pasture yielded the lines of ET with higher slopes as functions of the θ_{fc} and θ_{pwp} variations obtained with all PTF estimates. The found results supported us to choose some PTFs for modeling purposes in PRW, which is under a subtropical climate.

Acknowledgments

The authors wish to thank the Brazilian National Council for Scientific and Technological Development (CNPq) for the financial support and scholarships provided, and the Coordination for the Improvement of Higher Education Personnel - Brazil (CAPES), Finance Code 001, for scholarships.

References

- Aina, P.O., Periaswamy, S.P., 1985. Estimating available water-holding capacity of western Nigerian soils from soil texture and bulk density, using core and sieved samples. *Soil Science*, 140, 55–58.
- Alvarenga, L.A., Mello, C.R., Colombo, A., Cuartas, L.A., Bowling, L.C., 2016. Assessment of land cover change on the hydrology of a Brazilian headwater watershed using the distributed hydrology-soil vegetation model. *Catena*, 143, 7–17.
- Armindo, R.A., Wendroth, O., 2016. Physical soil structure evaluation based on hydraulic energy functions. *Soil Sci. Soc. Am. J.* 80, 1167–1180. <https://doi.org/10.2136/>

sssaj2016.03.0058.

Arruda, F.B., Zullo Junior, J., Oliveira, J.B., 1987. Parâmetros de solo para o cálculo da água disponível com base na textura do solo. *Revista Brasileira de Ciência do Solo*, 11, 11-15.

Baker, L., Ellison, D., 2008. The wisdom of crowds – ensembles and modules in environmental modelling. *Geoderma* 147, 1–7.

Barros, A.H.C., 2010. Desenvolvimento de funções de pedrotransferências e sua utilização em modelos agro-hidrológico. Doctoral thesis. University of São Paulo, Piracicaba/SP, Brazil.

Barros, A.H.C., De Jong van Lier, Q., Maia, A.H.N., Scarpere, F.V., 2013. Pedotransfer functions to estimate water retention parameters of soils in northeastern Brazil. *Revista Brasileira de Ciência do Solo*, 37, 379–391.

Bell, M.A., Keulen, H. van., 1995 Soil pedotransfer functions for four mexican soils. *Soil Science society of American Journal*, 59, 865–871.

Bell, M.A., Van Keulen, H., 1996. Effect of soil disturbance on pedotransfer function development for field capacity. *Soil Technology*, 8 (4), 321–329.

Blake, G.R., Hartge, K.H. Particle density. 1986. In: KLUTE, A., ed. *Methods of soil analysis*. 2.ed. Madison, Soil Science Society of America, 377-382.

Beskow, S., Timm, L.T., Tavares, V.E.Q., Caldeira, T.L. Aquino, L.S., 2016. Potential of the LASH model for water resources management in data-scarce basins: a case study of the Fragata River basin, southern Brazil. *Hydrological Sciences Journal*, 61 (14), 2567-2578.
<http://dx.doi.org/10.1080/02626667.2015.1133912>.

Bergström, S. 1992. The HBV Model–Its Structure and Applications. Report No 4, Swedish Meteorological and Hydrological Institute, Norrköping.

Botula, Y.D., 2013. Indirect methods to predict hydrophysical properties of soils of Lower Congo. Doctoral thesis. Ghent University, Belgium.

Botula, Y.D., Cornelis, W.M., Baert, G., Van Ranst, E., 2012. Evaluation of pedotransfer functions for predicting water retention of soils in Lower Congo (D.R. Congo). *Agricultural Water Management*, 111, 1–10. <http://dx.doi.org/10.1016/j.agwat.2012.04.006>.

Botula, Y.D., Van Ranst, E., Cornelis, W.M., 2014. Pedotransfer functions to predict water retention for soils of the humid tropics: A review. *Revista Brasileira de Ciência do Solo*, 38, 679-698.

Bouma, J., 1989. Using Soil Survey Data for Quantitative Land Evaluation. *Advances in Soil Science*, 9, 177–213.

- Botula, Y.D., 2013. Indirect methods to predict hydrophysical properties of soils of Lower Congo. Ghent University, Ghent, 236 p.
- Caldeira, T.L., Mello, C. R., Beskow, S., Timm, L.C., Viola, M.R., 2019. LASH hydrological model: An analysis focused on spatial discretization. *Catena*, 173, 183–193.
- Cambardella, C.A., Moorman, T.B., Novak, J.M., Parkin, T.B., Karlen, D.L., Turco, R.F., Konopka, A.E., 1994. Field-Scale Variability of Soil Properties in Central Iowa Soils. *Soil Science Society of America Journal*, 58, 1501–1511.
- Contreras, C.P., Bonilla, C.A., 2018. A comprehensive evaluation of pedotransfer functions for predicting soil water content in environmental modeling and ecosystem management. *Science of the Total Environment*, 644, 1580-1590.
- Costa, A. da, Albuquerque, J.A., Almeida, J.A.da, Costa, A.da, Luciano, R.V., 2013. Pedotransfer functions to estimate retention and availability of water in soils of the state of Santa Catarina, Brazil. *Revista Brasileira de Ciência do Solo*, 37, 889-910.
- Cressie, N., Hawkins, D.M., 1980. Robust estimation of the variogram: I. *Journal of the International Association for Mathematical Geology*, 12 (2), 115–125.
- De Jong van Lier, Q., Wendroth, O., Van Dam, J., 2015. Prediction of winter wheat yield with the SWAP model using pedotransfer functions: an evaluation of sensitivity, parameterization and prediction accuracy. *Agric. Water Manage. Sci. Direct.* 154, 29–42.
- Dijkerman, J.C., 1988. An Ustult-Aquult-Tropept Catena in Sierra Leone, West Africa, II. Land Qualities and Land. *Geoderma*, 42, 29-49.
- Dos Reis, A.M.H., Armindo, R.A., Durães, M.F., De Jong van Lier, Q., 2018. Evaluating pedotransfer functions of the Splintex model. *European Journal Soil Science*. 69, 685-697. <http://dx.doi.org/10.1111/ejss.12675>.
- Dongli, S., Qian, C., Timm, L.T., Beskow, S., Wei, H., Caldeira, T.L., 2017. Multi-scale correlations between soil hydraulic properties and associated factors along a Brazilian watershed transect. *Geoderma*, 286, 15–24. <http://dx.doi.org/10.1016/j.geoderma.2016.10.017>.
- Dourado-Neto, D., Nielsen, D.R., Hopmans, J.W., Reichardt, K., Bacchi, O.O.S., Lopes, P.P., 2001. Soil Water Retention Curve (SWRC). Version 3.0 beta.
- Freitas, K.S., Armindo, R.A., Pires, L.P., Filho, V.S., Ribeiro, S.J., 2019. A method to determine the soil bulk density of undisturbed samples with non-isodiametric shape. *Soil and Tillage Research*, 191, 344-352. <https://doi.org/10.1016/j.still.2019.04.007>
- Gee, G.W., Bauder, J.W., 1986. Particle-size analysis. in: Klute, A. (Eds.), *Methods of Soil Analysis. Part 1*, 2nd ed Agronomy Monograph. ASA-SSSA, Madison, WI, 9, 383–411.

Gupta, S.C., Larson, W.E., 1979. Estimating soil water retention characteristics from particle size distribution, organic matter percent, and bulk density. *Water Resources Research*, 15 (6), 1633–1635.

Gunarathna, M.H.J.P., Sakai, K., Nakandakari, T., Momii, K., Kumari, M.K.N., Amarasekara, M.G.T.S., 2019. Pedotransfer functions to estimate hydraulic properties of tropical Sri Lankan soils. *Soil & Tillage Research*, 190, 109–119.

Kuinchtner, A., Buriol, G.A., 2001. Clima do estado do Rio Grande do Sul segundo a classificação climática de the climate of Rio Grande do Sul state according to Köppen and Thornthwaite. *Disciplinarum Scientia. Série: Ciências Exatas*, 2 (1), 171–182.
<http://sites.unifra.br/Portals/36/tecnologicas/2001/clima.pdf>.

Lal, R., 1979. Physical properties and moisture retention characteristics of some Nigerian soils. *Geoderma*, 21, 209–223.

Li, Y., Chen, D., White, R.E., Zhu, A., Zhang, J., 2007. Estimating soil hydraulic properties of Fengqiu County soils in the North China Plain using pedo-transfer functions. *Geoderma*, 138, 261–271. <http://dx.doi.org/doi:10.1016/j.geoderma.2006.11.018>.

Medeiros, J.C., Cooper, M., Dalla Rosa, J., Grimaldi, M., Coquet, Y., 2014. Assessment of pedotransfer functions for estimating soil water retention curves for the amazon region. *Revista Brasileira de Ciência do Solo*, 38, 730–743.

Mello, C.R., Ávila, L.F., Lin, H., Terra, M.C.N.S., Chappell, N.A., 2019. Water balance in a neotropical forest catchment of southeastern Brazil. *Catena*, 173, 9–21.

Michelon, C.J., Carlesso, R., De Oliveira, Z.B., Knies, A.E., Petry, M.T., Martins, J.D., 2010. Funções de pedotransferência para estimativa da retenção de água em alguns solos do Rio Grande do Sul, 40 (4), 848–853.

Minasny, B., McBratney, A.B., Bristow, K.L., 1999. Comparison of different approaches to the development of pedotransfer functions for water-retention curves. *Geoderma*, 93, 225–253.

Minasny, B., Hartemink, A.E., 2011. Predicting soil properties in the tropics. *Earth-Science Reviews*, 106, 52–62, 2011. <http://dx.doi.org/10.1016/j.earscirev.2011.01.005>.

Mualem, Y. 1976. A new model for predicting the hydraulic conductivity of unsaturated porous media. *Water Resources Research*, 12, 513–522. <https://doi.org/10.1029/WR012i003p00513>

Nebel, Á.L.C, Timm, L.C., Cornelis, W., Gabriels, D., Reichardt, K., Aquino, L.S., Pauletto, E.A., Reinert, D.J., 2010. Pedotransfer functions related to spatial variability of water retention attributes for lowland soils. *Revista Brasileira de Ciência do Solo*, 34 (3), 669–680.

Nguyen, P.M., Le, K.V., Botula, Y.D., Cornelis, W.M., 2015. Evaluation of soil water retention pedotransfer functions for Vietnamese Mekong Delta soils. *Agricultural Water Management*, 158, 126-138.

Obalum, S.E.; Obi, M.E., 2012. Moisture characteristics and their point pedotransfer functions for coarse-textured tropical soils differing in structural degradation status. *Hydrological Processes*, 1-15.

Oliveira, L.B., Ribeiro, M.R, Jacomine, P.K.T., Rodrigues, J.J.V., Marques, F.A., 2002. Pedotransfer functions to predict soil water content at specific matric potentials in soils from Pernambuco state. *Revista Brasileira de Ciência do Solo*, 26, 315–323.

Otoni Filho, T.B., Leal, I.F., De Macedo, J.R., Reis, B.C.B., 2016. An Algebraic Pedotransfer Function to Calculate Standardized in situ Determined Field Capacity. *J. Agric. Sci.* 8, 158. <https://doi.org/10.5539/jas.v8n8p158>

Otoni, M.V., Otoni, T.B.F., Schaap, M.G., Lopes-Assad, M.L.R.C., Rotuno, O.C.F., 2018. Hydrophysical Database for Brazilian Soils (HYBRAS) and pedotransfer functions for water retention. *Vadose Zone Journal*, 17 (1), 1–17. <http://dx.doi.org/10.2136/vzj2017.05.0095>

Patil, N.G., Rajput, G.S., Nema, R.K., Singh, R.R., 2010. Calibration and evaluation of pedotransfer functions to estimate available water capacity of seasonally impounded shrink-swell soils of central India. *Archives of Agronomy and Soil Science*, 56, 535-538.

Pebesma, E.J., 2004. Multivariable geostatistics in S: The gstat package. *Computers and Geosciences*, 30, 683–691.

Peraza, J.E.S., 2003. Retenção de água e pedofunções para solos do Rio Grande do Sul. Santa Maria, Universidade Federal de Santa Maria, 118p. (Master's Thesis).

Perrotta, M.M., 2005. Processamento digital básico de imagens de sensores remotos ópticos para uso em mapeamento geológico – programa ENVI. Belo Horizonte: CPRM.

Pidgeon, J.D., 1972. The measurement and prediction of available water capacity of Ferrallitic soils in Uganda. *Soil Science Journal*, 23, 431-441.

Press, W. H., Teukolsky, S.A, Vetterling, W.T., Flannery, B.P., 2007. Numerical recipes: The art of scientific computing. Third Edition, in: Press, W.H., Teukolsky, S.A, Vetterling, W.T., Flannery, B.P. *Statistical Description of Data*, Chapter 14, Cambridge University Press, pp. 720-740.

Pringle, M.J., Romano, N., Minasny, B., Chirico, G.B., Lark, R.M., 2007. Spatial evaluation of pedotransfer functions using wavelet analysis. *Journal of Hydrology*, 333, 182–198.

R Core Team (2019). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. <http://www.R-project.org/>.

Rawls, W.J., Brakensiek, D.L., 1989. Estimation of soil water retention and hydraulic properties, in: Morel-Seytoux, H.J., (Eds.) *Unsaturated flow in hydrologic modeling -Theory and Practic*. Kluwer Academic Publishers., pp. 275-300.

Rawls, W.J., Brakensiek, D.L., Saxton, K.E., 1982. Estimation of soil water properties. *Transactions of the ASAE*, 25 (5), 1316–1320 and 1328.

Reynolds, W.D., Drury, C.F., Tan, C.S, Fox, C.A., Yang, X.M., 2009. Use of indicators and pore volume-function characteristics to quantify soil physical quality. *Geoderma*, 152, 252–263. <http://dx.doi.org/10.1016/j.geoderma.2009.06.009>.

Romano, N., Santini, A., 1997. Effectiveness of using pedo-transfer functions to quantify the spatial variability of soil water retention characteristics. *Journal of Hydrology*, 202, 137–157.

Rustanto, A., Booij, M.J., Wösten, H., Hoekstra, A. Y., 2017. Application and recalibration of soil water retention pedotransfer functions in a tropical upstream catchment: case study in Bengawan Solo, Indonesia. *Journal of Hydrology and Hydromechanics*, 65 (3), 307-320.

Saito, H., McKenna, S.A., Zimmerman, D.A, 2005. Geostatistical interpolation of object counts collected from multiple strip transects: Ordinary kriging versus finite domain kriging. *Stochastic Environmental Research and Risk Assessment*, 19, 71–85. <http://dx.doi.org/10.1007/s00477-004-0207-3>.

Dos Santos, W.J.R., Curi, N., Silva, S.H.G., De Araujo, E.F., Marques, J.J., 2013. Pedotransfer functions for water retention in different soil classes from the center-southern Rio Grande do Sul state. *Ciência e Agrotecnologia*, 37 (1), 49-60.

Schaap, M.G., Leij, F.J., van Genuchten, M.Th., 2001. ROSETTA: a computer program for estimating soil hydraulic parameters with hierarchical pedotransfer functions. *Journal of Hydrology*, 251, 163-176.

Silva, A.C., Armindo, R.A., Brito, A.S., Schaap, M.G., 2017. Splintex: A physically-based pedotransfer function for modeling soil hydraulic functions. *Soil Tillage Research*. 174, 261–272.

Silva, A.C., Armindo, R.A., Prevedello, C.L. 2020. Splintex 2.0: a physically-based model to estimate water retention and hydraulic conductivity parameters from soil physical data. *Computers and Eletronics in Agriculture*. 169, 1-10.

Soil Survey Staff., 2014. Keys to soil taxonomy. USDA, 12, pp. 410. http://www.nrcs.usda.gov/Internet/FSE_DOCUMENTS/nrcs142p2_051546.pdf.

Timm, L.C., Pires, L.F., Roveratti, R., Arthur, R.C.J., Reichardt, K., Oliveira, J.C.M., Bacchi, O.O.S., 2006. Field spatial and temporal patterns of soil water content and bulk density changes. *Sci. Agric.*, 63 (1), 55–64.

Tóth, B., Weynants, M., Nemes, A., Makó, A., Bilas, G., Tóth, G., 2015. New generation of hydraulic pedotransfer functions for Europe. *European Journal of Soil Science*, 66 (1), 226–238. <http://dx.doi.org/10.1111/ejss.12192>.

Touil, S.; Degré, A.; Chabaca, M.N., 2016. Transposability of pedotransfer functions for estimating water retention of Algerian soils. *Desalination and Water treatment*, 57, 5232–5240.

Tomasella, J., Hodnett, M.G., 2004. Pedotransfer functions for tropical soils. In: Pachepsky, Y.A. & Rawls, W.J., ed. *Development of pedotransfer functions in soil hydrology*. Amsterdam, Elsevier, p. 415–429.

Turek, M.E., Armindo, R.A., Wendroth, O., Dos Santos, I., 2018. Criteria for the estimation of field capacity and their implications for the bucket type model. *European Journal of Soil Science*, 70, 278–290. <http://dx.doi.org/10.1111/ejss.12747>

Turek, M.E., De Jong van Lier, Q., Armindo, R.A., 2020. Estimation and mapping of field capacity in Brazilian soils. *Geoderma*, 376, <https://doi.org/10.1016/j.geoderma.2020.114557>

Urach, F.L., 2007. Estimating soil water retention for irrigation. Federal University of Santa Maria (Master thesis), pp. 78.

Van Genuchten, M.T., 1980. A closed-form equation for predicting the hydraulic conductivity of unsaturated soils. *Soil Science society of American Journal*, 44, 892–898.

Vereecken, H., Maes, J., Feyen, J., Darius, P., 1989. Estimating the soil moisture retention characteristic from texture, bulk density, and carbon content. *Soil Science*, 148 (6), 389–403.

Vereecken, H., Weynants, M., Javaux, M., Pachepsky, Y., Schaap, M.G., van Genuchten, M.Th., 2010. Using pedotransfer functions to estimate the van Genuchten–Mualem soil hydraulic properties: a review. *Vadose Zone Journal*, 9, 1–26.

Vos, B.; Meirvenner, M.V.; Quataert, P.; Deckers, J.; Muys, B., 2005. Predictive quality of pedotransfer functions for estimating bulk density of forest soils. *Soil Science Society of America Journal*, 69, 500–510.

Wang, Z., Shi, W., 2018. Robust variogram estimation combined with isometric log-ratio transformation for improved accuracy of soil particle-size fraction mapping. *Geoderma*, 324, 56–66. <https://doi.org/10.1016/j.geoderma.2018.03.007>.

Wang, G., Zhang, Y., Yu, N., 2012. Prediction of soil water retention and available water of sandy soils using pedotransfer functions. *Procedia Engineering*, 37, 49–53.

Wilding, L.P., Drees, L.R., 1983. Spatial variability and pedology, in: Wilding, L.P., Smeck, N.E., Hall, G.F., (Eds.), *Pedogenesis and soil taxonomy: Concepts and interactions*. New York, Elsevier., pp. 83–116.

Wösten, J.H.M., Finke, P.A., Jansen, M.J.W., 1995. Comparison of class and continuous pedotransfer functions to generate soil hydraulic characteristics. *Geoderma*, 66, 227-237.

Wösten, J.H.M., Lilly, A., Nemes, A., Le Bas, C., 1999. Development and use of a database of hydraulic properties of European soils. *Geoderma*, 90, 169–185.

Wösten, J.H.M., Pachepsky, Y.A., Rawls, W.J., 2001. Pedotransfer functions: Bridging the gap between available basic soil data and missing soil hydraulic characteristics. *Journal of Hydrology*, 251, 123–150.

Zou, G., Li, Y., Wang, Y., Liu, D.L., Liu, X., Li, Y., Wu, J., 2016. Peto-transfer functions for estimating the hydraulic properties of paddy soils in subtropical central China. *Archives of Agronomy and Soil Science*, 62 (7), 982–993.
<http://Dx.Doi.Org/10.1080/03650340.2015.1109078>.

Table 1

Type of pedotransfer functions (PTFs) selected, types of climate, the origin of soils and input variables used.

PTFs	Type of PTFs	Climate	Origin of soils	Inputs variables	Output parameters
Vereecken <i>et al.</i> (1989)	Parametric	Temperate	Belgian	Sa, C, BD, OC	$\theta_s, \theta_r, \alpha, n$
Wösten <i>et al.</i> (1999)	Parametric	Temperate	Europe (HYPRES)	Si, C, BD, OM	θ_s, α, n
Rosetta (SSC)	Parametric	Temperate	USA and Europe	Sa, Si, C	$\theta_s, \theta_r, \alpha, n$
Barros <i>et al.</i> (2013)-4V	Parametric	Tropical	Northeastern Brazil	Sa, C, BD, OM, PtD	θ_r, α, n
Splintex 2.0-PTF1 Sim A	Parametric	-	physical-empirical model	PSD, BD, PtD.	θ_r, α, n , and $\theta_s = TP$
Splintex 2.0-PTF1 Sim B	Parametric	-	physical-empirical model	PSD, BD, PtD.	$\theta_s, \theta_r, \alpha, n$
Obalum and Obi (2012)	Point	Tropical	Southeastern Nigeria	Sa, C, OM	$\theta(h)$ point
Michelon <i>et al.</i> (2010)	Point	Subtropical	Rio Grande do Sul state, Brazil.	Si, C, FS, BD, PtD, Ma, Mi, TP	$\theta(h)$ point

SSC: sand + silt + clay; 4V: pedotransfer function with four soil variables; Sa: sand; Si: silt; C: clay; FS: fine sand, BD: bulk density; OC: organic carbon; OM: organic matter; PtD: particle density, Ma: macroporosity, Mi: microporosity, TP: total porosity, PSD: particle size distribution, Parametric: parameters of the van Genuchten-Mualem (1980) model with $m = 1 - 1/n$ a parametric restriction of Mualem (1976).

Table 2

Summary statistics of soil properties used as input variables to PTFs along the 25-km-transect established in the PRW

Soil properties	Min	Max	Mean	Std Dev	CV (%)	Skew	Kurt	K-S
Clay (%)	6.54	32.53	17.50	5.20	29.72	0.09	-0.24	0.05
Silt (%)	1.49	42.28	19.77	6.31	31.91	0.67	1.46	0.11
Sand (%)	35.70	91.57	62.73	9.72	15.50	-0.08	0.29	0.06
Fine sand (%)	4.56	21.31	17.07	2.86	16.75	-1.41	3.02	0.13
PtD (g cm ⁻³)	2.33	2.84	2.57	2.57	3.44	-0.13	1.13	0.06
BD (g cm ⁻³)	1.06	1.68	1.43	0.14	9.82	-0.58	0.08	0.08
OM (g kg ⁻¹)	32.93	55.18	39.68	4.20	10.59	0.87	1.52	0.08
Ma (cm ³ cm ⁻³)	0.008	0.332	0.163	0.08	47.97	0.08	-0.73	0.07
Mi (cm ³ cm ⁻³)	0.082	0.427	0.279	0.07	23.63	-0.09	0.10	0.06
TP (cm ³ cm ⁻³)	0.301	0.583	0.442	0.05	12.38	0.33	0.22	0.09

Min: minimum value; Max: maximum value, Mean: mean value; Std Dev: standard deviation; CV: coefficient of variation; Skew: skewness coefficient; Kurt: kurtosis coefficient; K-S: Kolmogorov-Smirnov test (critical value: 0.136, for *P-value* = 5%); PtD: particle density; BD: bulk density; OC: organic carbon content; OM: organic matter content, Ma: macroporosity, Mi: microporosity, TP: total porosity.

Table 3

Statistical metrics of the pedotransfer functions (PTFs) performance for estimating soil water content at -33 and -1,500 kPa

θ values	PTFs	RMSE (cm ³ cm ⁻³)	ME (cm ³ cm ⁻³)	SDE (cm ³ cm ⁻³)	MSE (cm ⁶ cm ⁻⁶)
$\theta(-33 \text{ kPa})$	Vereecken <i>et al.</i> (1989)	0.0553	-0.0095	0.0548	0.0031
	Wösten <i>et al.</i> (1999)	0.1354	0.1202	0.0626	0.0183
	Rosetta (SSC)	0.1177	0.1062	0.0510	0.0138
	Barros <i>et al.</i> (2013)-4V	0.0636	-0.0372	0.0519	0.0040
	Splintex 2.0-PTF1-SimA	0.1315	-0.1219	0.0496	0.0173
	Splintex 2.0-PTF1-SimB	0.1217	-0.1109	0.0502	0.0148
	Obalum and Obi (2012)	0.1665	0.1587	0.0507	0.0277
	Michelon <i>et al.</i> (2010)	0.0558	0.0186	0.0528	0.0031
$\theta(-1,500 \text{ kPa})$	Vereecken <i>et al.</i> (1989)	0.0610	-0.0339	0.0510	0.0037
	Wösten <i>et al.</i> (1999)	0.0510	0.0057	0.0509	0.0026
	Rosetta (SSC)	0.0638	-0.0398	0.0501	0.0041
	Barros <i>et al.</i> (2013)-4V	0.0580	-0.0328	0.0481	0.0034
	Splintex 2.0-PTF1-SimA	0.0994	-0.0854	0.0512	0.0099
	Splintex 2.0-PTF1-SimB	0.1526	-0.1402	0.0606	0.0233
	Obalum and Obi (2012)	0.0582	-0.0274	0.0516	0.0034
	Michelon <i>et al.</i> (2010)	0.0888	-0.0768	0.0449	0.0079

ME: mean error; SDE: standard deviation of estimation error; RMSE: root mean square error; MSE: mean square error (BVTO)

Table 4

Summary statistics for measured and estimated data by parametric-PTFs and Point-PTFs models for estimating soil water content at 0, -33, and -1500 kPa.

θ value	PTFs	Min	Max	Mean	Std Dev	CV	Skew	Kurt	K-S
$\theta(-33 \text{ kPa})$	Measured	0.060	0.353	0.227	0.06	27.90	-0.10	-0.65	0.06
	Vereecken <i>et al.</i> (1989)	0.093	0.356	0.217	0.05	24.51	0.12	-0.13	0.07
	Wösten <i>et al.</i> (1999)	0.245	0.462	0.347	0.04	11.83	0.82	0.56	0.11
	Rosetta (SSC)	0.270	0.399	0.333	0.02	7.02	-0.07	0.46	0.08
	Barros <i>et al.</i> (2013)-4V	0.074	0.301	0.190	0.05	25.44	0.08	-0.32	0.06
	Splintex 2.0-PTF1-SimA	0.043	0.187	0.105	0.03	31.27	0.33	-0.57	0.54
	Splintex 2.0-PTF1-SimB	0.033	0.221	0.116	0.04	37.06	0.41	-0.45	0.06
	Obalum and Obi (2012)	0.315	0.453	0.386	0.02	6.23	0.08	0.29	0.06
	Michelon <i>et al.</i> (2010)	0.051	0.393	0.246	0.07	26.64	-0.10	0.09	0.06
$\theta(-1,500 \text{ kPa})$	Measured	0.015	0.325	0.174	0.06	34.69	0.02	-0.31	0.05
	Vereecken <i>et al.</i> (1989)	0.075	0.235	0.140	0.03	23.18	0.23	-0.22	0.04
	Wösten <i>et al.</i> (1999)	0.047	0.283	0.180	0.04	22.61	0.05	0.74	0.09
	Rosetta (SSC)	0.058	0.191	0.134	0.03	19.40	-0.85	0.73	0.10
	Barros <i>et al.</i> (2013)-4V	0.041	0.227	0.141	0.04	28.76	-0.07	-0.44	0.07
	Splintex 2.0-PTF1-SimA	0.035	0.163	0.089	0.03	32.01	0.37	-0.45	0.07
	Splintex 2.0-PTF1-SimB	0.002	0.124	0.034	0.03	76.28	1.26	1.33	0.17
	Obalum and Obi (2012)	0.068	0.255	0.147	0.04	25.60	0.10	-0.24	0.05
	Michelon <i>et al.</i> (2010)	0.014	0.172	0.097	0.03	34.97	-0.23	-0.20	0.44

Min: minimum value; Max: maximum value, Mean: mean value; Std Dev: standard deviation; CV: coefficient of variation; Skew: skewness coefficient; Kurt: kurtoses coefficient; K-S: Kolmogorov-Smirnov test (critical value: 0.136, for $P\text{-value} < 5\%$).

Figure Captions

Fig. 1. Illustration of the 25-km-spatial transect established in the Pelotas River Watershed (PRW), located in Rio Grande do Sul state, Southern Brazil, presenting the Ponte Cordeiro de Farias outlet and the 100 sampling points.

Fig. 2. The textural domain of assessed PTFs and Pelotas River watershed

Fig. 3. Correlations between measured and estimated available water-AW along the 25km transect in the Pelotas River watershed.

Fig. 4. Boxplots of soil water contents at matric potentials of θ_{fc} (-33 kPa) (a) and θ_{pwp} (-1,500 kPa) (b) for the different land uses along the 25 km-Pelotas River watershed spatial transect.

Fig. 5. Influences of θ_{fc} and θ_{pwp} on measuring and estimating the actual evapotranspiration (ET) behavior from 0 to the maximum value in the land-uses found along the transect.

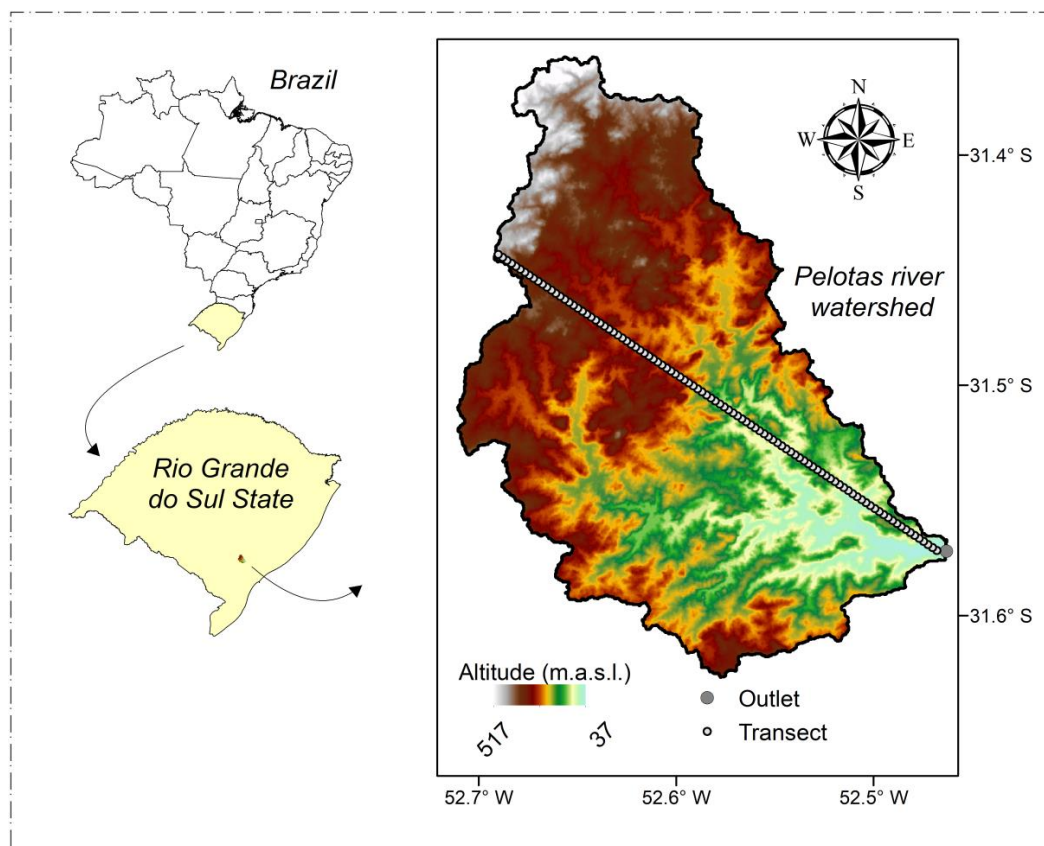


Fig. 1.

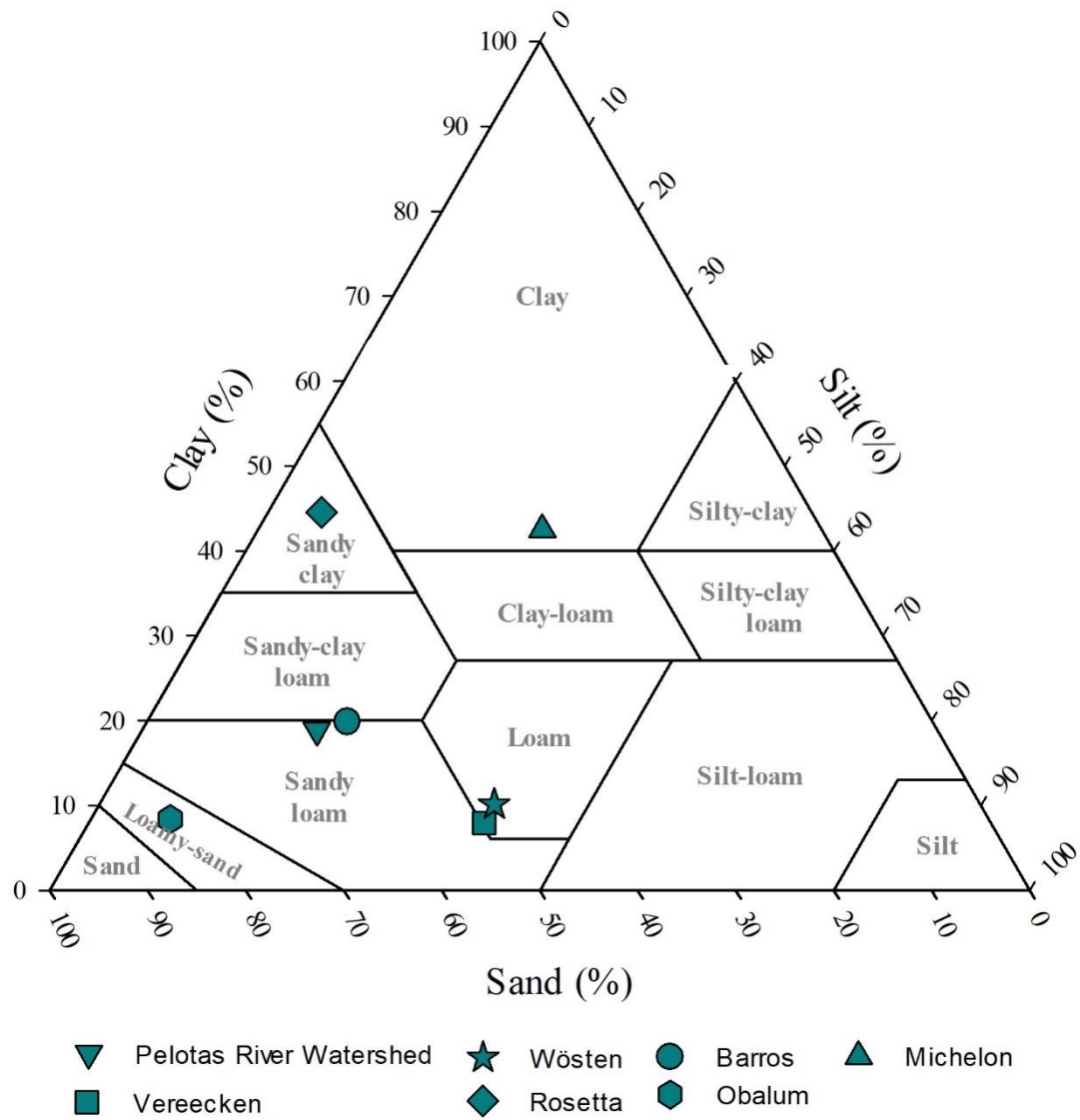


Fig. 2.

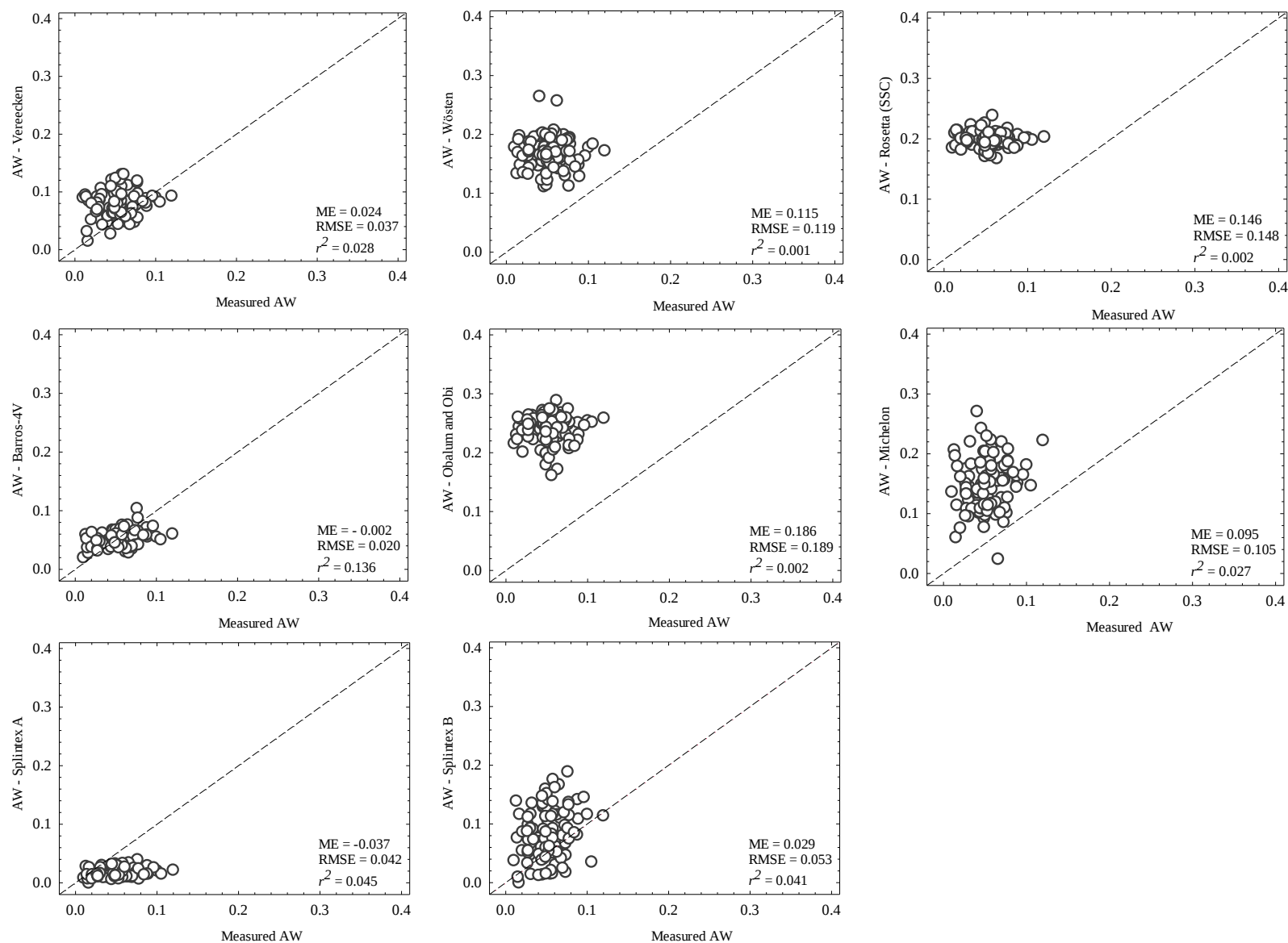


Fig. 3.

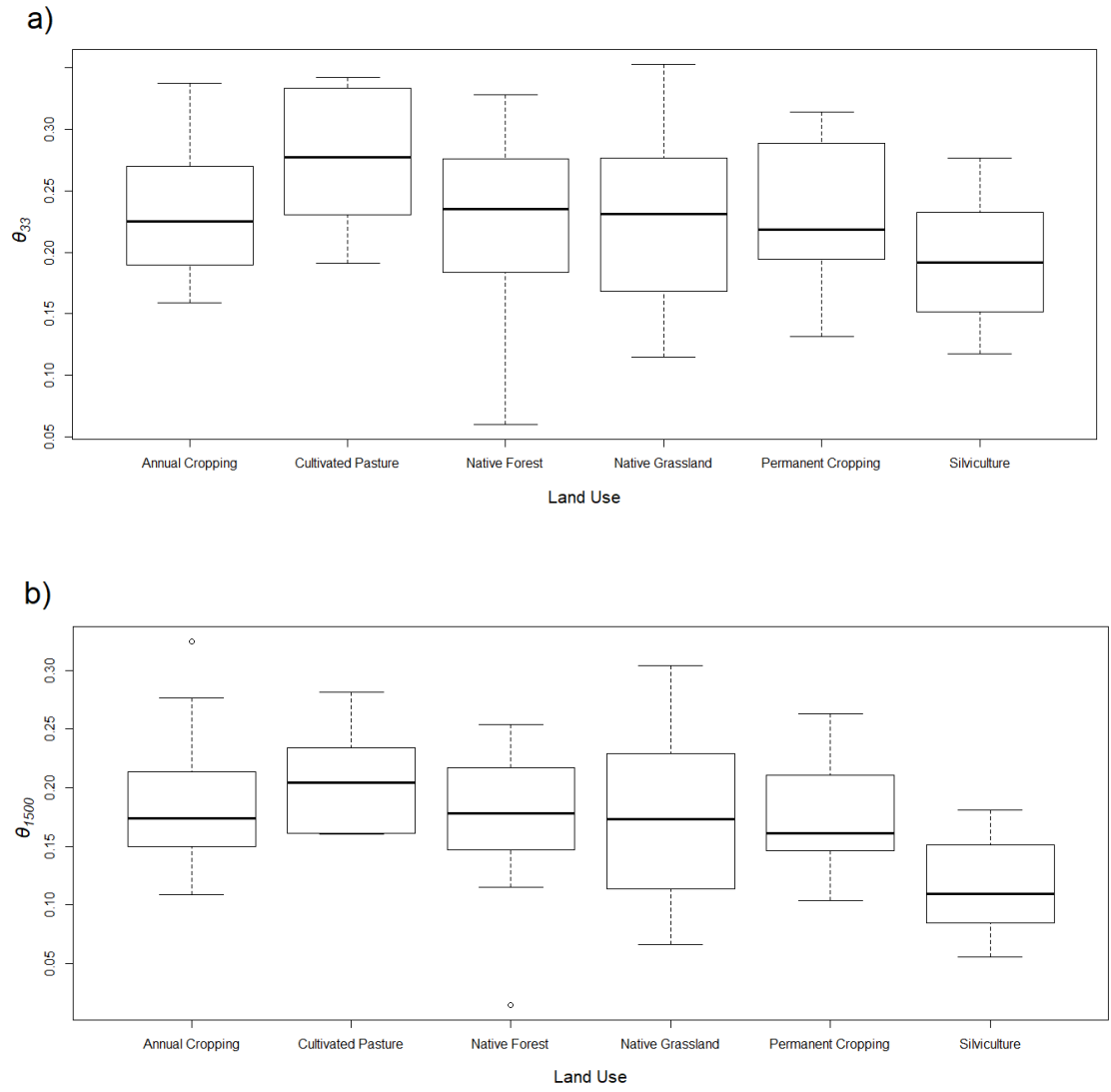
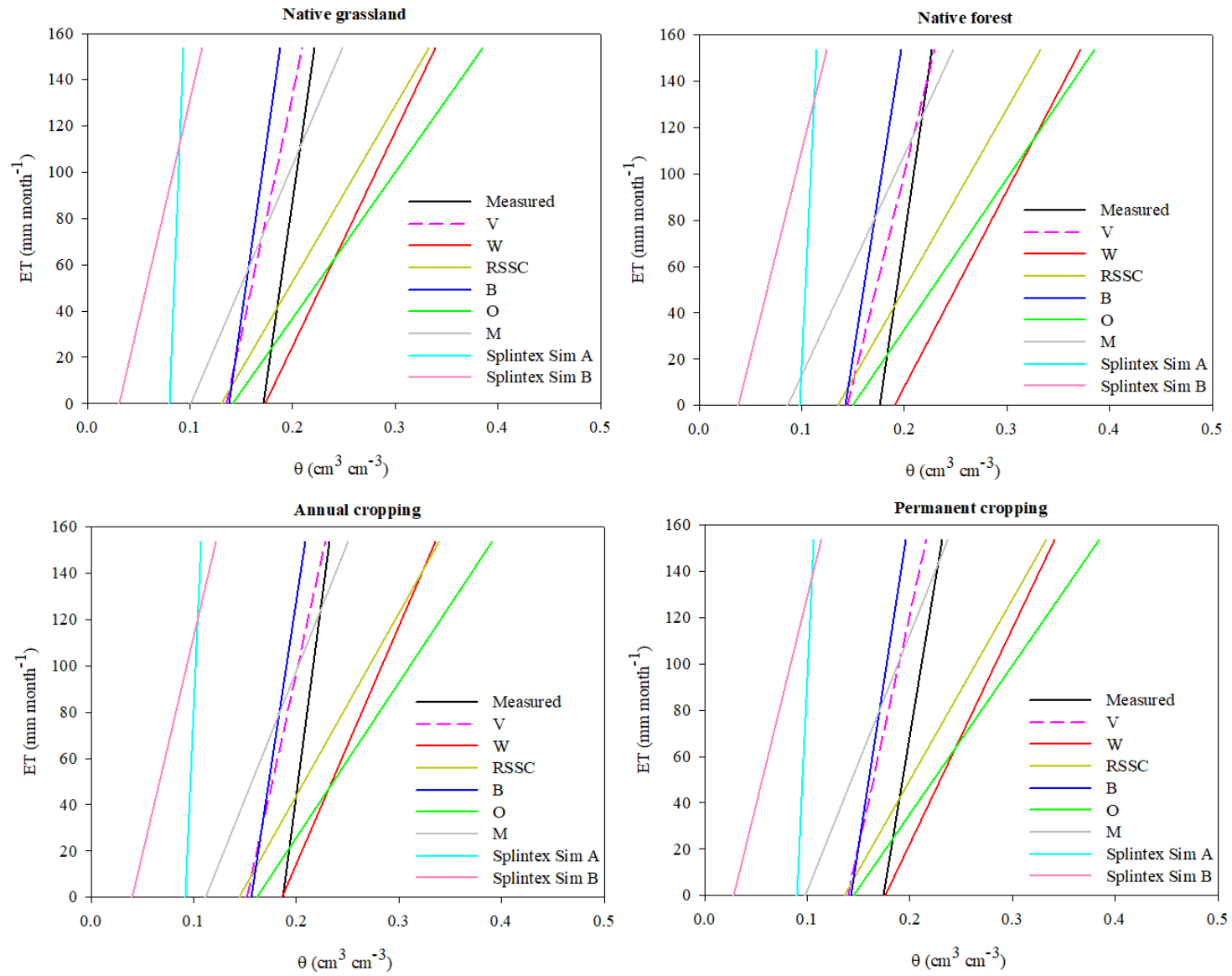


Fig. 4.



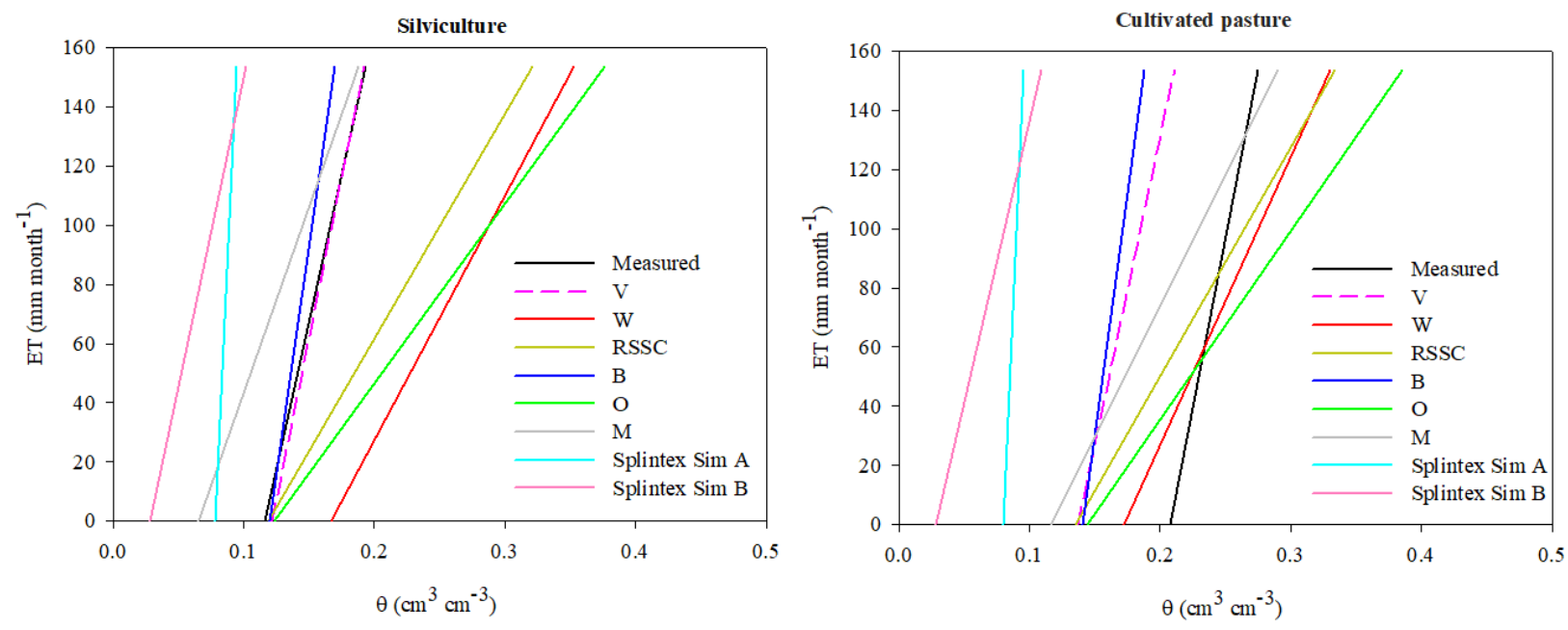


Fig. 5

5 Artigo 2

Random search and topographical attributes to estimate soil water content through pseudo-continuous pedotransfer function based on artificial neural network

Miguel David Fuentes-Guevara; Gabriel Almeida Gomes; Mauricio Fornalski Soares; Luciana Montebello de Oliveira; Ricardo Matsumura Araujo; Ulisses Brisolara Corrêa; Luís Carlos Timm.

Abstract

The pseudo-continuous pedotransfer function (PC-PTF), which is based on an artificial neural network (ANN), allows for the estimation of the water content in any soil matric potentials from a limited soil database. The ANN has hyper-parameters that influence the quality of these estimates and needs to be explored. Thus, the objective of this study was to assess the impact of the hyper-parameters, soil attributes, and topographical attributes on PC-PTF capacity to estimate the soil water content in a watershed. To achieve this, a soil database with several soils and topographical attributes, delivered from two watersheds located in the southern Rio Grande do Sul - Brazil, was created. The soil database was used to perform an ANN and to choose the best architecture. In this context, the random search technique was used to explore the best sets of ANN hyper-parameters, soil attributes, and topographical attributes. The random search technique allowed us to identify the best input attribute set that was composed of soil bulk density, soil organic carbon content, microporosity, clay, curvature, and plane curvature. However, the best ANN configuration, with the aim of estimating the soil water content in all matric potentials, was based on the hyper-parameters such as seed= 98.186, learning rate= 0.2, number of epochs= 5,000, and momentum= 0.2, along with obtaining performances on the test set of $r^2= 0,73$; $ME= 0,05 \text{ cm}^3 \text{ cm}^{-3}$ e $RMSE=$

0,06 cm⁻³ cm⁻³. Nonetheless, the ANN failed to estimate the soil water content for the matric potentials in an independent way. This suggests the need to continue research that further explores the relation among these attributes for a greater soil database.

Keywords: soil attributes; watershed; machine learning; hyper-parameters.

Introduction

Soil hydraulic properties comprehension is widely considered the most critical factor in agricultural and environmental sciences as it guides soil management decision-making. In addition, soil water content is a crucial input parameter of many hydrological models that aid in understanding the soil water dynamics in environmental systems, like agricultural areas and watersheds. However, measuring soil water content is difficult when it is quantified *in situ* or laboratory because it demands much time, specialized equipment, and high costs. For this reason, alternatives that contribute to estimating soil water must be explored since they are not easily available in the soil database where the study is developed.

Pedotransfer functions (PTFs) are an alternative tool to predict soil water content from basic soil properties. Currently, different categories of PTFs are used to estimate this variable. The most recent PTFs are the point and parametric referred to as Type 2 and Type 3 by Wösten *et al.* (2001). These PTFs estimate water content at the predefined potentials and parameters of the soil hydraulic model (e.g., Brooks and Corey (1964) or van Genuchten (1980) equations), respectively. These PTFs categories can be derived mainly by statistical regression and machine learning techniques (Vereecken *et al.* 2010). Remarkably, PTFs developed with machine learning (e.g., artificial neural networks) have demonstrated better performance in predicting soil water content (Minasny and McBratney, 2002; Wang *et al.*, 2012; D’Emilio *et al.*, 2018).

A severe limitation of generating Type 2 and Type 3 PTFs using artificial neural networks (ANNs) is the need for large soil databases, which is a problem in many countries and regions where soil hydraulic data is lacking. To solve this problem, an innovative approach was introduced by Haghverdi *et al.* (2012) to derive PTFs with the ANNs technique called pseudo-continuous PTF (PC-PTF). They claim that PC-PTF is an important tool because it enables us to increase the amount of data in the training set. Furthermore, the PC-PTF can use the matric potential to predict the soil water content at any matric potential.

PC-PTF has the potential to estimate soil water content at any established matric potential without the need to apply a specific equation and enables the user to generate a predictive model from limited data (Haghverdi *et al.*, 2012). Moreover, the authors stated that PC-PTF generated slightly better soil water content predictions when compared with soil water content predictions generated from point and parametric PTFs that used the same soil database. In light of the usefulness and accuracy of PC-PTF, they stated that PC-PTFs will need more focus and attention in the future.

A few years later, Haghverdi *et al.* (2014) claimed that the performance of PC-PTF developed with ANNs should be compared with the performance of PC-PTF developed with a different machine learning algorithm (MLA), as this comparison may yield findings that improve the overall effectiveness of PC-PTF. In this context, the authors investigated the effectiveness of another MLA called support vector machines (SVM). The authors found that the ANN-based PTF performed better than the SVM-based PTF. However, the main weakness in their study is that they do not attempt to explore the ANNs hyper-parameters since there are several options in the architecture of an ANN that can be explored to improve its performance.

According to Pham *et al.* (2019), many parameters in ANNs can be considered hyper-parameters. Some of these parameters include the learning rate and the activation function, along

with the number of neurons in the input layer and hidden layers. Several authors have established the influences of hyper-parameters in the optimization and improvement of the estimative performance of ANNs (Bergstra and Bengio, 2012; Géron, 2017; Pham *et al.*, 2019; Raschka, 2018). In view of this, the grid search method is the most widely used strategy for optimizing the hyper-parameters of the neural network architecture configuration. However, this method executes the search over a grid of trials based on their performance on the validation set, which makes this method more expensive in computational time. In the face of that challenge, random search is another method that can be explored to improve the optimization process of ANN hyper-parameters (Bergstra and Bengio, 2012).

It is important to realize that there are modifications in the methodologies introduced by Haghverdi *et al.* (2012) and Haghverdi *et al.* (2014) that can be explored, such as alterations in the hyper-parameters of the ANNs which enhance the estimations of the soil water content. Furthermore, the use of PC-PTF has not received much attention in recent years and the explorations of new soil properties as input data through these PTFs are scarce. Hence, this paper aims to investigate: (i) the impact of the hyper-parameters on the pseudo-continuous PTF capability to estimate the soil water content in a watershed, and (ii) the potential of *random search* to select new soil properties and topographical attributes as input data for the pseudo-continuous PTF.

Material and Methods

Soil database and experimental data

A soil database with 205 data points was generated from the junction of datasets from two watersheds located in the southern Rio Grande do Sul (RS) - Brazil, under a subtropical climate (Figure 1). One hundred points of soil data collected along a 25-km-spatial transect in the Pelotas River Watershed (PRW) composed the first dataset (Dongli *et al.*, 2017). The second dataset was

formed by data collected in the Ellert Creek Watershed (ECW), which is a headwater of the PRW. It is likely that the local hydrological events in ECW influence the hydrological behavior of the downstream PRW. The soil data in ECW was sampled in a 105-point grid (Soares *et al.*, 2020). From that point, the soil database with 205 data points was named PRW.

Insert figure 1

At all 205 sampling points, both disturbed and undisturbed samples were collected from the surface soil (0-20 cm). The disturbed samples were used to analyze the contents of sand, silt, and clay. These contents were measured by the USDA classification which utilizes the sedimentation method with sampling through pipette (Gee and Bauder, 1986). The organic carbon (OC) content was analyzed by the wet combustion method (Walkey-Black, 1934).

The undisturbed soil cores were sampled using a volumetric ring (a diameter of 5 cm and a height of 3 cm) to determine soil bulk density (BD), macroporosity, and microporosity (< -6 kPa). The soil water content (θ) was analyzed in each sample at seven matric potentials (h). After saturation, it was first measured at 0 kPa. Next, it was measured after applying matric potentials (h) of -1, -6, and -10 kPa with the use of a tension table. Lastly, it was measured after applying matric potential of -33, -100, and 1,500 with the use of Richard's pressure chamber apparatus.

The soil database was also composed of landscape attributes, also known as environmental variables by Wahren *et al.* (2016) and Carvalho *et al.* (2020), which represent local scale morphometry. Thus, we extracted six environmental variables from an acquired digital elevation model (DEM) using ArcGIS software. The DEM belonged to the ALOS sensor (Advanced Land Observing Satellite of Japan Aerospace Exploration Agency - JAXA) with 12.5 m spatial resolution. The environmental variables derived in all 205 sampling points were elevation, aspect,

soil slope, plan curvature, profile curvature, and curvature. These variables are most commonly derived in digital soil mapping (Chagas *et al.*, 2010; Bhering *et al.*, 2016; Chagas *et al.*, 2016; Giasson *et al.*, 2016; Wahren *et al.* 2016; Carvalho *et al.* 2020).

A partition was done in the soil database following the holdout method, such that two subsets were generated (Raschka, 2018). The first set contained 80% of data (training set) and the remaining 20% was used as a test set to evaluate the performance of all PC-PTF models.

Descriptive statistical analysis

Training and test sets were assessed via statistical analysis in order to observe the behavior of each variable studied in both sets. To analyze the sets, mean, standard deviation, minimum, and maximum values were calculated. The statistical analysis was done in the python programming language with the aid Sklearn metrics library.

Development of point PTFs with multiple linear regression

The soil database was submitted for analysis of Spearman's correlation coefficient to observe the relationships among all the attributes studied. Also, this analysis allows for the identification of the best predictors for estimating the soil water content on PRW. The training set was used to develop point PTFs through the multiple linear regression (MLR) method (Vereecken and Herbst, 2004). This method was chosen with the objective of comparing the capability of several MLR models to estimate the soil water content in the evaluated matric potentials ($h = 0$; -1; -6; -10; -33; -100; and -1,500 kPa) with several ANN models, since the MLR is widely used to develop point PTFs due to its ease of application (Gunarathna *et al.*, 2019).

Development of artificial neural network PC-PTF

This study utilized the Pseudo-continuous PTF (PC-PTF) approach, established by Haghverdi *et al.* (2012), because it is able to not only predict the soil water content in any matric potentials (h) determined, but also derive PTFs with limited data. When developing PC-PTF, the first step is to predefine the matric potential as the input data and the respective water content as the output data. Thereafter, the PC-PTF approach calls to increase the training dataset with a factor exact to the number of matric potentials (h) data used to determine the soil water content in each point in the dataset (Haghverdi *et al.*, 2014).

In this study, the PC-PTF settings utilized the machine learning algorithm ANN according to the methodology of Haghverdi *et al.* 2012 with modifications. The changes in the methodology were applied in the hyper-parameters and ANN architecture (e.g., number of neurons, hidden layers, learning rate, seed, and features explored in “input data”, momentum, and number of training epochs), which were done in open code in the language Python using Google colab and Jupyter tools. The attributes used as input data of the PC-PTF were matric potentials ($h = 0$; -1; -6; -10; -33; -100; and -1,500 kPa), sand, silt, and clay content, bulk density, macroporosity, microporosity, soil organic carbon, elevation, aspect, soil slope, plan curvature, profile curvature, and curvature. The input data were not normalized before the training ANN process.

Training and Regularization

During ANN training the original training set was subdivided into two subsets, a training subset and a validation subset. To obtain the validation subset, the train-test split function of Numpy library was utilized such that 10% of the original training set would be used for validation in a random way. The validation subset was then used to evaluate the training model and to control the issue of overfitting.

There are important factors to consider when aiming for the best ANN estimates while avoiding the issue of overfitting (Pham *et al.*, 2019). early stopping and batch normalization with PyTorch library help:

- The early stopping procedure was applied based on the validation set using specific metrics because of its simplicity and effectiveness in both improving the performance of ANNs and avoiding the overfitting issue (Xiong *et al.*, 2011; Madhyastha e Jain, 2019; Pham *et al.*, 2019).
- Batch normalization allows regularization of the input data in the ANN and helps the model avoid saturated nonlinearities because it can normalize each input data regardless of mean and variance changes during training (Srivastava *et al.*, 2014; Ioffe e Szegedy, 2015; Chollet, 2020). The batch normalization was applied in all ANN layers.

Design and hyper-parameters in the ANN

ANN are designed by optimizing the hyper-parameters during model training and generalization (Bergstra and Bengio, 2012). A hyper-parameter is a parameter that defines an adequate value prior to the learning process of a learning algorithm, and is different from the parameter that is obtained after the learning process (Ge *et al.*, 2018). In this study, the hyper-parameters considered to improve ANN performance were the number of neurons in the first layer (input data), the number of neurons in the hidden layer, the number of hidden layers, the epochs, the optimizer, the loss function, the activation function, momentum, the pseudorandom number generator seed, and the learning rate.

The number of neurons, layers, input and output data

In the initial PC-PTF design, the first layer contained the features explored in “input data” from the PRW soil database. This included seven soil attributes (sand, silt, and clay content, bulk density, macroporosity, microporosity, and soil organic carbon), six topographical variables (elevation, aspect, soil slope, plan curvature, profile curvature, and curvature), and one neuron that received all the matric potentials ($h = 0; -1; -6; -10; -33; -100; \text{ and } -1,500 \text{ kPa}$). The number of hidden layers varied between 1 and 3 layers. The number of neurons in the hidden layers varied between 5 and 35 neurons. These hyper-parameters were assessed according to the performance during hyper-parameters optimization in the ANN. The output layer was composed of one neuron, which generates the soil water content (θ) in the respective h selected.

Training epochs

Epoch is related to one complete transition through the full training set (Fang *et al.*, 2020). During the algorithm training, the network repeatedly cycles through all training set in order to adjust the weights at different nodes of the network until reaching convergence (Aggarwal, 2018). The number of necessary epochs was controlled by monitoring the training and validation loss values behavior throughout the training and by the early stopping technique.

Optimizer

An optimizer updates the network parameters to minimize the loss function value (Fang *et al.*, 2020). Among the most common optimizers is the *Stochastic gradient-descent* method, which helps minimize the loss function of prediction through achieving gradient-descent updates of the ANN weights for each minibatch (Aggarwal, 2018). Thus, this optimizer was applied during the training process of ANN.

Loss function

A loss function is a measure of compatibility between the prediction and true values. The purpose of parameter optimization is to minimize the loss value that can be calculated by a loss function (Fang et al., 2020). In this work, the MSE loss function was selected and expressed as follow:

$$MSE = \frac{1}{n} \sum_{i=0}^n (e_i - m_i)^2 \quad (1)$$

where n is the number of data instances used during training and validation, and e_i and m_i are, respectively, the predicted and true values. The monitoring of the loss function was conducted by the analysis of the curve in a graphical form through the matplotlib library.

Activation function

Activation functions are designed to convert linear relationships into nonlinear ones among input-output data (Minasny and McBratney, 2002; Fang *et al.*, 2020). Non-linear activation functions were utilized in all the layers. We used the Rectified Linear Unit (ReLU) function (Gloro and Bengio, 2010).

Pseudorandom number generator seed

The seed hyper-parameter defines the initialization point during the training with the objective of avoiding initialization in multiple local minimums in the loss surface, which makes it difficult to achieve model stability (Madhyastha e Jain, 2019). This parameter also was varied using random seeds when the training was initialized.

Learning rate

The learning rate is a hyper-parameter that determines how much the weights are updated up to network convergence, which is optimized with the validation set when many updates occur (Glorot and Bengio, 2010; Xu *et al.*, 2019; Fang *et al.*, 2020). Learning rate schedulers were used to explore the best suits values for our objective

Random search

Hyper-parameters are a key factor to achieving better learning algorithm performance. Hence, during neural network design, it is common to execute grid search experiments to explore the best hyper-parameter. Nevertheless, how this technique executes the search over a grid of trials based on their performance on the validation set, turn this method more expensive in computational time (e.g., our study would need about 2000 years to explore all the permutations of hyper-parameters mentioned above).

In the light of the foregoing, it is considered that random search emerges also as an optimization easier way of ANN hyper-parameters (Bergstra and Bengio, 2012). Due to the fact that this technique executes a random space search to select the possible hyper-parameter values into a bounded domain, as well as randomly taking some sampling points of these hyper-parameters in the assumed domain, expecting suboptimal results. In this study, hyper-parameters of PC-PTF (the number of neurons in the first layer = 14 - 3, momentum = 0.01; 0.02; 0.1; and 0.2; the pseudorandom number generator seed; and the learning rate = schedulers) were optimized by random search technique following Bergstra and Bengio (2012) methodology.

Model evaluation and performance criteria

The test set was used to verify the predictive capabilities of the models developed by PC-PTF in the different hyper-parameter architecture and the PTFs with RLM. Haghverdi *et al.* (2012) recommends analyzing these predictive capabilities by using combinations of selected statistics between measured and predicted soil water content data. The model evaluations were done while taking into account the statistical indices, such as coefficient of determination (r^2), Mean Error (ME), and the Root Mean Square Error (RMSE) expressed as follow:

$$ME = \frac{1}{n} \sum_{i=1}^n (e_i - m_i) \quad (2)$$

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (e_i - m_i)^2} \quad (3)$$

where n is the number of data instances used for modeling, and e_i and m_i are, respectively, the predicted and measured target data. In this study, each criterion was assessed separately and the RMSE metric was selected as the main criterion of comparison because it expresses the overall prediction error (Equation 3) (De Vos *et al.*, 2005). The evaluation of the estimated data behavior and the measured data in a graphical way were assessed with the library matplotlib library.

Results and discussion

Exploratory analysis of training and test datasets

Table 1 summarizes the descriptive statistics of both training and test datasets for each variable used as input data in the PC-PTF design. The characteristics of variables studied in both datasets are expressed in terms of means, standard deviations, minimum, and maximum values. These results show that both datasets are largely similar, which it could avoid systematic differences when the model evaluations are being executed on the test set while the soil water content in any matric potential is being estimated.

Insert table 1

Assessment of pedotransfer functions through multiple regression

The PTFs development through the MLR technique do not showed efficiency to estimate the soil water content in all matric potentials, which is observed in table 2. It was observed that the worst performances were found in the estimates θ_s and θ_{-1} with r^2 of -41.98 e 0.20 respectively, which could be an indicator that other techniques must be explored to estimate the soil water content efficiency for all matric potentials studied.

Insert table 2

The best performances of PTFs obtained from the MLR technique were found in the θ_{-33} , θ_{-100} e $\theta_{-1,500}$ potentials, showing that water retention of PRW soils in these determined potentials depends specifically on the attributes of microporosity and soil organic carbon. The presence of these attributes in the PTFs is also due to the highest correlation coefficients of the θ_{-33} , θ_{-100} e $\theta_{-1,500}$.

1,500 potentials, which is observed in Spearman's correlation matrix (Table 3). The dependence of water retention due to the microporosity was also reported by Auler *et al.* (2019), who evaluated the influence of several soil physical attributes for estimating water retention in diverse soil management systems through PTFs. These authors found optimal performances of PTFs with $r^2 > 0.70$ and affirmed that a large amount of water is retained in the small soil pores. Guanarathna *et al.* (2019), when evaluating the performance of developed PTFs to estimate the water retention in Sri Lanka soils, noticed that the generated PTFs with MLR to different input levels of attributes have improved their performance. These improvements were noticed when estimating θ_{-33} e $\theta_{-1,500}$ with the PTFs were used in the soil organic content as a predictor. This result was also observed in the present study, where the soil water content predictions between θ_{-33} e $\theta_{-1,500}$ were improved when the soil organic carbon was used as a predictor attribute.

Insert table 3

Performance of Pseudo-continuous pedotransfer function

The architecture more efficient of PC-PTF based on artificial neural networks to predict the soil water content for all matric potentials, was chosen during the evaluation performance of the ANN in the training process. During the assembly of ANN architecture, several structures were examined, and the best architectures were found in the one ANN with one input layer, two hidden layers, and one output layer. The first layer has seven neurons and one of them receives the matric potential to be estimated as input data. The two hidden layers have 20 and 10 neurons respectively, and the output layer has one neuron. The structure of the first layer, which receives the input attributes, was chosen according to the behavior of the variables set in the random search technique.

Therefore, the variables chosen as input parameters in the PC-PTF were those that optimized the assessed models and improved the soil water content predictions for all matric potentials.

The application of the random search technique also allowed us to find the set of hyper-parameters and the input attributes that improve the soil water content estimations for all matric potentials evaluated. Hence, among the ranges of hyper-parameters explored, it was possible to find which hyper-parameters values improved the ANN performance. The following hyper-parameters values were found as the best setting of ANN: learning rate= 0.2, moment= 0.2, number of epochs= 5,000, and seed= 98,186. This hyper-parameter setting was due to the optimum model generated by the ANN until now, which was chosen as the base to assess the ANN performance in the validation and testing phases. During the search of hyper-parameters and attributes input, more optimal for the ANN, was analyzed the loss function curve in the training process. Figure 2 shows the loss function behavior analyzed (MSE) in the present study for the best model of ANN chosen. This function was evaluated in relation to the number of epochs during the training process, realizing the oscillation absence on the error (MSE) by the end of this training.

Insert figure 2

The random search technique also allowed us to identify the input set that improved the soil water content estimates in all matric potentials. The attributes set with better performance were composed of soil bulk density, soil organic carbon, microporosity, clay, curvature, and plane curvature.

In the soil water retention estimative through PTFs is common that attributes like soil bulk density, soil organic carbon, microporosity, and clay are employed as input parameters, due to the intrinsic relationship with this soil hydraulic property (Minasny *et al.*, 1999; Schaap *et al.*, 2001;

Wösten *et al.*, 2001). Nevertheless, few searches have studied the topographical attributes as input parameters in models because these can have influences on soil retain water and water movement (Leij *et al.*, 2004; Famiglietti *et al.*, 1998; Western *et al.*, 1999; Pachepsky *et al.*, 2001). Leij *et al.* (2004) studying the relationship between the topographical attributes and the hydraulic properties concluded that this relationship exists because the soil formation processes are affected by several factors like vegetation, parent material, microclimate, moisture regime, and material transport, factors which are affected by topography. The same authors suggest that the following steps in the water retention research could use more topographical attributes when the data of other predictors do not exist or the availability be limited.

The ANN (PC-PTF) training was satisfactory in the performance of estimates for the majority of trained models throughout the iterations. The performance of the best model chosen during the training process is presented in table 3a, observing that the estimated and measured data curves showed proximity and followed the measured data trend. This specific model showed the following error metrics: $r^2 = 0.71$; $ME = 0.06 \text{ cm}^3 \text{ cm}^{-3}$; and $RMSE = 0.07 \text{ cm}^3 \text{ cm}^{-3}$.

Insert figure 3

The absence of overfitting in the training models was verified through the performance of the estimates conducted with the validation set. In Figure 3b are observed the model estimates with better performance assessed with the validation set. The error metrics generated with the validation set ($r^2 = 0.67$, $ME = 0.06 \text{ cm}^3 \text{ cm}^{-3}$, and $RMSE = 0.07 \text{ cm}^3 \text{ cm}^{-3}$) were very close to those obtained with the training set, proving that the ANN models do not have overfitting. Furthermore, it is noticed that estimated data followed the trend of measured data when the validation set was used.

When identifying that the generated models by the ANN do not present trends to the overfitting, it was decided to execute the ANN training without splitting the data internally into a training set and validation set. Consequently, a new training set was generated with all training data, executing the evaluation of ANN statistical performance with this new set. The loss function curve monitoring and the predictions conducted during the training with the new set showed a satisfactory behavior of the best model. The loss function during the new training did not present oscillations (Figure 4a) and followed the estimates trend of the soil water during the training process, showing similar statistical metrics ($r^2 = 0.73$, $ME = 0.05 \text{ cm}^3 \text{ cm}^{-3}$, and $RMSE = 0.07 \text{ cm}^3 \text{ cm}^{-3}$) to the training with the validation set (Figure 4b).

Insert figure 4

The model assessment with better performance generated by ANN with the new training set is shown in Table 4. This table shows that all the tested models presented the same input attributes (selected in the random search technique) but these models have the variation of a particular hyper-parameter "seed". This hyper-parameter variation was executed within a probabilistic exploration of the random search technique to find the best combinations among the hyper-parameters. The analysis allowed us to evaluate the influence of the seed variation on the estimates, confirming this influence and the randomness in the search by the minimum global point on the loss function curve during the ANN training process (Table 4).

Insert table 4

Within a random search by the best hyper-parameters, it is possible to see that the choice of seed value can influence the soil water content estimates with the ANN model chosen. In the chosen model, it is noted that the higher seed values tend to improve the performances of the estimates for water retention in all matric potentials during the training process (Table 4). For this reason, the generated model by PC-PTF that estimated better the soil water content for all potentials is the model that showed the following hyper-parameter setting: seed = 98.186; learning rate = 0.2; number of epochs = 5,000; and moment = 0.2 with statistical metrics of performance $r^2 = 0.73$; ME = $0.052 \text{ cm}^3 \text{ cm}^{-3}$; and RMSE = $0.065 \text{ cm}^3 \text{ cm}^{-3}$.

Regarding attributes chosen by the best model as input parameters (curvature and plane curvature, it is important to highlight the presence of topographical attributes (Table 4). From the random search conducted by the random search technique, it was possible to identify that those attributes have a relationship of dependency with the soil water retention. This relationship can not be identified when using more simple statistical techniques that only get to identify linear and non-linear relationships in one or two dimensions. Consequently, the ANN technique in association with the random search technique allows exploring relationships between soil physical attributes and topographical attributes that improve the soil water content estimates. One example where there is a lack of capacity of those techniques to identify the relationships between the soil attributes and topographical attributes is expressed in the results found in the present study, where is noted the absence of topographical attributes as a predictor in the developed PTFs with the MLR technique (Table 2) and the existence of low coefficients between the topographical attributes and the soil water content for all the evaluated potentials (Table 3).

A study carried out by Leij *et al.* (2014) examined the performance of developed PTFs to a watershed in Basilicata-Italy, using soil properties and topographical attributes. The author concluded that the topographical attributes can be linearly correlated with the hydraulic soil

properties and contribute new information to the predictions. In addition, the topographical attributes are widely available in digital elevation models that are used in the hydrologic modeling of watersheds. This availability allows the use of topographical attributes to help interpret and estimate better the hydraulic soil properties like the hydraulic conductivity and the water retention in complex soil systems. The same authors affirm that attributes like curvature are an important determinant of the lateral flow patterns, allowing those depressed areas are wetter in relation to the areas more elevated or plain that show differences in the sedimentation patterns. On the other hand, Pachepsky *et al.* (2001) found that the water content in specific matric potentials (-10 and 33 kPa) had a positive correlation with the curvature, indicating that the curvature divergence promotes soil water retention. In the same way, Leij *et al.* (2005) found a positive correlation between the residual soil water content (θ_r) and the curvature. In contrast to the other authors, Famiglietti *et al.* (1998) reported in their study a negative correlation between the surface water content and three types of curvature.

After, the ANN performance was tested with a data set unknown by the ANN (test set). When used the test set is perceptible that performance statistical metrics were equal for the majority of generated models by the ANN (Table 4). This performance confirms the ANN capacity to generalize the model and generate estimates more reliable with data unknown by the ANN. The generalization capacity of the model is observed in Figure 5, where the model estimates of the ANN with better performance are close to the measured data ($r^2 = 0.73$; $ME = 0.05 \text{ cm}^3 \text{ cm}^{-3}$, and $RMSE = 0.06 \text{ cm}^3 \text{ cm}^{-3}$) and follow the real data trend.

Insert figure 5

Once identified the best ANN model was needed to assess its performance to estimate the soil water content for all matric potential separately. Table 5 shows the performance statistical metrics for the best model chosen with the test set, noting that the model did not have the capacity to predict the soil water content for all matric potentials separately.

A hypothesis of the ANN's failure to estimate in an independent way the soil water content for each potential is due to the database size used in the ANN initial training. The network got to generalize very well the regression model for estimating the soil water using all the matric potentials as input data but failed when conducting the specific estimates that require a soil database with a higher amount of data.

Insert table 5

Conclusions

The use of the Pseudo-continuous pedotransfer functions based on artificial neural networks allowed us to find the relationship between the soil physical attributes (soil bulk density, soil organic carbon, microporosity, and clay) and the topographical attributes (curvature and plane curvature), confirming that relationships exist among those attributes in other dimensions that need to be studied. It is therefore suggested that new studies be conducted to explore the topographical attributes' influences on the soil water content estimates, using the pseudo-continuous pedotransfer functions based on artificial neural networks along with the random search technique. However, during the study of those relationships, it is needed to use a soil database with thousands of data from several topographical environments and climates that allow understanding the specificities of those relationships.

References

- Aggarwal CC. Neural networks and deep learning. Springer; 2018.
- Auler, A.C., Peres, L.F., Pineda, M.C, 2017. Influence of physical attributes and pedotransfer function for predicting water retention in management systems. *Revista Brasileira de Engenharia Agrícola e Ambiental* 21(11), 746-751. <http://dx.doi.org/10.1590/1807-1929/agriambi.v21n11p746-751>
- Bergstra, J., Bengio, Y., 2012. Random search for hyper-parameter optimization. *Journal of Machine Learning Research* 13, 281-305.
- Bhering, S.B., Chagas, C.D., Carvalho Junior, W., Pereira, N.R., Calderano Filho, B., Pinheiro, H.S.K., 2016. Digital mapping of sand, clay, and soil carbon by Random Forest models under different spatial resolutions. *Pesq Agropec Bras.* 51, 1359-1370. <https://doi.org/10.1590/s0100-204x2016000900035>
- Blake, G.R., Hartge, K.H. Particle density. 1986. In: KLUTE, A., ed. *Methods of soil analysis*. 2.ed. Madison, Soil Science Society of America, 377-382.
- Brooks, R.H., Corey, A.T., 1964. Properties of porous media affecting fluid flow. *Journal of Irrigation and Drainage* 92, 61–88.
- Carvalho Junior, W.de, Pereira, N.R., Fernandes Filho, E.I., Calderano Filho, B., Pinheiro, H.S.K., Chagas, C. da S., Bhering, S.B., Pereira, V.R., Lawall, S., 2020. Sample design effects on soil unit prediction with machine: randomness, uncertainty, and majority map. *Rev. Bras. Cienc. Solo*, 44:e0190120.
- Chagas C.S., Carvalho Junior W., Bhering S.B., Calderano Filho B., 2016. Spatial prediction of soil surface texture in a semiarid region using random forest and multiple linear regressions. *Catena* 139, 232-240. <https://doi.org/10.1016/j.catena.2016.01.001>
- Chagas, C.S., Fernandes Filho, E.I., Vieira, C.A.O., Schaefer, C.G.R., Carvalho Júnior, W. de., 2010. Topographic attributes and Landsat7 data in the digital soil mapping using neural networks. *Pesq. agropec. bras.* 45(5), p. 497-507.
- Chollet, F., 2020. Fundamentals of machine learning. In: *Deep Learning with Python*. 2. ed. Shelter Island, NY: Manning Publications Co. p. 93–116.
- De Vos, B., Meirvenner, M.V., Quataert, P., Deckers, J., Muys, B., 2005. Predictive quality of pedotransfer functions for estimating bulk density of forest soils. *Soil Science Society of America Journal*, 69, 500-510.
- D’Emilio, A., Aiello, R., Consoli, S., Vanella, D., Lovino, M., 2018. Artificial neural networks for predicting the water retention curve of Sicilian agricultural soils. *Water* 10 (1431), 1-13.

- Dongli, S., Qian, C., Timm, L.T., Beskow, S., Wei, H., Caldeira, T.L., Oliveria, L.M., 2017. Multi-scale correlations between soil hydraulic properties and associated factors along a Brazilian watershed transect. *Geoderma*, 286, 15–24. <http://dx.doi.org/10.1016/j.geoderma.2016.10.017>.
- Famiglietti, J. S., Rudnicki, J.W., Rodell, M. 1998. Variability in surface moisture content along a hillslope transect: Rattlesnake Hill, Texas, *J. Hydrol.* 210, 259– 281.
- Fang, Z., Wang, Y., Peng, L., Hong, H., 2020. Integration of convolutional neural network and conventional machine learning classifiers for landslide susceptibility mapping. *Computers and Geosciences* 139, 104470.
- Gee, G.W., Bauder, J.W., 1986. Particle-size analysis. in: Klute, A. (Eds.), *Methods of Soil Analysis*. Part 1, 2nd ed Agronomy Monograph. ASA-SSSA, Madison, WI, 9, 383–411.
- Ge, L., Hang, R., Liu, Y., Liu, Q., 2018. Comparing the performance of neural network and deep convolutional neural network in estimating soil moisture from satellite observations. *Remote Sensing* 10, 1327
- Géron, Aurélien, 2017. *Hands-on Machine Learning with Scikit-Learn& TensorFlow*.
- Giasson, E., Caten, A.T., Bagatini, T., Bonfatti, B., 2015. Instance selection in digital soil mapping: a study case in Rio Grande do Sul, Brazil. *Ciência Rural* 45(9), 1592-1598.
- Glorot, X., Bengio, Y., 2010. Understanding the difficulty of training deep feedforward neural networks. *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics, JMLR Workshop and Conference Proceedings* 9, 249-256.
- Gunarathna, M.H.J.P., Sakai, K., Nakandakari, T., Momii, K., Kumari, M.K.N., Amarasekara, M.G.T.S., 2019. Pedotransfer functions to estimate hydraulic properties of tropical Sri Lankan soils. *Soil & Tillage Research*, 190, 109-119.
- Haghverdi, A., Cornelis, W.M., Ghahraman, B., 2012. A pseudo-continuous neural network approach for developing water retention pedotransfer functions with limited data. *Journal of Hydrology* 442-442, 46-54.
- Haghverdi, A., Öztürk, H.s, Cornelis, W.M., 2014. Revisiting the pseudo continuous pedotransfer function concept: Impact of data quality and data mining method. *Geoderma* 226-227, 31-38
- Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction*. New York: Springer
- Ioffe, S., Szegedy, C., 2015. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. *arXiv:1502.03167*
- Leij, F.J., Romano, N., Palladino, M., Schaap, M.G., Coppola, A., 2004. Topographical attributes to predict soil hydraulic properties along a hillslope transect. *Water Resource Research* 40, 1-15.

- Madhyastha, P., Jain, R., 2019. On Model Stability as a Function of Random Seed. arXiv:1909.10447
- Minasny, B., McBratney, A.B., 2002. The *Neuro-m* method for fitting neural network parametric pedotransfer functions. *Soil. Sci. Soc. Am. J.* 66, 352-361.
- Minasny, B., A. B. McBratney, and K. L. Bristow, 1999. Comparison of different approaches to the development of pedotransfer functions for water-retention curves, *Geoderma* 93, 225– 253.
- Pachepsky, Y. A., Timlin, D, Rawls, W.J., 2001. Soil water retention as related to topographic variables, *Soil Sci. Soc. Am. J.* 65, 1787– 1795.
- Pham, K., Kim, D., Yoon, Y., Choi, H., 2019. Analysis of neural network based pedotransfer function for predicting soil water characteristic curve. *Geoderma* 351, 92-102
- Raschka, S., 2018. Model Evaluation, Model Selection, and Algorithm Selection in Machine Learning. arXiv:1811.12808, 49.
- Schaap, M.G., Leij, F.J., van Genuchten, M.T., 2001. Rosetta: A computer program for estimating soil hydraulic parameters with hierarchical pedotransfer functions, *Journal of Hydrology* 251, 163– 176.
- Soares, M.F., Centeno, L.N., Timm, L.C., Mello, C.R., Kaiser, D.R., Beskow, S., 2020. Identifying covariates to assess the spatial variability of saturated soil hydraulic conductivity using robust cokriging at the watershed scale. *Journal of Soil Science and Plant Nutrition* 20, (3), 1491-1502.
- Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R., 2014. Dropout: A simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.* 15(1), 1929–1958.
- van Genuchten, M.Th., 1980. A closed-form equation for predicting the hydraulic conductivity of unsaturated. *Soil Science Society of America Journal* 43, 892–898.
- Vereecken, H., Herbst, M., 2004. Statistical regression. In: Pachepsky, Y.A., Rawls, W.J. (Eds.), *Development of Pedotransfer Functions in Soil Hydrology*. Elsevier, Amsterdam, pp. 3–18.
- Vereecken, H., Weynants, M., Javaux, M., Pachepsky, Y., Schaap, M.G., van Genuchten, M.Th., 2010. Using pedotransfer functions to estimate the van Genuchten–Mualem soil hydraulic properties: a review. *Vadose Zone Journal* 9, 1–26.
- Wahren, F.T., Julich, S., Nunes, J.P., Gonzales-Pelayo, O., Hawtree, D., Feger, K.H., Keizer, J.J., 2016. Combining digital soil mapping and hydrological modeling in a data scarce watershed in north-central Portugal. *Geoderma* 264, 350-362.
- Walkley, A., Black, I.A., 1934. An examination of the Degtjareff method for determining soil organic matter and a proposed modification of the chromic acid titration method. *Soil Science* 37, 29-38.

Wang, G., Zhang, Y., Yu, N., 2012. Prediction of soil water retention and available water of sandy soils using pedotransfer functions. *Procedia Engineering* 31, 49-53.

Western, A.W., Grayson, R.B., Blöschl, G., Willgoose, G.R., McMahon, T.A., 1999. Observed spatial organization of soil moisture and its relation to terrain indices, *Water Resour. Res.* 35, 797– 810.

Wösten, J.H.M., Pachepsky, Y.A., Rawls, W.J., 2001. Pedotransfer functions: Bridging the gap between available basic soil data and missing soil hydraulic characteristics. *Journal of Hydrology* 251, 123–150.

Xiong, H.Y., Barash, Y., Frey, B.J., 2011. Bayesian prediction of tissue-regulated splicing using RNA sequence and cellular context. *Bioinformatics* 27(18), 2554-2562.

Xu, Z., Dai, A.M., Kemp, J., Metz, 2019. Learning an Adaptive Learning Rate Schedule. *arXiv:1909.09712*

Table 1.

Descriptive statistics of soil and environmental variables used in this study for training and sets.

Variables	Training set (n = 164)				Test set (n = 41)			
	Min	Max	Mean	SD	Min	Max	Mean	SD
Sand (%)	30.75	91.57	55.91	11.26	30.22	83.73	55.03	10.32
Silt (%)	1.49	50.96	24.79	7.97	9.30	35.95	25.12	7.46
Clay (%)	4.72	45.88	19.30	7.18	6.97	52.71	19.85	8.08
BD (g cm ⁻³)	0.81	1.73	1.40	0.18	1.10	1.67	1.44	0.12
Ma (cm ³ cm ⁻³)	0.009	0.537	0.224	0.10	0.018	0.316	0.196	0.09
Mi (cm ³ cm ⁻³)	0.073	0.475	0.241	0.07	0.149	0.425	0.249	0.07
OC (%)	0.13	3.20	1.58	0.82	0.22	2.90	1.46	0.87
Z (m)	48.0	406.0	279.3	100.91	50	348.3	275.1	99.50
Slope (%)	0.00	38.85	14.97	7.96	0.00	35.92	13.80	7.90
Aspect (cos(ϕ))	-1	359.9	185.3	94.18	-1.00	344.10	184.40	97.29
Cv (1/m)	-23.10	40.16	-0.05	8.48	-3.51	1.91	-0.02	0.98
PlCv (1/m)	-20.02	26.00	0.35	3.96	-1.15	1.53	0.09	0.57
PrCv (1/m)	-17.56	12.80	-0.26	2.74	-0.98	4.13	0.11	0.82
θ_s (cm ³ cm ⁻³)	0.226	0.695	0.443	0.08	0.270	0.620	0.443	0.08
θ_{-1} (cm ³ cm ⁻³)	0.185	0.638	0.381	0.07	0.227	0.490	0.382	0.06
θ_{-6} (cm ³ cm ⁻³)	0.119	0.522	0.265	0.07	0.155	0.459	0.275	0.07
θ_{-10} (cm ³ cm ⁻³)	0.107	0.483	0.237	0.07	0.128	0.443	0.248	0.07
θ_{-33} (cm ³ cm ⁻³)	0.060	0.413	0.189	0.07	0.086	0.353	0.202	0.07
θ_{-100} (cm ³ cm ⁻³)	0.027	0.375	0.164	0.07	0.065	0.313	0.177	0.07
$\theta_{-1,500}$ (cm ³ cm ⁻³)	0.015	0.339	0.136	0.06	0.047	0.304	0.150	0.07

Min: minimum value; Max: maximum value, Mean: mean value; SD: standard deviation; BD: bulk density; Ma: macroporosity, Mi: microporosity, OC: organic carbon content; Z: elevation; Cv: curvature; PlCv: plain curvature; PrCv: profile curvature; θ_s : Saturated soil water content; θ_{-1} , θ_{-6} , θ_{-10} , θ_{-33} , θ_{-100} , and $\theta_{-1,500}$: soil water content in -1, -6, -10, -33, -100 e -1,500 kPa, respectively.

Table 2.

Pedotransfer functions to estimate the soil water content in different matric potentials through multiple linear regression with test set.

Pedotransfer functions	Performance		
	ME	RMSE	r ²
$\theta_s = 0.629 - 0.171\text{BD} + 0.034\text{OC}$	0.438	0.530	-41.98
$\theta_{-1} = 0.254 + 0.524\text{Mi}$	0.045	0.057	0.20
$\theta_{-6} = 0.085 + 0.623\text{Mi} + 0.002\text{C}$	0.028	0.039	0.64
$\theta_{-10} = 0.059 + 0.641\text{Mi} + 0.001\text{C}$	0.028	0.037	0.69
$\theta_{-33} = 0.031 + 0.521\text{Mi} + 0.021\text{OC}$	0.026	0.033	0.74
$\theta_{-100} = 0.010 + 0.476\text{Mi} + 0.024\text{OC}$	0.027	0.034	0.74
$\theta_{-1,500} = -0.005 + 0.422\text{Mi} + 0.025\text{OC}$	0.028	0.036	0.72

BD: bulk density; OC: organic carbon content; Mi: microporosity; C: Clay; θ_s : Saturated soil water content ($\text{cm}^3 \text{ cm}^{-3}$); θ_{-1} , θ_{-6} , θ_{-10} , θ_{-33} , θ_{-100} , and $\theta_{-1,500}$: soil water content ($\text{cm}^3 \text{ cm}^{-3}$) in -1, -6, -10, -33, -100 e -1,500 kPa, respectively.

Table 3
Matrix of Spearman’s correlation coefficients (ρ) between all assessed variables

values spearman correlation coefficient (ρ) before land leveling																				
	Clay	Silt	Sand	BD	Ma	Mi	OC	Z	Slope	Aspect	Cv	PlCv	PrCv	θ_s	θ_{-1}	θ_{-6}	θ_{-10}	θ_{-33}	θ_{-100}	$\theta_{-1,500}$
Clay	1.0	0.130	-0.710**	-0.281**	-0.070	0.348**	0.114	0.277**	-0.090	0.028	0.139*	0.003	-0.100	-0.264**	0.385**	0.404**	0.382**	0.346**	0.332**	0.337**
Silt	-	1.0	-0.738*	-0.101	0.257**	-0.209**	-0.518**	0.455**	0.167*	0.277**	0.048	0.062	-0.108	-0.090	0.089	-0.050	-0.124	-0.216**	-0.253**	-0.276**
Sand	-	-	1.0	-0.286**	-0.186**	-0.046	0.321**	-0.522*	-0.049	-0.191*	-0.139*	-0.081	0.116	-0.125	-0.332**	-0.215**	-0.143*	-0.044	-0.007	0.006
BD	-	-	-	1.0	-0.515**	-0.100	-0.134	-0.184**	-0.034	0.023	-0.124	-0.106	0.020	-0.385**	-0.361**	-0.152*	-0.096	-0.041	-0.031	-0.054
Ma	-	-	-	-	1.0	-0.751**	-0.493**	0.511**	0.161*	0.167*	0.119	0.169*	-0.070	-0.031	-0.170*	-0.480**	-0.534**	-0.578**	-0.585**	-0.553**
Mi	-	-	-	-	-	1.0	0.673**	-0.477**	-0.145*	-0.202**	-0.072	-0.162*	0.099	0.379**	0.535**	0.751**	0.771**	0.778**	0.777**	0.756**
OC	-	-	-	-	-	-	1.0	-0.638**	-0.127	-0.264**	-0.075	-0.117	0.126	0.348**	0.264**	0.478**	0.536**	0.615**	0.647**	0.655**
Z	-	-	-	-	-	-	-	1.0	-0.044	-0.141*	0.088	0.092	-0.050	-0.147*	-0.109	-0.146*	-0.172*	-0.203**	-0.209**	-0.206**
Slope	-	-	-	-	-	-	-	-	1.0	-0.024	0.049	-0.025	-0.044	-0.038	-0.043	-0.153*	-0.179*	-0.194*	-0.197**	-0.189**
Aspect	-	-	-	-	-	-	-	-	-	1.0	0.088	0.092	-0.050	-0.147*	-0.109	-0.146*	-0.172*	-0.203**	-0.209**	-0.206**
Cv	-	-	-	-	-	-	-	-	-	-	1.0	0.685**	-0.773**	0.023	-0.032	-0.121	-0.134	-0.144*	-0.140*	0.132
PlCv	-	-	-	-	-	-	-	-	-	-	-	1.0	-0.392**	-0.081	-0.134	-0.180**	-0.183**	-0.182**	-0.179*	-0.166*
PrCv	-	-	-	-	-	-	-	-	-	-	-	-	1.0	0.092	0.110	0.165*	0.171*	0.174*	0.172*	0.166*
θ_s	-	-	-	-	-	-	-	-	-	-	-	-	-	1.0	0.788**	0.592**	0.552**	0.494**	0.469**	0.457**
θ_{-1}	-	-	-	-	-	-	-	-	-	-	-	-	-	-	1.0	0.829**	0.749**	0.635**	0.594**	0.578**
θ_{-6}	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	1.0	0.985**	0.932**	0.901**	0.862**
θ_{-10}	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	1.0	0.977**	0.955**	0.916**
θ_{-33}	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	1.0	0.994**	0.965**
θ_{-100}	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	1.0	0.984**
$\theta_{-1,500}$	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	1.0

BD: bulk density; Ma: macroporosity, Mi: microporosity, OC: organic carbon content; Z: elevation; Cv: curvature; PlCv: plain curvature; PrCv: profile curvature; θ_s : Saturated soil water content; θ_{-1} , θ_{-6} , θ_{-10} , θ_{-33} , θ_{-100} , and $\theta_{-1,500}$: soil water content in -1, -6, -10, -33, -100 e -1,500 kPa, respectively. *P < 0.05 and **P < 0.01. Values not indicated were not significant at 5%.

Table 4.

Performance of the Pseudo-continuous pedotransfer functions based on artificial neural networks to estimate the soil water content in any matric potentials for the best model.

Selected input atributes	Seed	Training set			Test set		
		ME	RMSE	r^2	ME	RMSE	r^2
<i>h</i> BD OC Mi C Curvature Plane curvature	98186	0.052	0.065	0.73	0.049	0.062	0.73
	81705	0.053	0.067	0.72	0.051	0.064	0.72
	8581	0.055	0.068	0.71	0.052	0.066	0.71
	42786	0.055	0.069	0.70	0.053	0.067	0.70
	878	0.055	0.069	0.70	0.053	0.067	0.70
	26915	0.056	0.069	0.70	0.053	0.067	0.70
	68097	0.056	0.070	0.69	0.053	0.067	0.69
	30093	0.069	0.086	0.53	0.068	0.084	0.52
	48417	0.079	0.097	0.41	0.080	0.100	0.32
	41938	0.080	0.099	0.39	0.087	0.107	0.23
	22738	0.093	0.114	0.19	0.089	0.109	0.19

h: matric potentials; BD: bulk density; OC: organic carbon content; Mi: microporosity; C: Clay; Curvature: surface curvature; Plane curvature: Plane curvature of the terrain. Hyper-parameters values of the ANN for the best model: seed= 98.186, learning rate= 0.2, number of epochs= 5,000 and moment= 0.2.

Table 5.

Performance of the Pseudo-continuous pedotransfer functions based on artificial neural networks to estimate the soil water content in specific matric potentials for the best model.

Soil water content	Test set		
	ME	RMSE	r ²
θ_s	0.063	0.079	0.05
θ_{-1}	0.045	0.059	0.16
θ_{-6}	0.042	0.053	0.35
θ_{-10}	0.046	0.059	0.20
θ_{-33}	0.059	0.059	0.23
θ_{-100}	0.059	0.057	0.25
$\theta_{-1,500}$	0.056	0.056	-0.05

Hyper-parameters values of the ANN for the best model: seed= 98.186, learning rate= 0.2, number of epochs= 5,000 and moment= 0.2.

Figure Captions

Fig 1. Watersheds maps indicating samples points collected along a spatial-transect in the Pelotas River Watershed (PRW) and a grid in the Ellert Creek Watershed (ECW).

Fig 2. Loss function monitoring expressed in MSE during the training epochs with the training set.

Fig 3. a) Comparison of the θ predictions vs measured data with training data b) Comparison of the θ predictions vs measured data with validation data.

Fig 4. a) Loss function monitoring expressed in MSE during the training epochs with the new training set b) Comparison of the θ predictions vs measured data during new training.

Fig 5. Performance of the θ predictions vs measured data with the test set.

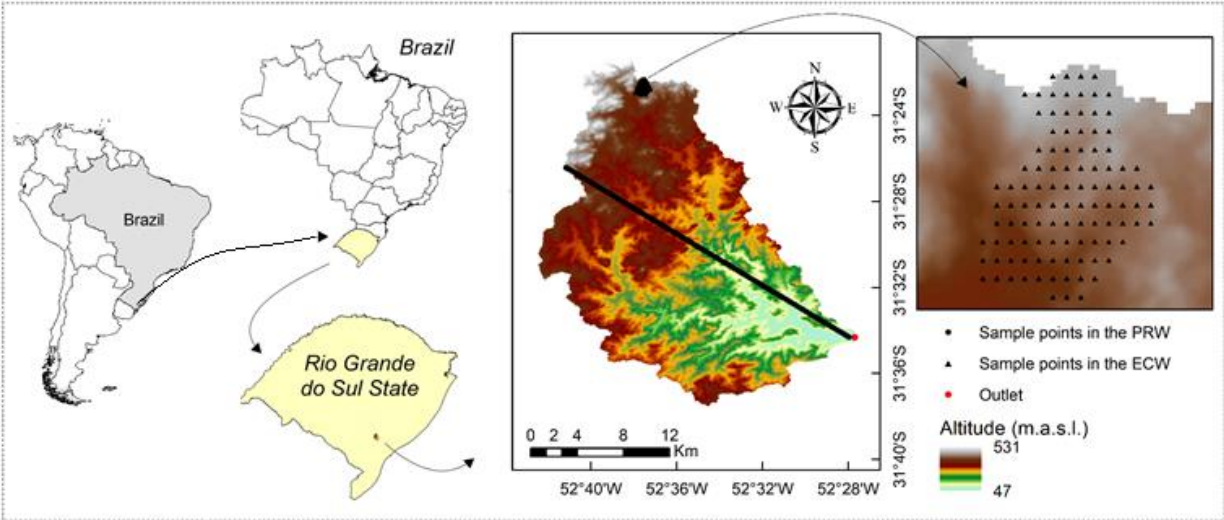


Fig. 1.

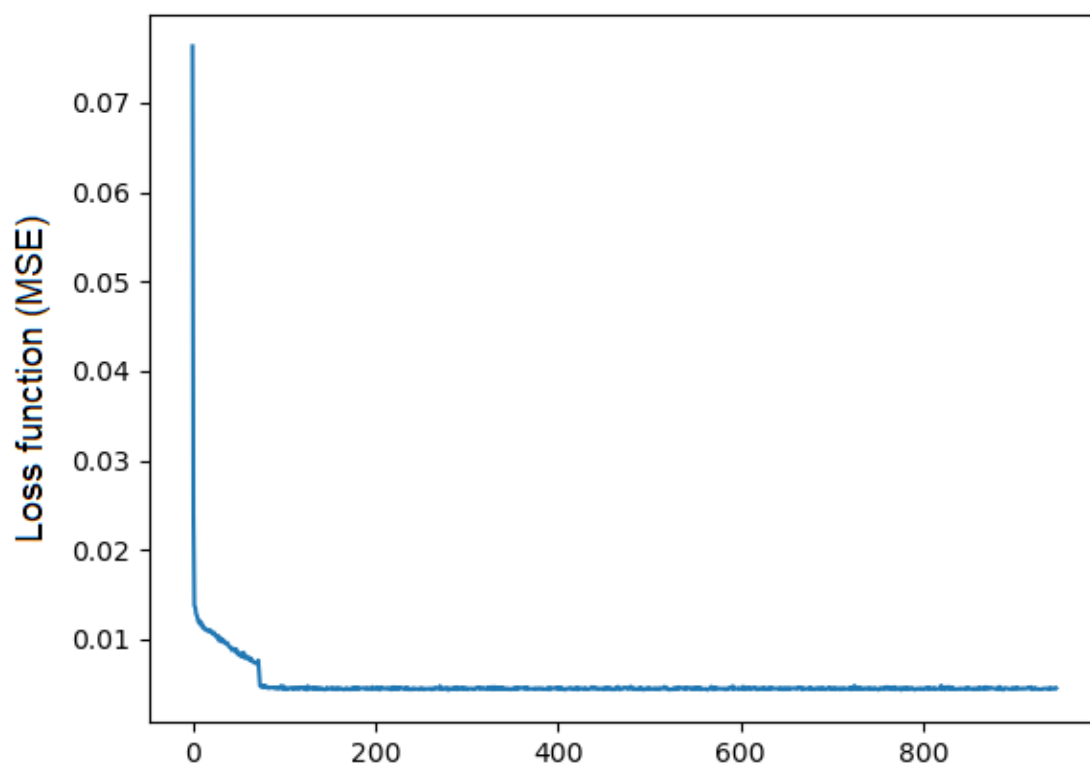


Fig. 2.

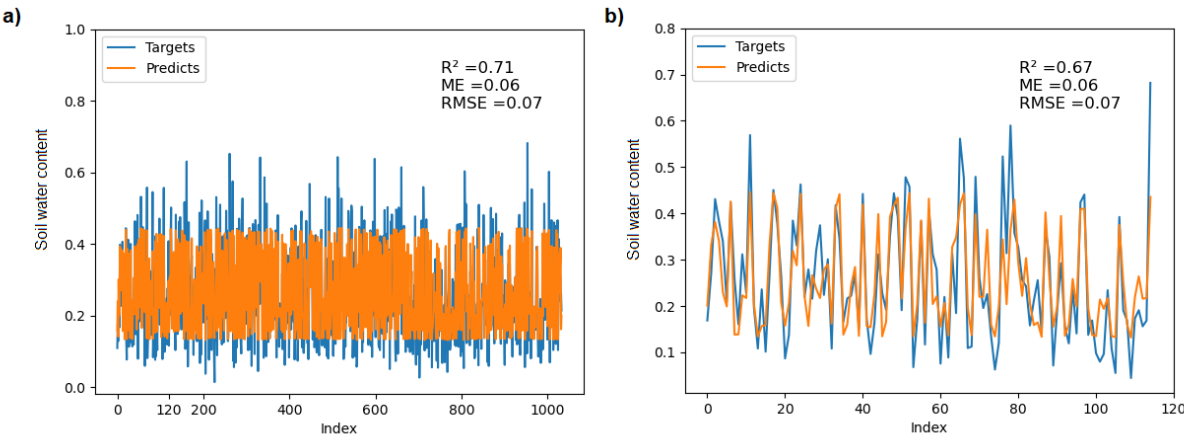


Fig. 3.

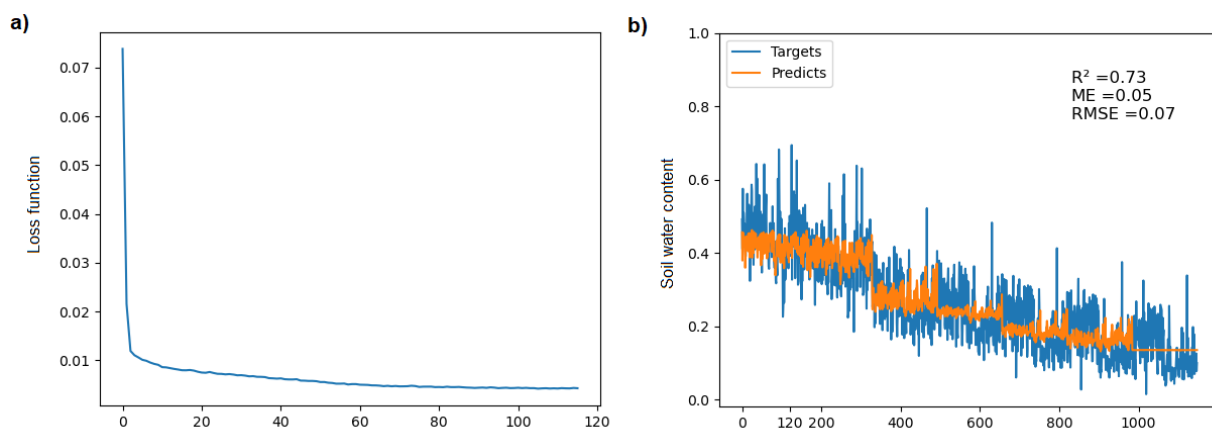


Fig. 4.

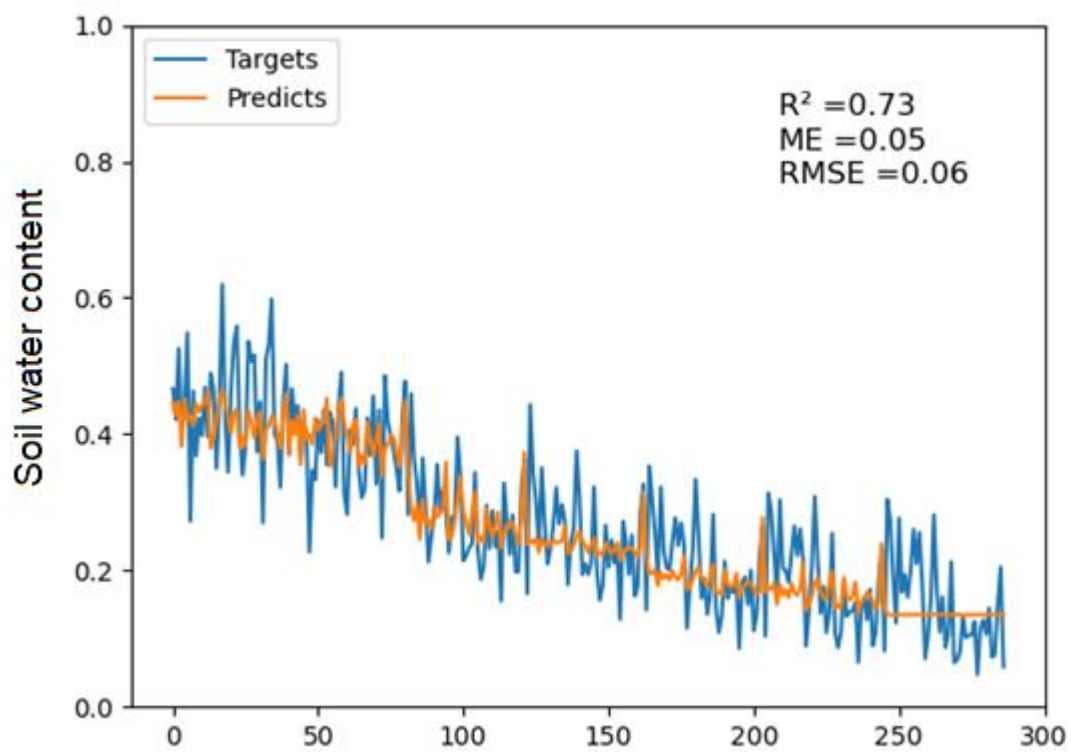


Fig. 5.

6 Considerações Finais

As funções de pedotransferência desenvolvidas em regiões de condições climáticas distintas às da bacia Arroio Pelotas demonstraram ter potencial em estimar a retenção de água no solo, confirmando suas possíveis transferibilidade e aplicabilidade. Portanto, elas têm potencial de serem usadas nas regiões agrícolas da bacia em projetos relacionados à irrigação e conservação de água no solo.

Os valores medidos de água no solo permitiram estimar as variações da evapotranspiração real como função das variações de θ_{fc} e θ_{pwp} e de acordo aos usos do solo, sendo as maiores inclinações dessas variações dadas pelos usos floresta nativa, silvicultura e pastagem cultivada, o qual indica taxas de evaporação menores nesses usos que apresentam uma maior cobertura vegetal. Além disso, as estimativas de água no solo e das variações de evapotranspiração total pelas funções de pedotransferência permitiram escolher quais têm maior potencial de utilização na modelagem hidrológica da bacia Arroio Pelotas baixo clima subtropical. Assim, este método permite identificar FPTs que são mais indicadas de usar em áreas da bacia que se requeiram ter conhecimento da retenção de água do solo em função da evapotranspiração real e do uso do solo.

A função de pedotransferência pseudo-contínua sob o enfoque de redes neurais artificiais e com o uso da técnica *random search* permitiu encontrar relações em outras dimensões entre os atributos físicos do solo, topográficos e a retenção de água no solo na bacia hidrográfica Arroio Pelotas. Portanto, sugere-se que sejam realizados novos estudos para explorar as influências dos atributos topográficos sobre as estimativas da retenção de água no solo, usando bases de dados maiores com dados de solo de diversos climas e ambientes topográficas que permitam entender as particularidades dessas relações.

Recomenda-se que sejam realizados trabalhos futuros como: uso de algoritmo genético e *random search* para a estimativa da retenção de água no solo, aplicação de rede neural convolucional na predição da retenção de água no solo, *random search* para a exploração de novos atributos topográficos e físico-hídricos e a sua influência na estimativa da retenção de água no solo, máquinas de vetores de suporte para a melhoria da estimativa de água no solo, *k-fold crossvalidation* na aleatorização dos conjuntos de

treino e de testes para melhoria das estimativas de retenção de água no solo, rede neural artificial para a estimativa da retenção de água no solo a partir do banco nacional de dados de solos brasileiros (HYBRAS), função de pedotransferência pseudo-contínua para estimativa da retenção de água no solo a partir do banco de dados HYBRAS.

Referências

AGGARWAL, C. C. **Neural Networks and Deep Learning**. [s.l.] Springer, Cham, 2018. 512 p.

ARAÚJO, L. **Solving XOR with a single Perceptron**. [s.l.: s.n.].

ARRUDA, F. B.; ZULLO, J. R.; OLIVEIRA, J. B. Parâmetros de solo para o cálculo de água disponível com base na textura do solo. **Revista Brasileira de Ciência do Solo**, v. 11, p. 11–15, 1987.

ARYA, L. M.; PARIS, J. F. A Physicoempirical Model to Predict the Soil Moisture Characteristic from Particle-Size Distribution and Bulk Density Data. **Soil Science Society of America Journal**, v. 45, p. 1023–1030, 1981.

ASCHONITIS, V. G.; KOSTOPOULOU, S. K.; ANTONOPOULOS, V. Z. Methodology to Assess the Effects of Rice Cultivation Under Flooded Conditions on van Genuchten's Model Parameters and Pore Size Distribution. **Transport in Porous Media**, v. 91, p. 861–876, 2012.

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.; CEZAR, L. I.; LIER, Q. D. J. Van; MAIA, A. D. 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.

BAYAT, H.; MAZAHARI, B.; MOHANTY, B. P. Estimating soil water characteristic curve using landscape features and soil thermal properties. **Soil and Tillage Research**, v. 189, p. 1–14, jun. 2019.

BAYAT, H.; NEYSHABURI, M. R.; MOHAMMADI, K.; NARIMAN-ZADEH, N.; IRANNEJAD, M.; GREGORY, A. S. Combination of artificial neural networks and fractal theory to predict soil water retention curve. **Computers and Electronics in Agriculture**, v. 92, p. 92–103, 2013.

BERGSTRA, J.; BENGIO, Y. Random Search for Hyper-Parameter Optimization Yoshua Bengio. **Journal of Machine Learning Research**, v. 13, p. 281–305, 2012. Disponível em: <<http://scikit-learn.sourceforge.net>>.

BORGESSEN, C. D.; SCHAAP, M. G. Point and parameter pedotransfer functions for water retention predictions for Danish soils. **Geoderma**, v. 127, p. 154–167, 2005.

BOTULA, Y. D.; CORNELIS, W. M.; BAERT, G.; VAN RANST, E. Evaluation of pedotransfer functions for predicting water retention of soils in Lower Congo (D.R. Congo). **Agricultural Water Management**, v. 111, p. 1–10, 2012.

BOTULA, Y.-D.; RANST, E. Van; CORNELIS, W. M. Pedotransfer functions to predict water retention for soils of the humid tropics: A review. **Revista Brasileira de Ciência do Solo**, v. 38, p. 679–698, 2014.

BOUMA, J. Using Soil Survey Data for Quantitative Land Evaluation. **Advances in Soil Science**, v. 9, p. 177–213, 1989.

BROWNLEE, J. **How to choose loss functions when training deep learning neural networks**. Disponível em: <<https://machinelearningmastery.com/how-to-choose-loss-functions-when-training-deep-learning-neural-networks/>>.

BRUAND, A. Utilizing mineralogical and chemical information in PTFs. **Developments in Soil Science**, v. 30, p. 153–158, 2004.

BUDIMAN, M.; MCBRATNEY, A. B.; MENDONÇA-SANTOS, M. de L.; SANTOS, H. G. dos. **Revisão sobre funções de pedotransferência (PTFs) e novos métodos de predição de classes e atributos do solo** Documentos n. 45 /Embrapa Solos. [s.l.: s.n.].

CAMPBELL, R. **Demystifying deep neural nets**. Disponível em: <<https://medium.com/@RosieCampbell/demystifying-deep-neural-nets-efb726eae941>>.

CASTELLINI, M.; IOVINO, M. Pedotransfer functions for estimating soil water retention curve of Sicilian soils. **Archives of Agronomy and Soil Science**, v. 65, n. 10, p. 1401–1416, 2019.

CHOLLET, F. Fundamentals of machine learning. In: **Deep Learning with Python**. 2. ed. Shelter Island, NY: Manning Publications Co., 2020. p. 93–116.

CORNELIS, W. M.; RONSYN, J.; VAN MEIRVENNE, M.; HARTMANN, R. Evaluation of Pedotransfer Functions for Predicting the Soil Moisture Retention Curve. **Soil Science Society of America Journal**, v. 65, p. 638–648, 2001.

COSTA, A. da; ALBUQUERQUE, J. A.; ALMEIDA, J. A. de; COSTA, A. da; LUCIANO, R. V. Pedotransfer functions to estimate retention and availability of water in soils of the state of Santa Catarina, Brazil. **Revista Brasileira de Ciência do Solo**, v. 37, p. 889–910, 2013.

DA SILVA, A. C.; ARMINDO, R. A.; BRITO, A. dos S.; SCHAAP, M. G. An Assessment of Pedotransfer Function Performance for the Estimation of Spatial Variability of Key Soil Hydraulic Properties. **Vadose Zone Journal**, v. 16, n. 9, p. 1–10, 2017.

DA SILVA, E.; CURI, N.; FERREIRA, M. M.; VOLPATO, M. M. L.; DOS SANTOS, W. J. R.; SILVA, S. H. G. Pedotransfer functions for water retention in the main soils from the Brazilian coastal plains. **Ciência e Agrotecnologia**, v. 39, n. 4, p. 331–338, 2015.

- DEISENROTH, M. P.; FAISAL, A. A.; ONG, C. S. **Mathematics for machine learning**. [s.l.] Cambridge: Cambridge University Press, 2020. 417 p.
- DIERCKX, J.; BELMANS, C.; PAUWELS, P. **SWATREX - a computer package for modelling the field water balance. Ref. Manual**. [s.l.: s.n.].
- DOBARCO, M. R.; COUSIN, I.; LE BAS, C.; MARTIN, M. P. Pedotransfer functions for predicting available water capacity in French soils, their applicability domain and associated uncertainty. **Geoderma**, v. 336, p. 81–95, 2019.
- DONATELLI, M.; WÖSTEN, J. H. M.; BELOCCHI, G. Methods to evaluate pedotransfer functions. **Developments in Soil Science**, v. 30, p. 357–411, 2004.
- DONGLI, S.; QIAN, C.; TIMM, L. C.; BESKOW, S.; WEI, H.; CALDEIRA, T. L.; DE OLIVEIRA, L. M. Multi-scale correlations between soil hydraulic properties and associated factors along a Brazilian watershed transect. **Geoderma**, v. 286, p. 15–24, 2017.
- ERTEL, W. **Introduction to Artificial Intelligence**. Oxford, UK: Springer, 2017. 365 p.
- ESPINO, A.; MALLANTS, D.; VANCLOOSTER, M.; FEYEN, J. Cautionary notes on the use of pedotransfer functions for estimating soil hydraulic properties. **Agricultural Water Management**, v. 29, p. 235–253, 1995.
- FAMIGLIETTI, J. S.; RUDNICKI, J. W.; RODELL, M. Variability in surface moisture content along a hillslope transect: Rattlesnake Hill, Texas. **Journal of Hydrology**, v. 210, n. 1–4, p. 259–281, 1998.
- GALLEGO, M. C. V. Algoritmo de búsqueda aleatoria pura aplicado al diseño de planes de muestreo doble. **Revista Universidad EAFIT**, v. 40, n. 135, p. 28–399, 2004.
- GIAROLA, N. F. B.; SILVA, A. P.; IMHOFF, S. Relações entre propriedades físicas e características de solos da região sul do Brasil. **Revista Brasileira de Ciência do Solo**, v. 26, p. 885–893, 2002.
- GUNARATHNA, M. H. J. P.; SAKAI, K.; NAKANDAKARI, T.; MOMII, K.; KUMARI, M. K. N.; AMARASEKARA, M. G. T. S. Pedotransfer functions to estimate hydraulic properties of tropical Sri Lankan soils. **Soil & Tillage Research**, v. 190, p. 109–119, 2019.
- HAGHVERDI, A.; CORNELIS, W. M.; GHAHRAMAN, B. A pseudo-continuous neural network approach for developing water retention pedotransfer functions with limited data. **Journal of Hydrology**, v. 442–443, p. 46–54, 2012.
- HAGHVERDI, A.; ÖZTÜRK, H. S.; CORNELIS, W. M. Revisiting the pseudo continuous pedotransfer function concept: Impact of data quality and data mining method. **Geoderma**, v. 226–227, n. 1, p. 31–38, 2014.
- HAYKIN, S. **Neural Networks: A Comprehensive Foundation**. India: Pearson Prentice Hall, 1994. 823 p.

- HERBST, M.; DIEKKRÜGER, B.; VEREECKEN, H. Geostatistical co-regionalization of soil hydraulic properties in a micro-scale catchment using terrain attributes. **Geoderma**, v. 132, p. 206–221, 2006.
- JAMES, G.; WITTEN, D.; HASTIE, T.; TIBSHIRANI, R. **An Introduction to Statistical Learning - with Applications in R**. New York: Springer, 2017. 419 p.
- JANA, R. B.; MOHANTY, B. P.; SHENG, Z. Upscaling soil hydraulic parameters in the picacho mountain region using bayesian neural networks. **Transactions of the ASABE**, v. 55, n. 2, p. 463–473, 2012.
- KOEKKOEK, E. J. W.; BOOLTINK, H. Neural network models to predict soil water retention. **European Journal of Soil Science**, v. 50, p. 489–495, 1999.
- LAL, R. Physical properties and moisture retention characteristics of some Nigerian soils. **Geoderma**, v. 21, p. 209–223, 1979.
- LI, X.; XU, X.; LIU, W.; XU, C.; ZHANG, R.; WANG, K. Prediction of profile soil moisture for one land use using measurements at a soil depth of other land uses in a karst depression. **Journal of Soils and Sediments**, v. 19, p. 1479–1489, 2018.
- LIBARDI, P. L. **Dinâmica da água no solo**. São Paulo: EDUSP, 2012. 352 p.
- MANZAN, J. R. G. **Análise de desempenho de redes neurais artificiais do tipo multilayer perceptron por meio do distanciamento dos pontos do espaço de saída**. 2016. Universidade Federal de Uberlândia, 2016.
- MCBRATNEY, A. B.; MINASNY, B.; CATTLE, S. R.; VERVOORT, R. W. From pedotransfer functions to soil inference systems. **Geoderma**, v. 109, n. 2002, p. 41–73, 2002.
- MCNEILL, S. J.; LILBURNE, L. R.; CARRICK, S.; WEBB, T. H.; CUTHILL, T. Pedotransfer functions for the soil water characteristics of New Zealand soils using S-map information. **Geoderma**, v. 326, p. 96–110, 2018.
- 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. de; 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. **Ciencia Rural**, v. 40, n. 4, p. 848–853, 2010.
- MINASNY, B.; HARTEMINK, A. E. Predicting soil properties in the tropics. **Earth-Science Reviews**, v. 106, p. 52–62, 2011.
- MINASNY, B.; MC BRATNEY, A. B.; MCKENZIE, N. J.; GRUNDY, M. J. Predicting soil properties using pedotransfer functions and environmental correlation. *In*: MCKENZIE,

- N. J.; GRUNDY, M. J.; WEBSTER, R.; RINGROSE-VOASE, A. J. **Guidelines for Surveying Soil and Land Resources Land Resources**. [s.l.] CSIRO Publishing, 2008. p. 349–367.
- MINASNY, B.; MCBRATNEY, A. B. Limited effect of organic matter on soil available water capacity. **European Journal of Soil Science**, v. 69, p. 39–47, 2018.
- MINASNY, B.; MCBRATNEY, Alex. B. The Neuro-m method for fitting neural network parametric pedotransfer functions. **Soil Science Society of America Journal**, v. 66, p. 352–361, 2002.
- MITCHELL, T. M. **Machine Learning**. 1. ed. New York: McGraw-Hill Science/Engineering/Math, 1997. v. 17432 p.
- MOREIRA DE MELO, T.; PEDROLLO, O. C. Artificial neural networks for estimating soil water retention curve using fitted and measured data. **Applied and Environmental Soil Science**, v. 2015, p. 1–16, 2015.
- MULAZZANI, R. P. **Desenvolvimento de pedotransferência física-empírica para modelagem da retenção de água no solo**. 2016. Universidade Federal de Santa Maria, 2016.
- NEBEL, Á. L. C.; TIMM, L. C.; CORNELIS, W.; GABRIELS, D.; REICHARDT, K.; AQUINO, L. S.; PAULETTO, E. A.; REINERT, D. J. Pedotransfer functions related to spatial variability of water retention attributes for lowland soils. **Revista Brasileira de Ciência do Solo**, v. 34, n. 3, p. 669–680, 2010.
- NEMES, A.; SCHAAP, M. G.; LEIJ, F. J.; WÖSTEN, J. H. M. Description of the unsaturated soil hydraulic database UNSODA version 2.0. **Journal of Hydrology**, v. 251, p. 151–162, 2001.
- NEMES, A.; SCHAAP, M. G.; WÖSTEN, J. H. M. Functional Evaluation of Pedotransfer Functions Derived from Different Scales of Data Collection. **Soil Science Society of America Journal**, v. 67, p. 1093–1102, 2003.
- NEMES, A.; WÖSTEN, J. H. M.; BOUMA, J.; VÁRALLYAY, G. Soil water balance scenario studies using predicted soil hydraulic parameters. **Hydrological Processes**, v. 20, p. 1075–1094, 2006.
- NGUYEN, P. M.; HAGHVERDI, A.; DE PUE, J.; BOTULA, Y. D.; LE, K. V.; WAEGEMAN, W.; CORNELIS, W. M. Comparison of statistical regression and data-mining techniques in estimating soil water retention of tropical delta soils. **Biosystems Engineering**, v. 153, p. 12–27, 2017.
- NGUYEN, P. M.; VAN LE, K.; BOTULA, Y. D.; CORNELIS, W. M. Evaluation of soil water retention pedotransfer functions for Vietnamese Mekong Delta soils. **Agricultural Water Management**, v. 158, p. 126–138, 2015.
- OLIVEIRA, L. M. De. **Relação entre atributos do solo aplicando a abordagem em Espaço de Estados em duas bacias hidrográficas na região sul do Rio Grande do Sul**. 2015. Universidade Federal de Pelotas, 2015.

- OLIVEIRA, L. B.; RIBEIRO, M. R.; JACOMINE, P. T. K.; RODRIGUES, J. J. V.; MARQUES, F. A. Funções de pedotransferência para predição de umidade retida a potenciais específicos em solos do estado do Pernambuco. v. 26, p. 315–323, 2002.
- OTTONI, M. V.; OTTONI FILHO, T. B.; SCHAAP, M. G.; LOPES-ASSAD, M. L. R. C.; ROTUNNO FILHO, O. C. Hydrophysical Database for Brazilian Soils (HYBRAS) and Pedotransfer Functions for Water Retention. **Vadose Zone Journal**, v. 17, n. 1, p. 1–17, 2018.
- OTTONI, M. V. **Sistema de classificação dos solos baseado na estrutura do espaço poroso**. 2017. Universidade Federal do Rio De Janeiro, 2017.
- PACHEPSKY, Y. A.; RAJKAI, K.; TÓTH, B. Pedotransfer in soil physics: Trends and outlook - A review. **Agrokemia es Talajtan**, v. 64, n. 2, p. 339–360, 2015.
- PACHEPSKY, Y. A.; RAWLS, W. J.; LIN, H. S. Hydropedology and pedotransfer functions. **Geoderma**, v. 131, p. 308–316, 2006.
- PATIL, N. G.; SINGH, S. K. Pedotransfer Functions for Estimating Soil Hydraulic Properties: A Review. **Pedosphere**, v. 26, n. 4, p. 417–430, 2016.
- PEREIRA, T. dos S. **Uso de inteligência artificial para estimativa da capacidade de suporte de carga do solo**. 2017. Universidade Federal de Santa Maria, 2017.
- PEREIRA, T. dos S.; ROBAINA, A. D.; PEITER, M. X.; TORRES, R. R.; BRUNING, J. The use of artificial intelligence for estimating soil resistance to penetration. **Engenharia Agrícola**, v. 38, n. 1, p. 142–148, 2018.
- PHAM, K.; KIM, D.; YOON, Y.; CHOI, H. Analysis of neural network based pedotransfer function for predicting soil water characteristic curve. **Geoderma**, v. 351, p. 92–102, 2019.
- PICCIAFUOCO, T.; MORBIDELLI, R.; FLAMMINI, A.; SALTALIPPI, C.; CORRADINI, C.; STRAUSS, P.; BLÖSCHL, G. A pedotransfer function for field-scale saturated hydraulic conductivity of a small watershed. **Vadose Zone Journal**, p. 1–15, 2019.
- PIDGEON, J. D. The measurement and prediction of available water capacity of ferrallitic soils in Uganda. **Journal of Soil Science**, v. 23, n. 4, p. 431–441, 1972.
- PINEDA, M. C.; VILORIA, J.; MARTÍNEZ-CASASNOVAS, J. A.; VALERA, A.; LOBO, D.; TIMM, L. C.; PIRES, L. F.; GABRIELS, D. Predicting soil water content at – 33 kPa by pedotransfer functions in stoniness 1 soils in northeast Venezuela. **Environmental Monitoring and Assessment**, v. 190, n. 161, p. 1–11, 2018.
- RAJKAI, K.; KABOS, S.; VAN GENUCHTEN, M. T. Estimating the water retention curve from soil properties: Comparison of linear, nonlinear and concomitant variable methods. **Soil and Tillage Research**, v. 79, p. 145–152, 2004.
- RASCHKA, S. Model Evaluation , Model Selection , and Algorithm Selection in Machine Learning. **arXiv:1811.12808**, p. 49, 2018.

- RASTGOU, M.; BAYAT, H.; MANSOORIZADEH, M.; GREGORY, A. S. Estimating the soil water retention curve : Comparison of multiple nonlinear regression approach and random forest data mining technique. **Computers and Electronics in Agriculture**, v. 174, p. 105502, 2020.
- RAWLS, W. J.; BRAKENSIEK, D. L.; SAXTON, K. E. Estimation of Soil Water Properties. **Transactions of the ASAE**, v. 25, p. 1316–1320, 1982.
- RAWLS, W. J.; PACHEPSKY, Y. A.; RITCHIE, J. C.; SOBECKI, T. M.; BLOODWORTH, H. Effect of soil organic carbon on soil water retention. **Geoderma**, v. 116, p. 61–76, 2003.
- REICHERT, J. M.; ALBUQUERQUE, J. A.; KAISER, D. R.; REINERT, D. J.; URACH, F. L.; CARLESSO, R. Estimation of water retention and availability in soils of rio grande do sul. **Revista Brasileira de Ciência do Solo**, v. 33, p. 1547–1560, 2009.
- REYNOLDS, W. D.; DRURY, C. F.; TAN, C. S.; FOX, C. A.; YANG, X. M. Use of indicators and pore volume-function characteristics to quantify soil physical quality. **Geoderma**, v. 152, p. 252–263, 2009.
- RICH, E. Artificial Intelligence. *In*: **Annals of Emergency Medicine**. New York: McGraw-Hill, 1983. p. 119–121.
- SANTOS, W. J. R. dos; CURI, N.; SILVA, S. H. G.; ARAÚJO, E. F. de; MARQUES, J. J. Pedotransfer functions for water retention in different soil classes from the center-southern Rio Grande Do Sul state. v. 37, n. 1, p. 49–60, 2013.
- SCHAAP, M. G.; LEIJ, F. J.; VAN GENUCHTEN, M. T. ROSETTA: a computer program for estimating soil hydraulic parameters with hierarchical pedotransfer functions. **Journal of Hydrology**, v. 251, p. 163–176, 2001.
- SCHAAP, M. G.; NEMES, A.; VAN GENUCHTEN, M. T. Comparison of models for indirect estimation of water retention and available water in surface soils. **Vadose Zone Journal**, v. 3, p. 1455–1463, 2004.
- SHI, L.; ÓLAFSSON, S. Nested partitions method for global optimization. **Operations Research**, v. 48, n. 3, p. 390–407, 2000.
- SILVA, A. C. da; ARMINDO, R. A.; BRITO, A. dos S.; SCHAAP, M. G. SPLINTEX: A physically-based pedotransfer function for modeling soil hydraulic functions. **Soil and Tillage Research**, v. 174, p. 261–272, 2017. Disponível em: <<http://dx.doi.org/10.1016/j.still.2017.07.011>>.
- SOARES, F. C.; ROBAINA, A. D.; PEITER, M. X.; RUSSI, J. L.; VIVAN, G. A. Redes neurais artificiais na estimativa da retenção de água do solo. **Ciência Rural**, v. 44, n. 2, p. 293–300, 2014.
- SOARES, F. S. **Uso de diferentes metodologias na geração de funções de pedotransferencia para a retenção de água em solos do Rio Grande do Sul**. 2013. UNIVERSIDADE FEDERAL DE SANTA MARIA, 2013.

SOARES, M. F. **Variabilidade espacial dos atributos físico-hídricos e do carbono orgânico do solo de uma bacia hidrográfica de cabeceira em Canguçu - RS**. 2018. Universidade Federal de Pelotas, 2018.

TOMASELLA, J.; HODNETT, M. Pedotransfer functions for tropical soils. **Developments in Soil Science**, v. 30, n. C, p. 415–429, 2004.

TOMASELLA, J.; HODNETT, M. G. Estimating soil water retention characteristics from limited data in Brazilian Amazonia. **Soil Science**, v. 163, n. 3, p. 190–202, 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.

TOMASELLA, J.; PACHEPSKY, Ya.; CRESTANA, S.; RAWLS, W. J. Comparison of Two Techniques to Develop Pedotransfer Functions for Water Retention. **Soil Science Society of America Journal**, v. 67, p. 1085–1092, 2003.

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, p. 226–238, 2015.

TRANter, G.; MCBRATNEY, A. B.; MINASNY, B. Using distance metrics to determine the appropriate domain of pedotransfer function predictions. **Geoderma**, v. 149, p. 421–425, 2009.

URACH, F. L. Estimativa da retenção de água em solos para fins de irrigação. **Universidade Federal de Santa Maria (Tese de mestrado)**, p. 78, 2007.

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

VAN LOOY, K.; BOUMA, J.; HERBST, M.; KOESTEL, J.; MINASNY, B.; MISHRA, U.; MONTZKA, C.; NEMES, A.; PACHEPSKY, Y. A.; PADARIAN, J.; SCHAAP, M. G.; TÓTH, B.; VERHOEF, A.; VANDERBORGHT, J.; VAN DER PLOEG, M. J.; WEIHERMÜLLER, L.; ZACHARIAS, S.; ZHANG, Y.; VEREECKEN, H. Pedotransfer Functions in Earth System Science: Challenges and Perspectives. **Reviews of Geophysics**, v. 55, p. 1199–1256, 2017.

VEREECKEN, H.; JAVAUX, M.; WEYNANTS, M.; PACHEPSKY, Y.; SCHAAP, M. G.; GENUCHTEN, V. Using pedotransfer functions to estimate the van genuchten- mualem soil hydraulic properties: A review. **Vadose Zone Journal**, v. 9, p. 795–820, 2010.

VEREECKEN, H.; MAES, J.; FEYEN, J. **Estimating unsaturated hydraulic conductivity from easily measured soil properties** *Soil Science* 1990.

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, n. 6, p. 389–403, 1989.

- WEIHERMÜLLER, L.; LEHMANN, P.; HERBST, M.; RAHMATI, M.; VERHOEF, A.; OR, D.; JACQUES, D.; VEREecken, H. Choice of Pedotransfer Functions Matters when Simulating Soil Water Balance Fluxes. **Journal of Advances in Modeling Earth Systems**, v. 13, n. 3, 1 mar. 2021.
- WÖSTEN, J. H. M.; FINKE, P. A.; JANSEN, M. J. W. Comparison of class and continuous pedotransfer functions to generate soil hydraulic characteristics. **Geoderma**, v. 66, p. 227–237, 1995.
- WÖSTEN, J. H. M.; LILLY, A.; NEMES, A.; LE BAS, C. Development and use of a database of hydraulic properties of European soils. **Geoderma**, v. 90, p. 169–185, 1999.
- ZHANG, Y.; SCHAAP, M. G. Weighted recalibration of the Rosetta pedotransfer model with improved estimates of hydraulic parameter distributions and summary statistics (Rosetta3). **Journal of Hydrology**, v. 547, p. 39–53, 2017.
- ZHAO, C.; SHAO, M.; JIA, X.; NASIR, M.; ZHANG, C. Using pedotransfer functions to estimate soil hydraulic conductivity in the Loess Plateau of China. **Catena**, v. 143, p. 1–6, 2016.
- ZIMMERMANN, E. D.; BASILE, P. A. Estimación de parámetros hidráulicos en suelos limosos mediante diferentes funciones de pedotransferencia. **Tecnología y Ciencias del Agua**, v. 2, n. 1, p. 99–116, 2011.
- ZOU, G.; LI, Y.; WANG, Y.; LIU, D. L.; LIU, X.; LI, Y.; WU, J. Pedo-transfer functions for estimating the hydraulic properties of paddy soils in subtropical central China. **Archives of Agronomy and Soil Science**, v. 62, n. 7, p. 982–993, 2016.

Anexos

Anexo A – Disponibilidade do repositório

Link de acesso ao repositório

https://drive.google.com/drive/folders/1mpM86FerqJDEZnc7dbUjghD1cXn8UhPj?usp=s_haring

Readme do repositório

RNA para estimar Umidade no solo

Este repositório contém o código utilizado para a exploração e treinamento de uma RNA que estime a umidade do solo a partir de algumas variáveis pré coletadas do solo.

Métricas utilizadas para a avaliação de modelos: R^2 , MAE, RMSE.

Ferramentas Utilizadas

1. Python
2. PyTorch
3. SkLearn
4. Pandas
5. Numpy
6. Matplotlib

Estrutura Output

...

└─ /Ano_mês_dia_hora_minuto (Em que o script foi iniciado)

| └─ /plots

| └─ /seed_(Numero da Seed)_lr_(Numero de Learning Rate)_momentum_(Numero de momentum)_epochs_(Numero de épocas)

| | └─ /variáveis selecionadas

| | | └─ /noVal

| | | | └─ /args_no_val.json

| | | | └─ /best_metrics_no_val.csv

| | | | └─ /best_model_no_val.pth

```

| | | | | — /Train_TargetsXOutputs.png
| | | | | — /Train_TargetsXOutputs(Losses).png
| | | | — /test_metrics.csv
| | | | — /Test_TargetsXPredicts.png

```

...

Conteúdo dos Arquivos

1. args_no_val.json: Contém as informações de hiper-parâmetros do modelo, assim como as colunas usadas no treino.
2. best_metrics_no_val.csv: Contém as métricas do melhor modelo encontrado com o conjunto de treino.
3. best_model_no_val.pth: Artefato do melhor modelo encontrado.
4. Train_TargetsXOutputs.png: Gráfico comparando a estimativa com o valor verdadeiro do conjunto de treino.
5. Train_TargetsXOutputs(Losses).png: Gráfico representando a variação da Loss ao longo das épocas com o conjunto de treino.
6. test_metrics.csv: Tabela com as métricas do modelo usando o conjunto de teste.
7. Test_TargetsXPredicts.png: gráfico comparando a estimativa com o valor verdadeiro do conjunto de teste.