CLUSTERIZAÇÃO DE FUNÇÕES DE PEDOTRANSFERÊNCIA PARA CONDUTIVIDADE HIDRÁULICA SATURADA: CARACTERIZAÇÃO METODOLÓGICA E DESEMPENHO

LUIS FERNANDO DA SILVA MARTINEZ[1]; GABRIEL BORGES DOS SANTOS[2]; OTTONI MARQUES MOURA DE LEON[3]; DANIELLE BRESSIANI[4], LUIS CARLOS TIMM[5]

[1]*Universidade Federal de Pelotas – engluisfernandomartinez@gmail.com*
[2]*Universidade Federal de Pelotas – gabrielqwsantos@gmail.com*
[3]*Universidade Federal de Pelotas – ottonibaixo@gmail.com*
[4]*Universidade Federal de Pelotas – daniebressiani@gmail.com*
[5]*Universidade Federal de Pelotas – lcartimm@yahoo.com.br*

1. INTRODUÇÃO

Uma pesquisa original deve trazer novas ideias, perspectivas, resultados metodológicos e ampliar o entendimento sobre diferentes assuntos (BASHITI, 2021). Entretanto, na atual era da informação, somos expostos diariamente a um volume expressivo de novas publicações científicas. Diante disso, torna-se necessário realizar levantamentos sistemáticos da produção existente em uma determinada área, a fim de esclarecer o estado do conhecimento, identificar caminhos para novos estudos e apontar lacunas ainda não exploradas.

Nesse contexto, a área de ciência do solo não é exceção. Propriedades hidráulicas do solo, como a condutividade hidráulica em solo saturado (Ksat), são dificilmente obtidas por medição direta, pelo custo elevado e pela complexidade metodológica inerentes à observação de tais parâmetros (ZHANG *et. al* 2018).

Para contornar essas limitações, pesquisadores têm desenvolvido e aperfeiçoado as chamadas funções de pedotransferência (PTFs). PTFs são conceituadas como equações ou modelos estatísticos/científicos capazes de estimar propriedades do solo de difícil obtenção a partir de informações mais acessíveis, como textura, matéria orgânica e densidade aparente (ZHANG; SCHAAP, 2019).

Desde os trabalhos pioneiros de BOUMA (1989) e VAN GENUCHTEN (1980), até avanços recentes que incorporam técnicas de aprendizado de máquina (ZHANG *et al*., 2018; HAGHVERDI *et al*., 2015), as PTFs têm se consolidado como ferramentas indispensáveis tanto para pesquisas quanto para aplicações práticas em manejo agrícola, engenharia e modelagem hidrológica. E a Ksat é uma das propriedades mais visadas para serem preditas.

A Ksat é um parâmetro que expressa a permeabilidade da água no solo em situação de saturação(OTTONI *et al*., 2025). E é uma propriedade fundamental para diversas finalidades, como projetos de irrigação, modelagem hidrológica e projetos de drenagem agrícola.

Este trabalho apresenta uma revisão bibliográfica sobre PTFs que estimam a Ksat, organizada para extrair e analisar *insights* relacionados às métricas de desempenho reportadas nos estudos, para tal foi desenvolvido um modelo de clusterização hierárquica.

2. METODOLOGIA

O levantamento de dados da revisão bibliográfica consistiu em um processo de seleção de artigos realizado em três etapas. A busca foi efetuada nas bases de dados Scopus e Web of Science, utilizando os termos 'soil saturated hydraulic

conductivity' ou 'saturated hydraulic conductivity' ou 'ks' ou 'k-sat' ou 'ksat' e 'pedotransfer functions' ou 'pedotransfer function' ou 'PTF' ou 'PTFs' nos campos de "título", "resumo" e "palavras-chave". O período de busca compreendeu os limites permitidos em cada plataforma, de 1960 a 2022 na Scopus e de 1945 a 2022 na Web of Science. Inicialmente, foram filtrados 597 artigos, e após a remoção de duplicatas e trabalhos fora do escopo do estudo, restaram 335 artigos para a análise (SANTOS, G. B., 2025). Desses 335 artigos, foram mantidos apenas os artigos que testaram seus modelos em um conjunto de dados diferente dos dados de treino, totalizando 82 amostras.

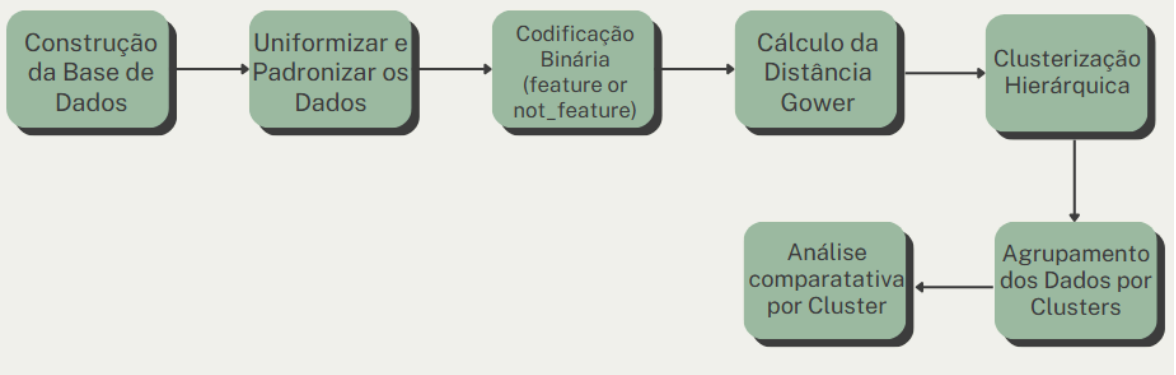
Para cada amostra foram selecionados os trabalhos científicos que possuíam as seguintes características: autoria, país, continente, clima, número de amostras, métodos de modelagem, preditores empregados e métricas para avaliação do desempenho das PTFs (R^2, RMSE). Para a realização da análise utilizou-se a linguagem Python e os dados foram coletados através da leitura de cada artigo.

Durante o pré-processamento, algumas colunas da planilha apresentavam múltiplas informações categóricas listadas em uma única célula, separadas por vírgulas (por exemplo, diferentes preditores ou métodos utilizados em um mesmo estudo). Para tornar esses dados compatíveis com a análise, cada elemento foi decomposto em variáveis binárias independentes (features), de modo que a presença fosse codificada como 1 e a ausência como 0. Além disso, para manter consistência na codificação, foram criadas colunas complementares com o prefixo “not_{variável}”, indicando explicitamente a não utilização de determinado preditor ou método em um artigo. Esse procedimento garantiu que todas as categorias estivessem representadas de forma uniforme, possibilitando o cálculo adequado da distância de Gower entre estudos com diferentes combinações de variáveis.

Tendo em vista o objetivo do presente estudo de agrupar artigos segundo características metodológicas e contextuais heterogêneas, optou-se pelo uso da distância de Gower, em substituição a abordagens tradicionais como o one-hot encoding (OHE) combinado a métricas euclidianas. Essa escolha se deve a três fatores principais: (i) o OHE pode gerar alta dimensionalidade e esparsidade quando aplicado a variáveis categóricas com muitas classes, reduzindo a interpretabilidade e aumentando o custo computacional; (ii) a distância euclidiana não é ideal para comparar vetores esparsos de categorias, pois atribui pesos artificiais à ausência simultânea de atributos; e (iii) a distância de Gower, ao contrário, foi especificamente concebida para lidar com variáveis mistas (categóricas e contínuas), permitindo atribuir a cada variável um peso proporcional e mantendo a comparabilidade entre diferentes tipos de dados (GOWER, 1971). Assim, assegura-se uma medida de similaridade mais robusta e adequada ao objetivo do trabalho.

A matriz de distância gerada alimentou uma clusterização (ou agrupamento) hierárquica aglomerativa e o corte em k clusters foi definido por inspeção do dendrograma e interpretabilidade dos grupos. Para comparar o desempenho entre clusters, foi realizado um agrupamento para mensurar o R^2 médio de cada cluster. Importante ressaltar que o R^2 não foi utilizado como dado de treino para que o modelo ficasse independete da variável de análise. Na Figura 1 apresentamos o fluxograma com as estapas de construção do modelo de agrupamento.

Figura 1 - Fluxograma das Etapas Metodológicas para Construção do Modelo



Fonte: Autores

3. RESULTADOS E DISCUSSÃO

A clusterização gerou 5 grupos interpretáveis: os grupos diferenciam-se principalmente por (i) riqueza de preditores (número/tipo), (ii) tamanho amostral e (iii) famílias de métodos ML/estatísticos. A Figura 2 mostra o tamanho de cada cluster e suas respectivas médias de R^2.

Figura 2 - Análise de R^2 por cluster

cluster	r2_test_mean	numero_artigos
0	0.586966	59
1	0.711375	16
2	0.948000	1
3	0.362000	5
4	0.400000	1

Fonte: Autores

Os clusters 2, 3, 4 obtiveram uma quantidade irrelevante de artigos, então, mesmo que mostrem qualquer resultado promissor, não possuem qualquer significância e poder preditivo. O cluster 1 obteve um número relativamente maior de amostras, com uma média de R^2 maior. O cluster 0 agregou a maioria das amostras, com 59 unidades, representando quase 72% do total.

Para que um cluster gere insights valiosos, é fundamental que ele agrupe dados com um perfil único e bem definido, cujas características o diferenciem claramente dos demais. A relevância desse cluster é confirmada quando ele apresenta um desempenho extremo (muito alto ou muito baixo) em uma métrica chave, que no nosso caso, foi o R^2. São esses extremos que permitem tirar conclusões significativas e direcionar ações.

4. CONCLUSÕES

Este trabalho apresenta um protocolo metodológico para a caracterização e agrupamento de estudos científicos, abordando o tema de funções de pedotransferência (PTFs). A metodologia inovadora proposta combina a distância de Gower com clusterização hierárquica. Nesse estudo apresentamos um fluxograma e passo a passo reprodutível de pré-processamento e sumarização. As principais contribuições e inovações são:

Proposição de um fluxo de trabalho replicável (padronização de metadados → Gower → agglomerative clustering → sumarização por cluster) que

facilita a comparação entre estudos heterogêneos sem depender de codificações excessivas.

Introdução do uso de clusterização de dados mistos como etapa conceitual para agrupar estudos metodologicamente semelhantes antes de comparar métricas de desempenho, reduzindo vieses decorrentes de comparações diretas entre estudos com características distintas.

Limitações metodológicas: o protocolo foi aplicado sobre uma base construída a partir de artigos com formatos variados, com extração manual; isso impõe restrições à generalização imediata do procedimento, mas a metodologia foi definida para ser extensível e automatizável. Outro fato importante é o número relativamente baixo de amostras, o que não permite tirar insights com significância.

Uma limitação relevante deste trabalho foi a utilização exclusiva do **coeficiente de determinação (R^2)** como variável de desempenho. Embora diversos artigos também reportem o **erro quadrático médio (RMSE)**, as unidades desse indicador variam conforme a propriedade hidráulica estimada (ex.: condutividade, umidade volumétrica, pressão mátrica), o que impossibilitou a padronização sem acesso aos dados originais. Dessa forma, a comparação entre estudos com base no RMSE se tornaria inviável ou até enganosa. Assim, optou-se por trabalhar apenas com o R^2, ciente de que essa escolha limita a análise a um único critério de ajuste e pode subestimar aspectos relacionados à magnitude absoluta dos erros.

Em próximos trabalhos será analisado se há alguma variável importante que tenha sido separada individualmente em algum cluster e que, devido a essa variável, seja ela de modelo ou variável preditora, o modelo tenha melhorado ou piorado significativamente.

5. AGRADECIMENTOS

O presente trabalho foi realizado com apoio: da FAPERGs e CNPq, assim como do projeto Universal 409280/21023-2.

6. REFERÊNCIAS BIBLIOGRÁFICAS

BOUWER, H. Groundwater Hydrology. New York: McGraw-Hill, 1978.

BOUMA, J. Using soil survey data for quantitative land evaluation. In: BEEK, K.J.; STEIBEL, J. (Eds.) Quantitative Land Evaluation Procedures. Wageningen: ILRI, 1989. Cap. 4, p. 177–213.

BASHITI, H. The importance of review articles & its prospects in scholarly literature. **Journal of Scientific Research and Reports**, London, v.27, n.5, p.1-5, 2021.

GOWER, J.C. A general coefficient of similarity and some of its properties. **Biometrics**, Washington, v.27, n.4, p.857-871, 1971.

VAN GENUCHTEN, M.T. A closed-form equation for predicting the hydraulic conductivity of unsaturated soils. **Soil Science Society of America** Journal, Madison, v.44, n.5, p.892-898, 1980.

ZHANG, Y.; SCHAAP, M. G. Estimation of saturated hydraulic conductivity with pedotransfer functions: A review. **Journal of Hydrology**, v. 575, p. 1011-1030, 2019.

ZHANG, Y.; SHANGGUAN, W.; ZHANG, H. Machine learning methods for soil hydraulic property prediction. **Geoderma**, Amsterdam, v.322, n.1, p.18-27, 2018.

SANTOS, G. B. Topographic and spectral predictors in modeling saturated soil hydraulic conductivity using machine learning. **Trabalho em elaboração**, 2025.