

UNIVERSIDADE ESTADUAL DE PONTA GROSSA
SETOR DE ENGENHARIAS, CIÊNCIAS AGRÁRIAS E DE TECNOLOGIA
PROGRAMA DE PÓS-GRADUAÇÃO EM AGRONOMIA

SANTOS HENRIQUE BRANT DIAS

MODELAGEM DE VARIÁVEIS HÍDRICAS POR MEIO DO SENSORIAMENTO
REMOTO ORBITAL E INTELIGÊNCIA ARTIFICIAL

PONTA GROSSA

2025

SANTOS HENRIQUE BRANT DIAS

**MODELAGEM DE VARIÁVEIS HÍDRICAS POR MEIO DO SENSORIAMENTO
REMOTO ORBITAL E INTELIGÊNCIA ARTIFICIAL**

Tese apresentada para obtenção do título de Doutor
em Agronomia na Universidade Estadual de Ponta
Grossa – Área de concentração: Ciência do Solo e
Recursos Ambientais.

Orientador: Prof. Dr. Fabrício Tondello Barbosa.

PONTA GROSSA

2025

D539

Dias, Santos Henrique Brant

Modelagem de variáveis hídricas por meio do sensoriamento remoto orbital e inteligência artificial / Santos Henrique Brant Dias. Ponta Grossa, 2025.
108 f.

Tese (Doutorado em Agronomia - Área de concentração: Ciência do Solo e Recursos Ambientais), Universidade Estadual de Ponta Grossa.

Orientador: Prof. Dr. Fabrício Tondello Barbosa

1. Evapotranspiração - referência. 2. Erosividade. 3. Machine learning. 4. Sensoriamento remoto. 5. Geoprocessamento. I. Barbosa, Fabrício Tondello. II. Universidade Estadual de Ponta Grossa. Ciência do Solo e Recursos Ambientais. III. T.

CDD: 631



CERTIFICADO DE APROVAÇÃO

Título da Tese: **“MODELAGEM DE VARIÁVEIS HÍDRICAS POR MEIO DO SENSORIAMENTO REMOTO ORBITAL E INTELIGÊNCIA ARTIFICIAL”.**

Nome: **SANTOS HENRIQUE BRANT DIAS**

Orientador: Prof. Dr. Fabrício Tondello Barbosa

Aprovado pela Comissão Examinadora:

Prof. Dr. Fabrício Tondello Barbosa (Orientador)

Prof. Dr. Eduardo Augusto Agnellos Barbosa

Prof. Dr. Daniel Ruiz Potma Gonçalves

Prof. Dr. Roberto Filgueiras

Prof. Dr. Rodrigo Vieira Luciano

Ponta Grossa 30 de agosto de 2022.



Documento assinado eletronicamente por **Fabrício Tondello Barbosa, Professor(a)**, em 14/09/2022, às 13:55, conforme Resolução UEPG CA 114/2018 e art. 1º, III, "b", da Lei 11.419/2006.



Documento assinado eletronicamente por **Eduardo Augusto Agnellos Barbosa, Professor(a)**, em 14/09/2022, às 13:57, conforme Resolução UEPG CA 114/2018 e art. 1º, III, "b", da Lei 11.419/2006.



Documento assinado eletronicamente por **Daniel Ruiz Potma Gonçalves, Professor(a)**, em 14/09/2022, às 14:28, conforme Resolução UEPG CA 114/2018 e art. 1º, III, "b", da Lei 11.419/2006.



Documento assinado eletronicamente por **ROBERTO FILGUEIRAS, Usuário Externo**, em 04/10/2022, às 14:47, conforme Resolução UEPG CA 114/2018 e art. 1º, III, "b", da Lei 11.419/2006.



Documento assinado eletronicamente por **Rodrigo Vieira Luciano, Usuário Externo**, em 04/10/2022, às 15:24, conforme Resolução UEPG CA 114/2018 e art. 1º, III, "b", da Lei 11.419/2006.



A autenticidade do documento pode ser conferida no site <https://sei.uepg.br/autenticidade> informando o código verificador **1098575** e o código CRC **0ABD311A**.

A todos os familiares e amigos que tanto acreditaram e me apoiaram, nos momentos mais difíceis, as minhas amadas tias do coração, ao meu grande avô, com quem aprendi a aproveitar ao máximo a vida. Aos meus pais, que tanto me prepararam a chegar aonde estou. Aos meus irmãos que sempre estão de prontidão. Em especial a minha avó Marieta (*in memoriam*), por todo o carinho e amor me dado em vida!

AGRADECIMENTOS

Todo o mérito deve ser atribuído às pessoas com quem convivi e compartilhei experiências pessoais e profissionais ao longo destes últimos anos. Gostaria de agradecer especialmente às orientações que recebi, principalmente pelo auxílio nos momentos de dúvida, tanto pessoais quanto profissionais.

Aos meus pais, Lilian Antunes Brant Dias e Wagner Danilo Santos Dias, meu irmão José dos Reis Dias Neto, minhas irmãs Fernanda Brant Dias e Sabrina Brant Dias, pelo amor incondicional e incentivo, contribuindo assim de forma significativa para o cumprimento desta etapa. A todos os meus familiares e amigos, pelo incentivo durante a realização do curso.

Ao meu amado avô José dos Reis Dias pelas excelentes conversas e estímulos, e in memoriam a Terezinha Auxiliadora Santos Dias, Marieta Antunes Brant e Sebastião Ferreira Brant.

As minhas amadas e queridas tias, Kivany, Daianny, Selma, Jannine e Rosana, que tanto me auxiliaram, amaram e me deram força em todos estes anos. Nos momentos em que mais precisava e pela grande preocupação e amor. Sem seus apoios, jamais conseguiria chegar aonde cheguei. Aos meus tios Alarico Neto, Devay e Milton Laidmer. Por todo o companheirismo, carinho e amor. Também queria agradecer o grande amor incondicional de irmão que tive dos meus primos, em especial Natan Dias Laidmer.

Ao meu orientador, e grande profissional, o Professor Dr. Fabrício Tondello Barbosa, pela amizade, confiança, paciência e a oportunidade de integrar o quadro discente de um excelente programa de pós-graduação, compartilhando comigo algumas de suas inúmeras ideias e grande parte de seu entusiasmo com a questão dos recursos hídricos e da vida. As mesmas palavras e agradecimentos, gostaria de oferecer ao meu orientador de mestrado, Fernando França da Cunha!

Ao meu grande amigo, “irmão” e parceiro Roberto Filgueiras, por todo o auxílio, preocupação, paciência e atenção.

A Prof^a Dra. Neyde Fabíola Balarezo Giarola pela amizade, carinho e por ter sido quase uma mãe nestes últimos 48 meses. Ao Prof. Dr. Eduardo Augusto Agnellos Barbosa, pela amizade, e os trabalhos desenvolvidos nestes últimos anos. Não tenho nem o que falar e como agradecer ao acolhimento, paciência e compreensão disponibilizada por estes professores.

A minha querida e amorosa namorada Mariane Beatriz Colman, por todo o carinho e auxílio.

“Um raciocínio lógico leva você de A a B. A imaginação leva você a qualquer lugar que você quiser.”

(Albert Einstein)

RESUMO

DIAS, Santos Henrique Brant. **Modelagem de variáveis hídricas por meio do sensoriamento remoto orbital e inteligência artificial**. 2025. Tese (Doutorado em Agronomia) – Universidade Estadual de Ponta Grossa, Ponta Grossa, 2025.

O solo e a água são recursos naturais fundamentais para a manutenção da vida na Terra, e a gestão adequada desses elementos é crucial para equilibrar as necessidades humanas com a preservação do meio ambiente. No ciclo hidrológico, a evapotranspiração é um dos principais componentes do balanço hídrico e seu conhecimento é essencial para o gerenciamento e conservação de bacias hidrográficas. Além disso, é utilizada no manejo da irrigação, visando maior eficiência no uso da água na agricultura e aumento da produtividade das culturas. A água, além de sustentar a vida, é a principal responsável pela erosão dos solos no Brasil, uma vez que as chuvas fornecem a energia necessária para os processos erosivos. O solo é um recurso natural não renovável em escala humana e as taxas de perda por erosão têm sido crescentes. A energia da chuva, conhecida como erosividade, é um dos principais fatores responsáveis pelas perdas de sedimentos em todo o mundo. A erosividade é utilizada em modelos de predição da erosão, como a Equação Universal de Perda de Solo (USLE) e a sua versão revisada (RUSLE), mas a obtenção dessas informações não é simples, sendo os valores geralmente pontuais e não extrapoláveis com segurança para outras áreas que necessitam de estimativas de perdas de solo. As técnicas tradicionais de espacialização, como krigagem, ponderação pelo inverso da distância ou polígonos de Thiessen, muitas vezes não representam adequadamente a variabilidade que ocorre nos ambientes. Por outro lado, modelos de machine learning, juntamente com produtos de sistemas de informação geográfica (SIG) e imagens orbitais, têm sido recentemente utilizados como ferramentas eficazes para monitorar e modelar diversas características da superfície terrestre. Neste trabalho, foram desenvolvidas ferramentas para modelagem das variáveis hídricas de evapotranspiração de referência (ET_o) e de erosividade das chuvas para todo o Brasil, utilizando sensoriamento remoto orbital e inteligência artificial. Os objetivos deste estudo foram: i) modelar a ET_o em tempo real para todo o Brasil, permitindo disponibilizar as imagens em até 10 minutos após o fim do dia; e ii) modelar e produzir um mapa de erosividade das chuvas para o Brasil, com resolução espacial de 30 segundos de arco (~ 1 km²). Para a modelagem da ET_o, foram utilizados produtos WorldClim, dados de radiação extraterrestre diária, médias e desvio padrão da evapotranspiração diária obtidos de estações meteorológicas do Instituto Nacional de Meteorologia (INMET), além de algoritmos de machine learning. Foram utilizados os indicadores de desempenho RMSE, nRMSE, MAPE, d e NSE para avaliar e selecionar o melhor modelo. O modelo que apresentou o melhor desempenho para ET_o foi o cubist, com MAPE de 17,4%, RMSE de 0,792, nRMSE de 57,3%, d de 0,883 e NSE de 0,634. Para modelagem da erosividade das chuvas, foram empregados dados provenientes de estudos científicos (erosividades determinadas por pluviógrafos), técnicas de SIG e modelagem com machine learning. Das 31 covariáveis pré-selecionadas, 7 foram utilizadas na modelagem, listadas em ordem de importância: longitude, radiação solar, precipitação anual, precipitação do trimestre mais frio, velocidade do vento, precipitação do trimestre mais quente e evapotranspiração de referência anual. Após 400 ciclos de treinamento e validação, o modelo que apresentou os melhores indicadores de desempenho para erosividade foi o Random Forest, utilizando os valores das medianas. Os índices obtidos foram: NSE de 0,5823, RMSE de 1.567,17 MJ mm ha⁻¹ h⁻¹ ano⁻¹, MAE de 1.135,90 MJ mm ha⁻¹ h⁻¹ ano⁻¹, nRMSE de 58.50%, ME de -17.76 MJ mm ha⁻¹ h⁻¹ ano⁻¹ e d de 0,8487. Conclui-se que o produto ET_o é extremamente útil e viável para estimativas da

evapotranspiração de referência em todo o Brasil. Além disso, foi possível utilizar variáveis climáticas e geográficas para construir mapas de erosividade das chuvas com níveis conhecidos de erro e precisão em uma escala nacional.

Palavras-chave: Evapotranspiração de referência; Erosividade; Aprendizado de máquina; Sensoriamento Remoto; Geoprocessamento.

ABSTRACT

DIAS, Santos Henrique Brant. **Modeling of hydrological variables using satellite remote sensing and artificial intelligence**. 2025. Thesis (Doctorate in Agronomy) – Universidade Estadual de Ponta Grossa, Ponta Grossa, 2025.

Soil and water are essential natural resources for sustaining life on Earth, and managing these elements properly is crucial to balancing human needs with environmental conservation. In the hydrological cycle, evapotranspiration is one of the main components of the water balance, and understanding it is key for watershed management and conservation. It is also used to guide irrigation practices, improving water efficiency in agriculture and increasing crop productivity. Besides supporting life, water is the primary cause of soil erosion in Brazil because rainfall supplies the energy for erosive processes. Soil is a non-renewable resource on a human time scale, and erosion rates have been increasing. Rainfall energy—known as erosivity—is a major factor driving sediment loss worldwide. Erosivity is used in erosion prediction models such as the Universal Soil Loss Equation (USLE) and its revised version (RUSLE), but obtaining these values is challenging, as they are generally measured at specific points and cannot be reliably extrapolated to other areas needing soil loss estimates. Traditional spatialization techniques, such as kriging, inverse distance weighting, or Thiessen polygons, often do not capture the variability of natural environments adequately. On the other hand, machine learning models, combined with geographic information system (GIS) products and satellite imagery, have recently been used as effective tools to monitor and model various features of the Earth's surface. In this study, we developed tools to model reference evapotranspiration (ET_o) and rainfall erosivity for the entire country of Brazil using satellite remote sensing and artificial intelligence. The study's objectives were to: (i) model ET_o in real time across Brazil, making images available within 10 minutes after the end of the day; and (ii) model and produce a rainfall erosivity map for Brazil with a spatial resolution of 30 arc-seconds (approximately 1 km²). For ET_o modeling, we used WorldClim products, daily extraterrestrial radiation data, and daily evapotranspiration means and standard deviations from meteorological stations of the National Institute of Meteorology (INMET), along with machine learning algorithms. Performance metrics (RMSE, nRMSE, MAPE, d, and NSE) were used to evaluate and select the best model. The Cubist model showed the best performance for ET_o, with a MAPE of 17.4%, RMSE of 0.792, nRMSE of 57.3%, d of 0.883, and NSE of 0.634. For rainfall erosivity modeling, we used data from scientific studies (erosivities measured by pluviographs), GIS techniques, and machine learning. Out of 31 pre-selected covariates, seven were used in the model—in order of importance: longitude, solar radiation, annual precipitation, precipitation in the coldest quarter, wind speed, precipitation in the hottest quarter, and annual reference evapotranspiration. After 400 cycles of training and validation, the Random Forest model (using median values) provided the best performance for rainfall erosivity, with an NSE of 0.5823, RMSE of 1,567.17 MJ mm ha⁻¹ h⁻¹ year⁻¹, MAE of 1,135.90 MJ mm ha⁻¹ h⁻¹ year⁻¹, nRMSE of 58.50%, ME of -17.76 MJ mm ha⁻¹ h⁻¹ year⁻¹, and d of 0.8487. We conclude that the ET_o product is highly useful and feasible for estimating reference evapotranspiration across Brazil. In addition, climatic and geographic variables can be effectively used to construct rainfall erosivity maps with known error margins and accuracy on a national scale.

Keywords: Reference evapotranspiration; Erosivity; Machine learning; Remote sensing; Geoprocessing.

LISTA DE FIGURAS

Figura 1 - Localização das estações meteorológicas do Instituto Nacional de Meteorologia (INMET).....	40
Figura 2 - Número de estações utilizadas para modelagem em cada data.	48
Figura 3 - Menores valores de MAE e RMSE e o maior de Rsquared para cada dia dos 365 dias que foram testados.	48
Figura 4 - Número de covariáveis selecionadas para melhor ajuste em cada data.....	49
Figura 5 - Número de vezes que a covariável foi selecionada (escala de 0 a 5) durante a avaliação dos 365 dias. O valor numérico acima do boxplot representa o número de dias que a variável foi selecionada.	50
Figura 6 - Valores observados e preditos de ETo para os diferentes modelos avaliados no estudo	52
Figura 7 - Boxplot dos índices utilizados para seleção do modelo. A - erro médio (mm/dia), B - erro absoluto médio, C - raiz do erro quadrático normalizado (%), D - índice de eficiência de Nash-Sutcliffe (adimensional), E - raiz do erro quadrático (mm/dia), F - índice de concordância (adimensional), G - porcentagem absoluta média do erro, H – porcentagem do erro médio	53
Figura 8 - Mapa de Evapotranspiração de referência no Brasil para o dia 11/07/2022	56
Figura 9 - Mapa de Evapotranspiração de referência no Brasil para o dia 12/07/2022	57
Figura 10 - Localização e altitude da área de interesse. Os pontos demarcados são os locais das erosividades calculadas pelos autores.	65
Figura 11 - Variação do ajuste (MAE - Mean Absolute Error e RMSE - Root Mean Squared Error) de acordo com o número de variáveis utilizadas em cada treino.	72
Figura 12 - Correlação de Pearson das variáveis selecionadas.	74
Figura 13 - Valores observados versus os valores preditos pelos diferentes modelos de <i>machine learning</i>	75
Figura 14 - Demonstração de um boxplot e obtenção dos outliers.	76
Figura 15 – Boxplot dos índices utilizados para seleção do modelo. A - erro médio (MJ mm / ano ha h), B - erro absoluto médio, C - raiz do erro quadrático normalizado (%), D - índice de eficiência de Nash-Sutcliffe (adimensional), E - raiz do erro quadrático (MJ mm / ano ha h), F - índice de concordância (adimensional).....	77
Figura 16 - Erosividade das chuvas (MJ mm year ⁻¹ ha ⁻¹ h ⁻¹) de acordo com os modelos de <i>machine learning</i> utilizando a classificação de Foster et al (1981).	80
Figura 17 - Erosividade das chuvas (MJ mm year ⁻¹ ha ⁻¹ h ⁻¹) para os diferentes modelos de <i>machine learning</i> , utilizando um corte de 3% dos maiores e menores valores.	81
Figura 18 - Desvio padrão da Erosividade das chuvas (MJ mm year ⁻¹ ha ⁻¹ h ⁻¹), na mesma escala do modelo definido como melhor modelo.....	82
Figura 19 - Mapa de erosividade das chuvas (MJ mm year ⁻¹ ha ⁻¹ h ⁻¹) (resolução espacial 30 segundos de arco), definido como padrão (modelo RF).....	83
Figura 20 - Erosividade no Brasil utilizando a krigagem.....	91

LISTA DE TABELAS

Tabela 1 - Parâmetros de estatística descritiva.....	29
Tabela 2 - Algoritmos de <i>machine learning</i> utilizados no estudo.....	42
Tabela 3 – Covariáveis disponíveis no WorldClim.....	43
Tabela 4 - Índices de avaliação de desempenho utilizados no estudo.....	45
Tabela 5 - Ordem de importância das covariáveis.	51
Tabela 6 - Algoritmos de <i>machine learning</i> utilizados no estudo.....	66
Tabela 7 - Covariáveis pré-selecionadas para modelagem da erosividade das chuvas.....	68
Tabela 8 - Índices de avaliação de desempenho.....	71
Tabela 9 - Seleção de covariáveis na modelagem. *descrição das covariáveis.	73
Tabela 10 - Classificação da erosividade anual das chuvas	79
Tabela 11 - Mediana dos índices estatísticos para os diferentes modelos de <i>machine learning</i> ...	89

SUMÁRIO

INTRODUÇÃO GERAL	14
CAPÍTULO 1 – REVISÃO BIBLIOGRÁFICA.....	17
1.1 SENSORIAMENTO REMOTO	17
1.2 INTELIGÊNCIA ARTIFICIAL E APRENDIZADO DE MÁQUINA (MACHINE LEARNING)	18
1.3 INTELIGÊNCIA ARTIFICIAL NO SENSORIAMENTO REMOTO	19
1.3.1 Modelagem de Imagens Orbitais com Inteligência Artificial	19
1.3.2 Sensoriamento Remoto e Inteligência Artificial no Estudo de Variáveis Hídricas	20
1.3.3 Componentes Necessários para Desenvolvimento de Pesquisa Utilizando Sensoriamento Remoto e Inteligência Artificial	21
1.4 BANCOS DE DADOS IN SITU DISPONÍVEIS.....	23
1.4.1 Instituto Nacional de Meteorologia (INMET).....	23
1.4.2 WorldClim.....	23
1.5 MÉTODOS DE PROCESSAMENTO DE DADOS.....	24
1.5.1 RF (Random Forest).....	24
1.5.2 SVM (Support Vector Machine)	24
1.5.3 LM (Linear Regression)	25
1.5.4 Cubist.....	25
1.5.5 PLS (Partial Least Squares).....	26
1.5.6 PCR (Principal Componentes Regression).....	26
1.5.7 GBM (Gradient Boosting Machine).....	26
1.5.8 ICR (Independent Component Regression)	27
1.5.9 BagEarth.....	27
1.5.10 gaussprRadial	28
1.5.11 SvmPoly	28
1.6 MÉTRICAS PARA AVALIAÇÃO DOS MODELOS.....	29
1.6.1 Estatística Descritiva	29
1.6.2 RMSE (Root Mean Squared Error).....	29
1.6.3 nRMSE (Normalized Root Mean Square Error)	30

1.6.4	R ² (Coeficiente de Determinação)	30
1.6.5	NSE (Nash-Sutcliffe Efficiency)	31
1.6.6	MAE (Mean Absolute Error)	31
1.6.7	ME (Mean Error)	32
1.6.8	D (Índice de Concordância de Willmott)	32
1.6.9	MAPE (Porcentagem Absoluta Média do Erro)	32
1.6.10	MPE (Erro de Porcentagem Média)	33
CAPÍTULO 2 – MAPA DIÁRIO DA EVAPOTRANSPIRAÇÃO DE REFERÊNCIA NO BRASIL EM TEMPO REAL		36
2.1	INTRODUÇÃO	37
2.2	MATERIAL E MÉTODOS	39
2.2.1	Área de Estudo e Obtenção dos Dados	39
2.2.2	Cálculo da Evapotranspiração	41
2.2.3	Algoritmos Utilizados para a Aprendizagem de Máquina (Machine Learning)	42
2.2.4	Variáveis e Covariáveis dos Modelos	42
2.2.4.1	WorldClim	42
2.2.4.2	Evapotranspiração de referência (ET _o) diária no Brasil	43
2.2.4.3	Radiação extraterrestre	44
2.2.5	Métricas de Avaliação dos Algoritmos/Modelos	44
2.2.5.1	Índice de desempenho	44
2.2.6	Treino e Validação do Modelo	46
2.3	RESULTADOS	47
2.3.1	Seleção das Covariáveis	49
2.3.2	Seleção do Modelo na Estimativa da ET _o	52
2.3.3	Determinação da ET _o	55
2.3.4	Utilização dos Dados	57
2.4	DISCUSSÃO	59
2.5	CONCLUSÕES	61
CAPÍTULO 3 - DETERMINAÇÃO DA EROSIVIDADE DAS CHUVAS NO BRASIL, UMA ABORDAGEM COM MACHINE LEARNING		62
3.1	INTRODUÇÃO	63

3.2	MATERIAL E MÉTODOS	65
3.2.1	Área de Estudo e Coleta de Dados	65
3.2.2	Algoritmos Utilizados nos Modelos de Aprendizagem de Máquinas	66
3.2.3	Covariáveis Utilizadas na Modelagem	67
3.2.4	Seleção das Covariáveis	69
3.2.5	Treino e Avaliação dos Modelos	69
3.2.6	Métricas de Avaliação dos Algoritmos/Modelos	69
3.2.6.1	Estatística descritiva	69
3.2.6.2	Índices de desempenho	70
3.3	RESULTADOS	72
3.3.1	Seleção das Covariáveis	72
3.3.2	Seleção do Modelo na Estimativa da Erosividade	74
3.3.3	Interpolação da Erosividade	79
3.3.4	Utilização dos Dados	84
3.4	DISCUSSÃO	85
3.4.1	Seleção das Covariáveis	87
3.4.2	Seleção do Modelo na Estimativa da Erosividade	88
3.4.3	Interpolação da Erosividade	89
3.4.4	Comparação do Mapa Desenvolvido com os já Disponíveis	90
3.4.5	Utilização dos Mapas de Erosividade	92
3.5	CONCLUSÕES	94
	REFERÊNCIAS	95

INTRODUÇÃO GERAL

O desenvolvimento da civilização depende da disponibilidade de recursos naturais, sendo o solo e a água os principais responsáveis por sustentar a vida na Terra e manter os ecossistemas que abrigam a flora e a fauna. A gestão sustentável destes recursos tem se tornado um foco de pesquisa e discussão, pois além de ser essencial para o avanço das sociedades, é importante para a preservação e o equilíbrio ambiental.

O manejo sustentável do solo e da água requer um planejamento cuidadoso, já que ambos são finitos em qualidade e quantidade. Para isso, é necessário o monitoramento adequado dos fatores que interferem na ciclagem e perda destes recursos. A taxa de formação dos solos é lenta e não sofre grandes alterações em curto espaço de tempo, no entanto as taxas de perda de solo por erosão são aceleradas pelas atividades antrópicas e são afetadas pelos fatores climáticos, como a erosividade das chuvas.

No balanço hídrico, a precipitação representa a principal entrada de água em um sistema, enquanto as saídas incluem o escoamento superficial e a evapotranspiração. É raro ser possível alterar a precipitação, mas é possível medir as entradas e saídas de água. Existem diversos dados de precipitação disponíveis globalmente e modelos para determinação do escoamento superficial, mas o monitoramento da evapotranspiração em grandes proporções da superfície terrestre ainda é limitado.

Estabelecer uma rede de monitoramento meteorológico em vastos territórios pode ser logisticamente e financeiramente inviável, e as variáveis podem não ser representadas com precisão quando as distâncias entre as estações são grandes. Embora a densidade de estações meteorológicas seja baixa, existem outras fontes de informação disponíveis para pesquisas científicas, como dados históricos e imagens de satélite, entre outras fontes, que tem sido cada vez maiores com o tempo. O volume de estudos para caracterização e entendimento dos processos e componentes da superfície terrestre é imenso, assim como os recursos humanos e capital necessários para realização desses trabalhos.

Atualmente, a obtenção de informações por meios tradicionais, como estações meteorológicas, tem sido dispendiosa em grandes escalas. A obtenção de dados de maneira espacial pode ser custosa, especialmente quando se necessita de uma alta frequência de reanálise, seja em proporções nacionais, estaduais ou mesmo em uma microbacia.

Muitos órgãos e instituições no país já realizam a coleta e análise de dados sistematicamente em diversos pontos e com frequência há algum tempo. Alguns exemplos são o Instituto Nacional de Meteorologia (INMET), que conta com aproximadamente 580 estações meteorológicas em todo o país, coletando dados horários de alta qualidade, e a Agência Nacional de Águas (ANA), que tem inúmeras estações com coleta de dados meteorológicos. Além dos órgãos públicos que disponibilizam informações de clima, água e solo, há também dados já publicados de diversos parâmetros por outras instituições de pesquisa, como os dados de erosividade das chuvas.

Algumas técnicas avançadas para coleta e processamento de dados têm-se tornado populares nos últimos anos, e dois exemplos são o uso de imagens coletadas por sensoriamento remoto orbital e as técnicas de aprendizagem de máquina (*machine learning*). Isso tem possibilitado a realização de estudos de classificação e estimativa de parâmetros de forma mais rápida e precisa.

A técnica de sensoriamento remoto orbital utiliza sensores passivos para coletar informação de refletância da superfície terrestre, a partir dos comprimentos de onda eletromagnéticos que incidem diretamente do Sol, obtendo basicamente as informações e características intrínsecas da superfície do local. A superfície terrestre tem sido imageada desde 1984 com auxílio do satélite Landsat 5. Atualmente, o número de sensores orbitando na Terra é muito grande, gerando um grande volume de informações diariamente.

As técnicas de inteligência artificial começaram a surgir na década de 1960 com simples algoritmos de busca, e atualmente têm se popularizado e utilizado para interpretação, classificação e estimativa de parâmetros que anteriormente eram inviáveis de serem feitos por métodos tradicionais.

A evapotranspiração de referência é um parâmetro importante no balanço hídrico do solo e na estimativa das perdas de água pela transpiração das plantas e evaporação do solo. É uma variável difícil de ser obtida pelo método padrão, que é o de Penman-Monteith FAO 56 (ALLEN et al., 1998), já que demanda a utilização de equipamentos de alto custo. No entanto, já pode ser determinada diariamente em cerca de 580 pontos em todo o Brasil.

Já a erosividade representa o potencial energético das chuvas de gerar perdas de solo por erosão hídrica pluvial. A erosividade é um dos fatores (fator R) da Equação Universal de Perda de Solo (USLE) e da sua versão revisada (RUSLE). Ambas as equações são mundialmente utilizadas

e estimam a erosão levando em consideração os diferentes componentes do processo erosivo, como a precipitação, características do solo, topografia e fatores afetados pelas atividades humanas, como uso/ocupação da Terra, cobertura e manejo do solo.

A evapotranspiração de um sistema e a erosividade das chuvas podem ser modeladas e estimadas em grandes escalas com os dados já existentes, utilizando técnicas combinadas de sensoriamento remoto e aprendizagem de máquina. Com isso, o objetivo geral deste trabalho é avaliar o potencial de modelagem da evapotranspiração de referência e da erosividade das chuvas no Brasil, disponibilizando produtos essenciais para o manejo e conservação do solo e da água.

CAPÍTULO 1 – REVISÃO BIBLIOGRÁFICA

1.1 SENSORIAMENTO REMOTO

O sensoriamento remoto é um método para obtenção de informações de objetos ou áreas sem contato direto, como, por exemplo, capturando informações da superfície da Terra e gerando imagens por meio de satélites ou aeronaves tripuladas e não tripuladas (KAHAKI et al., 2018; WEI et al., 2017). O uso de dados hiperespectrais tem aumentado significativamente devido ao surgimento de novos satélites e sensores avançados (KAVZOGLU et al., 2018). Atualmente, existe grande quantidade de satélites em órbita, tanto estatais quanto privados, e o número de plataformas e bancos de dados de imagens gratuitas disponíveis também tem crescido consideravelmente.

O sensoriamento remoto orbital teve início em 1968, quando a missão Apollo 9 produziu a primeira fotografia multi-espectral do espaço (GIARDINO, 2011). Em 23 de julho de 1972, a NASA lançou o programa ERTS, que consistia em uma série de satélites de tecnologia de recursos terrestres, que foram posteriormente renomeados para Landsat. Sete satélites Landsat foram lançados desde 1972, com seis deles alcançando a órbita e fornecendo imagens de todo o mundo (URI, 2017).

A constelação Landsat é uma ferramenta importante para estudos em sensoriamento remoto, desde o início de sua criação. Anteriormente, as imagens eram processadas pela NASA e vendidas em cópias impressas aos usuários. No entanto, desde 2008, o público tem acesso livre às imagens quase em tempo real por meio do US Geological Survey (USGS), sem custo e pela internet (URI, 2017).

Com o aumento do número de satélites, houve um aumento no volume de informações coletadas, gerando uma grande quantidade de dados. Esse volume é frequentemente referido como "big data". Em 2012, cerca de 2,5 quintilhões (10¹⁸) de bytes de dados foram coletados diariamente em todo o mundo (IBM, 2012; TOTH; JÓZKÓW, 2016).

1.2 INTELIGÊNCIA ARTIFICIAL E APRENDIZADO DE MÁQUINA (MACHINE LEARNING)

Com o grande volume de dados disponíveis, torna-se necessário utilizar técnicas mais avançadas para interpretá-los. Os algoritmos de inteligência artificial são altamente recomendados para essa finalidade, pois a interpretação direta é inviabilizada pelas interações e particularidades dos dados, como as interações espaciais. A técnica identifica e detecta as tendências nas informações.

O desenvolvimento de algoritmos de inteligência artificial teve início em 1943 (MCCULLOCH; PITTS, 1943), baseado no funcionamento de um neurônio (RUSSELL; NORVIG, 2016). Atualmente, esses modelos evoluíram para algoritmos mais complexos e são cada vez mais utilizados devido ao grande volume de informações geradas diariamente. O termo "big data" é excessivamente discutido tanto no cenário acadêmico quanto no comercial (GROVER et al., 2018).

A modelagem é uma técnica muito utilizada para associação e geração de informações (KIM; CHEN; GARCIA, 2016; LAMPERTI; ROVENTINI; SANI, 2018). Antes era feita exclusivamente com equações, muitas delas extensas, mas atualmente utilizam-se algoritmos complexos de inteligência artificial ou aprendizado de máquina. Essas técnicas se tornaram muito acessíveis ao público, principalmente após o surgimento do WEKA, que é uma coleção de algoritmos de aprendizado de máquina para tarefas de mineração de dados (HORNIK, 2018; HORNIK; BUCHTA; ZEILEIS, 2009; NAJI; ALI, 2018; WEKA, 2018), e de pacotes do software R, como o Caret (KUHN, 2008), que conta com 237 algoritmos de regressão e classificação.

O aprendizado de máquina (machine learning) é um dos ramos da inteligência artificial e é definido como a aplicação de técnicas estatísticas e algoritmos para identificar padrões em dados e fazer previsões a partir desses padrões. É comumente aplicado na classificação de sensoriamento remoto (COLLINS et al., 2018; GRAVES et al., 2016; HEYDARI; MOUNTRAKIS, 2018; ROGAN et al., 2008).

As técnicas de aprendizado de máquina têm melhor desempenho que os classificadores simples ao lidar com interações complexas entre cena, escala e agregação, melhorando a discriminação de classes em paisagens heterogêneas (COLLINS et al., 2018). Isso torna a técnica

uma ferramenta chave no desenvolvimento de pesquisa e construção do conhecimento para dados de sensoriamento remoto orbital.

1.3 INTELIGÊNCIA ARTIFICIAL NO SENSORIAMENTO REMOTO

A associação entre imagens de satélite e métodos de inteligência artificial não é tão recente. Estes, Sailer e Tinney (1986) propuseram técnicas para automatizar o processo de análise de imagens, com a inclusão de técnicas de inteligência artificial no desenvolvimento de sistemas de processamento de imagens.

Além do surgimento e popularização dos algoritmos de machine learning, o número de satélites que capturam imagens da superfície terrestre e oferecem dados abertos aumentou consideravelmente. Antigamente, a tecnologia para construção e lançamento de um satélite em órbita terrestre era restrita a empresas estatais, mas atualmente muitas empresas privadas possuem tecnologias avançadas que superam as das entidades estatais.

As possibilidades para o desenvolvimento de produtos e serviços com base em informações orbitais e algoritmos de machine learning são vastas e promissoras. A combinação entre sensoriamento remoto e inteligência artificial tem gerado resultados surpreendentes (ATASEVER; KESIKOGLU; OZKAN, 2016; DI SANTO et al., 2018; FARIDI; VERMA; MUKHERJEE, 2018; GONÇALVES et al., 2016; LARY et al., 2016; TIEN BUI et al., 2016) e é explorada em diversas áreas do conhecimento, com destaque na área de recursos hídricos, que abrange grandes extensões de território.

1.3.1 Modelagem de Imagens Orbitais com Inteligência Artificial

Nos recentes trabalhos publicados na comunidade científica, o uso de imagens orbitais e algoritmos de inteligência artificial têm sido amplamente explorado. Esses trabalhos tem fornecido informações importantes, como a localização da área de estudo, as imagens de satélite utilizadas, o banco de dados empregados para a obtenção dessas imagens, os dados de campo utilizados, os métodos de processamento e/ou algoritmos utilizados, e as métricas de avaliação dos algoritmos.

Essas informações são cruciais para o desenvolvimento de novos estudos e para a melhoria das técnicas de análise de imagens de satélite com o uso de inteligência artificial. Além disso, a

combinação dessas tecnologias tem mostrado resultados promissores em diversas áreas do conhecimento, como no monitoramento ambiental, agrícola e urbano.

Com a evolução tecnológica, tem sido possível o acesso a imagens de satélite com maior resolução e frequência, o que amplia as possibilidades de aplicação da inteligência artificial em diferentes áreas. Dessa forma, é importante a divulgação de informações detalhadas sobre as metodologias e técnicas utilizadas para que possam ser replicadas e aprimoradas por outros pesquisadores.

1.3.2 Sensoriamento Remoto e Inteligência Artificial no Estudo de Variáveis Hídricas

Atualmente, as pesquisas em sensoriamento remoto têm se concentrado na análise do balanço hídrico. Campos et al. (2016) obtiveram alta precisão na determinação da evapotranspiração para estimativa da umidade do solo, com erro médio quadrático (RMSE) de 0,48 mm/dia, utilizando imagens orbitais. Isso demonstra o grande potencial dos métodos baseados no uso de imagens de satélite para a estimativa da evapotranspiração.

Outra pesquisa relevante é o estudo de Bastiaanssen e Steduto (2017), que avaliaram a eficiência do uso da água por meio de imagens de satélite, determinando valores de eficiência pixel a pixel e propondo estimativas regionais para diferentes culturas. Essa abordagem oferece uma ferramenta útil na seleção de cultivares mais eficientes e adaptadas a cada região de produção agrícola.

O sensoriamento remoto é uma ferramenta poderosa para estudos que seriam inviáveis de serem realizados de outra forma, como a avaliação de mudanças hidrológicas causadas por fenômenos naturais que abrangem grandes áreas. Rodriguez et al. (2016) propuseram técnicas para avaliar as mudanças na hidrologia regional após terremotos. Esses estudos, entre outros, demonstram a viabilidade técnico-econômica do sensoriamento remoto em pesquisas, embora ainda seja necessário explorar e aprofundar as metodologias para maximizar o potencial dessa ferramenta.

Os estudos para determinação da evapotranspiração de referência (ET_o) vêm evoluindo com o uso de inteligência artificial combinada a dados de lisímetros, tendo como objetivo superar o método padrão da FAO 56. Embora esses métodos tenham produzido resultados satisfatórios para

locais pontuais, ainda não foram desenvolvidos modelos que abranjam todo um país ou que forneçam valores com frequência diária de ETo para os produtores rurais.

Landeras et al. (2017) exploraram o uso de inteligência artificial para criar modelos alternativos de estimativa de ETo em situações de baixa disponibilidade de dados meteorológicos, constatando que os métodos de machine learning apresentaram melhores resultados do que as equações tradicionais para dados faltantes. Abdullah e Malek (2016) também propuseram o uso de inteligência artificial para determinar a ETo, uma vez que a equação padrão de Penman-Monteith requer múltiplos elementos meteorológicos para sua aplicação. Embora os resultados sejam mais precisos, a disponibilidade de dados climáticos históricos amplos correlacionados com as propriedades geográficas de localização, demanda esforço, tempo e cuidado para cada projeto ou estudo de caso.

Khoshravesh, Sefidkouhi e Valipour (2017) utilizaram modelos não caixa preta, como o polinômio fracionário multivariante (MFP), a regressão robusta e a regressão bayesiana, para determinar a ETo. A equação de MFP apresentou os melhores resultados, atingindo R^2 de 0,98 e RMSE de 7,58 mm/mês, utilizando apenas dados de temperatura e radiação local.

Estudos indicam que o desempenho dos modelos de inteligência artificial/machine learning superam os métodos tradicionais na estimativa da ETo utilizando a equação de Penman-Monteith ou a medidas de campo reais (ABDULLAH; MALEK, 2016; CHOUBIN et al., 2018; KARANDISH; ŠIMŮNEK, 2016; KHOSHRAVESH; SEFIDKOUHI; VALIPOUR, 2017; LANDERAS et al., 2017; NAGHIBI; POURGHASEMI; DIXON, 2016; SARHADI et al., 2016; SHIRI et al., 2016; TIWARI; ADAMOWSKI, 2017). No entanto, para validar a precisão da estimativa de ETo, são necessários dados precisos de lisímetros. Embora o uso de inteligência artificial seja uma ferramenta promissora para aprimorar a determinação de ETo, ainda é necessário avançar nas metodologias para explorar todo o seu potencial.

1.3.3 Componentes Necessários para Desenvolvimento de Pesquisa Utilizando Sensoriamento Remoto e Inteligência Artificial

O desenvolvimento de pesquisas que integram sensoriamento remoto e inteligência artificial (IA) requerem uma abordagem multidisciplinar, que abrange desde a seleção e caracterização dos dados de sensores até a aplicação de algoritmos de aprendizado de máquina e

técnicas de processamento de imagens (SHARAD, 2021; YANG et al., 2024). A caracterização adequada dos dados, incluindo suas especificações técnicas e limitações, é crucial para garantir que os dados adquiridos sejam adequados para as análises subsequentes (MAHJOOBI; RAFIEI, 2024; SUN et al., 2022a).

Uma vez que os dados são adquiridos, o pré-processamento é uma etapa essencial para garantir a qualidade e a utilidade dos dados para análises posteriores. Para imagens de satélite sem calibração, inclui a calibração radiométrica, correção atmosférica e fusão de imagens, que são necessárias para transformar os dados brutos em informações utilizáveis (LI et al., 2024). A implementação de etapas de pré-processamento requer conhecimento técnico especializado, especialmente no caso de dados complexos, como os obtidos por SAR (MULLISSA et al., 2021).

A aplicação de algoritmos de aprendizado de máquina é um componente vital na análise de dados de sensoriamento remoto. Esses algoritmos permitem a extração de padrões e a classificação de dados, facilitando a identificação de características específicas em grandes volumes de dados (SHAIK; SHAIK; PRIYA, 2024). A utilização de técnicas de aprendizado profundo tem se mostrado particularmente eficaz na classificação de imagens e na detecção de mudanças ao longo do tempo, permitindo uma análise mais precisa e detalhada (ZHANG; HU; YU, 2021; ZHANG; ZHOU; LUO, 2022). Além disso, a integração de dados de múltiplas fontes, como dados coletados localmente e dados orbitais, afim de garantir que possa ter informações de qualidade alinhada com os dados locais e com o custo de obtenção das imagens orbitais, gratuitamente. O que melhora significativamente a acurácia das previsões e análises (GUO; HUANG; HOU, 2022; YANG et al., 2024).

A fusão de dados de diferentes sensores e plataformas pode proporcionar uma visão mais holística e precisa das condições ambientais (SHAIK; SHAIK; PRIYA, 2024; WANG; LI; LIN, 2022). A utilização de técnicas de IA para modelagem preditiva, por exemplo, pode ajudar a prever eventos climáticos extremos e a otimizar a gestão de recursos em tempo real (SHAIK; SHAIK; PRIYA, 2024; XU et al., 2024). Essa abordagem integrada não apenas melhora a eficiência das análises, mas também contribui para a tomada de decisões informadas em contextos de gestão ambiental e planejamento urbano (SABRI et al., 2021; YANG et al., 2024).

O desenvolvimento de pesquisas que integram sensoriamento remoto e inteligência artificial exige uma compreensão aprofundada dos componentes envolvidos na coleta e padronização dos dados. Desde a seleção de sensores e métodos de determinação dos dados, até os

melhores algoritmos de aprendizado de máquina. A interdependência entre essas etapas destaca a necessidade de uma abordagem colaborativa e multidisciplinar, para maximizar a eficácia das análises e a relevância dos resultados obtidos.

1.4 BANCOS DE DADOS IN SITU DISPONÍVEIS

1.4.1 Instituto Nacional de Meteorologia (INMET)

O Instituto Nacional de Meteorologia (INMET) é um órgão federal da administração direta do Ministério da Agricultura, Pecuária e Abastecimento (MAPA) do Brasil. Fundado em 18 de novembro de 1909, o instituto tem como missão fornecer informações meteorológicas por meio de monitoramento, análise e previsão do tempo e clima, visando contribuir para o desenvolvimento sustentável do país e minimizar os efeitos de desastres naturais, como inundações e secas extremas.

Atualmente, o INMET possui aproximadamente 580 estações automáticas distribuídas por todo o Brasil. Essas estações coletam dados meteorológicos a cada hora e disponibilizam informações históricas de um ano para download. Entre os dados coletados pelas estações estão a temperatura (instantânea, máxima e mínima), umidade (instantânea, máxima e mínima), ponto de orvalho (instantâneo, máximo e mínimo), pressão (instantânea, máxima e mínima), vento (velocidade, direção e rajada), radiação (kJ/m^2) e chuva (mm). Esses dados são utilizados para gerar previsões precisas e fornecer dados valiosos para diversas áreas, como agricultura, aviação, navegação, energia eólica, entre outras.

1.4.2 WorldClim

As variáveis do WorldClim foram desenvolvidas a partir do conjunto de estações meteorológicas da Organização Mundial de Meteorologia, como destacado por Fick e Hijmans (2017). O WorldClim disponibiliza diversas variáveis climáticas, que podem ser observadas na Tabela 3. Para a análise, foram utilizados dados com resoluções espaciais de 30 segundos, o que corresponde a cerca de 1 km^2 .

Para as variáveis meteorológicas, foi selecionado um valor para cada mês do ano, enquanto para as variáveis bioclimáticas, foi utilizada uma variável normal que caracteriza a região.

Mais informações sobre as variáveis disponíveis podem ser encontradas no site do WorldClim (<https://www.worldclim.org/data/worldclim21.html>).

1.5 MÉTODOS DE PROCESSAMENTO DE DADOS

1.5.1 RF (Random Forest)

O Random Forest (RF) é um modelo de machine learning utilizado tanto para classificação quanto para regressão. É capaz de lidar com redução de dimensionalidade, dados perdidos e valores discrepantes. Além disso, é resistente à multicolinearidade e soluciona interações complexas entre variáveis preditoras e explicadas (BROKAMP et al., 2017). De acordo com Collins et al. (2018), o RF, desenvolvido por Breiman (2001), é a técnica mais popular de aprendizado de máquina, frequentemente aplicada em sensoriamento remoto (BELGIU; DRĂGUȚ, 2016).

O RF é vantajoso em relação a outras técnicas de aprendizado de máquina devido aos seus excelentes resultados de classificação, velocidade de processamento (BELGIU; DRĂGUȚ, 2016; DU et al., 2015), capacidade de lidar com ruído e outliers em espaços de medição complexos (MELLOR et al., 2015; RODRIGUEZ-GALIANO et al., 2012) e habilidade de caracterizar interações complexas entre variáveis (CUTLER et al., 2007). Shendryk, Broich e Tulbure (2018) consideram o RF um dos mais populares e poderosos algoritmos de aprendizado de máquina.

1.5.2 SVM (Support Vector Machine)

O Support Vector Machine (SVM) é uma tecnologia avançada de reconhecimento de padrões baseada na teoria de Vapnik-Chervonenks (VC) (CARO, 2021). A dimensão VC é uma medida da capacidade dos algoritmos de classificação estatística, definida como a maior quantidade de pontos que o algoritmo pode separar com precisão. O algoritmo padrão do SVM converte um problema prático em um problema de programação quadrática convexa com restrições de desigualdade (GARAY TRINDADE et al., 2021).

O SVM é frequentemente utilizado em problemas de classificação, como identificação de imagens e reconhecimento de voz, e é particularmente eficaz quando os dados apresentam alta dimensionalidade (HUYNH; NGUYEN, 2023; LIU; LEU; HOLST, 2023). O SVM é capaz de

encontrar o hiperplano ótimo, que é a superfície de separação que maximiza a margem entre as classes (GONDIM et al., 2021). Esse método oferece uma alta capacidade de generalização e pode ser usado com dados não lineares por meio do uso de funções de kernel, que permitem que o SVM trabalhe em espaços de recursos de dimensão superior (BENAISSA et al., 2023).

1.5.3 LM (Linear Regression)

A regressão linear (LM) tem como objetivo encontrar um plano que minimize a soma dos quadrados dos erros (SSE) entre os dados observados e preditos. De acordo com Kuhn e Johnson (2013), a vantagem deste método é a fácil interpretação dos coeficientes gerados no modelo, além do custo operacional ser relativamente mais baixo em relação a outros métodos.

No entanto, a regressão linear é adequada apenas para dados que possuem uma relação linear entre as variáveis, e não é capaz de lidar com interações entre variáveis e na modelagem de comportamentos não lineares. Por isso, em alguns casos, é necessário utilizar outras técnicas de análise de dados.

1.5.4 Cubist

Um modelo amplamente utilizado no sensoriamento remoto é o cubist, que implementa um algoritmo de árvores de regressão acoplado com técnicas baseadas em instâncias e modelos para criar modelos de regressão multivariada a partir de dados de treinamento. O cubist é capaz de fazer a média de múltiplos modelos de regressão (KE et al., 2016).

O Cubist tem sido amplamente utilizado em diversas aplicações de sensoriamento remoto, como previsão de colheitas agrícolas, mapeamento de uso da terra e modelagem de mudanças climáticas (KE et al., 2016). Sua capacidade de modelar relações complexas entre variáveis faz dele um modelo de regressão eficiente em várias áreas de aplicação. Além disso, o Cubist oferece uma interface intuitiva e de fácil interpretação, o que o torna uma opção popular para pesquisadores e profissionais.

1.5.5 PLS (Partial Least Squares)

O modelo partial least squares (PLS) foi desenvolvido a partir dos dados do algoritmo Herman Wold's nonlinear interactive partial least squares (NIPALS), que busca linearizar modelos com parâmetros não lineares. Wold et al. (1983) adaptaram a metodologia de regressão do algoritmo NIPALS, para que ele pudesse realizar regressões com preditores correlacionados, chamando essa alteração de PLS (JAMES et al., 2013; KUHN; JOHNSON, 2013).

1.5.6 PCR (Principal Componentes Regression)

A Regressão por Componentes Principais (PCR) é uma técnica derivada da regressão linear múltipla (LM) que visa resolver o problema de alta correlação entre os preditores. Para isso, é realizada uma análise de componentes principais (PCA) nos preditores, que são pré-processados por esse método antes de serem utilizados na regressão. Essa técnica é especialmente recomendada quando há variáveis altamente correlacionadas entre si ou quando há mais variáveis do que observações. É importante ressaltar que o PCR mantém a interpretabilidade dos coeficientes gerados pelo modelo original de regressão linear, o que pode ser vantajoso em alguns casos (JAMES et al., 2013; KUHN; JOHNSON, 2013).

1.5.7 GBM (Gradient Boosting Machine)

O Gradient Boosting Machine (GBM) é uma técnica robusta de predição, que surgiu da ideia de modificar um preditor fraco para se tornar mais eficiente. O AdaBoost (Adaptive Boosting) foi a primeira realização dessa técnica, que trouxe avanços significativos no reconhecimento de padrões. O AdaBoost usa árvores de decisão como um dos seus preditores, que é considerado um preditor pouco robusto. Dessa forma, o AdaBoost e algoritmos relacionados foram reestruturados e passaram a ser conhecidos como Gradient Boosting Machine (FRIEDMAN, 2002).

O GBM pode ser utilizado para solucionar problemas de classificação e regressão. É um método iterativo que usa árvores de decisão como preditores fracos e realiza uma combinação linear deles para obter um modelo mais robusto. O algoritmo trabalha para minimizar o erro de predição adicionando sequencialmente novas árvores ao modelo.

Uma das principais vantagens do GBM é sua capacidade de lidar com dados faltantes, outliers e conjunto de dados desbalanceados, além de ser capaz de processar variáveis categóricas. No entanto, o GBM é mais sensível ao overfitting do que outros algoritmos, o que torna essencial realizar ajustes nos parâmetros para obter um modelo adequado.

1.5.8 ICR (Independent Component Regression)

O algoritmo Independent Component Regression (ICR) faz a decomposição dos dados de treinamento, realizando combinações lineares com as variáveis dependentes, de forma independente o máximo possível uns dos outros (COMON, 1994; TEMIZ et al., 2022).

Foi apresentado por Shao et al. (2006), e embora não traga uma novidade extremamente relevante, apresenta uma diferença em relação à técnica PCR (Principal Component Regression). De acordo com os autores, o ICR é muito semelhante ao PCR, mas utiliza os componentes independentes e sua matriz de coeficientes obtidos pela análise de componentes independentes na operação de regressão, ao invés dos componentes principais e matriz de pontuação obtidos pela análise de componentes principais (LI; WANG; ZHOU, 2015).

1.5.9 BagEarth

O algoritmo BagEarth é um algoritmo de aprendizado de máquina que utiliza bootstrap como técnica de reamostragem. De acordo com Kuhn e Johnson (2013) e Kulekçi et al. (2022), o modelo de encapsulamento fornece estimativas de precisão que se correlacionam bem com estimativas obtidas por meio de validação cruzada ou estimativas de conjuntos de teste.

A técnica combina a adaptativa multivariada com pontos de controle de regressão de curvas por meio da função earth (MILBORROW, 2019). O MARS é uma técnica de análise multivariável que utiliza uma regressão estatística geral (particionamento recursivo) para construir uma árvore de decisão que classifica corretamente membros da população (FRIEDMAN, 1991).

Para cada partição, é ajustada uma regressão linear múltipla por partes (MLR). A fim de evitar o overfitting, a quantidade de partição é controlada por validação cruzada, onde os pontos de dica são interrompidos quando o erro quadrático médio não diminui mais (ZHAO et al., 2022).

O bagEarth é uma abordagem útil para lidar com dados com múltiplas variáveis, permitindo ajustar o modelo de forma adaptativa às características específicas do conjunto de dados.

1.5.10 gaussprRadial

O modelo gaussprRadial é um algoritmo de aprendizado de máquina que utiliza um kernel de função de base radial para resolver problemas de classificação. Este algoritmo é projetado para transformar as características de entrada em um espaço de características de maior dimensão, o que permite a aplicação de métodos lineares em problemas que, à primeira vista, são não lineares (SUN et al., 2022b).

Essa capacidade de transformação é fundamental para a eficácia do gaussprRadial, pois possibilita que técnicas de classificação que normalmente se aplicam a dados lineares sejam utilizadas em contextos mais complexos (SUN et al., 2022b). No entanto, é importante notar que o gaussprRadial apresenta limitações, especialmente em relação a problemas de classificação que não podem ser adequadamente modelados como lineares, conforme discutido por diversos autores (SUN et al., 2022b).

1.5.11 SvmPoly

O modelo de aprendizado supervisionado Support Vector Machine com função polinomial (svmPoly) foi proposto por Vapnik (1999). O SVM busca encontrar um hiperplano em um espaço de alta dimensão para separar os dados utilizando o princípio de minimização de risco estrutural (SRM) com base na dimensão de Vapnik-Chervonenkis (VC). Essa abordagem visa minimizar o erro de generalização, considerando a complexidade do modelo e a quantidade de dados disponíveis para o treinamento (WU et al., 2018).

1.6 MÉTRICAS PARA AVALIAÇÃO DOS MODELOS

1.6.1 Estatística Descritiva

Na Tabela 1 são apresentados os principais parâmetros da estatística descritiva, que serão explorados neste trabalho para avaliação dos modelos.

Tabela 1 - Parâmetros de estatística descritiva

Parâmetro	Descrição
min	O valor mínimo de todos os pixels dentro de um objeto
max	O valor máximo de todos os pixels dentro de um objeto
avg	O valor médio de todos os pixels dentro de um objeto
std	O desvio padrão de todos os pixels dentro de um objeto
p25	O 25º percentil de todos os valores de pixel em um objeto
p50	O 50º percentil de todos os valores de pixel em um objeto (mediana)
p75	O 75º percentil de todos os valores de pixel em um objeto

Fonte: Adaptado de Fick; Hijmans (2017)

1.6.2 RMSE (Root Mean Squared Error)

O erro médio quadrático (RMSE) é uma medida comumente utilizada para avaliar a qualidade de modelos de regressão, pois fornece uma medida da magnitude média do erro através da diferença quadrática entre os dados estimados e observados. O RMSE varia de 0 a $+\infty$ e quantifica o erro com as mesmas unidades da variável original. Dessa forma, é possível comparar a magnitude dos erros do modelo com a magnitude das observações originais. O RMSE dá um maior peso aos grandes erros, tornando-o útil para evidenciar a presença de outliers, que são grandes erros que ocorrem de forma pontual. Essa medida é particularmente importante em aplicações em que grandes erros são indesejáveis, como em problemas de previsão de demanda ou de estoque (WILKS, 2011; WILLMOTT; MATSUURA, 2005).

Na fórmula para o cálculo do RMSE, onde O_i - valor observado; P_i - valor predito pelo modelo; n - número de dados.

$$RMSE = \sqrt{\frac{\sum_{i=1}^n (O_i - P_i)^2}{n}} \quad (1)$$

1.6.3 nRMSE (Normalized Root Mean Square Error)

A normalização do RMSE é uma técnica que facilita a comparação entre a eficiência de conjuntos de dados ou modelos com diferentes escalas. O RMSE normalizado é expresso em porcentagem e representa a porcentagem do erro em relação à variação da variável resposta. Dessa forma, valores menores indicam menor variação residual e, conseqüentemente, maior eficiência do modelo ou conjunto de dados avaliado.

Na fórmula para o cálculo do nRMSE, onde O_i - valor observado; P_i - valor predito pelo modelo; n - número de dados; $\bar{O}$ - valor médio observado.

$$\text{nRMSE} = 100 \times \frac{\sqrt{\frac{\sum_{i=1}^n (O_i - P_i)^2}{n}}}{\bar{O}} \quad (2)$$

1.6.4 R2 (Coeficiente de Determinação)

O coeficiente de determinação (r^2) é uma métrica que avalia a qualidade do ajuste de um modelo, indicando o quanto ele é capaz de explicar os dados observados. Esse índice é calculado a partir da proporção da variância total dos dados que é explicada pelo modelo. No entanto, é importante destacar que o r^2 não considera a falta de ajuste, que pode ser significativa, especialmente quando os valores observados e preditos não apresentam uma relação linear. Por essa razão, é importante utilizar outras métricas em conjunto com o r^2 para avaliar a qualidade do modelo, permitindo assim uma avaliação mais completa da capacidade do modelo em explicar os dados e identificar possíveis problemas.

Na fórmula para o cálculo do r^2 , onde O_i - valor observado; P_i - valor predito pelo modelo; n - número de dados; $\bar{O}$ - valor médio observado; $\bar{P}$ - valor médio predito pelo modelo.

$$r^2 = \frac{\left(\sum_{i=1}^n (P_i - \bar{P})(O_i - \bar{O}) \right)^2}{\left(\sum_{i=1}^n (P_i - \bar{P})^2 \right) \left(\sum_{i=1}^n (O_i - \bar{O})^2 \right)} \quad (3)$$

1.6.5 NSE (Nash-Sutcliffe Efficiency)

A eficiência de Nash-Sutcliffe (NSE) é uma medida de desempenho amplamente utilizada na hidrologia para avaliar a habilidade dos modelos hidrológicos em simular dados de vazão dos rios, sendo uma ferramenta importante para a gestão de recursos hídricos (MORIASI et al., 2007; WILLMOTT, 1982).

O NSE é calculado pela razão entre a soma dos quadrados das diferenças entre os dados observados e os dados simulados pelo modelo e a soma dos quadrados das diferenças entre os dados observados e a média dos dados observados. O índice varia de $-\infty$ a 1, sendo que o valor de 1 corresponde ao ajuste perfeito entre os dados estimados pelo modelo e os dados medidos. Portanto, quanto mais próximo de 1 o valor de NSE, melhor é a capacidade do modelo em simular os dados observados.

Na fórmula para o cálculo do NSE, onde O_i - valor observado; P_i - valor predito pelo modelo; n - número de dados; $\bar{O}$ - valor médio observado.

$$NSE = 1 - \frac{\sum_{i=1}^n (O_i - P_i)^2}{\sum_{i=1}^n (O_i - \bar{O})^2} \quad (4)$$

1.6.6 MAE (Mean Absolute Error)

O erro médio absoluto (MAE) é uma medida estatística utilizada para avaliar a precisão de um modelo. O MAE é calculado como a média das diferenças absolutas entre os valores preditos e observados. O resultado varia de 0 a $+\infty$ e é expresso nas mesmas unidades da variável original, o que facilita a interpretação dos resultados. O MAE é uma medida mais robusta do que o RMSE quando há presença de outliers, já que ele não eleva ao quadrado as diferenças, dando peso igual a todos os erros (CHAI; DRAXLER, 2014).

Na fórmula para o cálculo do MAE, onde O_i - valor observado; P_i - valor predito pelo modelo; n - número de dados.

$$MAE = \frac{1}{n} \sum_{i=1}^n |P_i - O_i| \quad (5)$$

1.6.7 ME (Mean Error)

O erro médio (ME) é uma medida de desempenho utilizada para verificar se os valores estimados pelo modelo são superestimados ou subestimados. Já o índice de concordância de Willmott (d) é uma métrica que varia de 0 a 1, onde 0 indica falta de concordância e 1 indica concordância perfeita entre os dados observados e estimados pelo modelo. O d é uma métrica amplamente utilizada em estudos hidrológicos, ambientais e climáticos para avaliar a precisão das previsões de modelos de simulação (CUNHA et al., 2017).

Na fórmula para o cálculo do ME, onde O_i - valor observado; P_i - valor predito pelo modelo; n - número de dados.

$$ME = \frac{1}{n} \sum_{i=1}^n (P_i - O_i) \quad (6)$$

1.6.8 D (Índice de Concordância de Willmott)

O Índice de Concordância de Willmott, é uma métrica estatística amplamente utilizada para avaliar a precisão de previsões em comparação com dados observados. Este índice é especialmente relevante em contextos onde a precisão das previsões é crítica, como em modelagens climáticas e avaliações de desempenho de modelos preditivos (TAO; XUE, 2024). Nos últimos anos, o Índice de Concordância de Willmott tem sido cada vez mais integrado em pesquisas que buscam validar modelos, o índice se destaca em estudos agrícolas (SILVA et al., 2024).

Na fórmula para o cálculo do D, onde O_i - valor observado; P_i - valor predito pelo modelo; n - número de dados; $\bar{O}$ - valor médio observado.

$$D = 1 - \frac{\sum_{i=1}^N (O_i - P_i)^2}{\sum_{i=1}^N (|P_i - \bar{O}| + |O_i - \bar{O}|)^2} \quad (7)$$

1.6.9 MAPE (Porcentagem Absoluta Média do Erro)

A Porcentagem Absoluta Média do Erro (MAPE) é uma métrica amplamente utilizada para avaliar a precisão de previsões em diversos campos, como economia, meteorologia e engenharia. Introduzido inicialmente como uma ferramenta para medir a precisão de previsões, o

MAPE se destaca por sua capacidade de expressar o erro em termos percentuais, facilitando a interpretação dos resultados.

A métrica é calculada como a média dos erros absolutos em relação aos valores reais, permitindo uma comparação direta entre diferentes conjuntos de dados e modelos preditivos (BHATT; SARANGI; KAMBLI, 2023). Em estudos recentes, o MAPE tem sido utilizado para avaliar a eficácia de modelos de previsão em várias aplicações, incluindo previsão de carga elétrica e análise de dados meteorológicos, onde a precisão das previsões é crucial (KHATOON et al., 2023; ZHU et al., 2023).

Na fórmula para o cálculo do MAPE, onde O_i - valor observado; P_i - valor predito pelo modelo; n - número de dados.

$$MAPE = \frac{1}{n} \times \sum_{i=1}^n \left| \frac{O_i - P_i}{O_i} \right| \quad (8)$$

1.6.10 MPE (Erro de Porcentagem Média)

O Erro de Porcentagem Média (MPE) é uma métrica estatística que quantifica a precisão de previsões, expressando a diferença entre valores previstos e observados em termos percentuais. O MPE foi inicialmente introduzido como uma ferramenta para avaliar a eficácia de previsões, e sua aplicação se expandiu ao longo dos anos, sendo utilizado em estudos que envolvem desde a previsão de preços de ações até a análise de dados ambientais (LANGENBUCHER et al., 2025). O MPE foi empregado em uma pesquisa sobre o tempo sedentário, onde a concordância entre diferentes métodos de medição foi avaliada, destacando a importância do MPE na validação de dados (BUCHAN; BAKER, 2023).

Na fórmula para o cálculo do MPE, onde O_i - valor observado; P_i - valor predito pelo modelo; n - número de dados.

$$MPE = \frac{100}{n} \times \sum_{i=1}^n \left(\frac{O_i - P_i}{O_i} \right) \quad (9)$$

OBJETIVOS

OBJETIVO GERAL

Avaliar o potencial de modelagem da evapotranspiração de referência e da erosividade das chuvas no Brasil por meio da utilização das técnicas combinadas de sensoriamento remoto e aprendizagem de máquina.

OBJETIVOS ESPECÍFICOS

1. Analisar o potencial de estimativa da evapotranspiração de referência (ET_o) e gerar um mapa diário de ET_o em tempo real com resolução espacial aproximada de 1 Km², para todo o Brasil, a fim de auxiliar no manejo de irrigação e balanço hídrico no solo em médias e grandes escalas.

2. Modelar e gerar um mapa de erosividade das chuvas para o Brasil, com resolução espacial de 30 segundos de arco (~ 1 km²), utilizando uma combinação de técnicas de SIG (sistema de informação geográfica), modelagem com aprendizado de máquina (machine learning) e produtos disponibilizados por outros autores. Serão selecionadas as melhores covariáveis e os melhores modelos para gerar o mapa final de erosividade.

HIPOTESES

- O uso combinado de sensoriamento remoto e aprendizagem de máquina pode melhorar significativamente a capacidade de modelagem da evapotranspiração de referência e da erosividade das chuvas no Brasil.
- Ao gerar um mapa diário de ETo em tempo real com resolução espacial aproximada de 1 km² e disponibilizar essas informações em uma página da internet para todo o Brasil, será possível fornecer informações úteis para o manejo de irrigação e balanço hídrico no solo em médias e grandes escalas, resultando em uma redução de desperdício de água e aumento da produtividade agrícola.
- Ao modelar o mapa de erosividade das chuvas para o Brasil com uma combinação de técnicas de SIG, aprendizado de máquina e produtos disponibilizados por outros autores, será possível obter uma estimativa precisa da erosividade das chuvas em todo o país, o que pode ser útil para prever a erosão do solo e orientar políticas de conservação ambiental.
- O uso de diferentes covariáveis e modelos de aprendizagem de máquina pode afetar significativamente a precisão das estimativas de evapotranspiração de referência e erosividade das chuvas. Portanto, o estudo pode ajudar a identificar as melhores covariáveis e modelos para cada finalidade.
- O desempenho do modelo de aprendizagem de máquina para a estimativa de evapotranspiração de referência pode ser afetado por fatores como a qualidade dos dados de entrada, a variabilidade climática, a influência de áreas urbanas e a topografia.

CAPÍTULO 2 – MAPA DIÁRIO DA EVAPOTRANSPIRAÇÃO DE REFERÊNCIA NO BRASIL EM TEMPO REAL

RESUMO

DIAS, Santos Henrique Brant. **Modelagem de variáveis hídricas por meio do sensoriamento remoto orbital e inteligência artificial**. 2025. 107f. Tese (Doutorado em Agronomia) – Universidade Estadual de Ponta Grossa, Ponta Grossa, 2025.

Este trabalho teve como objetivo modelar a evapotranspiração de referência em tempo real para todo o Brasil, a fim de auxiliar no manejo de irrigação e balanço hídrico do solo em médias e grandes escalas. Para isso, foram utilizados produtos de sistema de informações geográficas (SIG), imagens orbitais, dados históricos da radiação extraterrestre diária, média e desvio padrão da evapotranspiração diária, e algoritmos de machine learning. O modelo de Cubist apresentou o melhor desempenho, com um RMSE de 0,792, nRMSE de 57,3%, MAPE de 17,4%, D de 0,883 e NSE de 0,634. O resultado foi capaz de gerar um mapa diário de ETo em tempo real com resolução espacial aproximada de 1 Km², disponibilizado em página da internet para todo o Brasil em até 10 minutos após o fim do dia. Conclui-se que o produto desenvolvido é extremamente útil e viável para estimativas da evapotranspiração de referência em todo o país, sendo fundamental para o manejo e conservação de bacias hidrográficas e no manejo da irrigação. O uso inadequado da evapotranspiração pode resultar em perdas de água, de capital e de produtividade dos cultivos. Modelos de machine learning juntamente com produtos de sistema de informações geográficas (SIG) e imagens orbitais têm sido amplamente utilizados para monitorar e modelar inúmeras características da superfície, sendo que o tempo de disponibilização dos produtos é extremamente relevante para sua utilização prática.

Palavras-chave: Balanço hídrico; Manejo e conservação da água; *Machine learning*.

2.1 INTRODUÇÃO

A gestão apropriada da água no solo é cada vez mais importante, e seu papel se estende a uma ampla gama de estudos, desde a irrigação até a dinâmica das águas subterrâneas e o balanço hídrico das bacias hidrográficas (ANNELI VIRTANEN et al., 2016; JØRGENSEN; VILLHOLTH; REFSGAARD, 2017; KOUNDOURI; ROSETA-PALMA; ENGLEZOS, 2017). Além dos tradicionais balanços hídricos das culturas e do solo, a evapotranspiração pode ser usada para conservar tanto o solo quanto a água.

Muitos modelos utilizam a evapotranspiração para prever o desenvolvimento das plantas na cobertura do solo, prever o escoamento superficial e prevenir a erosão do solo. Também é possível utilizar a evapotranspiração para estimar a umidade do solo antes da chuva, o que ajuda a determinar a capacidade de infiltração da água (PRUSKI, 2009).

O uso inadequado da água na irrigação pode resultar em desperdícios de água, capital financeiro e redução da produtividade dos cultivos. Tais problemas podem ser evitados por meio do correto dimensionamento do projeto e da gestão da irrigação. Além disso, as frequentes crises hídricas estão tornando a água cada vez mais escassa em termos de quantidade e qualidade. Os cenários futuros preveem limitações crescentes na disponibilidade de água, o que aumenta a necessidade de uma gestão adequada e sustentável da água no solo, algo que muitas vezes é negligenciado pelos agricultores (DIAS et al., 2021; LEVIDOW et al., 2014).

O manejo adequado da água no solo é de extrema importância, e a estimativa da demanda hídrica é realizada através do balanço de água no solo, que utiliza a precipitação e a evapotranspiração como base (LIU et al., 2017). O método padrão de Penman-Monteith - FAO 56 é utilizado para o cálculo da evapotranspiração de referência (ET_o) e é recomendado pela Food Agriculture Organization (FAO) (ALLEN et al., 1998). Contudo, esse método depende fortemente das estações meteorológicas e de muitas variáveis do clima, o que pode torná-lo inviável em áreas distantes da estação, além do alto custo dos instrumentos meteorológicos.

Estudos têm demonstrado que modelos de inteligência artificial com uso de aprendizado de máquina apresentam um desempenho superior aos métodos tradicionais na estimativa da ET_o, principalmente quando utilizados em comparação com medidas de lisímetros (ABDULLAH; MALEK, 2016; CHOUBIN et al., 2018; KARANDISH; ŠIMŮNEK, 2016; KHOSHRAVESH; SEFIDKOUHI; VALIPOUR, 2017; LANDERAS et al., 2017; NAGHIBI; POURGHASEMI;

DIXON, 2016; SARHADI et al., 2016; SHIRI et al., 2016; TIWARI; ADAMOWSKI, 2017). Porém, é importante ressaltar que, para avaliar a precisão da estimativa de ETo, é necessário ter dados precisos dos lisímetros.

A instalação de estações meteorológicas em grande escala é inviável devido ao alto custo de aquisição e manutenção das estações, bem como a falta de acessibilidade em todos os locais e possíveis extravios que podem ocorrer com essas estações. Portanto, o sensoriamento remoto tem sido utilizado em estudos de evapotranspiração e manejo de água no solo em grandes áreas, de maneira espacial. Embora muitos trabalhos tenham sido realizados para determinar a evapotranspiração diária e a umidade atual do solo utilizando imagens de satélite, esse tipo de estudo ainda necessita de mais pesquisas para se aprimorar (ALTHOFF et al., 2020; HE et al., 2016; LAIOLO et al., 2016; SADEGHI et al., 2017).

Atualmente, há um procedimento alternativo para determinar a evapotranspiração de referência (ET_o) utilizando imagens de satélite, que emprega diferentes métodos de cálculo, como o Simple Algorithm for Evapotranspiration Retrieving (SAFER), Surface Energy Balance Algorithm for Land (SEBAL), Mapping EvapoTranspiration at high Resolution with Internalized Calibration (METRIC) e MODIS Global Evapotranspiration Project (MOD16), entre outros. Embora muitos desses métodos sejam baseados no modelo FAO 56, eles dependem da banda termal e de dados meteorológicos locais e não têm a mesma precisão que as estações meteorológicas com dados completos. No entanto, as imagens de satélite têm a vantagem da espacialização e abrangência de uma grande área, ao contrário das estações meteorológicas (TOUREIRO et al., 2017).

As técnicas de sensoriamento remoto combinadas com algoritmos de machine learning têm sido amplamente utilizadas em diversas áreas e recentemente aplicadas com sucesso na área dos recursos hídricos (CHOUBIN et al., 2018; KARANDISH; ŠIMŮNEK, 2016; NAGHIBI; POURGHASEMI; DIXON, 2016; SARHADI et al., 2016; SHIRI et al., 2016; TIWARI; ADAMOWSKI, 2017). Portanto, essas técnicas podem ser aplicadas para melhorar a determinação da ETo.

O sensoriamento remoto pode ser uma ferramenta ideal para a predição e calibração da ETo, com um enorme potencial de trabalho e exploração. Atualmente, algoritmos de inteligência artificial são utilizados para interpolar dados com dependência espacial. Os métodos tradicionais de interpolação, como a krigagem e o inverso da distância ponderada (IDW), são considerados

ultrapassados e não preveem as variabilidades do mundo real, como relevo, clima e vegetação, além de apresentarem problemas, como insensibilidade às mudanças bruscas, como de altitude.

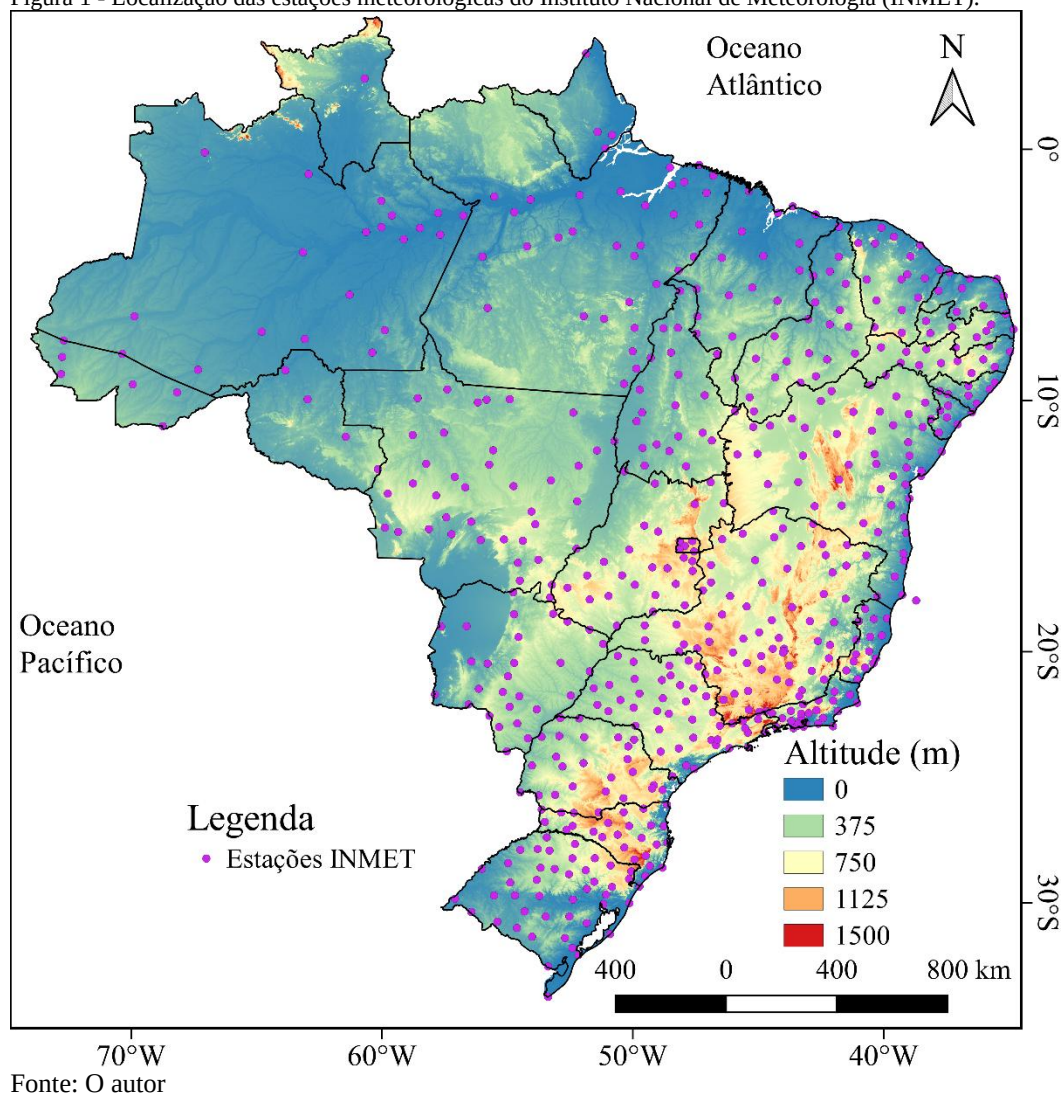
O presente trabalho tem como objetivos modelar a ETo, realizar diariamente uma imagem de ETo para todo o Brasil e disponibilizar o produto de forma simplificada e acessível, utilizando a rede de estações automáticas do Instituto Nacional de Meteorologia (INMET), conhecimentos em sensoriamento remoto, inteligência artificial, modelagem e linguagem de programação R, além do banco de dados WordClim (FICK; HIJMANS, 2017).

2.2 MATERIAL E MÉTODOS

2.2.1 Área de Estudo e Obtenção dos Dados

A área escolhida para o estudo é o Brasil, um país de dimensões continentais com grande variação em suas características morfológicas e climáticas. Os dados das estações meteorológicas automáticas de todo o país foram utilizados, totalizando cerca de 580 estações. Esses dados foram disponibilizados pelo Instituto Nacional de Meteorologia (INMET) (INMET, 2022) e processados no software R. A Figura 1 mostra a localização dessas estações.

Figura 1 - Localização das estações meteorológicas do Instituto Nacional de Meteorologia (INMET).



Embora a Figura 1 apresente 573 estações, a ETo foi calculada corretamente em apenas 157 a 261 estações entre as datas de 01/07/2021 a 01/07/2022. Essa diferença é atribuída a erros na coleta de dados, estações fora de funcionamento ou dados incorretos.

O INMET possui atualmente 573 estações meteorológicas automáticas em todo o país, que coletam dados meteorológicos horários e estão disponíveis para download com dados históricos de um ano. Essas estações coletam dados sobre temperatura do ar (°C) (instantânea, máxima e mínima), umidade relativa do ar (%) (instantânea, máxima e mínima), ponto de orvalho (°C) (instantânea, máxima e mínima), pressão atmosférica (hPa) (instantânea, máxima e mínima), vento (velocidade (m/s), direção (°) e rajada(m/s)), radiação solar (kJ/m²) e chuva (mm) (INMET, 2022).

2.2.2 Cálculo da Evapotranspiração

Nesta seção, descreve-se como foi feito o cálculo da evapotranspiração (ET₀) a partir dos dados obtidos das estações meteorológicas do INMET. Para esse cálculo, foi utilizada a equação padrão recomendada pela FAO, que considera uma superfície cultivada com grama hipotética com altura de 0,12 m, resistência aerodinâmica de 70 s m⁻¹ e albedo de 0,23 (ALLEN et al., 1998).

Os resultados obtidos a partir do cálculo da ET₀ são importantes para entender o comportamento do clima e da vegetação em diferentes regiões do país. A partir desses resultados, é possível realizar análises sobre a disponibilidade de água para o uso humano e para a produção agrícola, por exemplo.

Na fórmula para o cálculo da ET₀ - evapotranspiração de referência (mm d⁻¹), onde R_n - saldo de radiação na superfície, MJ m⁻² d⁻¹; G - fluxo de calor no solo, MJ m⁻² d⁻¹; t - temperatura média do ar, °C; u₂ - velocidade do vento a 2 m de altura, m s⁻¹; e_s - pressão de saturação de vapor, kPa; e_a - pressão de vapor atual do ar, kPa; Δ - declividade da curva de pressão de vapor de saturação, kPa °C⁻¹; γ - coeficiente psicrométrico, kPa °C⁻¹.

$$ET_0 = \frac{0,408 \Delta (R_n - G) + \gamma \frac{900}{t + 273} u_2 (e_s - e_a)}{\Delta + \gamma (1 + 0,34 u_2)} \quad (10)$$

2.2.3 Algoritmos Utilizados para a Aprendizagem de Máquina (Machine Learning)

Na Tabela 2 são apresentados os algoritmos de *machine learning* utilizados no estudo.

Tabela 2 - Algoritmos de *machine learning* utilizados no estudo

Modelo	Abreviação	Pacote R	Autor
Random Forest	RF	randomForest	Tin Kam Ho (1995)
Linear Regression	LM	Base R	
Cubist	cubist	Cubist	(ADAMS; STERLING, 1992; QUINLAN, 1993)
Partial Least Squares	PLS	pls	(GELADI; KOWALSKI, 1986; HELLAND, 1988)
Gradient Boosting Machine	GBM	gbm, plyr	(FRIEDMAN, 2001, 2002)
Independent Component Regression	ICR	fastICA	(WESTAD, 2005)
Bagged MARS	BagEarth	earth	(HASTIE; WRAPPER, 2021)
Gaussian Process with Radial Basis Function Kernel	gaussprRadial	kernlab	(KARATZOGLOU et al., 2004)
Support Vector Machines with Polynomial Kernel	svmPoly	kernlab	(KARATZOGLOU et al., 2004)

Fonte: O autor

2.2.4 Variáveis e Covariáveis dos Modelos

2.2.4.1 *WorldClim*

As variáveis do WorldClim foram desenvolvidas a partir do conjunto de estações meteorológicas da Organização Mundial de Meteorologia, como destacado por Fick e Hijmans (2017). O WorldClim disponibiliza diversas variáveis climáticas, que podem ser observadas na Tabela 3. Para a análise, foram utilizados dados com resoluções espaciais de 30 segundos, o que corresponde a cerca de 1 km².

Para as variáveis meteorológicas, foi selecionado um valor para cada mês do ano, enquanto para as variáveis bioclimáticas, foi utilizada uma variável normal que caracteriza a região.

Mais informações sobre as variáveis disponíveis podem ser encontradas no site do WorldClim (<https://www.worldclim.org/data/worldclim21.html>).

Tabela 3 – Covariáveis disponíveis no WorldClim

Abreviação	Descrição	Unidade	Autor
Bio1	Temperatura média anual	°C	Fick and Hijmans (2017)
Bio2	Faixa diurna média (média do mês (temperatura máxima - temperatura mínima))	°C	Fick and Hijmans (2017)
Bio3	Isotérmica (BIO2 / BIO7) ($\times 100$)	°C	Fick and Hijmans (2017)
Bio4	Sazonalidade da temperatura (desvio padrão $\times 100$)	°C	Fick and Hijmans (2017)
Bio5	Temperatura máxima do mês mais quente	°C	Fick and Hijmans (2017)
Bio6	Temperatura mínima do mês mais frio	°C	Fick and Hijmans (2017)
Bio7	Faixa anual de temperatura (BIO5-BIO6)	°C	Fick and Hijmans (2017)
Bio8	Temperatura média do quarto mais úmido	°C	Fick and Hijmans (2017)
Bio9	Temperatura média do quarto mais seco	°C	Fick and Hijmans (2017)
Bio10	Temperatura média do trimestre mais quente	°C	Fick and Hijmans (2017)
Bio11	Temperatura média do quarto mais frio	°C	Fick and Hijmans (2017)
Bio12	Precipitação anual	mm	Fick and Hijmans (2017)
Bio13	Precipitação do mês mais chuvoso	mm	Fick and Hijmans (2017)
Bio14	Precipitação do mês mais seco	mm	Fick and Hijmans (2017)
Bio15	Sazonalidade da precipitação (coeficiente de variação)	mm	Fick and Hijmans (2017)
Bio16	Precipitação do trimestre mais chuvoso	mm	Fick and Hijmans (2017)
Bio17	Precipitação do trimestre mais seco	mm	Fick and Hijmans (2017)
Bio18	Precipitação do trimestre mais quente	mm	Fick and Hijmans (2017)
Bio19	Precipitação do trimestre mais frio	mm	Fick and Hijmans (2017)
Altitude	Altitude	m	Fick and Hijmans (2017)
tmin	Temperatura mínima	°C	Fick and Hijmans (2017)
tmax	Temperatura máxima	°C	Fick and Hijmans (2017)
tavg	Temperatura média	°C	Fick and Hijmans (2017)
prec	Precipitação	mm	Fick and Hijmans (2017)
srad	Radiação solar	$\text{kJ m}^{-2} \text{ dia}^{-1}$	Fick and Hijmans (2017)
wind	Velocidade do vento	M s^{-1}	Fick and Hijmans (2017)

Fonte: O autor

2.2.4.2 Evapotranspiração de referência (ET_o) diária no Brasil

Em um trabalho inovador, Althoff et al. (2020) calcularam a evapotranspiração de referência (ET_o) diariamente para o Brasil no período de 01/06/2000 a 01/01/2020, totalizando 7.520 rasters. Os autores selecionaram o modelo cubist como o melhor desempenho, utilizando

849 estações e como covariáveis, o WorldClim e os dados da Medição Global de Precipitação (GPM) (ALTHOFF et al., 2020) (<https://data.mendeley.com/datasets/sstjw74ryh/4>).

A partir dessas imagens, foi realizada a média e desvio padrão de cada pixel para cada dia do ano, resultando em 366 imagens da ETo média e 366 imagens do desvio padrão da ETo.

2.2.4.3 Radiação extraterrestre

A radiação solar extraterrestre é a quantidade de radiação emitida pelo sol que chega à atmosfera terrestre, e é um fator importante para entender a evapotranspiração. Entretanto, a radiação que alcança a superfície terrestre é afetada pela interação com a atmosfera, o que resulta em diferentes níveis de radiação em cada região. Para o cálculo da radiação extraterrestre, os autores utilizaram as equações de 11 a 14.

Nas fórmulas para o cálculo do Q_o - Radiação solar extraterrestre ($\text{MJ.m}^{-2}.\text{d}^{-1}$), onde λ - Declinação solar ($^\circ$); dj - Dia juliano (dia); α - Latitude do pixel ($^\circ$); π - Constante PI (3,141592...); dD^2 - Constante solar da distância Terra sol; H_n - Ângulo zenital ($^\circ$).

$$\lambda = \alpha \times \sin\left(\frac{360}{365 \times (dj - 80)}\right) \times \frac{\pi}{180} \quad (11)$$

$$dD^2 = 1 + 0,033 \times \cos\left(\frac{360}{365} \times dj \times \frac{\pi}{180}\right) \quad (12)$$

$$H_n = \arccos\left(-\tan\left(\frac{\lambda \times \pi}{180}\right) \times \tan\left(\frac{\alpha \times \pi}{180}\right)\right) \times \frac{180}{\pi} \quad (13)$$

$$Q_o = \frac{37,6 \times dD^2 \times \pi \times \sin\left(\frac{\alpha \times \pi}{180}\right)}{180 \times H_n \times \sin\left(\frac{\lambda \times \pi}{180}\right)} + \cos\left(\frac{\lambda \times \pi}{180}\right) \times \cos\left(\frac{\alpha \times \pi}{180}\right) \times \sin\left(\frac{H_n \times \pi}{180}\right) \quad (14)$$

2.2.5 Métricas de Avaliação dos Algoritmos/Modelos

2.2.5.1 Índice de desempenho

Com o objetivo de avaliar o desempenho dos modelos durante as fases de calibração e validação, o presente estudo utilizou oito índices estatísticos, os quais são apresentados na Tabela 4. Para melhor observação dos resultados, foram confeccionados boxplots dos índices propostos. Os índices de avaliação selecionados foram: erro médio quadrático (RMSE), normalização do

RMSE (nRMSE), eficiência de Nash-Sutcliffe (NSE), índice de concordância Willmott (D), erro médio (ME), erro absoluto médio (MAE), porcentagem absoluta média do erro (MAPE) e porcentagem do erro médio (MPE). Importante destacar que essas equações têm sido utilizadas em inúmeros estudos para avaliação de modelos, como evidenciado por vários trabalhos (ASHRAFIAN et al., 2020; LI et al., 2020; MOHAMMED et al., 2020; MOUSAVI et al., 2020; YASEEN et al., 2020; ZAKHROUF et al., 2020).

Tabela 4 - Índices de avaliação de desempenho utilizados no estudo

Índices de avaliação	Definição
Erro médio quadrático	$RMSE = \sqrt{\frac{\sum_{i=1}^n (O_i - P_i)^2}{n}}$
Normalização do RMSE	$nRMSE = 100 \times \frac{\sqrt{\frac{\sum_{i=1}^n (O_i - P_i)^2}{n}}}{\bar{O}}$
Eficiência de Nash-Sutcliffe	$NSE = \left(1 - \frac{\sum_{i=1}^n (O_i - P_i)^2}{\sum_{i=1}^n (O_i - \bar{O})^2} \right) \times 100$
Erro médio absoluto	$MAE = \frac{1}{n} \sum_{i=1}^n P_i - O_i $
Erro médio	$ME = \frac{1}{n} \sum_{i=1}^n (P_i - O_i)$
Índice de concordância Willmott	$D = 1 - \frac{\sum_{i=1}^n (O_i - P_i)^2}{\sum_{i=1}^n (P_i - \bar{O} + O_i - \bar{O})^2}$
Porcentagem absoluta média do erro	$MAPE = \frac{1}{n} \times \sum_{i=1}^n \left \frac{O_i - P_i}{O_i} \right $
Erro de porcentagem média	$MPE = \frac{100}{n} \times \sum_{i=1}^n \left(\frac{O_i - P_i}{O_i} \right)$

Fonte: O autor

Em que: P_i - valor predito pelo modelo; O_i - valor observado; n - número de dados, $\bar{O}$ - valor médio observado; $\bar{P}$ - valor médio predito pelo modelo.

O erro médio quadrático (RMSE - root mean squared error) é um índice estatístico utilizado para medir a magnitude média do erro, ou seja, a diferença quadrática entre os dados estimados e observados. O RMSE varia de 0 a $+\infty$ e quantifica o erro com as mesmas unidades da variável original, dando um maior peso aos grandes erros e evidenciando a presença de outliers. Ele é particularmente útil quando grandes erros são indesejáveis (WILKS, 2011; WILLMOTT; MATSUURA, 2005).

A normalização do RMSE (nRMSE - normalized root mean square error) facilita a comparação entre a eficiência de conjuntos de dados ou modelos com diferentes escalas. Expresso em porcentagem, valores menores indicam menor variação residual (WILKS, 2011).

O coeficiente de determinação (r^2) é outra medida descritiva da qualidade do ajuste obtido pelo modelo. Ele indica o quanto o modelo é capaz de explicar os dados observados. No entanto, o índice não leva em consideração a falta de ajuste, que poderá ser grande mesmo que os valores observados e previstos não estejam linearmente relacionados. Dessa forma, o r^2 não deve ser considerado sozinho, mas sempre aliado a outras métricas.

A eficiência de Nash-Sutcliffe (NSE - Nash-Sutcliffe efficiency) é utilizada para avaliar o poder preditivo do modelo e varia de $-\infty$ a 1. Um valor de 1 corresponde ao ajuste perfeito entre os dados estimados pelo modelo com os dados medidos. A NSE é amplamente utilizada na avaliação de modelos hidrológicos (MORIASI et al., 2007; WILLMOTT, 1982).

O erro médio absoluto (MAE – Mean Absolute Error) é um índice que fornece um valor médio dos erros absolutos. Ele varia de 0 a $+\infty$ e quantifica o erro com as mesmas unidades da variável original. O erro do viés médio (MBE – Mean Bias Error) é indicado para visualizar as tendências de subestimativas ou superestimativas (CHAI; DRAXLER, 2014).

2.2.6 Treino e Validação do Modelo

Durante a realização dos treinamentos, foram empregados 90% dos dados, enquanto os 10% restantes foram reservados para validação e avaliação dos modelos. Tal divisão foi realizada manualmente e independente do treinamento dos modelos, a fim de ajustá-los de forma apropriada. É importante destacar que cada modelo tem a sua própria forma de separar os dados.

Conforme mencionado por Johnson et al. (2019), a separação adequada dos dados é crucial para garantir que o modelo seja capaz de generalizar bem em dados novos e não vistos anteriormente. Portanto, é importante que essa separação seja feita com cuidado e de maneira apropriada, a fim de evitar problemas como overfitting ou underfitting.

Além disso, é comum utilizar técnicas como cross-validation e bootstrap para auxiliar na seleção dos melhores modelos e evitar que os resultados obtidos sejam enviesados. Segundo Kim et al. (2020), essas técnicas têm sido amplamente utilizadas na comunidade científica e podem ser muito úteis na seleção de modelos com alta capacidade de generalização e baixa tendência a overfitting.

Em suma, a separação adequada dos dados é fundamental para garantir a qualidade dos modelos de machine learning. Por isso, é importante realizar essa tarefa com cuidado e atenção aos

detalhes, além de considerar o uso de técnicas adicionais para auxiliar na seleção dos melhores modelos.

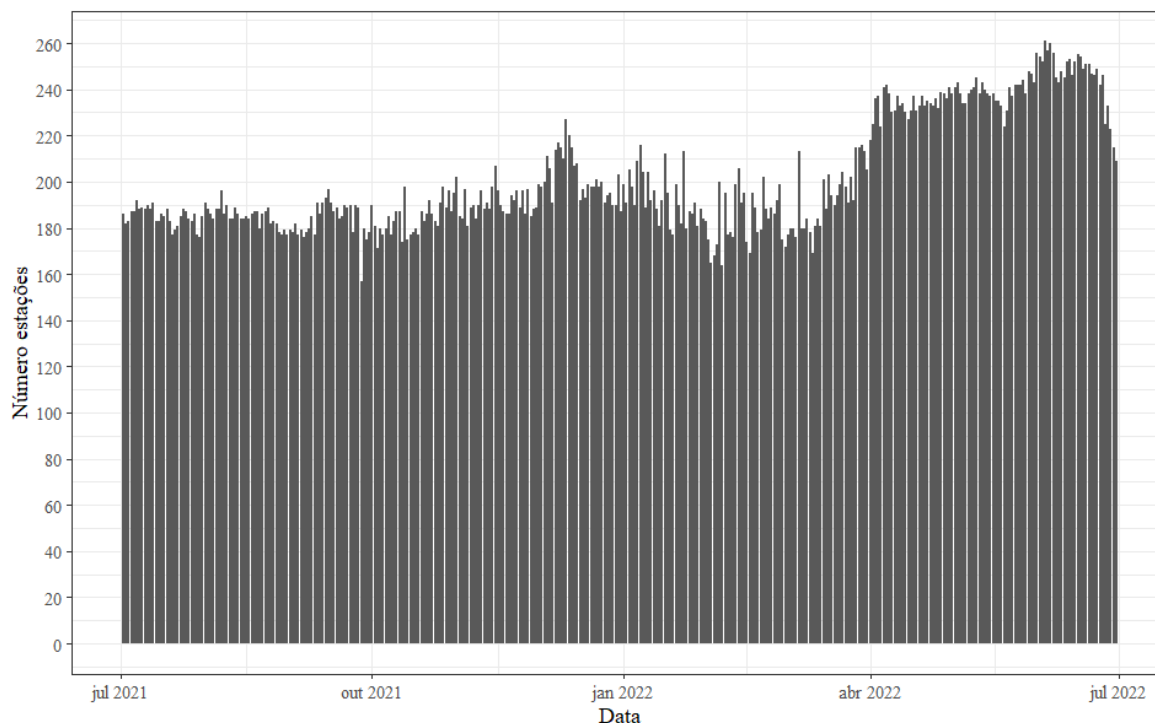
2.3 RESULTADOS

O trabalho em questão disponibilizou todos os códigos de desenvolvimento compilados em um pacote do R (<https://github.com/santoshbdias/MapEToBR>). O pacote foi otimizado ao máximo e requer apenas um comando para gerar o mapa de ETo após o carregamento das covariáveis. Todo o processamento demanda um tempo aproximado de 10 minutos, o que indica uma ótima performance computacional.

Para avaliar o desempenho do modelo, foram realizadas 365 simulações entre 01/07/2021 e 30/06/2022. A Figura 2 apresenta o número de estações meteorológicas utilizadas para o treinamento do modelo em cada data do ano avaliado. É importante mencionar que essa variação ocorre devido à disponibilidade de dados meteorológicos em cada estação.

Em suma, a compilação do código em um pacote do R e a sua otimização foram fundamentais para garantir a eficiência do modelo utilizado neste trabalho. Além disso, a realização de múltiplas simulações permitiu uma avaliação mais precisa do seu desempenho e robustez. As variações no número de estações utilizadas para treinamento do modelo em cada data do ano avaliado evidenciam a importância da disponibilidade de dados meteorológicos para a análise de ETo.

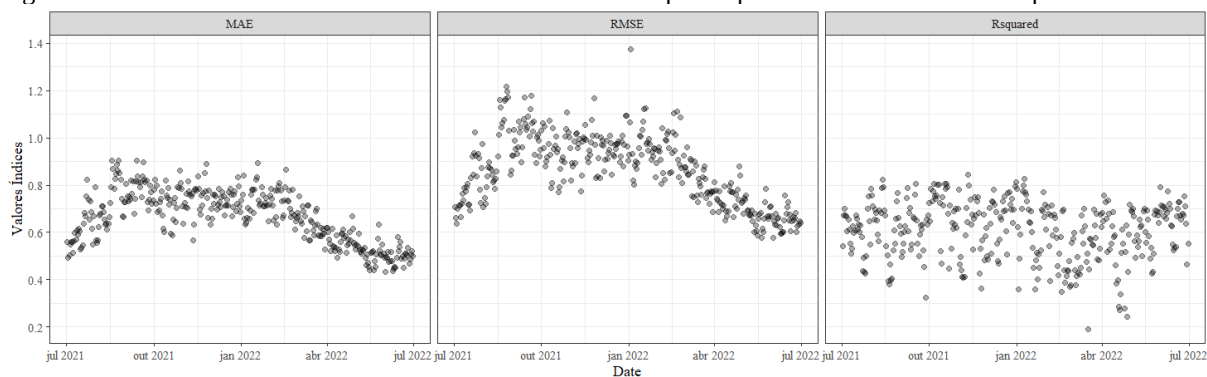
Figura 2 - Número de estações utilizadas para modelagem em cada data.



Fonte: O autor

Foram utilizadas diferentes quantidades de estações meteorológicas para a modelagem da evapotranspiração de referência (ET_o) em cada dia do período de simulação. O número mínimo de estações meteorológicas utilizadas foi de 157 no dia 27 de setembro de 2021, enquanto o número máximo foi de 261 estações no dia 4 de junho de 2022. Para cada data, foram realizadas as métricas MAE, RMSE e R², considerando de 1 a 32 covariáveis. Os resultados dos melhores índices obtidos para cada data podem ser visualizados na Figura 3.

Figura 3 - Menores valores de MAE e RMSE e o maior de Rsquared para cada dia dos 365 dias que foram testados.

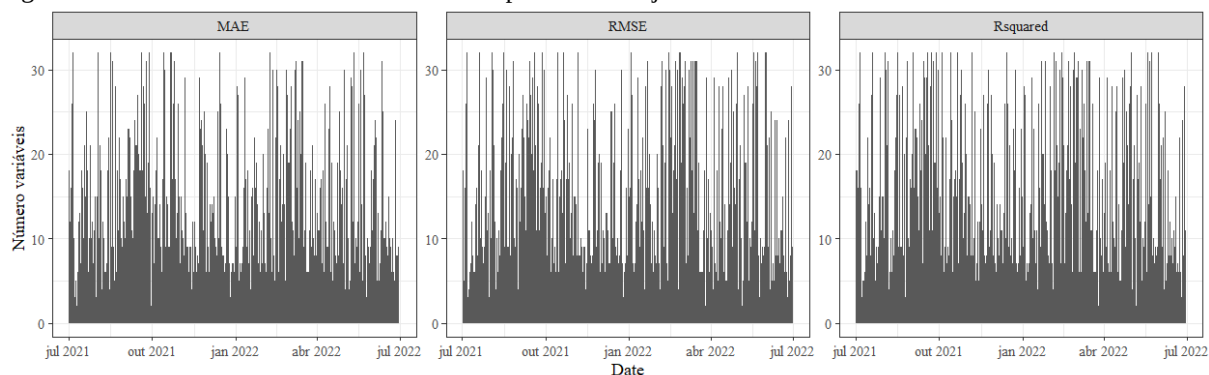


Fonte: O autor

2.3.1 Seleção das Covariáveis

No contexto de modelagem, as covariáveis são os fatores que têm maior influência sobre uma determinada variável em um dia específico. A seleção das melhores covariáveis é crucial para obter resultados precisos em um modelo. Após a análise dos índices de desempenho, foram escolhidos os conjuntos de covariáveis mais adequados para cada dia do ano. Essa seleção foi feita com base nos dados históricos disponíveis, e as covariáveis escolhidas foram aquelas que tiveram maior correlação com a variável em questão em cada dia específico. Na Figura 4 pode ser observado o número de covariáveis utilizadas em cada dia do ano.

Figura 4 - Número de covariáveis selecionadas para melhor ajuste em cada data.

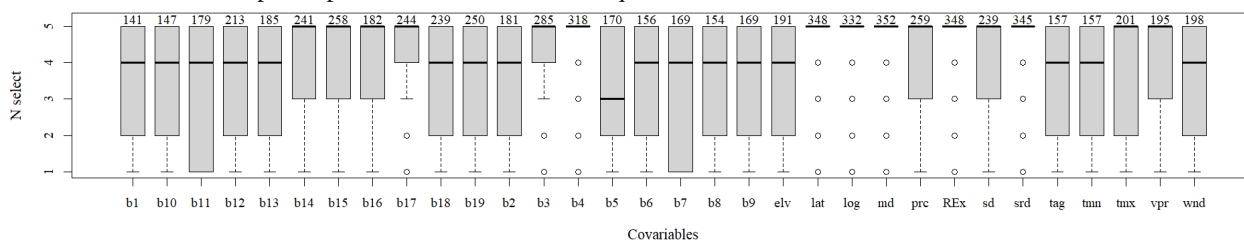


Fonte: O autor

O gráfico apresentado na Figura 5 mostra o boxplot que representa o número de vezes que cada conjunto de covariáveis foi selecionado como o melhor para cada dia do ano. A escala de uso varia de 1 a 5 e é possível observar que a média normal da evapotranspiração diária (md) é a covariável mais frequente na escolha do modelo. Em segundo lugar, temos a latitude (lat), seguida pela radiação extraterrestre (REx) e radiação Wordclim (srd), respectivamente. As demais covariáveis apresentaram um uso menos frequente no modelo.

É importante ressaltar que a escolha do conjunto de covariáveis varia de acordo com a época do ano e as condições climáticas específicas de cada período. Dessa forma, os resultados apresentados na Figura 5 podem ser úteis para aprimorar o processo de seleção de variáveis em estudos futuros relacionados à modelagem da evapotranspiração diária.

Figura 5 - Número de vezes que a covariável foi selecionada (escala de 0 a 5) durante a avaliação dos 365 dias. O valor numérico acima do boxplot representa o número de dias que a variável foi selecionada.



Fonte: O autor

A Tabela 5 apresenta a lista das principais covariáveis selecionadas em ordem de importância. A seleção foi realizada utilizando o comando `recursive_feature_elimination`, presente no pacote `labgeo` (FERNANDES FILHO, 2020).

Essas covariáveis foram escolhidas com base em sua relevância para o modelo, de forma a maximizar sua capacidade preditiva. Dentre as covariáveis selecionadas, podemos citar a média da evapotranspiração (`md`), a radiação solar global (`srd`) e a longitude e latitude, que apresentaram grande influência no desempenho do modelo.

Tabela 5 - Ordem de importância das covariáveis.

Covariável	Descrição	Ajuste
md	Média da evapotranspiração de referência diária	12,9846
srd	Radiação solar	11,6361
log	Longitude	10,4536
lat	Latitude	9,7338
REx	Radiação solar extraterrestre	9,6951
b4	Sazonalidade da temperatura (desvio padrão $\times$ 100)	8,9810
b3	Isotérmica (BIO2 / BIO7) ($\times$ 100)	8,3876
prc	Precipitação	7,9592
sd	Desvio padrão da evapotranspiração de referência diária	7,8710
b18	Precipitação do trimestre mais quente	7,6409
b15	Sazonalidade da precipitação (coeficiente de variação)	7,6321
b17	Precipitação do quarto mais seco	7,5653
b14	Precipitação do mês mais seco	7,5016
b19	Precipitação do trimestre mais frio	7,3432
wnd	Velocidade do vento	6,9358
b12	Precipitação anual	6,8648
b16	Precipitação do bairro mais chuvoso	6,7839
vpr	Pressão de vapor	6,5066
tmx	Temperatura máxima	6,4694
b2	Faixa Diurna Média (Média do mês (temperatura máxima - temperatura mínima))	6,4612
b13	Precipitação do mês mais chuvoso	6,4414
elv	Altitude	6,1818
b5	Temperatura máxima do mês mais quente	6,0200
b7	Faixa anual de temperatura (BIO5-BIO6)	5,9382
b11	Temperatura média do quarto mais frio	5,9090
b6	Temperatura mínima do mês mais frio	5,6435
b9	Temperatura média do quarto mais seco	5,6410
tmn	Temperatura mínima	5,5237
b8	Temperatura média do quarto mais úmido	5,4746
tag	Temperatura média	5,3395
b10	Temperatura média do trimestre mais quente	5,3080
b1	Temperatura média anual	5,2993

Fonte: O autor

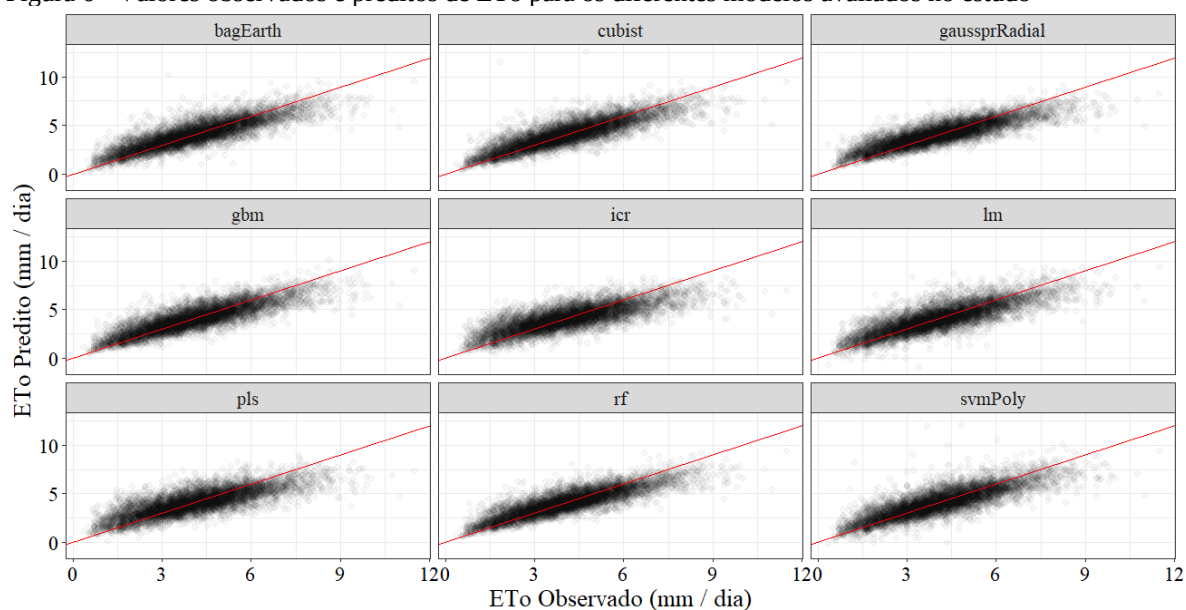
Em ordem de prioridade foram selecionadas as covariáveis: md, REx, lat, srd, log e b4. Os resultados observados pelo boxplot foram diretamente proporcionais ao ajuste da Tabela 5.

2.3.2 Seleção do Modelo na Estimativa da ETo

O uso de modelos de *machine learning* para interpolação de valores tem se tornado cada vez mais comum, proporcionando modelagens mais precisas e eficientes. Na Figura 6, são apresentados os valores observados e preditos dos diferentes modelos avaliados neste trabalho.

É possível observar que os modelos apresentaram desempenhos variados na predição dos valores, o que pode ser atribuído à escolha das covariáveis, aos parâmetros de ajuste de cada modelo. Dentre os modelos avaliados, destaca-se o modelo de Regressão Linear (RL), que apresentou bons resultados na predição dos valores, com pouca dispersão em relação aos valores observados.

Figura 6 - Valores observados e preditos de ETo para os diferentes modelos avaliados no estudo



Fonte: O autor

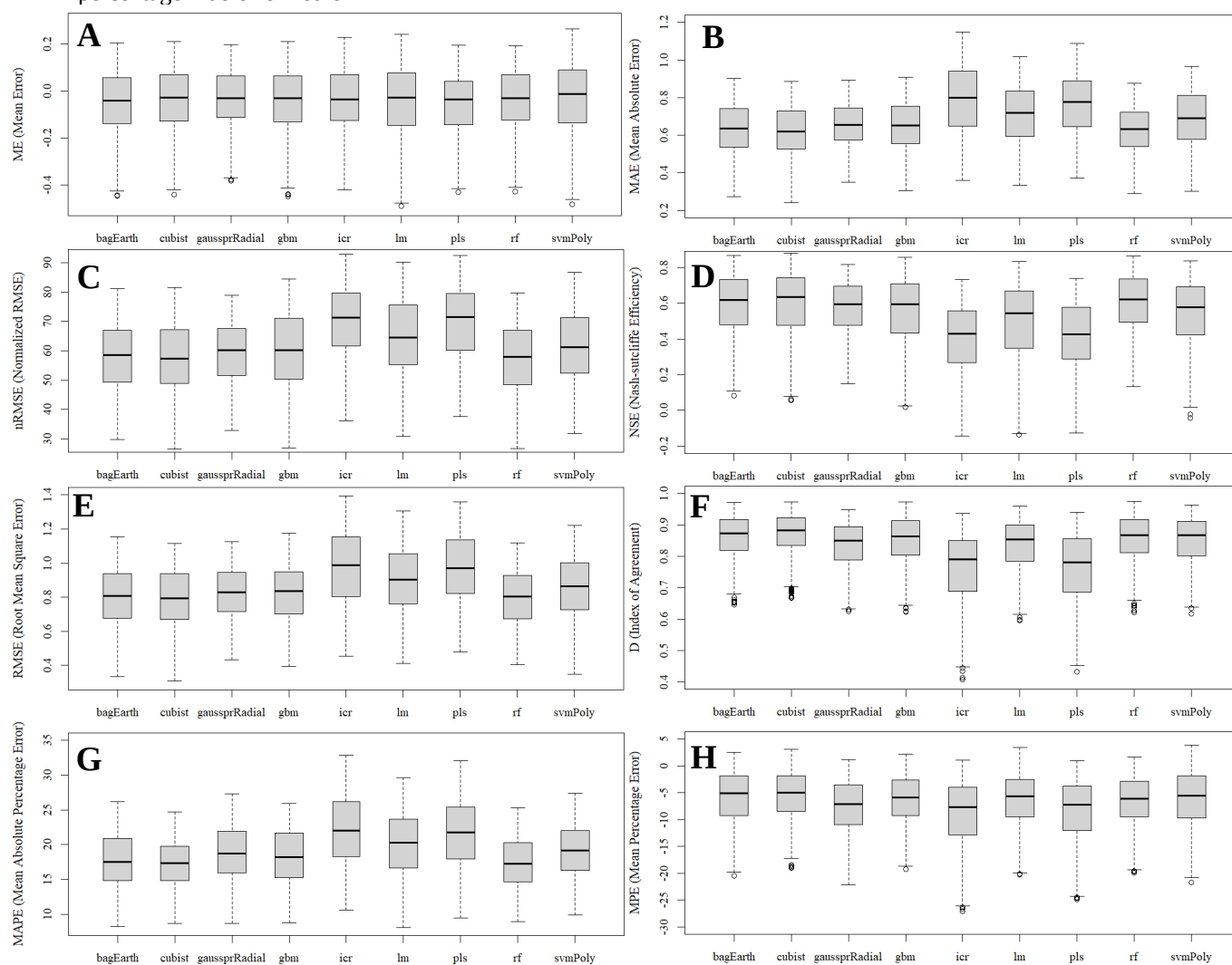
Para avaliar o desempenho dos diferentes modelos, foram utilizados os índices de avaliação descritos na seção de Material e Métodos. A Figura 7 apresenta os boxplots referentes aos oito índices de desempenho para os nove modelos de machine learning testados.

É possível observar que os modelos apresentaram resultados distintos para cada um dos índices avaliados. Dentre os modelos testados, destaca-se a Regressão Linear (RL) e o Random Forest (RF), que apresentaram bons resultados na maioria dos índices avaliados. Por outro lado, o

modelo Independent Component Regression (icr) apresentou um dos piores resultados na maioria dos índices.

Cada um dos índices de desempenho avalia uma característica específica do modelo, como precisão, acurácia, sensibilidade, especificidade, entre outras. Dessa forma, é importante considerar cada índice individualmente e avaliar o modelo de acordo com as necessidades do estudo.

Figura 7 - Boxplot dos índices utilizados para seleção do modelo. A - erro médio (mm/dia), B - erro absoluto médio, C - raiz do erro quadrático normalizado (%), D - índice de eficiência de Nash-Sutcliffe (adimensional), E - raiz do erro quadrático (mm/dia), F - índice de concordância (adimensional), G - porcentagem absoluta média do erro, H - porcentagem do erro médio



O objetivo de utilizar o ME é verificar se os modelos superestimam ou subestimam as estimativas. Os resultados indicam que todos os modelos subestimaram a ETo, conforme observado

nas medianas dos modelos: bagEarth (-0,0412), cubist (-0,0292), gaussprRadial (-0,0315), gbm (-0,0314), icr (-0,0362), lm (-0,0271), pls (-0,0367), rf (-0,0308), svmPoly (-0,0132).

Ao analisar a porcentagem do ME (MPE), a mediana dos modelos foi: bagEarth (-5,06), cubist (-4,97), gaussprRadial (-7,05), gbm (-5,83), icr (-7,66), lm (-5,68), pls (-7,24), rf (-6,07), svmPoly (-5,56).

O MAE é utilizado para observar o erro dos modelos, e a mediana MAE dos modelos foram: bagEarth (0,635), cubist (0,619), gaussprRadial (0,654), gbm (0,652), icr (0,798), lm (0,720), pls (0,777), rf (0,632), svmPoly (0,690). Por sua vez, o MAPE, que é a porcentagem do MAE, fornece uma maneira mais fácil de observar, e a mediana dos modelos foram: bagEarth (17,5), cubist (17,4), gaussprRadial (18,7), gbm (18,2), icr (22,0), lm (20,3), pls (21,7), rf (17,3), svmPoly (19,1). Ao analisar os valores de MAPE, o modelo que apresentou o melhor desempenho foi o random forest (rf), com cubist e bagEarth muito próximos.

O RMSE, que é semelhante ao MAE, penaliza valores de outlier. A mediana do RMSE para os modelos foram: bagEarth (0,808), cubist (0,792), gaussprRadial (0,829), gbm (0,836), icr (0,988), lm (0,902), pls (0,969), rf (0,803), svmPoly (0,864). Para a porcentagem do RMSE, que é o nRMSE, a mediana dos modelos foram: bagEarth (58,6), cubist (57,3), gaussprRadial (60,3), gbm (60,3), icr (71,3), lm (64,6), pls (71,5), rf (57,9), svmPoly (61,2). Nesse indicador, o cubist apresentou o melhor desempenho, com rf e bagEarth muito próximos.

Os índices D e NSE variam de 0 a 1 e representam o ajuste do modelo para calcular o valor observado. As medianas do índice D foram: bagEarth (0,873), cubist (0,883), gaussprRadial (0,850), gbm (0,865), icr (0,790), lm (0,853), pls (0,782), rf (0,868), svmPoly (0,867). Para o índice D, o modelo com melhor desempenho foi o cubist, com bagEarth e rf muito próximos. Para o índice NSE, as medianas foram: bagEarth com 0,618, cubist com 0,634, gaussprRadial com 0,595, gbm com 0,593, icr com 0,428, lm com 0,544, pls com 0,425, rf com 0,621 e svmPoly com 0,578.

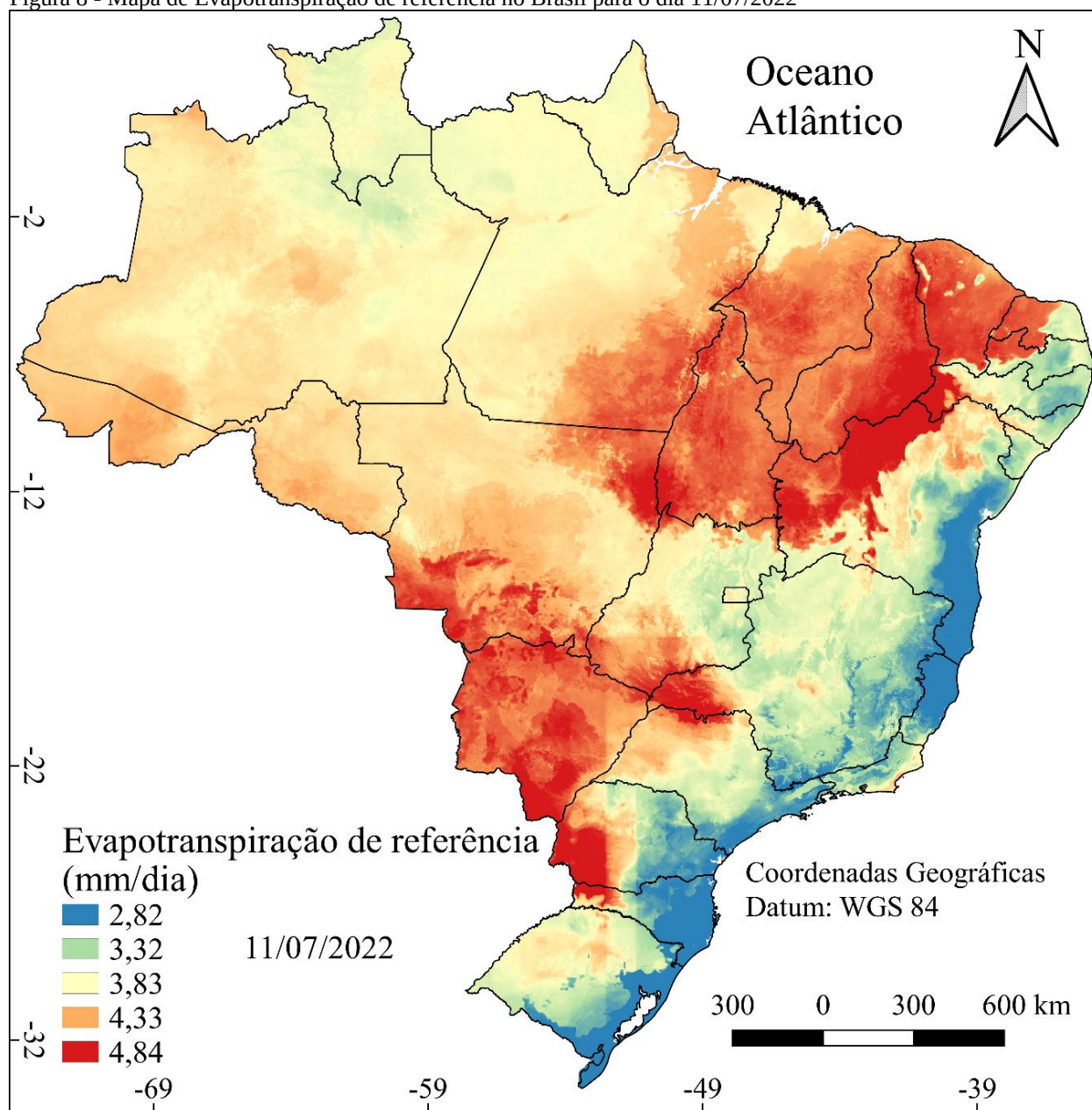
Entre os modelos testados, o Cubist apresentou o melhor desempenho, seguido de perto pelo Random Forest (rf). Esses dois modelos são amplamente utilizados para a modelagem de fatores hidrológicos. Ao avaliar os quatro índices utilizados na seleção dos modelos (MAPE, nRMSE, D e NSE), o Cubist obteve melhor desempenho em dois deles, o que reforça a sua superioridade em relação aos demais modelos avaliados.

2.3.3 Determinação da ETo

Após baixar os dados horários das estações meteorológicas do INMET, foram realizados os cálculos de mínima, máxima, média e soma para cada variável meteorológica, a fim de obter os dados diários. Foram obtidos os valores mínimos de temperatura e umidade, bem como os valores máximos de temperatura e umidade. A temperatura média, velocidade do vento, pressão atmosférica e radiação foram obtidos pela média diária. A evapotranspiração de referência (ET₀) foi calculada para cada estação usando o método FAO 56.

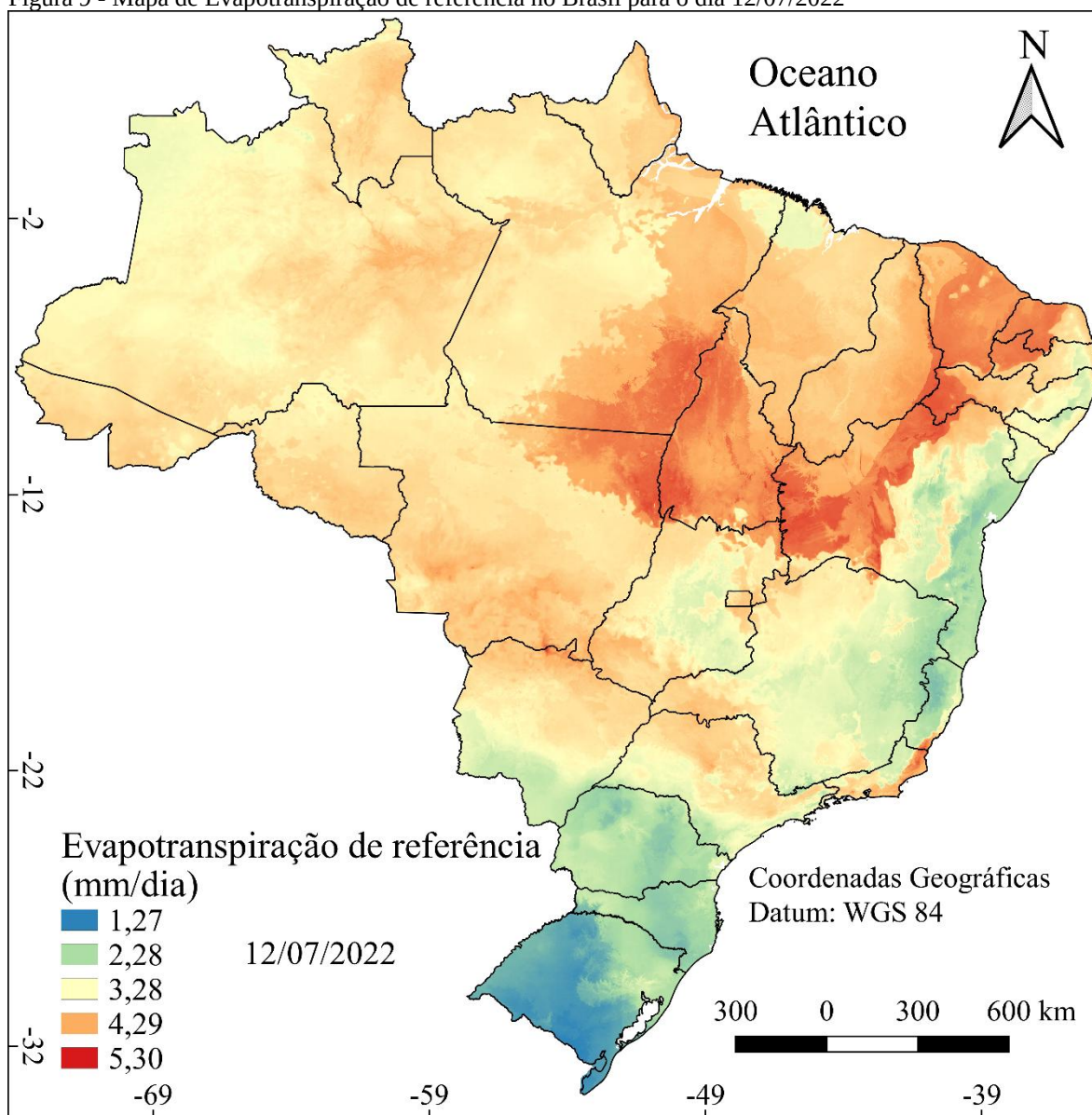
Em seguida, o modelo Cubist foi treinado e aplicado aos mapas das covariáveis. Como exemplo, as Figura 8 e 9 apresentam a evapotranspiração de referência para todo o Brasil nos dias 11 e 12 de julho de 2022, respectivamente.

Figura 8 - Mapa de Evapotranspiração de referência no Brasil para o dia 11/07/2022



Fonte: O autor

Figura 9 - Mapa de Evapotranspiração de referência no Brasil para o dia 12/07/2022



Fonte: O autor

2.3.4 Utilização dos Dados

O pacote confeccionado neste capítulo, pode ser acessado nas seguintes páginas:
<https://github.com/santoshbdias/MapETtoBR> ou em
https://www.santoshbdias.com.br/trabalhos_publicados/R_pacotes

Para realizar a execução das funções do pacote, pode-se seguir a página do GitHub, ou utilizar os seguintes comandos:

```

install.packages('devtools')
devtools::install_github("santoshbdias/MapEToBR")

#Para limpar todas as variáveis e memória ocupada pelo R
rm(list = ls()); gc(); removeTmpFiles(h=0)

#Instalar/carregar todos os pacotes necessários
if(!require("pacman")) install.packages("pacman");pacman::p_load(
  raster, rgdal, Terra, MapEToBR)

#Base stacks rasters criado pelo código Create_Base_Raster.R

#Rasters de Latitude, Longitude, Elevação WorldClim e Bioclimaticas WorldClim
stacktotal <- raster::stack(dir(path = 'D:/OneDrive/Doutorado/Tese/Base_Dados_BR/bspredTotal',
  pattern = ".tif", full.names = TRUE, recursive = T))

#Rasters de dados mensais do WorldClim: temperatura mínima, temperatura máxima, temperatura média,
precipitação, radiação solar, velocidade do vento, pressão de vapor de água.
stackmonth <- raster::stack(dir(path = 'D:/OneDrive/Doutorado/Tese/Base_Dados_BR/bspredMes',
  pattern = ".tif", full.names = TRUE, recursive = T))

#Rasters diários, como média e desvio padrão da ETo e radiação extraterrestre
stackday <- raster::stack(dir(path = 'D:/OneDrive/Doutorado/Tese/Base_Dados_BR/bspredDay',
  pattern = ".tif", full.names = TRUE, recursive = T))

#Carregar shapefile Brasil
shapflbr <- readOGR(dsn = 'D:/OneDrive/Doutorado/Tese/Base_Dados_BR/Shape_Brasil/Shape_Brasil.shp')

#Modelagem do raster de ETo para o Brasil
#Usar o formato de data = "2022-06-14"
#Modelos do caret
RastEToBR <- pred.raster.EToBR((Sys.Date()-1), stacktotal, stackmonth, stackday, shapflbr, 'cubist')

#RasterE gerado pela função pred.raster.EToBR()
#Latitude e longitude pode ser um conjunto de dados (c()), ou coluna de um data frame df$Latitude e df$Longitude
ETos <- ExtrValCoord(RastEToBR, Latitude, Longitude)

#Outras funções:
#Usar o formato de data = "2022-06-14"
#Dados da ETo do Brasil para a data de ontem
datfram <- ETo_BR(date=Sys.Date()-1)

#Selecionar rasters do dia selecionado e criar raster da radiação extraterrestre para o determinado dia
stackday <- inc.sec.dy.rad.extr(datfram,stackday,shapflbr)

#Conferir rasters selecionados.
plot(stackday)

#Extração dos valores dos rasters para as estações com dados neste dia. Arquivo para fazer treino dos modelos de AI
dftv <- ExtrValRast(datfram,stackday,stackmonth,stacktotal)

```

2.4 DISCUSSÃO

O trabalho atual traz uma metodologia/ferramenta para realização do cálculo da ETo em tempo real, que pode ser realizada por qualquer outro pesquisador. Até o momento não foi publicado nenhum trabalho trazendo metodologias de cálculo da ETo em escalas tão grandes em tempo real.

Inúmeros trabalhos de evapotranspiração diária tem sido desenvolvidos em todo o mundo (AGHELPOUR; NOROOZ-VALASHEDI, 2022; ALTHOFF et al., 2020; CHEN et al., 2020; EL-KENAWY et al., 2022; WANG et al., 2022; WU et al., 2021). Althoff et al. (2020) desenvolveram a ETo diária para todo o Brasil utilizando as estações do INMET, estações automáticas e convencionais, contando com um grande volume de dados observados. Também utilizaram imagens de precipitação do Tropical Rainfall Measuring Mission (TRMM). Os autores concluíram que o modelo cubist obteve os melhores desempenhos para realização da modelagem, mesmo modelo definido como padrão neste trabalho.

Este trabalho apresenta uma nova metodologia/ferramenta para o cálculo da Evapotranspiração de Referência (ETo) em tempo real, uma importante variável climática utilizada na agricultura para o planejamento e manejo dos cultivos. Até o momento, nenhum estudo foi publicado trazendo uma metodologia para o cálculo da ETo em escalas tão grandes e em tempo real, o que torna este trabalho uma contribuição importante para a comunidade científica.

A evapotranspiração diária é um tema amplamente estudado em todo o mundo, como evidenciado por diversos trabalhos (AGHELPOUR; NOROOZ-VALASHEDI, 2022; ALTHOFF et al., 2020; CHEN et al., 2020; EL-KENAWY et al., 2022; WANG et al., 2022; WU et al., 2021). Althoff et al. (2020) desenvolveram um modelo de ETo diária para todo o Brasil utilizando as estações meteorológicas do INMET, estações automáticas e convencionais, além de dados observados de precipitação do Tropical Rainfall Measuring Mission (TRMM). Os autores concluíram que o modelo cubist obteve o melhor desempenho na modelagem da ETo, o que foi considerado como o modelo padrão neste trabalho.

Wang et al. (2022) utilizaram modelos de aprendizado de máquina para modelar a ETo mensal no Paquistão, com base em dados de 22 estações meteorológicas. Os autores concluíram que os modelos baseados em árvores apresentaram um desempenho significativamente superior em relação às redes neurais e às multifunções. No entanto, os autores utilizaram krigagem para

espacializar os dados de ETo, enquanto a metodologia proposta neste trabalho utiliza um algoritmo diferente. Outros autores, como Allies et al. . (2022), propuseram a fusão de várias imagens de satélite para obter um produto de ETo diário para todo o mundo, utilizando um modelo diferente do padrão FAO56 adotado neste trabalho.

A metodologia proposta neste estudo utiliza dados de estações meteorológicas em tempo real e uma técnica de interpolação por meio de krigagem para estimar a ETo em qualquer ponto da região estudada. A krigagem é uma técnica de interpolação espacial que utiliza a variância dos dados para estimar o valor em locais não observados. A técnica é conhecida por ser eficiente na estimativa de variáveis climáticas em áreas com poucas estações meteorológicas, o que pode ajudar em áreas onde a rede de estações é limitada.

A ferramenta de cálculo da ETo em tempo real proposta neste estudo pode ser facilmente utilizada por outros pesquisadores, pois é baseada em algoritmos abertos, além de fornecer resultados precisos e atualizados em intervalos regulares de tempo. Isso pode ajudar no planejamento e manejo da irrigação em áreas agrícolas, aumentando a produtividade dos cultivos e reduzindo o uso desnecessário de água.

Diversas pesquisas têm se concentrado na previsão da ETo (AGHELPOUR; NOROOZ-VALASHEDI, 2022; DIMITRIADOU; NIKOLAKOPOULOS, 2022; EL-KENAWY et al., 2022). No entanto, ainda não foram realizados trabalhos com o intuito de determinar a ETo em tempo real, de modo a produzir um mapa da evapotranspiração de referência para uma região completa no final do dia.

Vários estudos, como o de Altalib, Mahmood e Al-Ogaidi (2021), têm proposto a regionalização da ETo com base em dados históricos. Alguns desses trabalhos utilizaram interpoladores tradicionais, como a krigagem ou a ponderação do inverso da distância, sem levar em consideração a variabilidade regional. Além disso, a maioria desses autores calculou a ETo mensalmente, o que pode não ser útil para certas finalidades, como o manejo da água em bacias hidrográficas ou o gerenciamento da irrigação.

Baratto et al. (2022) utilizaram o modelo Random Forest, um modelo de árvore similar ao Cubist, para especializar valores de ETo obtidos por estações meteorológicas. Eles compararam o método com técnicas tradicionais de espacialização, como a ponderação da distância angular (ADW) e a ponderação da distância inversa (IDW). No entanto, os autores usaram dados históricos e limitados a algumas bacias. O número de estações meteorológicas utilizadas foi uma grande

limitação desse e de outros estudos. Wang et al. (2022) empregaram 22 estações para o Paquistão, enquanto Altalib, Mahmood e Al-Ogaidi (2021) usaram somente 23 estações para todo o Iraque. Até o momento, nenhum estudo propôs um monitoramento em tempo real da ETo para uma região continental como o Brasil.

2.5 CONCLUSÕES

Em resumo, o modelo cubist mostrou-se como a melhor opção para a modelagem da evapotranspiração de referência (ETo) no Brasil, apresentando um desempenho superior em relação aos demais modelos avaliados. As covariáveis selecionadas para compor o modelo foram capazes de explicar de forma significativa a ETo, e dentre elas, a média da ETo diária (md), radiação solar (srd), longitude (log), latitude (lat), radiação solar extraterrestre (REx), sazonalidade da temperatura (desvio padrão x 100) (b4), isotérmica (BIO2/BIO7) (x 100) (b3), precipitação (prc) e desvio padrão da ETo diária (sd) foram as mais importantes.

Portanto, é possível concluir que os modelos de machine learning (de árvore como o cubist) são ferramentas extremamente úteis e viável para a estimativa da ETo em todo o território brasileiro, sendo o único trabalho com foco na determinação da ETo diária em tempo real. Com base nos resultados obtidos, pode-se inferir que o modelo apresenta uma alta capacidade de previsão e pode ser utilizado em diversas áreas, como agricultura, gestão de recursos hídricos e meio ambiente. É importante ressaltar que, apesar da eficácia do modelo cubist, é fundamental a realização de novos estudos para aprimorar e validar o modelo em diferentes regiões do país.

CAPÍTULO 3 - DETERMINAÇÃO DA EROSIVIDADE DAS CHUVAS NO BRASIL, UMA ABORDAGEM COM MACHINE LEARNING

RESUMO

DIAS, Santos Henrique Brant. **Modelagem de variáveis hídricas por meio do sensoriamento remoto orbital e inteligência artificial**. 2025. 107f. Tese (Doutorado em Agronomia) – Universidade Estadual de Ponta Grossa, Ponta Grossa, 2025.

O solo é um recurso natural vital para a manutenção da vida, no entanto, as atividades humanas aceleraram a taxa de perda de solo por erosão. A erosividade das chuvas é um dos principais causadores de perda de sedimentos e nutrientes em todo o mundo. As equações USLE ou RUSLE usam a erosividade das chuvas para estimar as perdas de solo, mas a obtenção dos dados é complexa e os valores são pontuais, dificultando sua aplicação em outras regiões. Algumas técnicas de espacialização, como krigagem, IDW ou polígonos de Thiessen, têm sido usadas para extrapolar esses dados, mas elas não consideram a variabilidade real do local que afeta o índice de erosividade. O objetivo deste estudo foi criar um mapa de erosividade das chuvas para o Brasil, com uma resolução espacial de 30 segundos de arco ($\sim 1 \text{ km}^2$). Para isso, foram utilizados produtos disponibilizados por outros artigos, técnicas de SIG e modelagem com machine learning. Dos 31 fatores pré-selecionados, 8 foram usados na modelagem em ordem de importância: longitude, radiação solar, precipitação anual (BIO12), precipitação do trimestre mais frio (BIO19), velocidade do vento, precipitação do trimestre mais quente (BIO18) e evapotranspiração de referência anual. O modelo foi treinado e validado 400 vezes, e o Random Forest obteve os melhores indicadores de desempenho. Com as medianas, os índices de desempenho do modelo foram: NSE de 0,5823, RMSE de $1.567,17 \text{ MJ mm ha}^{-1} \text{ h}^{-1} \text{ ano}^{-1}$, MAE de $1.135,90 \text{ MJ mm ha}^{-1} \text{ h}^{-1} \text{ ano}^{-1}$, nRMSE de 58,50%, ME de $-17,76 \text{ MJ mm ha}^{-1} \text{ h}^{-1} \text{ ano}^{-1}$ e índice D de 0,8487. O mapa de erosividade gerado é útil e promissor para obter estimativas precisas da erosividade das chuvas no Brasil, proporcionando melhorias significativas em relação aos produtos já existentes. A modelagem realizada neste trabalho teve um erro menor em comparação aos trabalhos já publicados, o que aumenta a confiabilidade do modelo e sua aplicabilidade em diferentes regiões do país. Portanto, o estudo destaca a importância da modelagem de erosividade das chuvas para a conservação do solo e a sustentabilidade ambiental. A utilização de técnicas de SIG e machine learning permite obter informações mais precisas e detalhadas sobre a erosividade das chuvas, o que pode contribuir para a adoção de medidas efetivas para a prevenção da erosão do solo. Além disso, o modelo desenvolvido neste estudo pode ser uma ferramenta valiosa para apoiar a tomada de decisões em várias áreas, incluindo a agricultura, a construção civil, a gestão de recursos hídricos e a preservação ambiental.

Palavras-chave: Random forest, Sistema de Informação Geográfica, Perdas de solo

3.1 INTRODUÇÃO

De acordo com Löbmann et al. (2022), o solo é um recurso natural fundamental para a sustentabilidade da vida na Terra. No entanto, é considerado um recurso não renovável ou renovável em uma escala temporal muito grande, já que sua formação é extremamente lenta e as taxas de perda têm sido muito superiores à capacidade de renovação (WUEPPER; BORRELLI; FINGER, 2020). Apesar da importância do controle da erosão, a erosão global do solo tem aumentado constantemente (BORRELLI et al., 2017).

As chuvas podem se tornar um fator perturbador, devido à sua força erosiva, que é expressa como erosividade das chuvas (BELLOCCHI; DIODATO, 2020). A erosividade é um dos principais fatores responsáveis pelas perdas de sedimentos e nutrientes em todo o mundo.

A erosividade das chuvas é medida pelo potencial energético das chuvas em gerar perdas de solo por erosão hídrica pluvial (DA SILVA, 2004). A erosividade é um dos fatores (fator R) da Equação Universal de Perda de Solo (USLE - Universal Soil Loss Equation) e de sua versão revisada (RUSLE - Revised Universal Soil Loss Equation) (WISCHMEIER; SMITH, 1978; RENARD, 1997). Ambas as equações são amplamente utilizadas em todo o mundo para estimar a erosão, levando em consideração os diferentes componentes do processo erosivo (DONG et al., 2022), incluindo fatores relacionados à precipitação, características do solo, topografia e uso/ocupação da terra. Essa equação é representada desta forma:

$$A = RKLSCP$$

Onde A é a perda média de solo ($t\ ha^{-1}\ ano^{-1}$), R é a erosividade anual das chuvas ($MJ\ mm\ ha^{-1}\ h^{-1}\ ano^{-1}$), K é a erodibilidade do solo ($t\ ha\ h\ ha^{-1}\ MJ^{-1}\ mm^{-1}$), L é o fator de comprimento da rampa (-), S é o fator da inclinação da rampa (-), C é o fator da cobertura e manejo do solo (-) e P é o fator das práticas conservacionistas auxiliares (-).

A erosividade é definida como a interação entre a energia cinética das gotas de chuva e a superfície do solo (HUDSON, 1971, 1981; WISCHMEIER; SMITH, 1978). Esse fenômeno está relacionado à desagregação e transporte das partículas superficiais do solo, que aumentam proporcionalmente à energia cinética e intensidade da chuva, mantendo as demais variáveis constantes (DA SILVA, 2004).

Para prever os processos erosivos do solo e propor medidas de conservação, é fundamental considerar a erosividade das chuvas como um dos principais parâmetros em modelos de predição

(PANAGOS et al., 2017). É importante destacar que as previsões em larga escala são essenciais, já que a erosão do solo é uma ameaça significativa à sustentabilidade do solo e, conseqüentemente, à segurança da água e dos alimentos.

Entretanto, como apontado pelos autores do método original (WISCHMEIER; SMITH, 1978), o método requer o uso de pluviógrafos para medir o fator R, que é baseado na variação instantânea da precipitação em eventos de chuva. Essa informação é pontual e difícil de obter e processar, o que limita os estudos regionais sobre o poder erosivo dos padrões de chuva (BELLOCCHI; DIODATO, 2020).

No entanto, é importante destacar que os valores pontuais de erosividade não podem ser aplicados em todas as áreas que requerem estimativas das perdas de solo. Para contornar esse problema, alguns autores realizaram a espacialização desses dados com o objetivo de determinar ou estimar os valores de erosividade em locais sem dados pluviográficos (DE MELLO et al., 2020; OLIVEIRA; WENDLAND; NEARING, 2013).

No entanto, essas espacializações não levam em consideração as variações reais do local, como precipitação, relevo, altitude e clima em geral. Como a erosividade da chuva é altamente dependente desses fatores, as interpolações, como krigagem, inverso da potência das distâncias (IDW) e polígonos de Thiessen, não conseguem capturar essas variações locais.

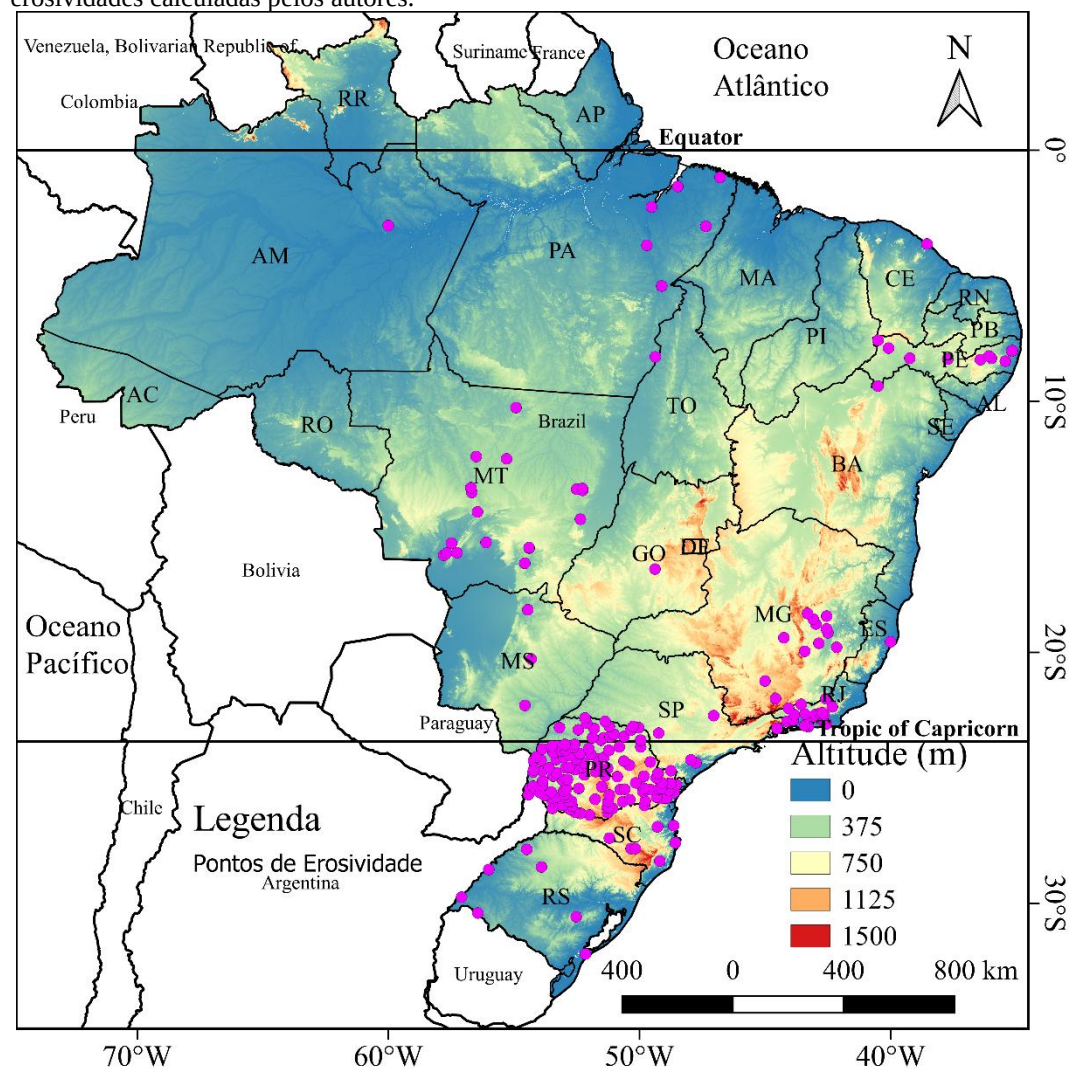
Assim, o objetivo deste estudo foi modelar um mapa de erosividade das chuvas para o Brasil, com uma resolução espacial de 30 segundos de arco ($\sim 1 \text{ km}^2$). Para alcançar esse objetivo, uma combinação de técnicas de SIG (sistema de informação geográfica), modelagem com machine learning e produtos disponibilizados por outros autores foram usados. As melhores covariáveis e modelos foram selecionados para gerar o mapa final.

3.2 MATERIAL E MÉTODOS

3.2.1 Área de Estudo e Coleta de Dados

O Brasil é a área deste estudo, a qual tem proporções continentais e apresenta grande variabilidade em suas características morfológicas e climáticas. Na Figura 10 pode ser observado a sua posição em latitude e longitude, as divisões políticas dos estados e a variabilidade da altitude.

Figura 10 - Localização e altitude da área de interesse. Os pontos demarcados são os locais das erosividades calculadas pelos autores.



Através da Figura 10, pode-se identificar os pontos de erosividade encontrados na pesquisa bibliográfica, que estão destacados em rosa. Foram selecionados 256 valores de erosividade provenientes de três publicações principais. Parte significativa desses dados utilizados no estudo foram compilados por Oliveira, Wendland e Nearing (2013). Os demais valores de erosividade foram estimados no estado do Paraná, conforme apontam as pesquisas já realizadas (RUFINO; BISCAIA; MERTEN, 1993; WALTRICK, 2010).

Importante destacar que todos os dados coletados são provenientes de cálculos de erosividade por pluviográficos e são considerados escassos.

3.2.2 Algoritmos Utilizados nos Modelos de Aprendizagem de Máquinas

Na Tabela 6 são apresentados os algoritmos de *machine learning* utilizados no estudo.

Tabela 6 - Algoritmos de *machine learning* utilizados no estudo

Modelo	Abreviação	Pacote R	Autor
Random Forest	RF	randomForest	Tin Kam Ho (1995)
Linear Regression	LM	Base R	
Cubist	cubist	Cubist	(ADAMS; STERLING, 1992; QUINLAN, 1993)
Partial Least Squares	PLS	pls	(GELADI; KOWALSKI, 1986; HELAND, 1988)
Gradient Boosting Machine	GBM	gbm, plyr	(FRIEDMAN, 2001, 2002)
Independent Component Regression	ICR	fastICA	(WESTAD, 2005)
Bagged MARS	BagEarth	earth	(HASTIE; WRAPPER, 2021)
Gaussian Process with Radial Basis Function Kernel	gaussprRadial	kernlab	(KARATZOGLOU et al., 2004)
Support Vector Machines with Polynomial Kernel	svmPoly	kernlab	(KARATZOGLOU et al., 2004)

Fonte: O autor

3.2.3 Covariáveis Utilizadas na Modelagem

Existem dois tipos de modelagem comumente utilizados em machine learning: classificação e regressão. A modelagem de classificação é recomendada para variáveis qualitativas, enquanto a de regressão é adequada para variáveis quantitativas. Além disso, para realizar essas modelagens, são necessárias pelo menos duas covariáveis para serem usadas como base dos modelos. Neste trabalho, trinta (30) covariáveis foram pré-selecionadas para alimentar os modelos, conforme apresentado na Tabela 7.

Tabela 7 - Covariáveis pré-selecionadas para modelagem da erosividade das chuvas

Abreviação	Descrição	Unidade	Autor
Bio1	Temperatura média anual	°C	Fick and Hijmans (2017)
Bio2	Faixa diurna média (média do mês (temperatura máxima - temperatura mínima))	°C	Fick and Hijmans (2017)
Bio3	Isotérmica (BIO2 / BIO7) ($\times 100$)	°C	Fick and Hijmans (2017)
Bio4	Sazonalidade da temperatura (desvio padrão $\times 100$)	°C	Fick and Hijmans (2017)
Bio5	Temperatura máxima do mês mais quente	°C	Fick and Hijmans (2017)
Bio6	Temperatura mínima do mês mais frio	°C	Fick and Hijmans (2017)
Bio7	Faixa anual de temperatura (BIO5-BIO6)	°C	Fick and Hijmans (2017)
Bio8	Temperatura média do quarto mais úmido	°C	Fick and Hijmans (2017)
Bio9	Temperatura média do quarto mais seco	°C	Fick and Hijmans (2017)
Bio10	Temperatura média do trimestre mais quente	°C	Fick and Hijmans (2017)
Bio11	Temperatura média do quarto mais frio	°C	Fick and Hijmans (2017)
Bio12	Precipitação anual	mm	Fick and Hijmans (2017)
Bio13	Precipitação do mês mais chuvoso	mm	Fick and Hijmans (2017)
Bio14	Precipitação do mês mais seco	mm	Fick and Hijmans (2017)
Bio15	Sazonalidade da precipitação (coeficiente de variação)	mm	Fick and Hijmans (2017)
Bio16	Precipitação do trimestre mais chuvoso	mm	Fick and Hijmans (2017)
Bio17	Precipitação do trimestre mais seco	mm	Fick and Hijmans (2017)
Bio18	Precipitação do trimestre mais quente	mm	Fick and Hijmans (2017)
Bio19	Precipitação do trimestre mais frio	mm	Fick and Hijmans (2017)
Altitude	Altitude	m	Fick and Hijmans (2017)
tmin	Temperatura mínima	°C	Fick and Hijmans (2017)
tmax	Temperatura máxima	°C	Fick and Hijmans (2017)
tavg	Temperatura média	°C	Fick and Hijmans (2017)
prec	Precipitação	mm	Fick and Hijmans (2017)
srad	Radiação solar	$\text{kJ m}^{-2} \text{dia}^{-1}$	Fick and Hijmans (2017)
wind	Velocidade do vento	M s^{-1}	Fick and Hijmans (2017)
vapr	Pressão de vapor de água	kPa	Fick and Hijmans (2017)
Lat	Valor numérico da latitude no pixel	°	Autoral
Long	Valor numérico da longitude no pixel	°	Autoral
Etobr	Normal climatológica da evapotranspiração de referência	mm	Dias et al. (2021)

Fonte: O autor

As variáveis utilizadas como covariáveis neste estudo foram obtidas a partir do conjunto de estações meteorológicas da Organização Mundial de Meteorologia e estão disponíveis no banco de dados WorldClim (FICK; HIJMANS, 2017). Os rasters de longitude e latitude foram derivados do raster de altitude do WorldClim. A variável EToBR, por sua vez, foi obtida por meio da correção

do produto MOD16, que fornece a evapotranspiração potencial, transformando-a em evapotranspiração de referência (DIAS et al., 2021).

3.2.4 Seleção das Covariáveis

Para selecionar as covariáveis mais relevantes nos modelos, utilizou-se a função `rfe` do pacote `Caret` do software R (R CORE TEAM, 2024; WING et al., 2017), que emprega o método Recursive Feature Elimination (RFE) (SINGH; MAHAJAN; MANSOTRA, 2022). Esse método utiliza o algoritmo random forest para ranquear as variáveis mais promissoras. Foram realizados testes utilizando entre uma (1) a trinta (30) das variáveis, onde o algoritmo foi treinado e retornou o R^2 e/ou RMSE para cada treino. O procedimento foi realizado cinco (5) vezes de forma aleatória, separando os dados em treino e validação.

3.2.5 Treino e Avaliação dos Modelos

Os modelos foram treinados e validados utilizando os pacotes `labgeo` e `Caret` do software R (FERNANDES FILHO, 2020; R CORE TEAM, 2024). Após a seleção das covariáveis mais importantes para o modelo, foram realizados 400 treinos e validações. Em cada treino, 85% dos dados foram usados para o treinamento e 15% para a validação, de forma aleatória, utilizando o comando `set.seed( )` para garantir a aleatoriedade. Para executar os algoritmos, utilizou-se o comando `run_models()` para todos os algoritmos mencionados anteriormente, e em seguida foram realizados testes de validação usando o comando `run_models_performance()`.

3.2.6 Métricas de Avaliação dos Algoritmos/Modelos

3.2.6.1 Estatística descritiva

A estatística descritiva consiste em técnicas cujo objetivo é organizar, apresentar, descrever, analisar, resumir, interpretar e explorar dados numéricos, permitindo adquirir um senso crítico sobre eles. Entre as medidas de tendência central mais utilizadas, destacam-se a média aritmética e a mediana. A comparação da média com a mediana fornece uma medida da assimetria

da distribuição. Entretanto, a mediana é considerada a medida de tendência central mais realista de um conjunto de dados, uma vez que valores de outliers podem mascarar os verdadeiros valores de tendência central.

3.2.6.2 *Índices de desempenho*

Foram utilizados nove índices estatísticos para avaliar o desempenho dos modelos durante as fases de calibração e validação, conforme apresentados na Tabela 8. Para uma melhor visualização dos resultados, foram confeccionados boxplots dos índices propostos. Entre os índices de avaliação estão: Erro médio quadrático (RMSE); normalização do RMSE ou (nRMSE); coeficiente de determinação (r^2); eficiência de Nash-Sutcliffe (NSE); erro médio (ME); índice de concordância Willmott (WI). Tais equações têm sido amplamente utilizadas em diversos trabalhos para avaliação de modelos (ASHRAFIAN et al., 2020; LI et al., 2020; MOHAMMED et al., 2020; MOUSAVI et al., 2020; YASEEN et al., 2020; ZAKHROUF et al., 2020).

Tabela 8 - Índices de avaliação de desempenho

Índices de avaliação	Definição
Erro médio quadrático	$RMSE = \sqrt{\frac{\sum_{i=1}^n (O_i - P_i)^2}{n}}$
Normalização do RMSE	$nRMSE = 100 \times \frac{\sqrt{\frac{\sum_{i=1}^n (O_i - P_i)^2}{n}}}{\bar{O}}$
Eficiência de Nash-Sutcliffe	$NSE = \left(1 - \frac{\sum_{i=1}^n (O_i - P_i)^2}{\sum_{i=1}^n (O_i - \bar{O})^2} \right) \times 100$
Erro médio absoluto	$MAE = \frac{1}{n} \sum_{i=1}^n P_i - O_i $
Erro médio	$ME = \frac{1}{n} \sum_{i=1}^n (P_i - O_i)$
Índice de concordância Willmott	$WI = 1 - \frac{\sum_{i=1}^n (O_i - P_i)^2}{\sum_{i=1}^n (P_i - \bar{O} + O_i - \bar{O})^2}$
Coefficiente de determinação	$r^2 = \frac{(\sum_{i=1}^n (P_i - \bar{P})(O_i - \bar{O}))^2}{(\sum_{i=1}^n (P_i - \bar{P})^2)(\sum_{i=1}^n (O_i - \bar{O})^2)}$
Porcentagem absoluta média do erro	$MAPE = \frac{1}{n} \times \sum_{i=1}^n \left \frac{O_i - P_i}{O_i} \right $
Erro de porcentagem média	$MPE = \frac{100}{n} \times \sum_{i=1}^n \left(\frac{O_i - P_i}{O_i} \right)$

Fonte: O autor

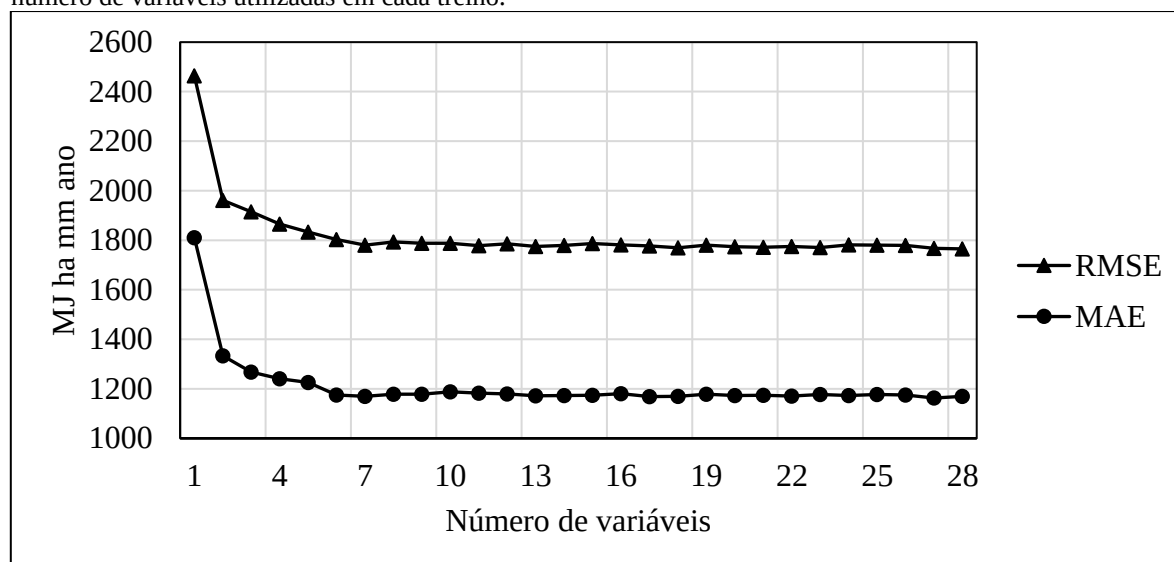
*Em que: P_i - valor predito pelo modelo; O_i - valor observado; n - número de dados, $\bar{O}$ - valor médio observado; $\bar{P}$ - valor médio predito pelo modelo.

3.3 RESULTADOS

3.3.1 Seleção das Covariáveis

Para garantir que o modelo não seja prejudicado por um grande número de covariáveis e reduzir a possibilidade de ruídos, é importante limitar o número de covariáveis. A função Recursive feature elimination, presente no pacote Caret, foi utilizada para esse fim (R CORE TEAM, 2024; WING et al., 2017). Na Figura 11, é possível visualizar o ajuste do modelo para cada incremento de covariáveis durante a utilização dessa função.

Figura 11 - Variação do ajuste (MAE - Mean Absolute Error e RMSE - Root Mean Squared Error) de acordo com o número de variáveis utilizadas em cada treino.



Fonte: O autor

Analisando a Figura 11, percebe-se que o erro do modelo diminui consideravelmente a partir de cinco (5) covariáveis, e depois se estabiliza. Entretanto, o menor erro é alcançado utilizando oito (8) covariáveis, sem apresentar ganho adicional de ajuste ao aumentar o número de covariáveis para além de sete e até trinta.

Quanto à importância das variáveis na determinação da erosividade das chuvas, as mais relevantes em ordem decrescente foram: I) Bio12 - precipitação anual; II) longitude; III) Bio19 - precipitação do trimestre mais frio; IV) Bio3 - Isotérmica (BIO2 / BIO7) ($\times 100$); V) Bio_18 - precipitação do trimestre mais quente; VI) radiação solar ($\text{kJ/m}^2\text{dia}$); VII) Wind - velocidade do vento (m/s); e VIII) Bio15 - Sazonalidade da precipitação (coeficiente de variação).

Na etapa de seleção das variáveis, utilizou-se a função Recursive feature elimination do pacote Caret, que testou a importância de cada variável no modelo cinco (5) vezes. O resultado foi o ranqueamento de oito (8) variáveis mais importantes, como apresentado na Tabela 9.

Tabela 9 - Seleção de covariáveis na modelagem. *descrição das covariáveis.

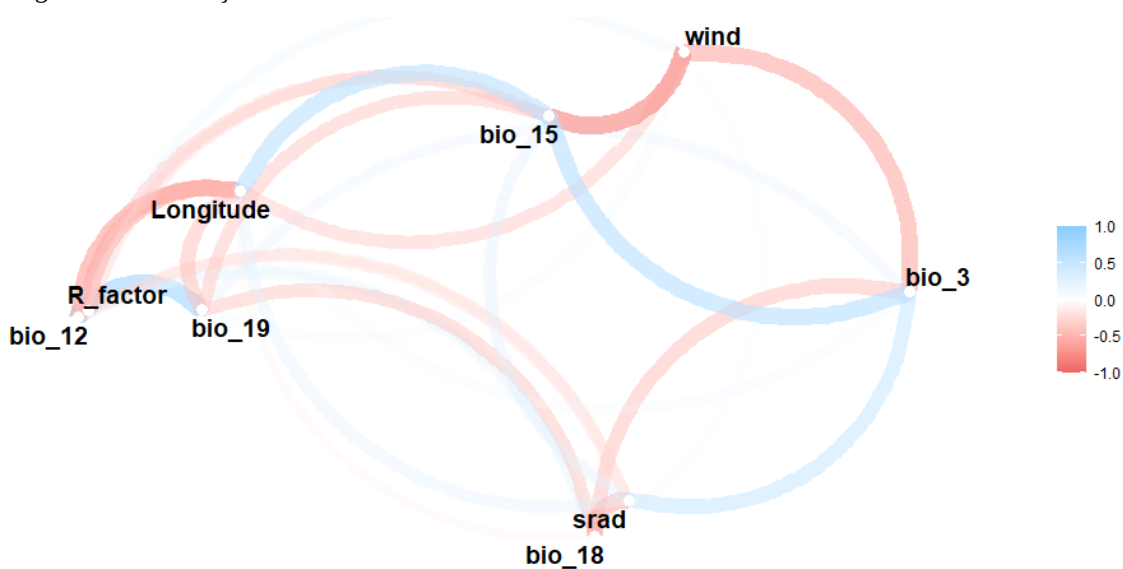
N	Covariável	Var_sel	n	over
I	Bio_12	7	5	24,2830324
II	Longitude	7	5	20,097053
III	Bio_19	7	5	13,9389298
IV	Bio_3	7	3	12,0657095
V	Bio_18	7	2	11,95446
VI	srad	7	3	11,944843
VII	wind	7	3	11,3296214
VIII	Bio_15	7	4	11,3194389
	Latitude	7	2	10,8046043
	Bio_4	7	1	10,3279608
	Bio_17	7	2	10,2232239

Fonte: O autor

* Bio12 - precipitação anual; Bio19 - precipitação do trimestre mais frio; Bio3 - Isotérmica (BIO2 / BIO7) ($\times 100$); Bio18 - Precipitação do trimestre mais quente; srad - radiação solar; wind - velocidade do vento; Bio15 - Sazonalidade da precipitação (coeficiente de variação); Bio4 - Sazonalidade da temperatura (desvio padrão $\times 100$); Bio17 - Precipitação do trimestre mais seco.

Apesar de o modelo ter o melhor desempenho utilizando 8 covariáveis, em diferentes seleções de 7 covariáveis (realizadas 5 vezes). Conforme destacado por Riquetti et al. (2020), a seleção de covariáveis significativas que não apresentam covariância ou correlação é de grande importância para a qualidade estatística do modelo final. Observando a Figura 12, podemos verificar que isso é bastante relevante para o modelo atual, pois mesmo que as covariáveis não apresentem correlação com o fator da modelagem, elas podem ser utilizadas para corrigir a variabilidade de outras covariáveis utilizadas no modelo.

Figura 12 - Correlação de Pearson das variáveis selecionadas.



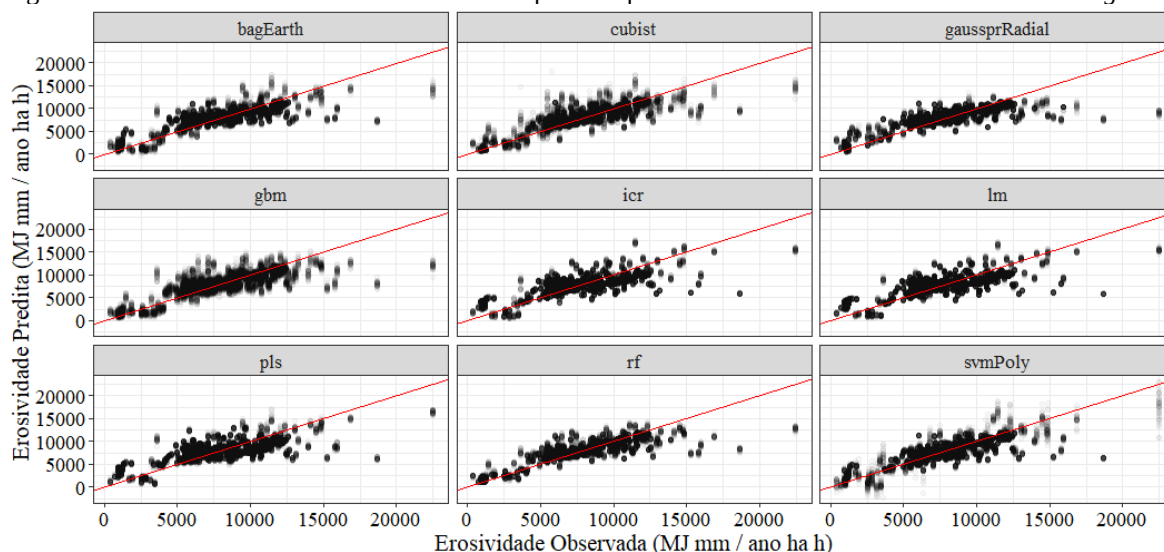
Fonte: O autor

A partir da Figura 12, é possível notar que algumas covariáveis, como Bio_12, Bio_19 e Longitude, possuem uma forte correlação com o fator R de erosividade das chuvas. Por outro lado, outras variáveis, como a velocidade do vento, não foi observado na Figura 12 correlação positiva ou negativa com a erosividade das chuvas. No entanto, mesmo essas variáveis sem correlação direta com o fator R podem ser úteis para corrigir as demais variáveis do modelo. É importante destacar a importância da utilização de covariáveis que sejam significativas e não apresentem covariância ou correlação, conforme destacado por Riquetti et al. (2020). Isso contribui para a qualidade estatística final do modelo.

3.3.2 Seleção do Modelo na Estimativa da Erosividade

A aplicação de modelos de aprendizado de máquina para a interpolação dos valores calculados de erosividade é fundamental para produzir resultados precisos em larga escala. Estudos como o de Vantas, Sidiropoulos e Loukas (2020), que apresentam a erosividade anual máxima, são relevantes para a comunidade, mas muitas vezes realizam interpolações por métodos tradicionais que não representam com precisão as características reais do local e dos valores calculados, bem como na disponibilização do produto gerado. A Figura 13 apresenta os valores preditos e observados durante o treinamento dos diferentes modelos de aprendizado de máquina utilizados neste estudo.

Figura 13 - Valores observados versus os valores preditos pelos diferentes modelos de *machine learning*.



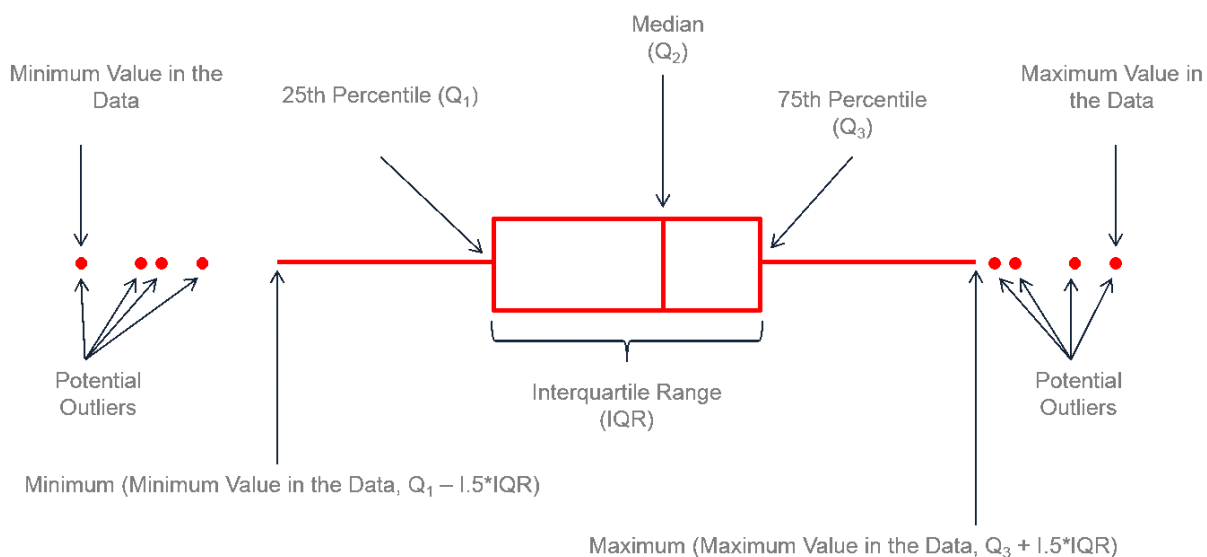
Fonte: O autor

Inicialmente, a aplicação de modelos de machine learning para a interpolação de valores de erosividade é crucial para fornecer produtos eficientes em grande escala. Apesar disso, muitas vezes, trabalhos como o de Lee et al. (2022), que apresentam a erosividade anual máxima, acabam usando métodos tradicionais de interpolação, que não representam as características reais do local ou dos valores calculados, nem na disponibilização do produto gerado. Na Figura 13, pode-se verificar os valores preditos e observados no treino dos diferentes modelos de machine learning utilizados neste trabalho.

Não há, à primeira vista, valores discrepantes na Figura 13. Contudo, os modelos pls e icr superestimam valores observados próximos a 5.000 MJ mm/ano ha h. Para selecionar o melhor modelo, foram utilizados os boxplots dos índices, que incluíram o erro médio - ME (MJ mm ano⁻¹ ha⁻¹ h⁻¹), o erro absoluto médio - MAE (MJ mm ano⁻¹ ha⁻¹ h⁻¹), a raiz do erro quadrático normalizado (nRMSE) (%), o índice de eficiência de Nash-Sutcliffe - NSE (adimensional), a raiz do erro quadrático - RMSE (MJ mm ano⁻¹ ha⁻¹ h⁻¹) e o índice de concordância - D (adimensional).

No entanto, o gráfico do tipo boxplot apresenta uma grande desvantagem na visualização, que são os valores de outliers. Os outliers são valores extremos que não representam o conjunto de dados analisados. Neste artigo, os valores de outlier serão retirados para selecionar o modelo com os maiores níveis de confiança. Para retirar os valores de outlier, consideraram-se valores menores que $q1 - 1,5 * (q3 - q1)$ e valores maiores que $q3 + 1,5 * (q3 - q1)$, sendo $q1$ o valor que separa o primeiro quartil e $q3$ o valor que separa o terceiro quartil, representado na Figura 14 a seguir:

Figura 14 - Demonstração de um boxplot e obtenção dos outliers.

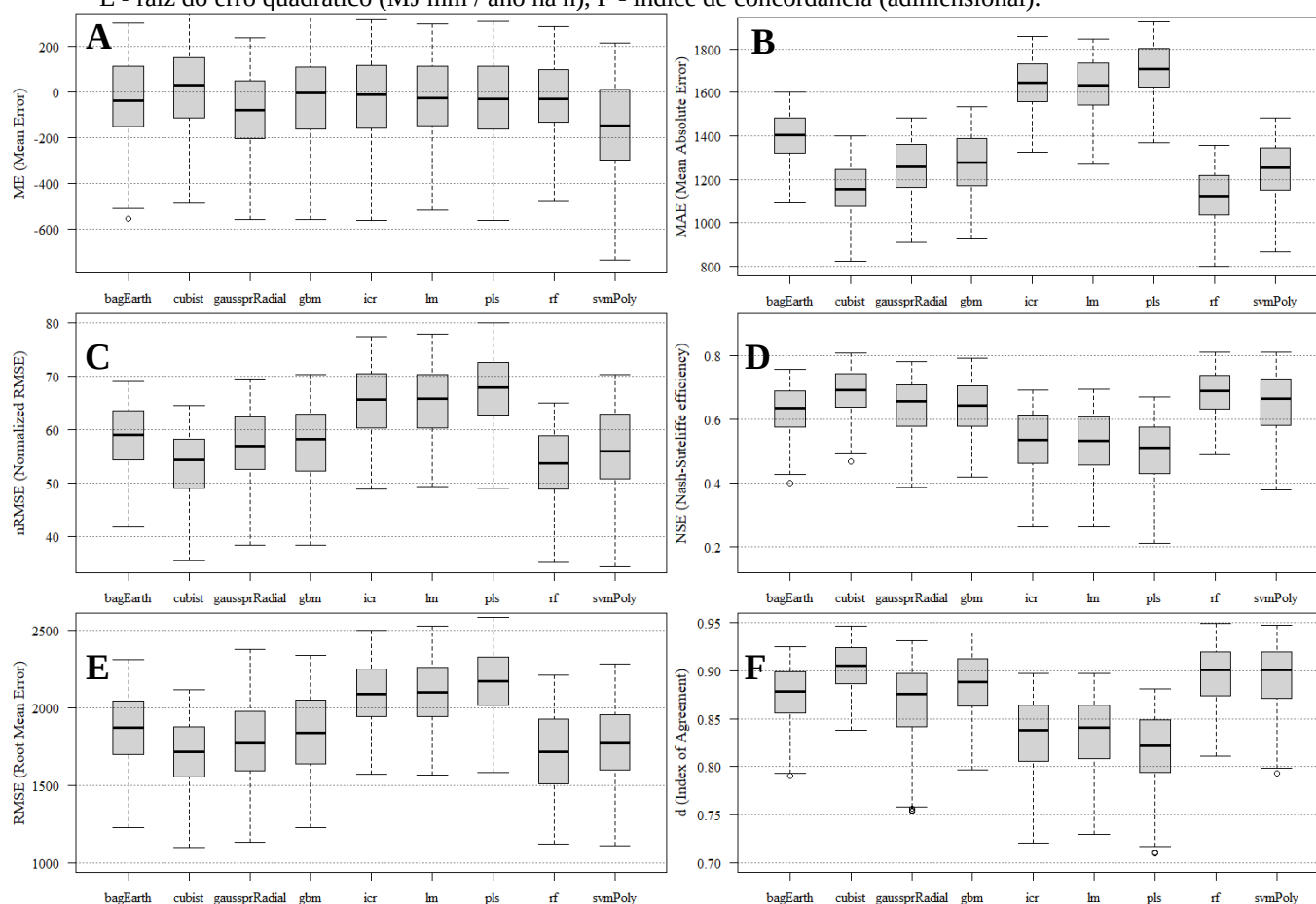


Fonte: O autor

Após a remoção dos valores discrepantes positivos e negativos, foram criados os boxplots, que podem ser visualizados na Figura 15. Apesar do esquema utilizado para a retirada dos outliers, ainda há alguns valores discrepantes que podem ser observados na figura abaixo.

http://127.0.0.1:26951/graphics/plot_zoom_png?width=1278&height=553

Figura 15 – Boxplot dos índices utilizados para seleção do modelo. A - erro médio (MJ mm / ano ha h), B - erro absoluto médio, C - raiz do erro quadrático normalizado (%), D - índice de eficiência de Nash-Sutcliffe (adimensional), E - raiz do erro quadrático (MJ mm / ano ha h), F - índice de concordância (adimensional).



Fonte: O autor

A partir da análise dos boxplots dos índices utilizados pelos diferentes modelos, a Figura 15 mostra que houve um comportamento bastante semelhante entre eles. Para avaliar qual o melhor modelo, a mediana foi utilizada, visto que é um dos melhores índices de medida central. Conforme os resultados obtidos, os valores de mediana do índice NSE - eficiência de Nash-Sutcliffe foram 0,692 para o cubist, 0,643 para gbm, 0,534 para icr, 0,533 para lm, 0,510 para pls, 0,634 para bagEarth, 0,657 para gaussprRadial, 0,664 para svmPoly e 0,687 para rf. Os modelos com melhor desempenho neste índice foram o rf, o svmPoly e o cubist.

Os valores de mediana do índice RMSE - raiz média do erro ao quadrado foram 1.718 para o cubist, 1.840 para gbm, 2.090 para icr, 2.103 para lm, 2.174 para pls, 1.871 para bagEarth, 1.770 para gaussprRadial, 1.773 para svmPoly e 1.715 para rf. Neste caso, os modelos com melhor desempenho foram o rf, cubist e gaussprRadial juntamente com svmPoly. Quanto ao índice MAE,

os valores de mediana foram 1.156 para o cubist, 1.276 para gbm, 1.645 para icr, 1.634 para lm, 1.710 para pls, 1.259 para gaussprRadial, 1.253 para svmPoly, 1.405 para bagEarth e 1.122 para rf. Os modelos com melhor desempenho neste requisito foram rf, cubist e gbm juntamente com svmPoly.

Em relação ao nRMSE - raiz média do erro ao quadrado normalizado, os valores de mediana foram 54,3% para o cubist, 58,3% para gbm, 65,7% para icr, 65,8% para lm, 68,0% para pls, 59,1% para bagEarth, 57,0% para gaussprRadial, 56,0% para svmPoly e 53,7% para rf. Neste caso, os melhores modelos foram rf, cubist e svmPoly. Por sua vez, os valores de mediana do ME - erro médio foram 29,3 para o cubist, -2,9 para gbm, -11,1 para icr, -27,3 para lm, -30,8 para pls, -78,1 para gaussprRadial, -147,0 para svmPoly, -36,0 para bagEarth e -30,5 para rf. O erro médio não é utilizado para selecionar um determinado modelo, mas é utilizado para observar se o modelo subestima ou superestima o valor da erosividade. No geral, o cubist superestima e os demais subestimam o valor de erosividade.

Já em relação ao índice D - índice de concordância, os valores de mediana foram 0,906 para o cubist, 0,888 para gbm, 0,838 para icr, 0,841 para lm, 0,822 para pls, 0,876 para gaussprRadial, 0,901 para svmPoly, 0,879 para bagEarth e 0,901 para rf. Neste caso, o cubist apresentou o melhor desempenho, seguido pelo rf e pelo svmPoly. Os modelos com maior potencial foram o cubist, rf e svmPoly. Em geral, o modelo cubist superou todos os modelos e foi selecionado para fazer a estimativa da erosividade em todo o Brasil.

Embora o modelo linear (lm) não tenha se destacado entre os modelos com melhor desempenho, é uma opção muito simples e fácil de reproduzir. Os resultados obtidos foram NSE de 0,533, RMSE de 2.103, MAE de 1.634, nRMSE de 65,8% e D de 0,841. A equação de erosividade abaixo apresenta a estimativa da erosividade resultante do modelo lm:

Na fórmula para o cálculo da Erosividade, onde Bio12 - Precipitação anual (mm); Bio18 - Precipitação do trimestre mais quente (mm); Bio19 - Precipitação do trimestre mais frio (mm); Longitude - Valor numérico da longitude no pixel (°); sradi - Radiação solar ($\text{kJ m}^{-2} \text{ dia}^{-1}$); Bio3 - Isotérmica ($\text{BIO2} / \text{BIO7} \times 100$) (°C); Bio15 - Sazonalidade da precipitação (coeficiente de variação) (mm).

$$\text{Erosividade (MJ mm ha}^{-1} \text{ h}^{-1} \text{ ano}^{-1}) = -0,009502 + 3,198 \times (\text{Bio}_{12}) - 150,6 \times (\text{Longitude}) \\ + 3,524 \times (\text{Bio}_{19}) + 22,57 \times (\text{Bio}_{3}) + 5,478 \times (\text{Bio}_{18}) - 0,0007489 \times (\text{srad}) + 7,379 \times (\text{Bio}_{15})$$

3.3.3 Interpolação da Erosividade

Diversos estudos têm sido conduzidos em diferentes partes do mundo para mapear a erosividade do solo. Wischmeier e Smith (1978) publicaram o mapa isoerodente para os Estados Unidos, Leprun (1981) desenvolveu o mapa isoerodente para a região nordeste do Brasil, Bertoni e Lombardi Neto (2005) para o estado de São Paulo (Brasil) e Hu et al. (2000) para a República da Coreia.

O mapa isoerodente é uma fonte importante de informação sobre o potencial de erosão para uma região. Para padronizar informações sobre erosão no Brasil, foram estabelecidas algumas diretrizes, incluindo a erodibilidade do solo, a erosividade das chuvas e outros fatores relacionados à erosão, conforme afirmado por Pereira et al. (1994). Esses autores também destacaram a importância de estudar e avaliar os padrões de distribuição da erosividade das chuvas e os possíveis efeitos sobre sua variabilidade e capacidade de causar erosão do solo.

No entanto, no Brasil, ainda são utilizadas interpolações de valores obtidos com equações lineares, como os mapas publicados por Da Silva (2004). Observa-se claramente na Figura 16 que equações lineares genéricas não são uma boa escolha na estimativa da erosividade. Apesar das limitações da abordagem e técnicas apresentadas pelo autor, ele foi o pioneiro em disponibilizar mapas de erosividade para todo o Brasil.

Para facilitar o entendimento e visualização dos mapas de erosividade, a Tabela 10 foi criada por Foster et al. (1981) e modificada por Carvalho (1994).

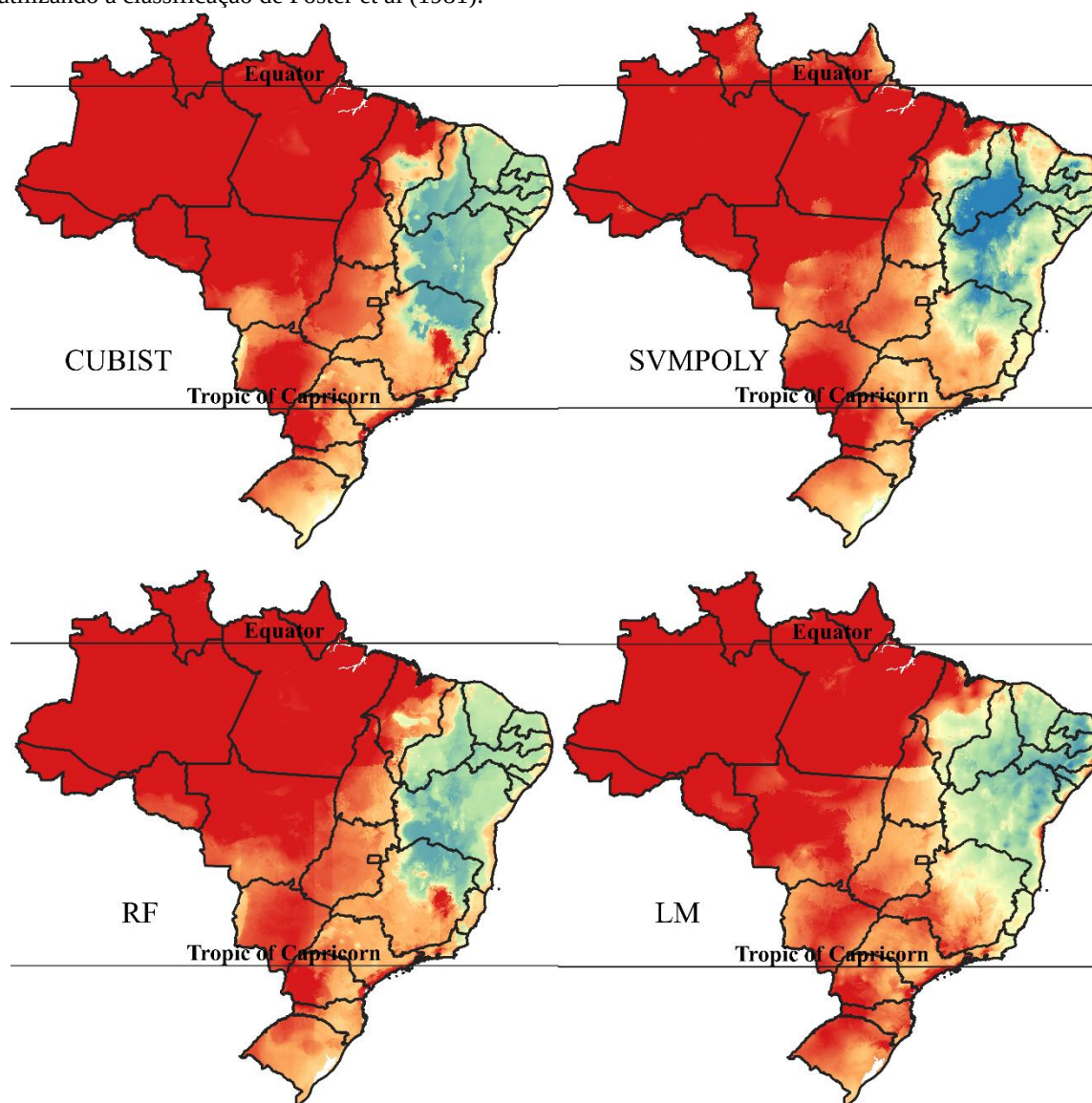
Tabela 10 - Classificação da erosividade anual das chuvas

Erosividade (MJ mm year⁻¹ ha⁻¹ h⁻¹)	Classes de erosividade
R ≤ 2452	Baixa erosividade
2452 < R ≤ 4905	Média erosividade
4905 < R ≤ 7357	Média a forte erosividade
7357 < R ≤ 9810	Forte erosividade
R > 9810	Muito forte erosividade

Fonte: Foster et al. (1981), modificado por Carvalho (1994).

Na Figura 16 pode ser observado a classificação da Tabela 8 nos mapas gerados com os modelos mais promissores.

Figura 16 - Erosividade das chuvas ($\text{MJ mm year}^{-1} \text{ ha}^{-1} \text{ h}^{-1}$) de acordo com os modelos de *machine learning* utilizando a classificação de Foster et al (1981).



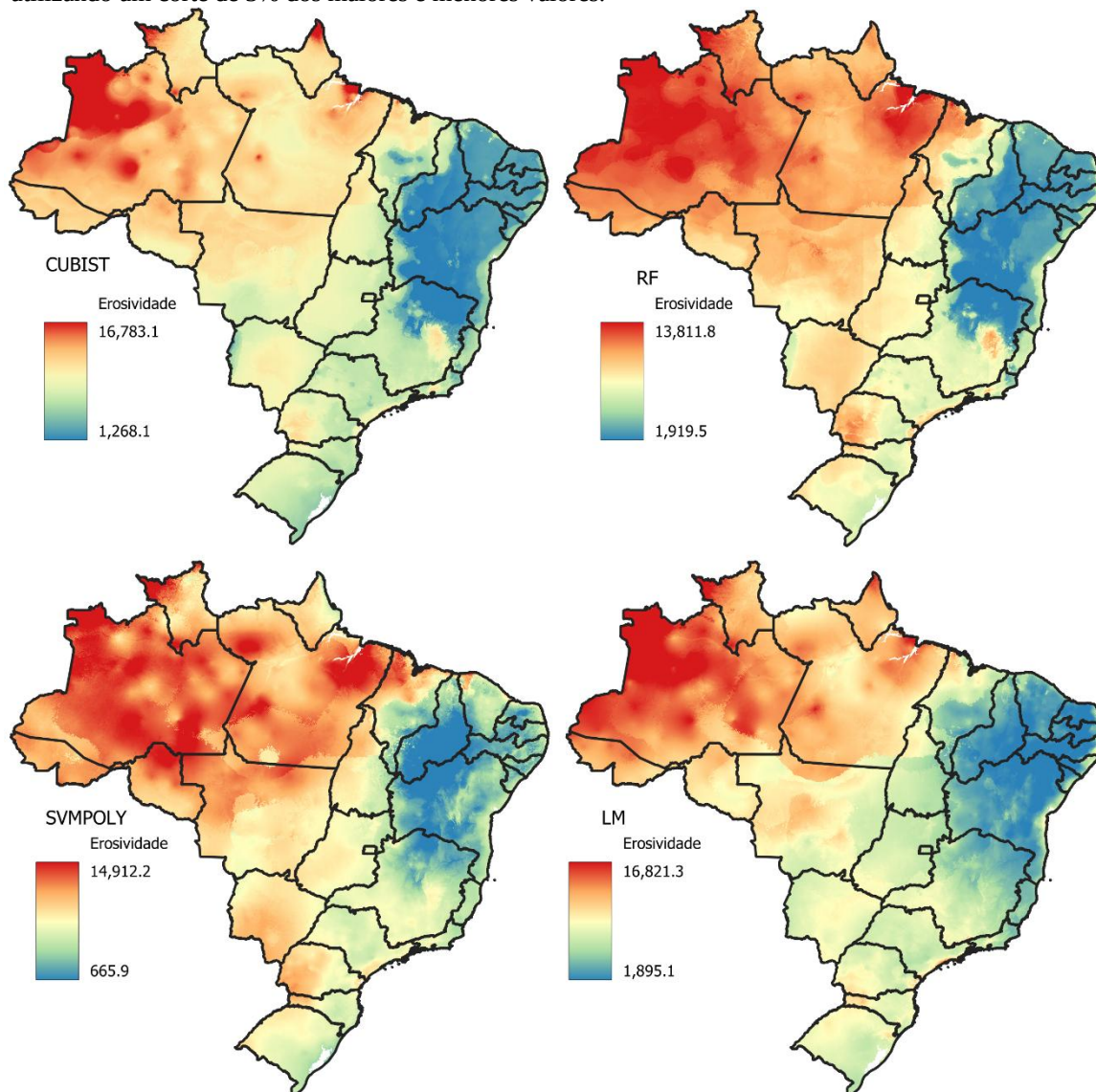
Erosividade ($\text{MJ mm} / \text{ano ha h}$)

- ≤ 2452 - Baixa erosividade
- 2452-4905 - Média erosividade
- 4905-7357 - Média a forte erosividade
- 7357-9810 - Forte erosividade
- >9810 - Muito forte erosividade

Fonte: O autor

Conforme ilustrado na Figura 16, a legenda única é útil para comparar os mapas dos diferentes modelos, mas não permite identificar a escala dos valores de erosividade para cada modelo. Para resolver esse problema, na Figura 17 (abaixo) são apresentados os mapas para os modelos com escalas de legendas individualizadas em quantis.

Figura 17 - Erosividade das chuvas ($\text{MJ mm year}^{-1} \text{ ha}^{-1} \text{ h}^{-1}$) para os diferentes modelos de *machine learning*, utilizando um corte de 3% dos maiores e menores valores.

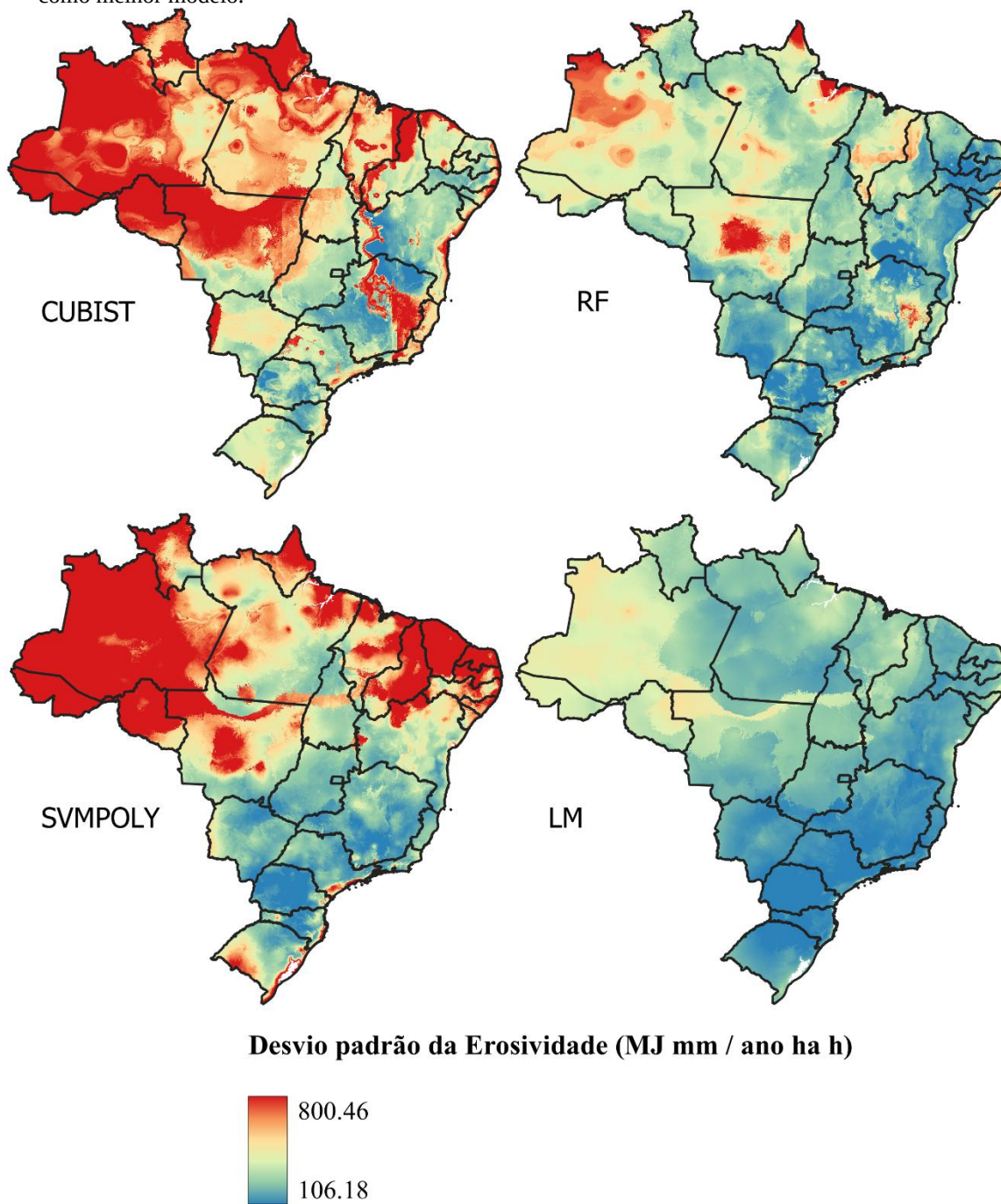


Fonte: O autor

A partir da Figura 17, é possível identificar facilmente as diferenças entre os modelos. Por exemplo, no modelo linear (lm), os valores são claramente superestimados, enquanto no mapa do modelo Cubist é possível observar a superestimativa pelos valores máximos da legenda. Já o

modelo gbm apresentou ruídos na estimativa, principalmente na região amazônica. No entanto, esse comportamento era esperado, considerando a mediana dos índices.

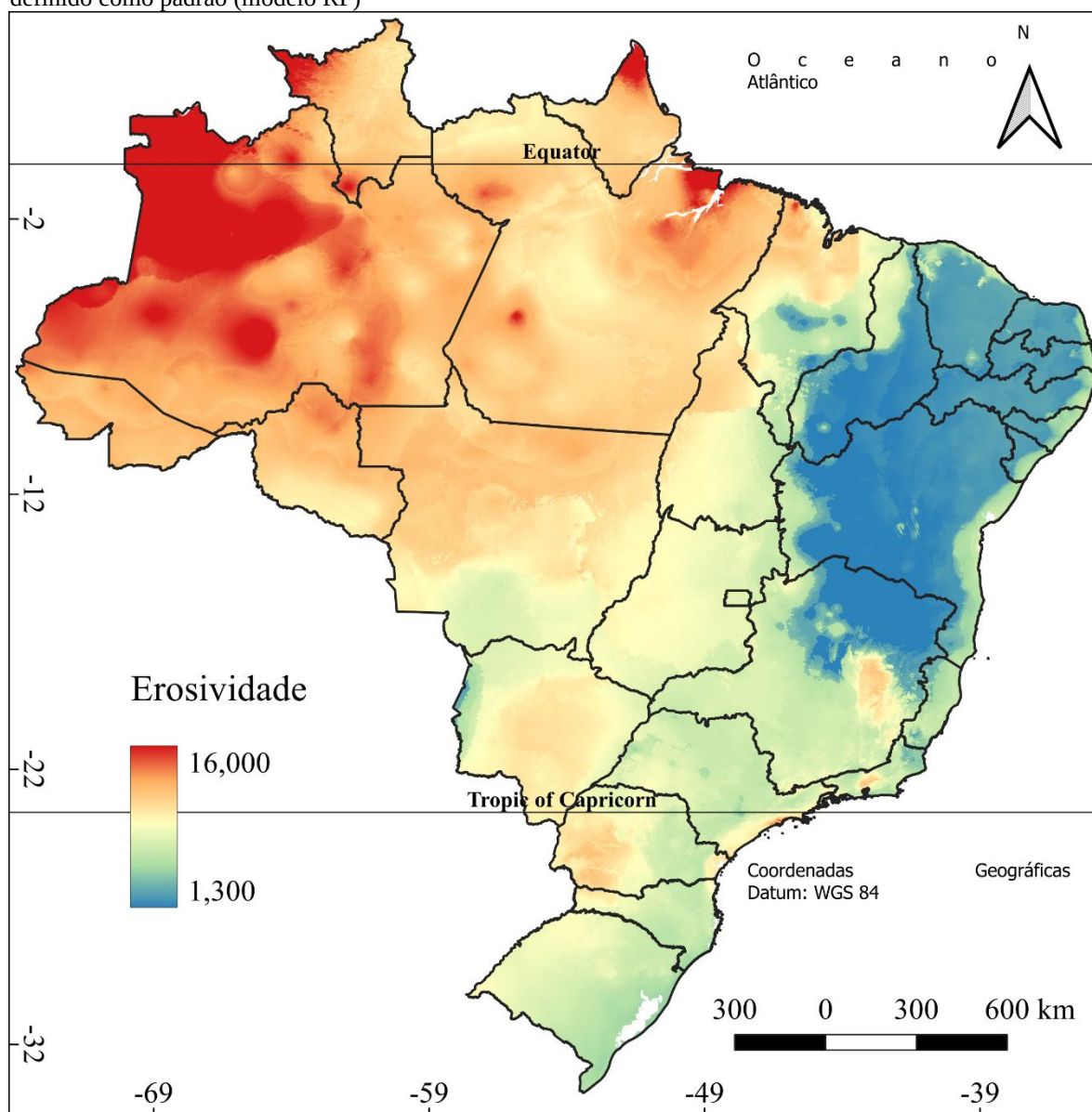
Figura 18 - Desvio padrão da Erosividade das chuvas ($\text{MJ mm year}^{-1} \text{ ha}^{-1} \text{ h}^{-1}$), na mesma escala do modelo definido como melhor modelo.



Fonte: O autor

Para estimar o erro do modelo, utilizou-se o desvio padrão. Para isso, realizou-se a modelagem 100 vezes, usando 90% dos dados para treinamento. A cada treinamento, foram retirados aleatoriamente 10% dos dados, de forma que os dados retirados eram diferentes a cada treinamento. Na Figura 19, pode-se observar a estimativa de erro utilizando o modelo random forest (rf), que foi definido como o melhor modelo para a estimativa da erosividade no Brasil.

Figura 19 - Mapa de erosividade das chuvas (MJ mm year-1 ha-1 h-1) (resolução espacial 30 segundos de arco), definido como padrão (modelo RF)



Fonte: O autor

Inicialmente, observou-se que o mapa de erosividade elaborado apresentou uma distribuição adequada dos valores para todo o território brasileiro.

3.3.4 Utilização dos Dados

Para realizar o download dos dados utilizados para realização da modelagem, ou do mapa final em formato raster.tiff, pode ser acessado as seguintes páginas: <https://zenodo.org/record/3987714> ou em https://www.santoshbdias.com.br/produtos_sig_sr

Para realizar a extração dos valores em uma determinada coordenada, será disponibilizado posteriormente uma plataforma no site descrito acima. O procedimento pode ser feito diretamente pelo software R, utilizando os seguintes comandos:

```
install.packages(c('devtools','Terra'))
devtools::install_github("santoshbdias/MapEToBR")
erosivity <- Terra::rast("./Erosivity.tif")
#Latitude e longitude pode ser um conjunto de dados (c()), ou coluna de um data frame df$Latitude e df$Longitude
Data_eros <- ExtrValCoord(erosivity, Latitude, Longitude)
```

3.4 DISCUSSÃO

A erosividade das chuvas é um tema de grande importância em todo o mundo, e muitos trabalhos vêm sendo realizados para gerar produtos do fator R. Diversos estudos já foram realizados em diferentes partes do globo, como na Europa, onde Boardman e Poesen (2006) e Panagos et al. (2015) desenvolveram mapas de erosividade das chuvas para todo o continente. Na África alguns autores (DIODATO; KNIGHT; BELLOCCHI, 2013; VRIELING; HOEDJES; VAN DER VELDE, 2014) realizaram trabalhos de estimativa da erosividade. Na China, Zhu e Yu (2015) modelaram mapas de erosividade para a China continental, assim como Qin et al. (2016).

Na Ásia Central, Duulatov et al. (2019) confeccionaram mapas de erosividade, enquanto no Japão, Shiono et al. (2013) e Teng et al. (2016) desenvolveram mapas de erosividade para todo o país e para toda a Austrália, respectivamente. Outros autores foram mais ambiciosos e criaram mapas de erosividade das chuvas para todo o mundo (NAIPAL et al., 2015; PANAGOS et al., 2017; YANG et al., 2003).

Hoyos, Waylen e Jaramillo (2005) desenvolveram mapas de erosividade dos Andes colombianos. Mapas feitos também para o Chile Central (BONILLA; VIDAL, 2011) e no Tibet também foi realizado o mesmo trabalho (GU et al., 2020). Klik et al. (2015) desenvolveram mapas de erosividade para a Nova Zelândia. No centro-oeste da Argentina os mapas foram realizados por Rojas e Conde (1985).

A relevância do estudo sobre erosividade do solo tem sido demonstrada por diversos trabalhos recentes, como os citados anteriormente. No entanto, muitos desses estudos tiveram uma abordagem simplista, o que pode gerar maiores erros na estimativa da erosividade em locais onde não há medições diretas.

Atualmente, as técnicas de SIG e machine learning estão bem desenvolvidas e eficientes, permitindo uma análise mais precisa dos dados. Além disso, a programação em software R tem facilitado o acesso aos algoritmos, análises e manipulação de dados SIG.

Panagos et al. (2017) utilizaram uma abordagem de regressão gaussiana de processos e um modelo de machine learning que empregou os mapas do WorldClim (FICK; HIJMANS, 2017) como covariáveis. Nachtergaele et al. (2010) e Yang et al. (2003) usaram chuvas anuais para estimar a erosividade, enquanto todos os outros modelos publicados apresentaram um viés significativo na inclinação, interceptação ou em ambos.

Riquetti et al. (2020) estimaram a erosividade atual e futura para a América do Sul, utilizando um modelo semelhante ao de Panagos et al. (2017), mas ajustando um procedimento de regressão stepwise com o auxílio do Statistical Analyse System (SAS). Diferentemente de Panagos et al. (2017), os autores incluíram variáveis geográficas, como altitude, longitude e latitude, juntamente com o modelo. Vantas, Sidiropoulos e Loukas (2020) utilizaram a mesma abordagem que Riquetti et al. (2020), porém aplicando o modelo random forest para a Grécia, o que aumentou significativamente a importância e eficiência dos mapas.

A relação entre a quantidade de chuva e a erosividade não é direta, e muitos autores desconsideraram esse fato na estimativa da erosividade (NAIPAL et al., 2015), o que pode levar a uma superestimação da erosividade das chuvas. Da Silva (2004) apresentou um trabalho classificando a erosividade em grandes escalas para o Brasil, usando cinco equações para cinco grandes regiões. No entanto, esse método não considera a variabilidade dentro de cada grande região.

Oliveira, Wendland e Nearing (2013) utilizaram erosividades calculadas de várias publicações já realizadas no Brasil. Embora tenham utilizado o método da krigagem na interpolação, considerado um método robusto, ele não leva em consideração a variabilidade existente ao longo dos pontos.

Panagos et al. (2017) afirmam que, embora a erosão hídrica seja a principal causa de degradação do solo em todo o mundo, os padrões globais de erosividade das chuvas permanecem pouco quantificados e as estimativas têm grandes incertezas. Eles destacam a importância das chuvas intensas e eventos extremos para as mudanças climáticas, a economia e a sociedade.

De Souza et al. (2022) realizaram um estudo de predição espaço-temporal da erosividade da chuva usando aprendizado de máquina no sudeste do Brasil. Eles também utilizaram o modelo random forest para determinar a erosividade das chuvas para o estado de Minas Gerais. Os autores destacam a importância de métodos mais precisos para a estimativa da erosividade, especialmente em um país com grande variabilidade climática como o Brasil.

Em suma, os estudos recentes destacam a importância de abordagens mais precisas e sofisticadas para estimar a erosividade do solo, utilizando técnicas avançadas de SIG e machine learning. Essas técnicas têm demonstrado serem eficientes para prever a erosividade em locais onde não há medições diretas, o que pode contribuir significativamente para o manejo e conservação do solo.

É interessante notar que esses trabalhos de mapeamento da erosividade das chuvas são importantes para diversas áreas, como a agricultura e a gestão de recursos hídricos, e ajudam a entender a erosão do solo e seus impactos no ambiente. Esses estudos também permitem a identificação de áreas mais vulneráveis à erosão, auxiliando na tomada de decisões para a preservação dessas áreas e a redução dos impactos negativos da erosão. Pode fornecer informações valiosas para o planejamento de uso da terra e manejo de bacias hidrográficas.

A modelagem da erosividade das chuvas é uma área em constante desenvolvimento, e cada vez mais estudos vêm sendo realizados em diferentes partes do mundo. A disponibilidade desses mapas de erosividade pode ser útil para os tomadores de decisão e profissionais que atuam em áreas relacionadas à gestão ambiental, permitindo a adoção de medidas de preservação e manejo adequado do solo.

3.4.1 Seleção das Covariáveis

A predição da erosividade das chuvas no Brasil está diretamente relacionada com a longitude, sendo esta a variável mais relevante, conforme destacado por Da Silva (2004) e outros autores. O autor mencionado encontrou uma correlação positiva entre a longitude e a erosividade, com um coeficiente de determinação (r^2) de 0,484, indicando que quanto menor a longitude, menor é a erosividade das chuvas. Isso acontece porque a longitude influencia diretamente a intensidade e erosividade das chuvas, tanto na maritimidade quanto na continentalidade. Além disso, a proximidade de corpos hídricos volumosos também pode afetar a intensidade e poder erosivo das chuvas.

Outro fator importante que influencia na intensidade das chuvas é a radiação (Srad), que regula todos os processos atmosféricos da Terra, incluindo os movimentos dos ventos e a evapotranspiração das águas nos mares terrestres. Esses dois fatores são cruciais na intensidade das chuvas convectivas, que ocorrem principalmente no verão e possuem maior poder energético e potencial de erosão, como observado por outros autores (PANAGOS et al., 2017; ZIPSER et al., 2006). Esses autores apontam que a erosividade das chuvas está associada com eventos extremos de chuva, e definem chuvas intensas com base na velocidade vertical convectiva da chuva.

A precipitação do trimestre mais frio (Bio_19) influencia o volume e ritmo de precipitação no inverno, que tende a ser de chuvas mais brandas, menos volumosas e menos erosivas.

Entretanto, devido às proporções continentais da modelagem realizada, as precipitações se comportam de maneira diferente em diferentes regiões do Brasil, e as chuvas de inverno tendem a ser mais erosivas em algumas regiões do que em outras.

A precipitação anual (Bio_12), é o fator mais diretamente relacionado com a erosividade das chuvas, conforme relatado por diversos autores que elaboraram equações de erosividade a partir desses valores. O volume de precipitação anual tem um impacto significativo na erosividade anual, de forma que quanto maior for o volume de precipitação, maior será o poder de erosão das chuvas.

Além disso, a precipitação do trimestre mais quente (Bio_18) é a variável mais correlacionada com a erosividade anual das chuvas. Esse tipo de chuva, conhecido como convectiva ou de verão, é caracterizado pelo alto volume de chuva em um curto período, o que pode levar a uma maior erosão hídrica. De acordo com Diodato et al. (2019), eventos catastróficos causados por tempestades são comuns durante o verão no Mediterrâneo (DIAKAKIS; PRISKOS; SKORDOULIS, 2018; DIODATO et al., 2019; SALVATI et al., 2018). No Brasil, esses eventos também são comuns durante o verão, quando predominam as chuvas convectivas.

No entanto, como pode ser visto na Figura 12, algumas covariáveis não apresentam correlação direta com o fator R de erosividade das chuvas. No entanto, essas variáveis podem ser usadas pelo modelo para corrigir variáveis com correlação direta com a erosividade. Por exemplo, embora o vento não apresente correlação direta com a erosividade, ele tem uma correlação positiva baixa com a radiação e a Bio_19 e uma correlação negativa com a longitude.

Em suma, os fatores listados acima são importantes na predição da erosividade das chuvas no Brasil, e sua influência pode variar de acordo com a região geográfica. Além disso, eventos extremos de chuva que são altamente erosivos são muitas vezes imprevisíveis.

3.4.2 Seleção do Modelo na Estimativa da Erosividade

De acordo com Da Silva (2004), em algumas regiões do Brasil, equações lineares ou exponenciais têm sido utilizadas com eficiência na estimativa da erosividade a partir de dados pluviométricos, como mostram as pesquisas de De OLIVEIRA (1988), LEPRUN (1981) e RUFINO et al. (1993). Por outro lado, para as ilhas de Cabo Verde, Mannaerts e Gabriels (2000) encontraram relações significativas utilizando a equação exponencial, considerando chuvas com intensidade superior a 9 mm.

Em estudo mais recente, De Souza et al. (2022) elegeram o modelo random forest como aquele com melhor desempenho para determinar a erosividade das chuvas no estado de Minas Gerais. Contudo, os autores identificaram índices semelhantes para os modelos random forest, cubist e regressão linear. Apesar disso, o random forest foi o escolhido para determinar a erosividade, devido ao melhor ajuste proporcionado por esse modelo.

A tabela 11 apresentada abaixo exibe as medianas dos diferentes índices selecionados pelos autores.

Tabela 11 - Mediana dos índices estatísticos para os diferentes modelos de *machine learning*.

Modelo	RMSE	NSE	D	nRMSE	MAE	ME
bagEarth	2191,562	0,4250	0,7868	74,75	1505,44	35,6465
cubist	2314,653	0,3962	0,7919	76,60	1486,64	125,6015
gaussprRadial	2210,148	0,4471	0,7572	73,35	1503,92	26,8347
gbm	2183,285	0,4423	0,8041	73,60	1424,13	34,5429
icr	2652,416	0,1789	0,5734	89,35	1913,18	0,0442
Lm	2649,111	0,1752	0,5851	89,55	1911,73	-20,3485
pls	2626,264	0,1906	0,5832	88,70	1909,76	-5,6390
Rf	2053,773	0,5011	0,8113	69,65	1302,78	57,7726
svmPoly	2455,084	0,3107	0,7472	81,90	1623,96	-91,0208

Fonte: O autor

3.4.3 Interpolação da Erosividade

A tendência observada neste estudo, de aumento dos valores anuais de erosividade de leste a oeste, já havia sido verificada em estudos anteriores, como o de Leprun (1981), para a região nordeste do Brasil, o de Da Silva (1999), para o estado do Tocantins, e o de Da Silva (2004), que considerou todo o território brasileiro. Entretanto, apesar da existência dessas correlações, os autores não modelaram essa tendência da mesma forma que foi encontrada neste trabalho.

Nesses estudos anteriores, foram utilizados interpoladores como mapas de Isoerodente, Krigagem e Interpolação de Ponderação por Distância Inversa (IDW), entre outros. Esses métodos são úteis para visualizar a distribuição da erosividade no espaço, mas não levam em conta as tendências geográficas ou as variações sazonais, como as identificadas neste estudo.

Dessa forma, os resultados obtidos neste trabalho destacam a importância de se considerar a tendência geográfica na modelagem da erosividade, o que pode ser útil para a gestão e

conservação dos solos em diferentes regiões. Além disso, os resultados reforçam a necessidade de combinar diferentes técnicas de análise espacial para obter informações mais precisas sobre a distribuição da erosividade no espaço.

3.4.4 Comparação do Mapa Desenvolvido com os já Disponíveis

Embora existam várias estimativas e produtos de erosividade das chuvas em todo o mundo, apenas dois produtos se assemelham ao desenvolvido neste estudo. No entanto, qualquer melhoria nos ajustes justifica a elaboração de um novo produto para a estimativa da erosividade.

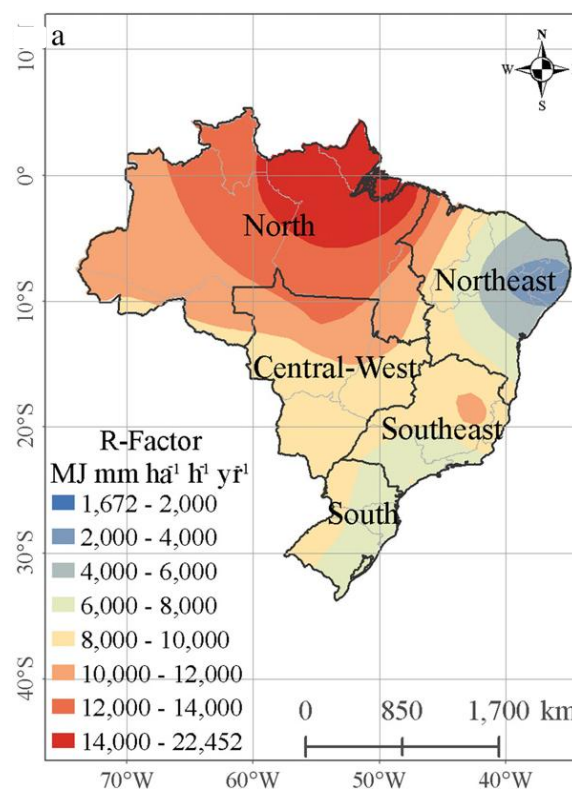
O produto desenvolvido por Panagos et al. (2017) utilizou um banco de dados de erosividade anual extremamente grande para gerar o produto de erosividade global a 30 segundos de arco (~1 km), semelhante ao desenvolvido neste estudo. No entanto, eles utilizaram a regressão de processo gaussiana (GPR) para modelar a erosividade.

Antes da modelagem apresentada neste estudo, existiam dois produtos promissores para a erosividade. Oliveira et al. (2013) elaboraram um mapa de erosividade, utilizando a erosividade de outros autores interpolados por meio da krigagem. Na Figura 20, é possível observar o mapa elaborado pelos autores.

Embora o mapa de Oliveira et al. (2013) forneça informações úteis sobre a erosividade em diferentes regiões, a sua precisão pode ser limitada pela qualidade e disponibilidade dos dados utilizados na interpolação. Além disso, a modelagem de Panagos et al. (2017) pode ser limitada pela resolução espacial do produto, que pode não ser suficiente para a tomada de decisões de manejo de solo em nível local.

Portanto, o produto desenvolvido neste estudo, que combina diferentes variáveis meteorológicas para a estimativa da erosividade em alta resolução espacial, pode fornecer informações mais precisas e úteis para a gestão e conservação do solo em diferentes regiões.

Figura 20 - Erosividade no Brasil utilizando a krigagem..



Fonte: Adaptado de Oliveira et al. (2013)

A figura acima ilustra que a krigagem, embora seja um método de interpolação robusto, não leva em conta a variabilidade real do terreno, nem variáveis como a de precipitação, radiação e longitude. Panagos et al. (2017) estimaram a erosividade global utilizando o kernel gaussiano da função de base radial (RBF) e calcularam valores de erosividade para algumas estações e outros valores publicados por outros autores. No Brasil, utilizaram os dados disponibilizados por Oliveira et al. (2013). Embora o modelo RBF seja robusto, alguns modelos podem apresentar melhores resultados na modelagem. No presente estudo, foram obtidos melhores resultados, devido à maior quantidade de valores de erosividade e ao uso de covariáveis não consideradas por Panagos et al. (2017).

O presente estudo considerou 31 covariáveis, das quais oito tiveram melhor potencial para os modelos. Muitas dessas variáveis não foram consideradas por Panagos et al. (2017), como a longitude, radiação solar, velocidade do vento e a normal da evapotranspiração de referência anual. Essas covariáveis melhoraram a estimativa do modelo atual. Além disso, o número de estações

consideradas neste estudo para o Brasil é consideravelmente maior do que as utilizadas por Panagos et al. (2017).

No entanto, mesmo com o uso das covariáveis deste estudo, o modelo RBF não é tão robusto na modelagem de variáveis climáticas quanto o random forest ou o GBM (Stochastic Gradient Boosting). A Tabela 11 ilustra que certas abordagens em alguns modelos não são tão eficientes na maioria das vezes para a predição de variáveis agrometeorológicas.

Para comparar o algoritmo utilizado por Panagos et al. (2017), foi realizado um treino com os dados do presente estudo e utilizando também o modelo gaussprRadial. A Tabela 11 apresenta os índices de desempenho para o modelo de gauss e os modelos utilizados neste estudo. Os treinos e validações foram realizados 100 vezes da mesma forma que neste estudo, com um número maior de estações com erosividade e oito covariáveis, condições melhores do que as utilizadas por Panagos et al. (2017) no Brasil. Os autores obtiveram RMSE de $1.629 \text{ MJ mm ha}^{-1} \text{ h}^{-1} \text{ ano}^{-1}$ para a América do Sul, valores muito próximos aos encontrados neste treino.

3.4.5 Utilização dos Mapas de Erosividade

Panagos et al. (2017) destacam a importância do uso da erosividade como um produto na análise das mudanças climáticas. As mudanças climáticas têm sido motivo de grande preocupação, especialmente para a comunidade científica e a sociedade em geral. Decisões baseadas em dados meteorológicos precisos e com o mínimo de erros são fundamentais (FREEBAIRN; ZILLMAN, 2002). Mesmo pequenos ajustes nos modelos justificam a criação de novos produtos.

A erosividade é essencialmente a energia das chuvas, sendo que as chuvas mais intensas causam os maiores prejuízos relacionados à erosão dos solos. Esses eventos afetam os processos hidrológicos e erosivos, que são cruciais para a tomada de decisão de práticas de conservação do solo e da água na adaptação da agricultura às mudanças climáticas (RAMON et al., 2017).

No entanto, como explica Diodato et al. (2019), a erosividade por si só não é suficiente para definir a erosão ou as práticas culturais adequadas, já que diferentes formas de relevo podem reagir de maneira diferente aos eventos. A estimativa da erosividade é útil para técnicos e cientistas conservacionistas entenderem o potencial de dano das chuvas nos locais mais críticos, permitindo a implementação de precauções para minimizar a erosão do solo nessas áreas (DA SILVA, 2004).

A erosividade também ajuda os engenheiros civis em projetos de construção de edifícios, estradas, barragens e tubulações (BELLOCCHI; DIODATO, 2020).

Mapas com escalas continentais e alta resolução espacial ajudam a compreender os processos básicos e prever condições variáveis, o que auxilia na compreensão dos custos (financeiros e ambientais) de negligenciar eventos de chuvas intensas (erosividade) em relação aos sistemas de conservação do solo e planejamento do uso da Terra. Esses eventos alteram a estrutura do solo, erodindo as camadas superficiais do solo e levando uma grande parte do carbono orgânico e nutrientes do solo, causando graves consequências ecológicas (BELLOCCHI; DIODATO, 2020).

Conforme apontado por Borrelli et al. (2017), a negligência em relação à erosividade das chuvas pode ser grave, não apenas causando perda de fertilidade e degradação do solo, mas também efeitos como a sedimentação, assoreamento e eutrofização de vias aquáticas ou inundações inadequadas (BOARDMAN; POESEN, 2006).

A principal utilização da erosividade é na previsão da erosão do solo com modelos como o USLE (WISCHMEIER; SMITH, 1978) ou o modelo RUSLE (RENARD, 1997), que podem fornecer estimativas razoavelmente precisas.

3.5 CONCLUSÕES

Os resultados obtidos neste estudo demonstram que é possível utilizar variáveis climáticas e geográficas para construir mapas de erosividade das chuvas com níveis conhecidos de erro e precisão para uma escala nacional e continental. Esses mapas gerados podem ser utilizados em estudos de modelagem e planejamento de uso do solo, contribuindo assim para a preservação ambiental.

A determinação da erosividade anual para o Brasil utilizando regressão linear associada a um conjunto de dados covariáveis foi satisfatória. Contudo, o modelo que proporcionou maiores incrementos de ajuste foi o Random Forest. Isso sugere que modelos mais complexos podem ser mais adequados para estimar a erosividade anual.

As covariáveis utilizadas são explicativas para prever a erosividade anual, mas inserir covariáveis específicas para a estação de verão pode gerar melhores previsões. As covariáveis mais importantes encontradas para prever a erosividade anual foram longitude, radiação solar, precipitação do trimestre mais frio (Bio_19), precipitação anual (Bio_12), velocidade do vento e precipitação do trimestre mais quente (Bio_18). É importante destacar que a inclusão de outras covariáveis pode melhorar ainda mais a precisão do modelo.

O produto gerado por este estudo é extremamente útil e promissor para obtenção da estimativa da erosividade das chuvas no Brasil, uma vez que traz incrementos significativos nos produtos já existentes. A modelagem da erosividade neste trabalho teve um erro menor quando comparado aos trabalhos já publicados. No entanto, é importante ressaltar que o modelo pode ser ajustado com o incremento de novas estações na estimativa da erosividade, o que pode aumentar ainda mais a precisão do modelo e a sua aplicabilidade em outras regiões.

Em suma, este estudo representa um importante avanço na modelagem da erosividade das chuvas no Brasil e pode contribuir para um melhor planejamento do uso do solo e preservação ambiental.

REFERÊNCIAS

- ABDULLAH, S. S.; MALEK, M. A. Empirical Penman-Monteith equation and artificial intelligence techniques in predicting reference evapotranspiration: a review. **International Journal of Water**, v. 10, n. 1, p. 55, 2016.
- ADAMS, A.; STERLING, L. **AI '92: Proceedings of the 5th Australian Joint Conference on Artificial Intelligence**. AI '92. **Anais...** Em: 5TH AUSTRALIAN JOINT CONFERENCE ON ARTIFICIAL INTELLIGENCE. Hobart, Tasmania: WORLD SCIENTIFIC, dez. 1992. Disponível em: <<http://www.worldscientific.com/doi/abs/10.1142/9789814536271>>. Acesso em: 9 ago. 2020
- AGHELPOUR, P.; NOROOZ-VALASHEDI, R. Predicting daily reference evapotranspiration rates in a humid region, comparison of seven various data-based predictor models. **Stochastic Environmental Research and Risk Assessment**, 17 jun. 2022.
- ALLEN, R. G. et al. **Crop evapotranspiration: guidelines for computing crop water requirements**. Rome: Food and Agriculture Organization of the United Nations, 1998.
- ALLIES, A. et al. A remote sensing data fusion method for continuous daily evapotranspiration mapping at kilometric scale in Sahelian areas. **Journal of Hydrology**, v. 607, p. 127504, abr. 2022.
- ALTALIB, A. A.; MAHMOOD, M. T.; AL-OGAIDI, A. A. Mapping Reference Evapotranspiration for Iraq Using FAO-Penman-Monteith Method. **Agricultural Engineering International: CIGR Journal**, v. 23, n. 2, 2021.
- ALTHOFF, D. et al. ETo-Brazil: A daily gridded reference evapotranspiration dataset for Brazil (2000 – 2018). **Water Resources Research**, 19 jun. 2020.
- ANNELI VIRTANEN, S. et al. **Potential of controlled drainage and sub-irrigation to manipulate groundwater table for mitigating acid loadings in Finnish acid sulfate soils**. . Em: 2016 10TH INTERNATIONAL DRAINAGE SYMPOSIUM CONFERENCE, 6-9 SEPTEMBER 2016, MINNEAPOLIS, MINNESOTA. ASABE, 2016. Disponível em: <<http://elibrary.asabe.org/abstract.asp?aid=47401&t=1>>
- ASHRAFIAN, A. et al. Compressive strength of Foamed Cellular Lightweight Concrete simulation: New development of hybrid artificial intelligence model. **Construction and Building Materials**, v. 230, p. 117048, jan. 2020.
- ATASEVER, U. H.; KESIKOGLU, M. H.; OZKAN, C. A New Artificial Intelligence Optimization Method for PCA Based Unsupervised Change Detection of Remote Sensing Image Data. **Neural Network World**, v. 26, n. 2, p. 141–154, 2016.
- BARATTO, P. F. B. et al. Random forest for spatialization of daily evapotranspiration (ET₀) in watersheds in the Atlantic Forest. **Environmental Monitoring and Assessment**, v. 194, n. 6, p. 449, jun. 2022.

BASTIAANSSEN, W. G. M.; STEDUTO, P. The water productivity score (WPS) at global and regional level: Methodology and first results from remote sensing measurements of wheat, rice and maize. **Science of The Total Environment**, v. 575, p. 595–611, jan. 2017.

BELGIU, M.; DRĂGUȚ, L. Random forest in remote sensing: A review of applications and future directions. **ISPRS Journal of Photogrammetry and Remote Sensing**, v. 114, p. 24–31, abr. 2016.

BELLOCCHI, G.; DIODATO, N. Rainfall Erosivity in Soil Erosion Processes. **Water**, v. 12, n. 3, p. 722, 6 mar. 2020.

BENAISSA, A. et al. Transformative Progress in Document Digitization: An In-Depth Exploration of Machine and Deep Learning Models for Character Recognition. **Data and Metadata**, v. 2, p. 174, 27 dez. 2023.

BERTONI, J. **Conservação do solo**. São Paulo: Icone, 2005.

BHATT, K.; SARANGI, R. K.; KAMBLI, M. COMPARATIVE ANALYSIS OF OCEAN COLOR FORECASTING MODELS: A CASE STUDY ON INDIAN SOUTHERN PENINSULA USING SARIMA AND DEEP LEARNING APPROACHES. **International Research Journal of Modernization in Engineering Technology and Science**, 8 dez. 2023.

BOARDMAN, J.; POESEN, J. Soil Erosion in Europe: Major Processes, Causes and Consequences. Em: BOARDMAN, J.; POESEN, J. (Eds.). **Soil Erosion in Europe**. Chichester, UK: John Wiley & Sons, Ltd, 2006. p. 477–487.

BONILLA, C. A.; VIDAL, K. L. Rainfall erosivity in Central Chile. **Journal of Hydrology**, v. 410, n. 1, p. 126–133, 15 nov. 2011.

BORRELLI, P. et al. An assessment of the global impact of 21st century land use change on soil erosion. **Nature Communications**, v. 8, n. 1, p. 2013, 8 dez. 2017.

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

BROKAMP, C. et al. Exposure assessment models for elemental components of particulate matter in an urban environment: A comparison of regression and random forest approaches. **Atmospheric Environment**, v. 151, p. 1–11, fev. 2017.

BUCHAN, D. S.; BAKER, J. S. Development and Evaluation of Sedentary Time Cut-Points for the activPAL in Adults Using the GGIR R-Package. **International Journal of Environmental Research and Public Health**, v. 20, n. 3, p. 2293, 27 jan. 2023.

CARO, M. C. Binary classification with classical instances and quantum labels. **Quantum Machine Intelligence**, v. 3, n. 1, p. 18, jun. 2021.

CARVALHO, N. DE O. **Hidrossedimentologia prática**. Rio de Janeiro - RJ: Embrapa Pantanal (CPAP), 1994. v. 372

CHAI, T.; DRAXLER, R. R. Root mean square error (RMSE) or mean absolute error (MAE)? – Arguments against avoiding RMSE in the literature. **Geoscientific Model Development**, v. 7, n. 3, p. 1247–1250, 30 jun. 2014.

CHEN, Z. et al. Estimating daily reference evapotranspiration based on limited meteorological data using deep learning and classical machine learning methods. **Journal of Hydrology**, v. 591, p. 125286, dez. 2020.

CHOUBIN, B. et al. River suspended sediment modelling using the CART model: A comparative study of machine learning techniques. **Science of The Total Environment**, v. 615, p. 272–281, fev. 2018.

COLLINS, L. et al. The utility of Random Forests for wildfire severity mapping. **Remote Sensing of Environment**, v. 216, p. 374–384, out. 2018.

COMON, P. Independent component analysis, A new concept? **Signal Processing**, v. 36, n. 3, p. 287–314, abr. 1994.

CUNHA, F. F. DA et al. PERFORMANCE OF ESTIMATIVE MODELS FOR DAILY REFERENCE EVAPOTRANSPIRATION IN THE CITY OF CASSILÂNDIA, BRAZIL. **Engenharia Agrícola**, v. 37, n. 1, p. 173–184, fev. 2017.

CUTLER, D. R. et al. RANDOM FORESTS FOR CLASSIFICATION IN ECOLOGY. **Ecology**, v. 88, n. 11, p. 2783–2792, nov. 2007.

DA SILVA, A. M. Aplicações de técnicas de geoprocessamento no estudo das relações entre os processos erosivos e sedimentológicos de bacia hidrográfica. 1999.

DA SILVA, A. M. Rainfall erosivity map for Brazil. **CATENA**, v. 57, n. 3, p. 251–259, ago. 2004.

DE MELLO, C. R. et al. Daily rainfall erosivity as an indicator for natural disasters: assessment in mountainous regions of southeastern Brazil. **Natural Hazards**, 9 maio 2020.

DE SOUZA, C. M. P. et al. Spatiotemporal prediction of rainfall erosivity by machine learning in southeastern Brazil. **Geocarto International**, p. 1–19, 18 abr. 2022.

DI SANTO, K. G. et al. Active demand side management for households in smart grids using optimization and artificial intelligence. **Measurement**, v. 115, p. 152–161, fev. 2018.

DIAKAKIS, M.; PRISKOS, G.; SKORDOULIS, M. Public perception of flood risk in flash flood prone areas of Eastern Mediterranean: The case of Attica Region in Greece. **International Journal of Disaster Risk Reduction**, v. 28, p. 404–413, jun. 2018.

DIAS, S. H. B. et al. Reference evapotranspiration of Brazil modeled with machine learning techniques and remote sensing. **PLOS ONE**, v. 16, n. 2, p. e0245834, 9 fev. 2021.

DIMITRIADOU, S.; NIKOLAKOPOULOS, K. G. Artificial Neural Networks for the Prediction of the Reference Evapotranspiration of the Peloponnese Peninsula, Greece. **Water**, v. 14, n. 13, p. 2027, 24 jun. 2022.

DIODATO, N. et al. Communicating Hydrological Hazard-Prone Areas in Italy With Geospatial Probability Maps. **Frontiers in Environmental Science**, v. 7, p. 193, 12 dez. 2019.

DIODATO, N.; KNIGHT, J.; BELLOCCHI, G. Reduced complexity model for assessing patterns of rainfall erosivity in Africa. **Global and Planetary Change**, v. 100, p. 183–193, 1 jan. 2013.

DONG, L. et al. An optimized method for extracting slope length in RUSLE from raster digital elevation. **CATENA**, v. 209, p. 105818, fev. 2022.

DU, P. et al. Random Forest and Rotation Forest for fully polarized SAR image classification using polarimetric and spatial features. **ISPRS Journal of Photogrammetry and Remote Sensing**, v. 105, p. 38–53, jul. 2015.

DUULATOV, E. et al. Projected Rainfall Erosivity Over Central Asia Based on CMIP5 Climate Models. **Water**, v. 11, n. 5, p. 897, maio 2019.

EL-KENAWY, E.-S. M. et al. Improved weighted ensemble learning for predicting the daily reference evapotranspiration under the semi-arid climate conditions. **Environmental Science and Pollution Research**, 22 jun. 2022.

ESTES, J. E.; SAILER, C.; TINNEY, L. R. APPLICATIONS OF ARTIFICIAL INTELLIGENCE TECHNIQUES TO REMOTE SENSING*. **The Professional Geographer**, v. 38, n. 2, p. 133–141, maio 1986.

FARIDI, M.; VERMA, S.; MUKHERJEE, S. Integration of GIS, Spatial Data Mining, and Fuzzy Logic for Agricultural Intelligence. Em: PANT, M. et al. (Eds.). **Soft Computing: Theories and Applications**. Singapore: Springer Singapore, 2018. v. 583p. 171–183.

FERNANDES FILHO, E. I. **labgeo: Collection of functions to fit models with emphasis in land use and soil mapping**. , 2020. Disponível em: <<https://github.com/elpidiofilho/labgeo>>

FICK, S. E.; HIJMANS, R. J. WorldClim 2: new 1-km spatial resolution climate surfaces for global land areas: NEW CLIMATE SURFACES FOR GLOBAL LAND AREAS. **International Journal of Climatology**, v. 37, n. 12, p. 4302–4315, out. 2017.

FOSTER, G. R. et al. Conversion of the universal soil loss equation to SI metric units. **Journal of Soil and Water Conservation**, v. 36, n. 6, p. 355, 1 nov. 1981.

FREEBAIRN, J. W.; ZILLMAN, J. W. Economic benefits of meteorological services. **Meteorological Applications**, v. 9, n. 1, p. 33–44, 2002.

FRIEDMAN, J. H. Multivariate adaptive regression splines. **The annals of statistics**, v. 19, n. 1, p. 1–67, 1991.

FRIEDMAN, J. H. Greedy Function Approximation: A Gradient Boosting Machine. **The Annals of Statistics**, v. 29, n. 5, p. 1189–1232, 2001.

FRIEDMAN, J. H. Stochastic gradient boosting. **Computational Statistics & Data Analysis**, v. 38, n. 4, p. 367–378, fev. 2002.

GARAY TRINDADE, M. et al. Effects of thermal neutron radiation on a hardware-implemented machine learning algorithm. **Microelectronics Reliability**, v. 116, p. 114022, jan. 2021.

GELADI, P.; KOWALSKI, B. R. Partial least-squares regression: a tutorial. **Analytica Chimica Acta**, v. 185, p. 1–17, 1986.

GIARDINO, M. J. A history of NASA remote sensing contributions to archaeology. **Journal of Archaeological Science**, v. 38, n. 9, p. 2003–2009, set. 2011.

GONÇALVES, W. G. E et al. Classificação de estratos florestais utilizando redes neurais artificiais e dados de sensoriamento remoto. **Ambiente e Água - An Interdisciplinary Journal of Applied Science**, v. 11, n. 3, p. 612, 23 jun. 2016.

GONDIM, J. et al. **Support Vector Machines in Smile detection: A comparison of auto-tuning standard processes in Gaussian kernel**. Anais do XVII Workshop de Visão Computacional (WVC 2021). **Anais...** Em: WORKSHOP DE VISÃO COMPUTACIONAL. Brasil: Sociedade Brasileira de Computação - SBC, 22 nov. 2021. Disponível em: <<https://sol.sbc.org.br/index.php/wvc/article/view/18900>>. Acesso em: 15 fev. 2025

GRAVES, S. et al. Tree Species Abundance Predictions in a Tropical Agricultural Landscape with a Supervised Classification Model and Imbalanced Data. **Remote Sensing**, v. 8, n. 2, p. 161, 19 fev. 2016.

GROVER, V. et al. Creating Strategic Business Value from Big Data Analytics: A Research Framework. **Journal of Management Information Systems**, v. 35, n. 2, p. 388–423, 3 abr. 2018.

GU, Z. et al. Spatial and Temporal Patterns of Rainfall Erosivity in the Tibetan Plateau. **Water**, v. 12, n. 1, p. 200, jan. 2020.

GUO, J.; HUANG, C.; HOU, J. A Scalable Computing Resources System for Remote Sensing Big Data Processing Using GeoPySpark Based on Spark on K8s. **Remote Sensing**, v. 14, n. 3, p. 521, 22 jan. 2022.

HASTIE, S. M. D. FROM MDA:MARS BY T.; WRAPPER, R. T. U. A. M. F. UTILITIES WITH T. L. LEAPS. **earth: Multivariate Adaptive Regression Splines**. [s.l: s.n.].

HE, M. et al. Satellite detection of soil moisture related water stress impacts on ecosystem productivity using the MODIS-based photochemical reflectance index. **Remote Sensing of Environment**, v. 186, p. 173–183, dez. 2016.

HELLAND, I. S. On the structure of partial least squares regression. **Communications in Statistics - Simulation and Computation**, v. 17, n. 2, p. 581–607, jan. 1988.

HEYDARI, S. S.; MOUNTRAKIS, G. Effect of classifier selection, reference sample size, reference class distribution and scene heterogeneity in per-pixel classification accuracy using 26 Landsat sites. **Remote Sensing of Environment**, v. 204, p. 648–658, jan. 2018.

HORNIK, K. RWeKajars: R/Weka Interface Jars. R package version 3.9. 1-3. 2018.

HORNIK, K.; BUCHTA, C.; ZEILEIS, A. Open-source machine learning: R meets Weka. **Computational Statistics**, v. 24, n. 2, p. 225–232, maio 2009.

HOYOS, N.; WAYLEN, P. R.; JARAMILLO, Á. Seasonal and spatial patterns of erosivity in a tropical watershed of the Colombian Andes. **Journal of Hydrology**, v. 314, n. 1, p. 177–191, 25 nov. 2005.

HU, Q. et al. Rainfall erosivity in the Republic of Korea. **Journal of Soil and Water Conservation**, v. 55, n. 2, p. 115, 1 abr. 2000.

HUDSON, N. **Soil conservation**. London: Batsford, 1971.

HUDSON, N. **Soil Conservation**. 2nd ed ed. Ithaca, N.Y: Cornell University Press, 1981.

HUYNH, P. H.; NGUYEN, V. H. A Novel Ensemble of Support Vector Machines for Improving Medical Data Classification. **Engineering Innovations**, v. 4, p. 47–66, 15 fev. 2023.

IBM. **Big Data Analytics | IBM Analytics**. Disponível em: <<https://www.ibm.com/analytics/hadoop/big-data-analytics>>. Acesso em: 25 jul. 2018.

INMET. **Instituto Nacional de Meteorologia**. Dados meteorológicos. Disponível em: <<http://www.inmet.gov.br>>. Acesso em: 2 fev. 2021.

ISIDRO CAMPOS et al. Estimation of total available water in the soil layer by integrating actual evapotranspiration data in a remote sensing-driven soil water balance. **Journal of Hydrology**, v. 534, p. 427–439, mar. 2016.

JAMES, G. et al. (EDS.). **An introduction to statistical learning: with applications in R**. New York: Springer, 2013.

JOHNSON, R.; ZHANG, T. Overfitting in machine learning: what it is and how to prevent it. **ACM SIGKDD Explorations Newsletter**, v. 21, n. 2, p. 22–29, 2019.

JØRGENSEN, L. F.; VILLHOLTH, K. G.; REFSGAARD, J. C. Groundwater management and protection in Denmark: a review of pre-conditions, advances and challenges. **International Journal of Water Resources Development**, v. 33, n. 6, p. 868–889, 2 nov. 2017.

KAHAKI, S. M. M. et al. Geometric feature descriptor and dissimilarity-based registration of remotely sensed imagery. **PLOS ONE**, v. 13, n. 7, p. e0200676, 19 jul. 2018.

KARANDISH, F.; ŠIMŮNEK, J. A comparison of numerical and machine-learning modeling of soil water content with limited input data. **Journal of Hydrology**, v. 543, p. 892–909, dez. 2016.

KARATZOGLOU, A. et al. **kernlab** - An S4 Package for Kernel Methods in R. **Journal of Statistical Software**, v. 11, n. 9, 2004.

KAVZOGLU, T. et al. Dimensionality Reduction and Classification of Hyperspectral Images Using Object-Based Image Analysis. **Journal of the Indian Society of Remote Sensing**, 19 jul. 2018.

KE, Y. et al. Downscaling of MODIS One Kilometer Evapotranspiration Using Landsat-8 Data and Machine Learning Approaches. **Remote Sensing**, v. 8, n. 3, p. 215, 8 mar. 2016.

KHATOON, S. et al. Comparison of fuzzy time series, ANN and wavelet techniques for short term load forecasting. **International Journal of Power Electronics and Drive Systems (IJPEDS)**, v. 14, n. 2, p. 1260, 1 jun. 2023.

KHOSHRAVESH, M.; SEFIDKOUHI, M. A. G.; VALIPOUR, M. Estimation of reference evapotranspiration using multivariate fractional polynomial, Bayesian regression, and robust regression models in three arid environments. **Applied Water Science**, v. 7, n. 4, p. 1911–1922, jul. 2017.

KIM, J. S.; CHEN, J.; GARCIA, H. E. Modeling, control, and dynamic performance analysis of a reverse osmosis desalination plant integrated within hybrid energy systems. **Energy**, v. 112, p. 52–66, out. 2016.

KIM, Y.; KWON, O.; LEE, K. A comprehensive study on feature selection methods for sentiment analysis. **Information Sciences**, v. 509, p. 350–364, 2020.

KLIK, A. et al. Spatial and temporal distribution of rainfall erosivity in New Zealand. **Soil Research**, v. 53, n. 7, p. 815–825, 17 nov. 2015.

KOUNDOURI, P.; ROSETA-PALMA, C.; ENGLEZOS, N. Out of Sight, Not Out of Mind: Developments in Economic Models of Groundwater Management. **International Review of Environmental and Resource Economics**, v. 11, n. 1, p. 55–96, 24 out. 2017.

KUHN, M. **Caret package**. : Journal of Statistical Software. Journal of Statistical Software. 28(5), , 2008. Disponível em: <<https://cran.r-project.org/web/packages/caret/index.html>>

KUHN, M.; JOHNSON, K. **Applied Predictive Modeling**. New York, NY: Springer New York, 2013.

KULEKÇI, M. et al. Usefulness of MARS and Bagging MARS Algorithms in Prediction of Honey Production in Beekeeping Enterprises from Elazığ Province of Turkey. **Pakistan Journal of Zoology**, v. 54, n. 3, 2022.

LAILOLO, P. et al. Impact of different satellite soil moisture products on the predictions of a continuous distributed hydrological model. **International Journal of Applied Earth Observation and Geoinformation**, v. 48, p. 131–145, jun. 2016.

LAMPERTI, F.; ROVENTINI, A.; SANI, A. Agent-based model calibration using machine learning surrogates. **Journal of Economic Dynamics and Control**, v. 90, p. 366–389, maio 2018.

LANDERAS, G. et al. New alternatives for reference evapotranspiration estimation in West Africa using limited weather data and ancillary data supply strategies. **Theoretical and Applied Climatology**, 12 abr. 2017.

LANGENBUCHER, A. et al. Performance of a simplified strategy for formula constant optimisation in intraocular lens power calculation. **Acta Ophthalmologica**, v. 103, n. 1, fev. 2025.

LARY, D. J. et al. Machine learning in geosciences and remote sensing. **Geoscience Frontiers**, v. 7, n. 1, p. 3–10, jan. 2016.

LEE, S. et al. Estimation of rainfall erosivity factor in Italy and Switzerland using Bayesian optimization based machine learning models. **CATENA**, v. 211, p. 105957, abr. 2022.

LEPRUN, J. A erosao, a conservação e o manejo do solo no nordeste brasileiro : balanço, diagnostico e novas linhas de pesquisas. **Recife: Ministério do Interior. SUDENE, Brasil**, 1981.

LEVIDOW, L. et al. Improving water-efficient irrigation: Prospects and difficulties of innovative practices. **Agricultural Water Management**, v. 146, p. 84–94, dez. 2014.

LI, Q. et al. GANs-LSTM model for soil temperature estimation from meteorological: A new approach. **IEEE Access**, p. 1–1, 2020.

LI, Z. et al. **Process and system for monitoring the progress of transmission line construction and troubleshooting unplanned operations based on satellite remote sensing**. (H. Kolivand, A. J. Moshayedi, Eds.)Fourth International Conference on Computer Graphics, Image, and Virtualization (ICCGIV 2024). **Anais...** Em: FOURTH INTERNATIONAL CONFERENCE ON COMPUTER GRAPHICS, IMAGE AND VIRTUALIZATION (ICCGIV 2024). Chengdu, China: SPIE, 9 out. 2024. Disponível em: <<https://www.spiedigitallibrary.org/conference-proceedings-of-spie/13288/3045303/Process-and-system-for-monitoring-the-progress-of-transmission-line/10.1117/12.3045303.full>>. Acesso em: 16 fev. 2025

LIU, J.-X.; LEU, J.-S.; HOLST, S. Stock price movement prediction based on Stocktwits investor sentiment using FinBERT and ensemble SVM. **PeerJ Computer Science**, v. 9, p. e1403, 7 jun. 2023.

LIU, X. et al. Comparison of 16 models for reference crop evapotranspiration against weighing lysimeter measurement. **Agricultural Water Management**, v. 184, p. 145–155, abr. 2017.

LÖBMANN, M. T. et al. Systems knowledge for sustainable soil and land management. **Science of The Total Environment**, v. 822, p. 153389, maio 2022.

MAHJOobi, E.; RAFIEI, F. Exploring the Potential of Geospatial Data: An In-Depth Investigation. Em: M. ABDALLA, R. (Ed.). **Exploring Remote Sensing - Methods and Applications**. [s.l.] IntechOpen, 2024.

MCCULLOCH, W. S.; PITTS, W. A logical calculus of the ideas immanent in nervous activity. **The bulletin of mathematical biophysics**, v. 5, n. 4, p. 115–133, 1943.

MELLOR, A. et al. Exploring issues of training data imbalance and mislabelling on random forest performance for large area land cover classification using the ensemble margin. **ISPRS Journal of Photogrammetry and Remote Sensing**, v. 105, p. 155–168, jul. 2015.

MILBORROW, M. S. Package ‘earth’. **R Software package**, 2019.

MOHAMMED, M. et al. Shallow Foundation Settlement Quantification: Application of Hybridized Adaptive Neuro-Fuzzy Inference System Model. **Advances in Civil Engineering**, v. 2020, p. 1–14, 22 fev. 2020.

MORIASI, D. N. et al. Model evaluation guidelines for systematic quantification of accuracy in watershed simulations. **Transactions of the ASABE**, v. 50, n. 3, p. 885–900, 2007.

MOUSAVI, S. N. et al. Evaluation of pressure fluctuations coefficient along the USBR Type II stilling basin using experimental results and AI models. **ISH Journal of Hydraulic Engineering**, p. 1–8, 1 abr. 2020.

MULLISSA, A. et al. Sentinel-1 SAR Backscatter Analysis Ready Data Preparation in Google Earth Engine. **Remote Sensing**, v. 13, n. 10, p. 1954, 17 maio 2021.

NACHTERGAELE, F. et al. Global land degradation information system (GLADIS). **Beta version. An information database for land degradation assessment at global level. Land degradation assessment in drylands technical report**, v. 17, 2010.

NAGHIBI, S. A.; POURGHASEMI, H. R.; DIXON, B. GIS-based groundwater potential mapping using boosted regression tree, classification and regression tree, and random forest machine learning models in Iran. **Environmental Monitoring and Assessment**, v. 188, n. 1, jan. 2016.

NAIPAL, V. et al. Improving the global applicability of the RUSLE model – adjustment of the topographical and rainfall erosivity factors. **Geoscientific Model Development**, v. 8, n. 9, p. 2893–2913, 15 set. 2015.

NAJI, H. I.; ALI, R. H. **Qualitative analysis of cost risk using WEKA program**. IEEE, jan. 2018. Disponível em: <<https://ieeexplore.ieee.org/document/8340566/>>. Acesso em: 25 jul. 2018

OLIVEIRA, P. T. S.; WENDLAND, E.; NEARING, M. A. Rainfall erosivity in Brazil: A review. **CATENA**, v. 100, p. 139–147, jan. 2013.

PANAGOS, P. et al. Rainfall erosivity in Europe. **Science of The Total Environment**, v. 511, p. 801–814, 1 abr. 2015.

PANAGOS, P. et al. Global rainfall erosivity assessment based on high-temporal resolution rainfall records. **Scientific Reports**, v. 7, n. 1, p. 4175, 23 jun. 2017.

QIN, W. et al. Spatial distribution and temporal trends of rainfall erosivity in mainland China for 1951–2010. **CATENA**, v. 147, p. 177–186, 1 dez. 2016.

QUINLAN, J. R. **Combining instance-based and model-based learning**. . Em: PROCEEDINGS OF THE TENTH INTERNATIONAL CONFERENCE ON MACHINE LEARNING. 1993.

R CORE TEAM. **R: A Language and Environment for Statistical Computing**. Vienna, Austria: R Foundation for Statistical Computing, 2024.

RAMON, R. et al. Kinetic energy estimation by rainfall intensity and its usefulness in predicting hydrosedimentological variables in a small rural catchment in southern Brazil. **CATENA**, Achieving sustainability through conservation in a changing world. v. 148, p. 176–184, 1 jan. 2017.

RENARD, K. G. (ED.). **Predicting soil erosion by water: a guide to conservation planning with the revised universal soil loss equation (RUSLE)**. Washington, D. C: [s.n.].

RIQUETTI, N. B. et al. Rainfall erosivity in South America: Current patterns and future perspectives. **Science of The Total Environment**, v. 724, p. 138315, 1 jul. 2020.

RODRIGUEZ, J. et al. Food, water, and fault lines: Remote sensing opportunities for earthquake-response management of agricultural water. **Science of The Total Environment**, v. 565, p. 1020–1027, set. 2016.

RODRIGUEZ-GALIANO, V. et al. Downscaling Landsat 7 ETM+ thermal imagery using land surface temperature and NDVI images. **International Journal of Applied Earth Observation and Geoinformation**, v. 18, p. 515–527, 1 ago. 2012.

ROGAN, J. et al. Mapping land-cover modifications over large areas: A comparison of machine learning algorithms. **Remote Sensing of Environment**, v. 112, n. 5, p. 2272–2283, maio 2008.

ROJAS, A.; CONDE, A. Estimación del factor R de la ecuación universal de pérdidas de suelo para el centro-este de la República Argentina. **Ciencia del Suelo**, v. 3, n. 1–2, p. 85–94, 1985.

RUFINO, R. L.; BISCAIA, R. M.; MERTEN, G. Determinação do potencial erosivo da chuva do estado do Paraná, através de pluviometria: terceira aproximação. **Revista brasileira de Ciência do Solo**, v. 17, n. 3, p. 439–444, 1993.

RUSSELL, S. J.; NORVIG, P. **Artificial intelligence: a modern approach**. [s.l.] Malaysia; Pearson Education Limited, 2016.

SABRI, Y. et al. Cloud Computing in Remote Sensing: Big Data Remote Sensing Knowledge Discovery and Information Analysis. **International Journal of Advanced Computer Science and Applications**, v. 12, n. 5, 2021.

SADEGHI, M. et al. The optical trapezoid model: A novel approach to remote sensing of soil moisture applied to Sentinel-2 and Landsat-8 observations. **Remote Sensing of Environment**, v. 198, p. 52–68, set. 2017.

SALVATI, P. et al. Gender, age and circumstances analysis of flood and landslide fatalities in Italy. **Science of The Total Environment**, v. 610–611, p. 867–879, jan. 2018.

SARHADI, A. et al. Water resources climate change projections using supervised nonlinear and multivariate soft computing techniques. **Journal of Hydrology**, v. 536, p. 119–132, maio 2016.

SHAIK, A. S.; SHAIK, N.; PRIYA, DR. C. K. Predictive Modeling in Remote Sensing Using Machine Learning Algorithms. **International Journal of Current Science Research and Review**, v. 07, n. 06, 21 jun. 2024.

SHAO, X. et al. A new regression method based on independent component analysis. **Talanta**, v. 69, n. 3, p. 676–680, 15 maio 2006.

SHARAD, W. The development of the earth remote sensing from satellite. **MECHANICS OF GYROSCOPIC SYSTEMS**, n. 40, p. 46–54, 26 dez. 2021.

SHENDRYK, I.; BROICH, M.; TULBURE, M. G. Multi-sensor airborne and satellite data for upscaling tree number information in a structurally complex eucalypt forest. **International Journal of Applied Earth Observation and Geoinformation**, v. 73, p. 397–406, dez. 2018.

SHIONO, T. et al. Expected impacts of climate change on rainfall erosivity of farmlands in Japan. **Ecological Engineering**, Soil Bio- and Eco-Engineering: The Use of Vegetation to Improve Slope Stability. v. 61, p. 678–689, 1 dez. 2013.

SHIRI, J. et al. Prediction of Water-Level in the Urmia Lake Using the Extreme Learning Machine Approach. **Water Resources Management**, v. 30, n. 14, p. 5217–5229, nov. 2016.

SILVA, E. R. D. et al. Determinação da equação de intensidade, duração e frequência para precipitações pluviométricas máximas de Porto Velho, RO: Determination of the intensity, duration, and frequency equation for maximum rainfall in Porto Velho, RO. **Brazilian Journal of Animal and Environmental Research**, v. 7, n. 1, p. 730–739, 23 mar. 2024.

SINGH, K.; MAHAJAN, A.; MANSOTRA, V. Using Recursive Feature Elimination and Fisher Score with Convolutional Neural Network for Identifying Port Scan Attempts. Em: ZHANG, Y.-D. et al. (Eds.). **Smart Trends in Computing and Communications**. Lecture Notes in Networks and Systems. Singapore: Springer Singapore, 2022. v. 286p. 551–560.

SUN, C. et al. A review of remote sensing for potato traits characterization in precision agriculture. **Frontiers in Plant Science**, v. 13, p. 871859, 18 jul. 2022a.

SUN, L. et al. Optimization of Cervical Cancer Screening: A Stacking-Integrated Machine Learning Algorithm Based on Demographic, Behavioral, and Clinical Factors. **Frontiers in Oncology**, v. 12, p. 821453, 15 fev. 2022b.

TAO, L.; XUE, X. AN IMPROVED RANDOM FOREST MODEL TO PREDICT BOND STRENGTH OF FRP-TO-CONCRETE. **JOURNAL OF CIVIL ENGINEERING AND MANAGEMENT**, v. 30, n. 6, p. 520–535, 5 jul. 2024.

TEMIZ, S. et al. Regeneration of Lithium-ion battery impedance using a novel machine learning framework and minimal empirical data. **Journal of Energy Storage**, v. 52, p. 105022, ago. 2022.

TENG, H. et al. Assimilating satellite imagery and visible–near infrared spectroscopy to model and map soil loss by water erosion in Australia. **Environmental Modelling & Software**, v. 77, p. 156–167, 1 mar. 2016.

TIEN BUI, D. et al. Hybrid artificial intelligence approach based on neural fuzzy inference model and metaheuristic optimization for flood susceptibility modeling in a high-frequency tropical cyclone area using GIS. **Journal of Hydrology**, v. 540, p. 317–330, set. 2016.

TIN KAM HO. **Random decision forests**. Proceedings of 3rd International Conference on Document Analysis and Recognition. **Anais...** Em: 3RD INTERNATIONAL CONFERENCE ON DOCUMENT ANALYSIS AND RECOGNITION. Montreal, Que., Canada: IEEE Comput. Soc. Press, 1995. Disponível em: <<http://ieeexplore.ieee.org/document/598994/>>. Acesso em: 9 ago. 2020

TIWARI, M. K.; ADAMOWSKI, J. F. An ensemble wavelet bootstrap machine learning approach to water demand forecasting: a case study in the city of Calgary, Canada. **Urban Water Journal**, v. 14, n. 2, p. 185–201, 7 fev. 2017.

TOTH, C.; JÓZKÓW, G. Remote sensing platforms and sensors: A survey. **ISPRS Journal of Photogrammetry and Remote Sensing**, v. 115, p. 22–36, maio 2016.

TOUREIRO, C. et al. Irrigation management with remote sensing: Evaluating irrigation requirement for maize under Mediterranean climate condition. **Agricultural Water Management**, v. 184, p. 211–220, abr. 2017.

URI, J. Landsat at 45: How it Changed the Way We See the Earth. 2017.

VANTAS, K.; SIDIROPOULOS, E.; LOUKAS, A. Estimating Current and Future Rainfall Erosivity in Greece Using Regional Climate Models and Spatial Quantile Regression Forests. **Water**, v. 12, n. 3, p. 687, mar. 2020.

VAPNIK, V. **The nature of statistical learning theory**. [s.l.] Springer science & business media, 1999.

VRIELING, A.; HOEDJES, J. C. B.; VAN DER VELDE, M. Towards large-scale monitoring of soil erosion in Africa: Accounting for the dynamics of rainfall erosivity. **Global and Planetary Change**, v. 115, p. 33–43, 1 abr. 2014.

WALTRICK, P. C. **EROSIVIDADE DE CHUVAS NO PARANÁ: ATUALIZAÇÃO, INFLUÊNCIA DO “EL NIÑO” E “LA NIÑA” E ESTIMATIVA PARA CENÁRIOS CLIMÁTICOS FUTUROS**. Dissertação de Mestrado—Curitiba-PR: Universidade Federal do Paraná, 2010.

WANG, J. et al. Development of Monthly Reference Evapotranspiration Machine Learning Models and Mapping of Pakistan—A Comparative Study. **Water**, v. 14, n. 10, p. 1666, 23 maio 2022.

WANG, J.; LI, M.; LIN, P. Evaluation Model of Regional Comprehensive Disaster Reduction Capacity under Complex Environment. **Journal of Environmental and Public Health**, v. 2022, n. 1, p. 1593536, jan. 2022.

WEI, Z. et al. A Small UAV Based Multi-Temporal Image Registration for Dynamic Agricultural Terrace Monitoring. **Remote Sensing**, v. 9, n. 9, p. 904, 31 ago. 2017.

WEKA. **Weka 3 - Data Mining with Open Source Machine Learning Software in Java**. Disponível em: <<https://www.cs.waikato.ac.nz/~ml/weka/>>. Acesso em: 25 jul. 2018.

WESTAD, F. Independent component analysis and regression applied on sensory data. **Journal of Chemometrics**, v. 19, n. 3, p. 171–179, mar. 2005.

WILKS, D. S. **Statistical methods in the atmospheric sciences**. Oxford; Waltham, MA: Academic Press, 2011.

WILLMOTT, C. J. Some comments on the evaluation of model performance. **Bulletin American Meteorological Society**, v. 30, p. 1309–1310, 1982.

WILLMOTT, C.; MATSUURA, K. Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. **Climate Research**, v. 30, p. 79–82, 2005.

WING, M. K. C. FROM J. et al. **caret: Classification and Regression Training**. 5. ed. Stuttgart, Germany: Journal of statistical software, 2017. v. 28

WISCHMEIER, W. H.; SMITH, D. D. **Predicting rainfall erosion losses: a guide to conservation planning**. Washington, DC: Department of Agriculture, Science and Education Administration, 1978.

WOLD, S.; MARTENS, H.; WOLD, H. The multivariate calibration problem in chemistry solved by the PLS method. Em: **Matrix pencils**. [s.l.] Springer, 1983. p. 286–293.

WU, T. et al. Evaluation of stacking and blending ensemble learning methods for estimating daily reference evapotranspiration. **Computers and Electronics in Agriculture**, v. 184, p. 106039, maio 2021.

WU, W. et al. A comparison of support vector machines, artificial neural network and classification tree for identifying soil texture classes in southwest China. **Computers and Electronics in Agriculture**, v. 144, p. 86–93, jan. 2018.

WUEPPER, D.; BORRELLI, P.; FINGER, R. Countries and the global rate of soil erosion. **Nature Sustainability**, v. 3, n. 1, p. 51–55, jan. 2020.

XU, C. et al. **Application of remote sensing big data in provincial natural resource management: a case of Hainan Province**. (B. Zou, Y. Cui, Eds.) International Conference on Remote Sensing, Mapping, and Geographic Information Systems (RSMG 2024). **Anais...** Em: SECOND INTERNATIONAL CONFERENCE ON REMOTE SENSING, MAPPING, AND GEOGRAPHIC INFORMATION SYSTEMS (RSMG 2024). Zhengzhou, China: SPIE, 27 nov. 2024. Disponível em: <<https://www.spiedigitallibrary.org/conference-proceedings-of-spie/13402/3048670/Application-of-remote-sensing-big-data-in-provincial-natural-resource/10.1117/12.3048670.full>>. Acesso em: 16 fev. 2025

YANG, D. et al. Global potential soil erosion with reference to land use and climate changes. **Hydrological Processes**, v. 17, n. 14, p. 2913–2928, 2003.

YANG, Y. et al. Identification and Analysis of the Geohazards Located in an Alpine Valley Based on Multi-Source Remote Sensing Data. **Sensors**, v. 24, n. 13, p. 4057, 21 jun. 2024.

YASEEN, Z. M. et al. Prediction of evaporation in arid and semi-arid regions: a comparative study using different machine learning models. **Engineering Applications of Computational Fluid Mechanics**, v. 14, n. 1, p. 70–89, 1 jan. 2020.

ZAKHROUF, M. et al. Novel hybrid approaches based on evolutionary strategy for streamflow forecasting in the Chellif River, Algeria. **Acta Geophysica**, v. 68, n. 1, p. 167–180, fev. 2020.

ZHANG, K.; HU, C.; YU, H. Remote Sensing Image Land Classification Based on Deep Learning. **Scientific Programming**, v. 2021, p. 1–12, 24 dez. 2021.

ZHANG, X.; ZHOU, Y.; LUO, J. Deep learning for processing and analysis of remote sensing big data: a technical review. **Big Earth Data**, v. 6, n. 4, p. 527–560, 2 out. 2022.

ZHAO, D. et al. Clay content mapping and uncertainty estimation using weighted model averaging. **CATENA**, v. 209, p. 105791, fev. 2022.

ZHU, H. et al. Study of the influence of meteorological factors on HFMD and prediction based on the LSTM algorithm in Fuzhou, China. **BMC Infectious Diseases**, v. 23, n. 1, p. 299, 5 maio 2023.

ZHU, Z.; YU, B. Validation of Rainfall Erosivity Estimators for Mainland China. 2015.

ZIPSER, E. J. et al. Where are the most intense thunderstorms on earth? **Bulletin of the American Meteorological Society**, v. 87, n. 8, p. 1057–1072, 1 ago. 2006.