

## MAPEAMENTO DE RIOS EM IMAGENS RGB COM APRENDIZAGEM DE MÁQUINA SUPERVISIONADA.

### *RIVER MAPPING IN SPATIAL HIGH RESOLUTION IMAGES USING MACHINE LEARNING ALGORITHMS.*

 Mariany Kerriany Gonçalves de Souza <sup>A</sup>

 Mayara Maezano Faita Pinheiro <sup>A</sup>

 Danielle Elis Garcia Furuya <sup>A</sup>

 Lucas Prado Osco <sup>A</sup>

 José Marcato Júnior <sup>B</sup>

 Wesley Nunes Gonçalves <sup>B</sup>

 Ana Paula Marques Ramos <sup>C</sup>

<sup>A</sup> Universidade do Oeste Paulista (UNOESTE), Presidente Prudente, SP, Brasil

<sup>B</sup> Universidade Federal do Mato Grosso do Sul (UFMS), Campo Grande, MS, Brasil

<sup>C</sup> Universidade Estadual Paulista (UNESP), Presidente Prudente, SP, Brasil

Recebido em: 14/04//2022 | 31/07/2023    DOI: 10.12957/tamoios.2024.66545

Correspondência para: Ana Paula Marques Ramos (anaramos@unoeste.br)

### Resumo

O mapeamento de rios tem uma grande importância para estudos ambientais, especialmente no que diz respeito à proteção e conservação de recursos naturais. Rios são elementos cruciais para a manutenção da vida e dos ecossistemas, por isso o mapeamento preciso dessas características é essencial. Neste trabalho, o objetivo foi mapear cursos d'água em imagens RGB usando algoritmos de aprendizado de máquina supervisionado - *Random Forest* (RF) e *Support Vector Machine* (SVM). O estudo de caso foi conduzido na região da 22ª Unidade de Gerenciamento de Recursos Hídricos do Pontal do Paranapanema, em São Paulo, utilizando imagens aéreas RGB de alta resolução espacial (1 metro). O método incluiu a preparação dos dados, treinamento e validação dos algoritmos de aprendizado de máquina, usando métricas quantitativas como o F1-score, e uma avaliação qualitativa dos resultados. Embora ambos os algoritmos tenham apresentado desempenho elevado, o SVM mostrou-se superior em termos de F1-score médio. No entanto, o RF destacou-se ao lidar com regiões mais complexas das imagens RGB, como áreas secas dentro dos rios, apresentando menos falsos negativos. Esta pesquisa abre caminho para o uso eficiente de técnicas de aprendizado de máquina na análise de imagens RGB de alta resolução espacial para mapear rios, com potencial para auxiliar estudos de impacto ambiental e na gestão de recursos hídricos.

**Palavras-chave:** Hidrografia, aprendizagem rasa de máquina, classificação de imagens, impacto ambiental.

### Abstract

The mapping of rivers holds great importance for environmental studies, especially regarding protecting and conserving natural resources. Rivers are crucial for maintaining life and ecosystems, making the accurate mapping of these features essential. This work aimed to map watercourses in RGB images using supervised machine learning algorithms - *Random Forest* (RF) and *Support Vector Machine* (SVM). The case study was conducted in the region of the 22nd Water Resource Management Unit of Pontal do Paranapanema, in São Paulo, using high spatial resolution RGB aerial images (1 meter). The method included data preparation, training, and validation of the machine learning algorithms, using quantitative metrics such as the F1-score and a qualitative evaluation of the results. Although both algorithms presented high performance, SVM proved superior in the average F1-score. However, RF stood out when dealing with more complex regions of the RGB images, such as dry areas within rivers, presenting fewer false negatives. This research paves the way for the efficient use of machine learning techniques in analyzing high spatial resolution RGB images to map rivers, with the potential to assist in environmental impact studies and water resource management.





**Keywords:** Hydrography, shallow machine learning, image classification, environmental impact.

## INTRODUÇÃO

A proteção de recursos hídricos é um elemento central nas chamadas emergências ambientais do século XXI. O consenso sobre a importância desta temática pode ser facilmente avaliado em documentos de natureza diversa, especialmente aqueles que esboçam cenários e estratégias futuras. Como componente essencial para a vida humana e para a dinâmica de todos os sistemas ambientais, a água, em seus diversos estados, determina as características zonais e as tipologias dos ecossistemas, entendidos como conjuntos de biótopos e biocenoses em todas as escalas espaciais (TOMASONI et al. 2009). A disponibilidade de recursos hídricos no planeta representa um dos desafios mais complexos da atualidade, dada a elevada e alarmante escassez deste recurso natural essencial à vida na terra. Este é um tema de interesse que, há 25 anos, no Brasil, foi estabelecida a Lei nº 9.433, conhecida como Política Nacional de Recursos Hídricos (BRASIL, 1997), para estabelecer ações de uso e preservação desses recursos.

Os diversos tipos de corpos d'água superficiais, incluindo lagos, rios e reservatórios, desempenham um papel crucial na geração de energia na terra, e a falta desses recursos afeta profundamente as paisagens naturais e as atividades humanas (JIANG et al. 2018). No entanto, um gerenciamento efetivo desses recursos depende de informações geográficas sobre como estão distribuídos no espaço. Com essas informações, é possível analisar a dinâmica de interação entre o meio natural e antrópico com tais recursos, obtendo uma compreensão mais bem definida dos processos que ocorrem no meio físico e biótico. O mapeamento dessas características é favorecido pelo uso de Geotecnologias, como o uso de dados de sensoriamento remoto e técnicas de geoprocessamento aplicadas em ambiente GIS (*Geographic Information System*), minimizando a necessidade de levantamentos de campo, que muitas vezes são caros e exaustivos, para cartografar essas características.

Com o passar dos anos, métodos de classificação de imagens cada vez mais avançados têm sido desenvolvidos para melhorar a eficácia e eficiência na extração de informações de dados de sensoriamento remoto. É neste contexto que a Aprendizagem de Máquina ganha espaço no processamento de dados espaciais, visando produzir informação geoespacial de interesse para diversas áreas, incluindo a ambiental. A Aprendizagem de Máquina (*Machine Learning* – ML), subárea da Inteligência Artificial subdividida em aprendizagem rasa e profunda (*Deep Learning*), foi originalmente desenvolvida para o processamento de imagens na área médica (GOODFELLOW et al. 2016). No entanto, devido às suas potencialidades, vem sendo empregada com sucesso para o processamento de imagens de outras áreas, como o sensoriamento remoto (OSCO et al. 2021). Esta é considerada uma alternativa contemporânea para a classificação de imagens de sensoriamento remoto, sendo, portanto, pertinente testá-la em projetos de planejamento e monitoramento ambiental.

Na literatura, existem vários trabalhos que utilizam técnicas de ML para o mapeamento de águas superficiais com dados de sensoriamento remoto. Como exemplo, o trabalho de Nascimento et al. (2013) comparou o desempenho do algoritmo *Support Vector Machine* (SVM) com o de um classificador estatístico tradicional, o algoritmo de Máxima Verossimilhança (MaxVer), para classificar uma região de reservatórios de água no Brasil usando imagens CBERS-2, sensor CCD. Esses autores constataram uma diferença



considerável em acurácia (71% para o SVM, e 49% do MaxVer) na tarefa, com o SVM superior em 22%. Outro estudo, Martins et al. (2016), demonstrou a capacidade do SVM em mapear o uso e a cobertura do solo na região ao redor de um reservatório hidrelétrico no estado do Rio de Janeiro. Neste experimento, os autores avaliaram a melhor parametrização para o SVM na tarefa proposta, testando funções do tipo linear, polinomial e de base radial (RBF), em imagens TM/Landsat-5. O SVM com a função RBF foi o mais eficaz em termos de acurácia, segundo os autores.

Testes com algoritmos de ML para mapeamento de águas superficiais também foram conduzidos com imagens do satélite Landsat-8, sensor OLI, como demonstrado no trabalho de Jiang et al. (2018). Estes obtiveram uma acurácia de 98% na segmentação de massas de água usando o índice NDWI (Índice de Água por Diferença Normalizada) processado com o algoritmo SVM. Nesse contexto, é possível afirmar que os algoritmos de ML apresentam uma acurácia efetiva na tarefa de mapear corpos hídricos em imagens orbitais multiespectrais de média resolução espacial, como as imagens CBERS e Landsat. No entanto, ainda é necessário explorar se há um aumento na acurácia ao aplicar tais algoritmos em imagens RGB, isto é, de menor resolução espectral, mas de alta resolução espacial para a mesma tarefa. Se essa hipótese for confirmada, teríamos a possibilidade de mapear esses recursos com um nível maior de detalhe, expandindo as possibilidades de uso desses dados pelos órgãos responsáveis pela proteção das águas.

Diante disso, o objetivo deste trabalho é caracterizar o desempenho de algoritmos rasos de ML para mapear cursos d'água de médio porte, em termos de largura do corpo hídrico, em imagens RGB de alta resolução espacial. Este é um estudo inédito na região oeste do estado de São Paulo, que enfrenta sérios problemas de proteção das águas devido ao desmatamento da Mata Atlântica, resultando em um *déficit* de áreas de preservação permanente. A definição de estratégias para o mapeamento, monitoramento e recuperação desses recursos é uma tarefa recorrente do Comitê de Bacias Hidrográficas do Pontal do Paranapanema (CBH-PP).

## METODOLOGIA

### Área de Estudo

A área de estudo é localizada na 22ª Unidade de Gerenciamento de Recursos Hídricos do Pontal do Paranapanema (UGRHI-22), composta por sete importantes bacias hidrográficas e que abrange 26 municípios no oeste do estado de São Paulo. Os principais rios dessa unidade são o Paraná, Paranapanema, Santo Anastácio e Pirapozinho.

Para a realização dos experimentos, utilizamos o *software* GIS ESRI ArcGIS Pro 2.8, que possui vários algoritmos de aprendizagem de máquina. O conjunto de dados empregados é composto por imagens RGB de alta resolução espacial (1 metro), provenientes do projeto Mapeia-SP (EMPLASA, 2014). Este projeto forneceu imagens aéreas do estado de São Paulo com uma distância de amostra no solo (*Ground Sample Distance* - GSD) de 0,45 m, obtidas em 2011. Após a etapa de ortorretificação, as imagens foram cedidas ao público com uma resolução de 1 metro, o que permite o mapeamento de características em grande escala, como 1:10.000.

Na Figura 1, as imagens usadas para treinamento e teste dos algoritmos são destacadas em vermelho e azul, respectivamente. Importante ressaltar que as cenas selecionadas abrangem partes dos dois maiores rios da região: o Rio Paranapanema, que divide os estados



de São Paulo e Paraná, e o Rio Paraná, que divide os estados de São Paulo e Mato Grosso do Sul. Estes rios são considerados largos comparados a outros cursos d'água na UGRHI-22.

### Preparação dos dados

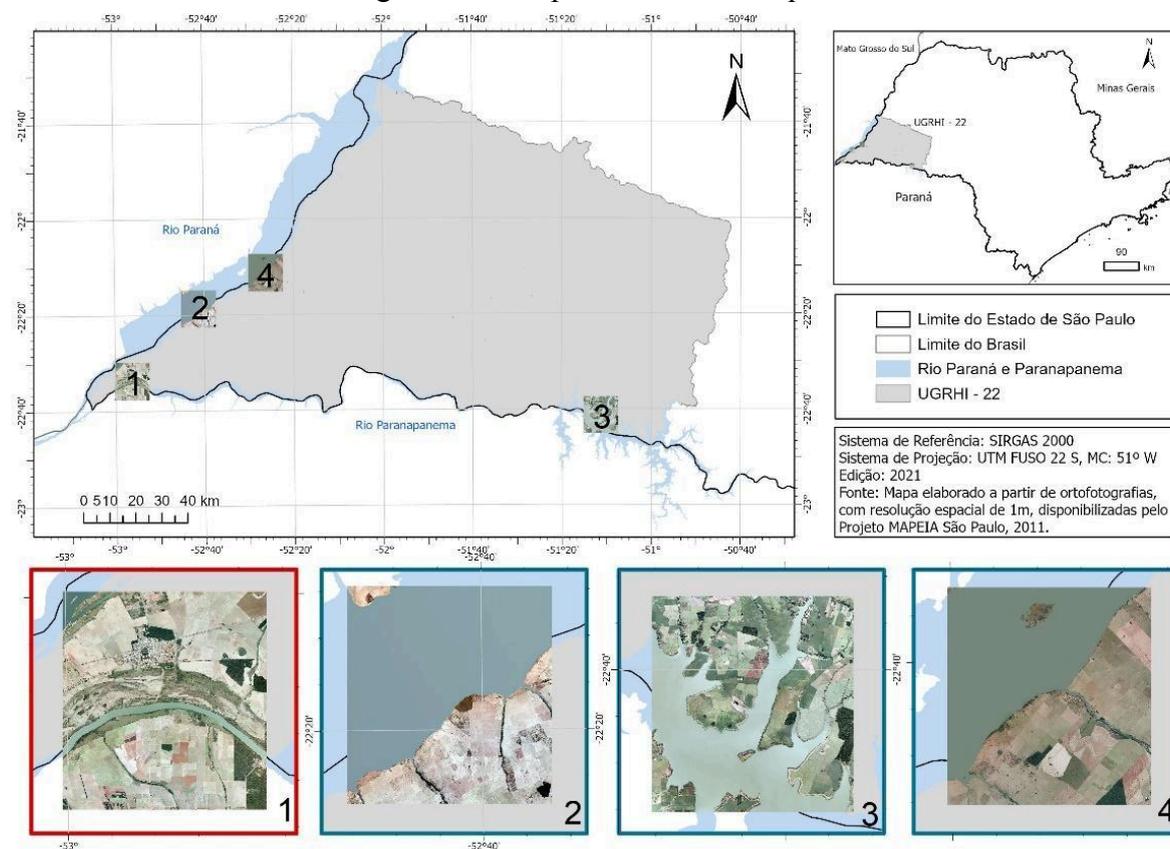
Os algoritmos utilizados no estudo de caso são o *Random Forest* e o *Support Vector Machine*. Estes são alguns dos algoritmos mais utilizados na classificação de imagens de sensoriamento remoto (JAIN et al. 2020) e que estão disponíveis no *software* GIS ESRI ArcGIS Pro. Três ortofotos (Figura 1) foram selecionadas para teste de desempenho desses algoritmos e uma para treino. Há duas razões para o número de imagens de teste ser maior que o de treino em nosso experimento. Primeiro, porque o real desempenho de um algoritmo de ML é medido nos dados de teste, ou seja, nos conjuntos de dados não vistos pelos algoritmos durante a tarefa de aprendizagem, neste caso, o reconhecimento de rios em imagens RGB. A segunda razão decorre, em geral, da alta demanda computacional exigida pela tarefa de treinamento do algoritmo. A quantidade de dados a ser processada pode tornar a tarefa inviável para uma máquina convencional, como a usada no presente estudo, que possui 4GB de memória RAM, processador Intel Core i5 e placa de vídeo AMD R7 com 2GB. Uma alternativa, portanto, é treinar o algoritmo com menos dados, mas garantindo que estes sejam representativos para a tarefa.

Ressalta-se que foi considerado nessa estratégia de definição da quantidade de dados para treino e teste a afirmativa de Goodfellow et al. (2016), que diz que embora a utilização de uma quantidade adequada de dados de teste seja importante para a caracterização do desempenho de um algoritmo de aprendizado de máquina, é crucial garantir que haja um equilíbrio adequado entre os dados de treino e teste. Isto é para evitar o sobreajuste, que pode comprometer a capacidade do modelo de generalizar para novos dados. No estudo de caso, esta representatividade significa oferecer ao algoritmo exemplos de rios que caracterizam aqueles que se pretende mapear na área. A Tabela 1 apresenta a nomenclatura (usada pela própria EMPLASA) das imagens usadas no treinamento e teste dos algoritmos.

Para a realização dos experimentos, os dados de referência, também chamados de verdades terrestres (*ground truth*), foram obtidos pelo processo de vetorização das imagens RGB. Neste caso, os dados de referência são as regiões de rios. Neste caso, a anotação (vetorização) das feições ocorreu na escala de visualização de 1:3.000 no GIS ArcGIS Pro 2.8, usando a geometria polígono. Esta foi a primitiva geométrica adequada em função da largura dos corpos d'água observada nas imagens. Com isso, o processo de anotação resultou em duas classes de informação: “rio” e “não rio”. A segunda classe foi obtida pela subtração do polígono que envolve toda a área da ortofoto, ou parte dela representada pelo polígono em vermelho na Figura 2, e as feições de “rio” vetorizadas. Essa estratégia foi necessária, pois os algoritmos de ML requerem exemplos das duas classes, “rio” (Classe 1) e “não rio” (Classe 2), para desenvolver o processo de aprendizagem.



**Figura 1.** Ilustração das amostras de imagens utilizadas para o treinamento e a avaliação dos algoritmos de aprendizado de máquina.



Fonte: Autores.

**Tabela 1.** Identificação das ortofotos utilizadas nos procedimentos de treinamento e teste dos algoritmos de aprendizado de máquina.

Algoritmo	Nomenclatura das ortofotos	
	Treino	Teste
<i>Random Forest - RF</i>	orto_yavi1no	orto_bi1so
<i>Support Vector Machine - SVM</i>		orto_bvi2so
		yaiiii4no
		orto_bi1so
	orto_bvi2so	
		yaiiii4no

O problema em questão, mapear rios em imagens, é influenciado pelo aspecto de desbalanceamento entre as classes (rio e não rio), uma vez que há mais regiões da classe “não rio” nas imagens. Na estratégia de treinamento, isso precisa ser considerado, pois impacta diretamente e negativamente o desempenho de um algoritmo de ML (GOODFELLOW et al. 2016). Para contornar esse problema a estratégia definida foi trabalhar somente com 50 metros para além de cada margem dos rios. Esta área de influência foi gerada pela construção de uma faixa de distância de 50 metros (um *buffer*) calculada sobre o vetor de rios. Assim,



essa faixa de 50 m foi usada como máscara para o recorte das imagens RGB. A Figura 2 exemplifica o antes e o depois do processo descrito sobre a faixa de distância de 50m.

**Figura 2** Comparação da imagem RGB (orto\_yavi1no) antes e após a implementação da faixa de buffer de 50 metros.



Fonte: Autores.

### Processamento dos dados

No campo do sensoriamento remoto, o mapeamento de rios em imagens é um problema de classificação de imagens. De acordo com Jensen (2014), a classificação de imagens pode ser feita por pixel, por região ou orientada a objetos. Na classificação por pixel, cada pixel é classificado de forma independente com base em suas características espectrais. A classificação por região, em contraste, agrupa e classifica conjuntos de pixels com base na semelhança espectral e proximidade espacial. Por último, a classificação orientada a objetos incorpora um nível adicional de complexidade, classificando conjuntos de pixels ("objetos") com base não apenas nas características espectrais, mas também em atributos como forma, tamanho, textura e contexto espacial. Cada método tem suas vantagens e limitações, dependendo das necessidades específicas do projeto de sensoriamento remoto.

No processo de classificação de imagens, particularmente na classificação por região e na classificação orientada a objetos, uma etapa prévia à classificação é essencial e a segmentação da imagem. Existem muitos métodos de segmentação, como *Watershed*, que é uma técnica de segmentação de imagens baseada em morfologia matemática (VINCENT; SOILLE, 1991); *K-Means*, que é uma técnica de agrupamento e que também pode ser usada para a segmentação de imagens, sendo a ideia básica dividir os pixels da imagem em K grupos (ou segmentos), onde cada pixel é atribuído ao grupo com a média mais próxima (no espaço de cores ou espectral) (CELEBI et al. 2013); e o *MeanShift*, que é um algoritmo de segmentação não-paramétrico, amplamente usado no processamento de imagens de sensoriamento remoto para a delimitação de regiões de interesse na imagem (COMANICIU; MEER, 2002). Este algoritmo possui três principais parâmetros que devem ser ajustados. O primeiro é o detalhe espectral: este parâmetro controla o grau de similaridade espectral necessária para que dois pixels sejam considerados como pertencentes à mesma região. Quanto maior o valor, maior a variação espectral permitida dentro de uma região. O segundo é o detalhe espacial: este parâmetro controla o grau de proximidade espacial necessária para



que dois pixels sejam considerados como pertencentes à mesma região. Quanto maior o valor, maior a dispersão espacial permitida dentro de uma região. O terceiro é o tamanho mínimo das regiões: este parâmetro determina o tamanho mínimo que uma região deve ter para ser mantida como uma entidade distinta, sendo as regiões menores do que esse tamanho fundido com regiões vizinhas (COMANICIU; MEER, 2002).

Neste trabalho, optou-se por segmentar as imagens pelo *MeanShift* e, após vários testes empíricos, definiu-se seguintes valores para cada um dos seus parâmetros. No detalhe espectral, considerou-se o valor 17, que varia de 1 a 20; para o detalhe espacial, utilizou-se o valor 10, que varia de 1 a 20; e o tamanho mínimo dos segmentos, em pixel, foi de 150, assim, como se trata de uma ortofoto de 1m de resolução, cada região terá pelo menos 150 m<sup>2</sup>. Após a aplicação do *MeanShift*, cada segmento ou "região" da imagem é tratado como uma unidade básica para análises posteriores. Neste trabalho, realizou-se uma classificação por regiões, portanto, estas foram classificadas com base nas características espectrais médias dos segmentos gerados pelo *MeanShift*, isto é, o vetor médio de características espectrais foi usado como entrada para os classificadores *Random Forest* ou *Support Vector Machine*. A Figura 3 apresenta parte da imagem RGB (a orto\_yavi1no) antes e o depois da segmentação pelo método *MeanShift*.

**Figura 3.** Representação do resultado da segmentação da imagem RGB, especificamente a orto\_yavi1no, utilizando o algoritmo *MeanShift*.



Fonte: Autores.

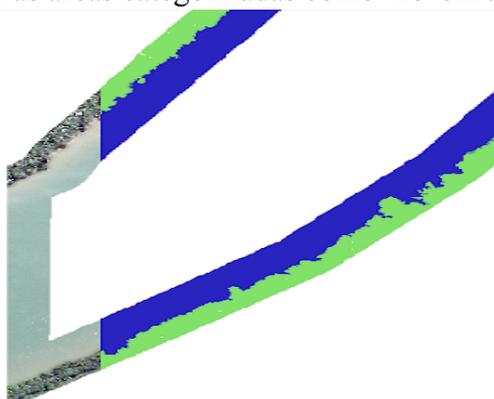
### Treinamento dos algoritmos de aprendizagem de máquina

Cada algoritmo de aprendizado de máquina supervisionado foi treinado cinco vezes no GIS ArcGIS Pro 2.8, usando as regiões segmentadas pelo *MeanShift*. Isso permitiu a análise da média de desempenho dos modelos por meio da matriz de confusão. Para o treinamento do algoritmo *Random Forest* (RF), foram considerados valores para os parâmetros de número máximo de árvores e profundidade máxima de árvores (BREIMAN, 2001). Os valores utilizados foram 400 árvores e profundidade máxima de 50. No entanto, é importante notar que esses valores podem variar dependendo do problema que está sendo resolvido, bem como das características e do volume dos dados disponíveis. O RF é um algoritmo que combina vários preditores baseados em árvores de decisão. Cada árvore é treinada em um subconjunto distinto dos dados de treinamento, criado por amostragem *bootstrap*, que seleciona aleatoriamente uma amostra dos dados com substituição. Assim, cada árvore tem uma visão



ligeiramente diferente dos dados. O RF, então, faz previsões baseadas na combinação (média, no caso de regressão; moda, no caso de classificação) das previsões das árvores individuais (BREIMAN, 2001). A Figura 4 ilustra o resultado do treinamento com RF, representando a classe 'rio' na cor azul, e a classe 'não rio' na cor verde.

**Figura 4.** Apresentação qualitativa dos resultados obtidos pelo algoritmo *Random Forest* nos dados de treinamento, com as áreas categorizadas como 'rio' em azul e 'não rio' em verde.



Fonte: Autores.

Para o treinamento do algoritmo *Support Vector Machine* (SVM), este tenta encontrar um hiperplano que melhor separe as classes de dados (“rio” x “não rio”) em um espaço de características  $n$ -dimensional, onde  $n$  é o número de características de cada amostra (CORTES; VAPNIK, 1995). No estudo de caso, cada amostra é um polígono gerado pelo *MeanShift*, e as características incluem a média espectral e a cor convergida de cada polígono, portanto, um espaço bidimensional. Cada polígono é representado como um ponto neste espaço bidimensional, onde os valores da média espectral e da cor convergida são tratados como coordenadas. O SVM utiliza funções kernel, como o kernel linear ou o *Radial Basis Function* (RBF), para transformar o espaço de características e traçar um hiperplano que separe as classes de dados (neste estudo, 'rio' e 'não rio') com a maior margem possível. A classificação é então realizada com base na posição relativa dos polígonos em relação a este hiperplano (CORTES; VAPNIK, 1995). No ArcGIS Pro é utilizado o RBF como função kernel padrão para o algoritmo SVM na análise de classificação de imagens.

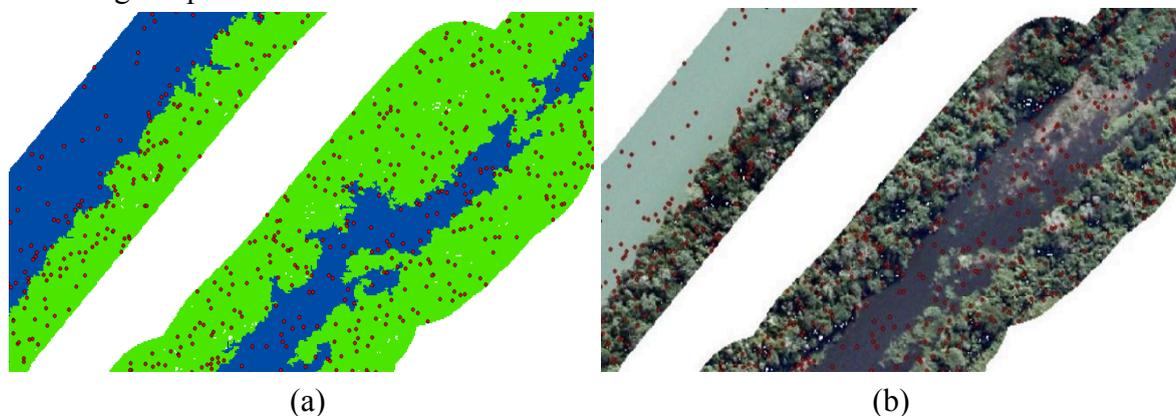
### Validação dos algoritmos de aprendizagem de máquina

A performance dos algoritmos supervisionados, RF e SVM, foi avaliada tanto nos dados de treinamento quanto nos dados de teste. Para essa avaliação, foram utilizados 10.000 pontos distribuídos ao longo da área de estudo (Figura 5), seguindo uma estratégia de amostragem estratificada proporcional ao tamanho das classes 1 e 2. Métricas de desempenho, para tarefas de classificação, como acurácia, sensibilidade (ou recall), precisão e medida F1 - score, foram calculadas a partir da matriz de confusão (Figura 6).

**Figura 5.** Ilustração da distribuição de pontos de validação usados para avaliar o desempenho dos algoritmos nos dados de teste, apresentada sobre a imagem RGB. Em (a), os pontos de



validação são distribuídos sobre a imagem RGB que foi classificada usando o algoritmo SVM. Em (b), os mesmos pontos de validação são mostrados, mas sobrepostos à imagem RGB original para referência.



Fonte: Autores.

Figura 6. Esquema ilustrativo da matriz de confusão.

		Detectada	
		Sim	Não
Real	Sim	Verdadeiro Positivo (VP)	Falso Negativo (FN)
	Não	Falso Positivo (FP)	Verdadeiro Negativo (VN)

Fonte: Autores.

Verdadeiros Positivos (VP): número de feições corretamente classificadas como pertencentes à classe "rio". Falsos Negativos (FN): feições que foram incorretamente classificadas como pertencentes à classe "não rio", mas que na realidade são da classe "rio". Falsos Positivos (FP): feições que foram incorretamente classificadas como pertencentes à classe "rio", mas que na realidade são da classe "não rio". Verdadeiros Negativos (VN): número de feições corretamente classificadas como pertencentes à classe "não rio".

A acurácia é uma métrica utilizada para avaliar a performance de um modelo de classificação. Ela representa a proporção de predições corretas (tanto positivas quanto negativas) em relação ao total de observações (SOKOLOVA; LAPALME, 2009). Em outras palavras, é a razão entre o número total de predições corretas e o número total de predições. Para calcular a acurácia, adiciona-se o número de verdadeiros positivos e verdadeiros negativos e divide-se pelo total de observações (SOKOLOVA; LAPALME, 2009).

$$\text{acurácia} = \frac{VP + VN}{VP + FN + VN + FP}$$

A sensibilidade (também conhecida como *recall* ou revocação) é uma métrica que avalia a proporção de Verdadeiros Positivos (VP) em relação à soma de Verdadeiros Positivos (VP) e Falsos Negativos (FN). Essa métrica indica a porcentagem de resultados positivos



corretamente identificados pelo modelo dentre todos os resultados que deveriam ter sido identificados como positivos. Se um classificador é perfeito, sua precisão é igual ao seu recall, e ambos têm valor 1 (SOKOLOVA; LAPALME, 2009).

$$sensibilidade = \frac{VP}{VP + FN}$$

A precisão (*Precision*) é uma métrica que avalia a proporção de Verdadeiros Positivos (VP) em relação à soma de Verdadeiros Positivos (VP) e Falsos Positivos (FP). Essa métrica mede a acurácia das classificações positivas feitas pelo modelo, indicando a porcentagem de identificações corretas de positivos dentre todas as classificações de positivos realizadas pelo modelo (SOKOLOVA; LAPALME, 2009).

$$precisão = \frac{VP}{VP + FP}$$

A medida F1-score é uma média harmônica entre precisão e revocação, fornecendo um único valor que indica a qualidade geral do modelo. Esta métrica é particularmente útil para conjuntos de dados com classes desproporcionais, pois considera tanto falsos positivos quanto falsos negativos na sua cálculo (SOKOLOVA; LAPALME, 2009). O F1-score varia entre 0 e 1, sendo que um valor de 1 indica uma precisão e revocação perfeitas. A métrica penaliza modelos que possuem um desempenho muito ruim em qualquer uma dessas duas medidas (precisão ou revocação) (SOKOLOVA; LAPALME, 2009).

$$f1 = 2 * \frac{precisão * sensibilidade}{precisão + sensibilidade}$$

A avaliação de desempenho é fundamental em tarefas de classificação para garantir a eficácia do modelo. Métricas como acurácia, precisão, recall e F1-score fornecem uma visão detalhada do desempenho do modelo, ajudando a entender a qualidade das previsões positivas, a capacidade de evitar falsos negativos e a proporção de acertos em geral, além de fornecer uma avaliação equilibrada em situações de classes desproporcionais (SOKOLOVA; LAPALME, 2009).

## RESULTADOS E DISCUSSÃO

Os algoritmos *Random Forest* e *Support Vector Machine* demonstraram um desempenho elevado na tarefa de classificação de rios em imagens RGB de alta resolução espacial. Como esperado, o desempenho dos modelos foi superior no conjunto de treinamento em comparação com o conjunto de teste (Tabelas 2 e 3). Isso se deve ao fato de que os algoritmos foram treinados nesse conjunto e, portanto, ajustaram seus parâmetros para minimizar o erro nele. Contudo, para garantir uma avaliação justa e evitar *overfitting*, os modelos foram validados em um conjunto de dados distinto do utilizado para treinamento, isto é, foram validados no conjunto de teste, proporcionando uma estimativa mais realista de seu desempenho.

**Tabela 2.** Desempenho dos algoritmos de aprendizado de máquina no conjunto de treinamento para a classificação de cursos d'água em imagens RGB.

Método	Treino	Iterações	Acurácia	F1-Score	Precisão	Recall
<i>Random Forest</i>	orto_yavi1no	1º	0,9018	0,9100	0,8398	0,9930
		2º	0,9018	0,9100	0,8398	0,9930
		3º	0,9019	0,9101	0,8400	0,9930
		4º	0,8982	0,9070	0,8348	0,9929
		5º	0,9037	0,9116	0,8428	0,9926
		<b>Média</b>	0,901	0,910	0,8390	0,9930
<i>Support Vector Machine</i>	orto_yavi1no	1º	0,9011	0,9083	0,8464	0,9800
		2º	0,9095	0,9164	0,8519	0,9914
		3º	0,9095	0,9164	0,8519	0,9914
		4º	0,9095	0,9164	0,8519	0,9914
		5º	0,9095	0,9164	0,8519	0,9914
		<b>Média</b>	0,908	0,915	0,851	0,989

**Tabela 3.** Performance dos algoritmos de aprendizado de máquina no conjunto de testes.

Método	Teste	Acurácia	F1-Score	Precisão	Recall
<i>Random Forest</i>	orto_bi1so	0,7915	0,8264	0,7079	0,9924
	orto_bvi2so	0,8453	0,8664	0,7670	0,9954
	yaiii4no	0,7909	0,8260	0,7077	0,9918
	<b>Média</b>	0,809	0,840	0,728	0,9930
<i>Support Vector Machine</i>	orto_bi1so	0,8785	0,8886	0,8203	0,9694
	orto_bvi2so	0,9196	0,9246	0,8767	0,9780
	yaiii4no	0,8775	0,8897	0,8155	0,9788
	<b>Média</b>	0,892	0,901	0,838	0,975

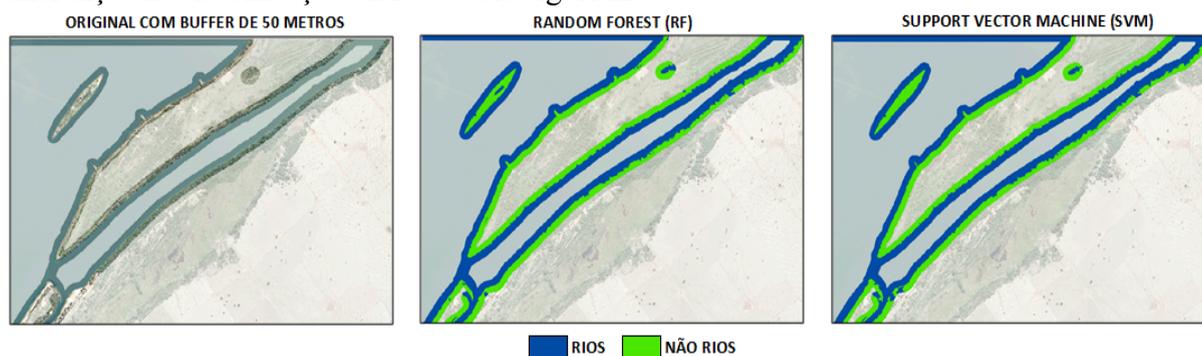
A análise quantitativa do desempenho dos algoritmos nos dados de teste (Tabela 3) destaca o SVM como o modelo mais eficaz para a tarefa proposta, demonstrando um F1-score médio superior a 90%. Esse resultado está alinhado com estudos anteriores de classificação de imagens de alta resolução espacial. Por exemplo, Silva et al. (2020), ao classificar o uso e a ocupação do solo utilizando imagens AisaFENIX com resolução de 1,1 m, obtiveram desempenhos mais satisfatórios com o SVM, que exibiu valores do índice Kappa e da acurácia global superiores aos obtidos pelo RF.

A Tabela 3 revela que o principal desafio enfrentado pelo RF foi o mapeamento incorreto de áreas de fundo (ex: "não rio") como "rio", provavelmente devido à similaridade na resposta espectral entre essas classes. Este fator resultou em uma grande quantidade de falsos positivos, refletida pelo baixo valor de precisão. Esta tendência para falsos positivos na classificação de rios com RF também foi observada por Bazzan et al. (2022), onde o algoritmo erroneamente classificou como água sombras de edificações verticalizadas, causando a superestimação da classe água em áreas urbanas. De forma similar, Cruz e

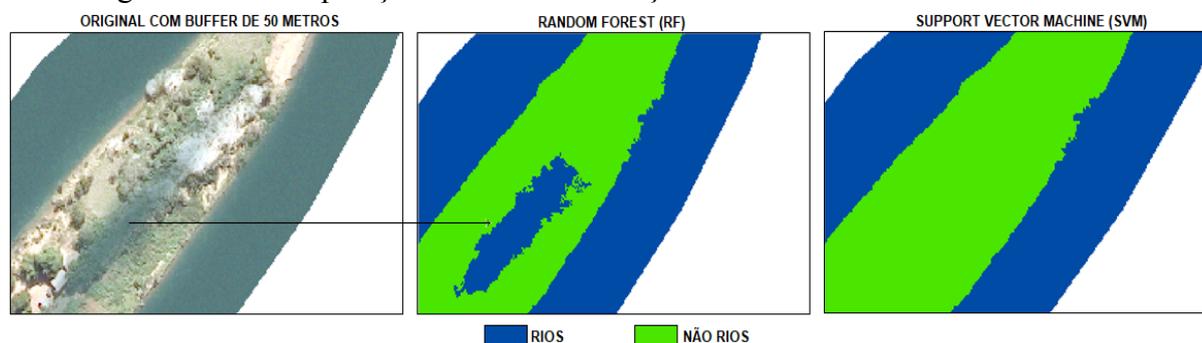


Oliveira (2021) notaram uma superestimação da presença de corpos hídricos na bacia hidrográfica pelo algoritmo, além de erros na classificação em áreas aquáticas de menor dimensão e nas bordas do Lago de Furnas, com áreas queimadas sendo erroneamente mapeadas como corpos hídricos. A resposta espectral escura destas áreas, semelhante à dos corpos hídricos na composição usada para o estudo, foi apontada como causa destes falsos positivos. Contudo, é possível reduzir ou até eliminar estes falsos positivos na classificação de rios ao aumentar as variáveis preditoras e coletar mais amostras durante o processo de classificação da imagem (Cruz e Oliveira, 2021; Bazzan et al., 2022). O RF, assim como o SVM, demonstrou um bom desempenho no mapeamento da classe de maior interesse, "rio". Consequentemente, gerou poucos falsos negativos e atingiu um alto valor de sensibilidade (recall), como ilustrado na Tabela 3. Isso indica a sua capacidade de corretamente identificar a classe "rio", minimizando a possibilidade de deixar de mapeá-la.

**Figura 7.** Representação visual dos resultados obtidos a partir da classificação das imagens RGB, utilizando os algoritmos *Random Forest* e *Support Vector Machine*. A figura ilustra as diferenças na classificação entre os dois algoritmos.



**Figura 8.** Representação gráfica da análise de falsos-negativos ocorridos durante a classificação das imagens RGB. A figura ilustra as localizações e/ou a frequência dos falsos-negativos em comparação com as classificações corretas.



Em situações de complexidade elevada, como a presença de “pequenas ilhas” (Figura 8), o algoritmo *Random Forest* superou o *Support Vector Machine* na redução de falsos-negativos, conforme pode ser visualizado na análise qualitativa (Figuras 7 e 8). Este



aspecto do RF é de grande relevância para o estudo em questão, uma vez que um modelo com alta incidência de falsos-negativos subestimaria a quantidade de recursos hídricos disponíveis na área mapeada. A comparação dos valores de recall, com 99,3% para o RF e 98,9% para o SVM, comprova numericamente essa superioridade (Tabela 3). Presumindo a utilização deste mapeamento para determinar locais de coleta de amostras de água em campo, por exemplo, áreas marcadas como falsos-negativos poderiam ser erroneamente desconsideradas pela equipe.

O presente estudo forneceu evidências da eficácia dos algoritmos *Random Forest* e *Support Vector Machine* no mapeamento de corpos hídricos de médio porte a partir de imagens RGB de alta resolução espacial. Embora o SVM tenha demonstrado um desempenho superior no que diz respeito ao F1-score médio, o RF se destacou na capacidade de mapear áreas complexas com precisão, a exemplo das "ilhas", resultando em menos falsos-negativos. Este atributo é fundamental para este tipo de mapeamento, onde a subestimação de recursos hídricos poderia ter implicações práticas, como a inadequada seleção de locais de coleta de amostras de água. Entretanto, o RF enfrentou desafios ao classificar áreas não-riverinas como "rio", ocasionando um número elevado de falsos-positivos. Acreditamos que aumentar as variáveis preditoras e coletar mais amostras durante o processo de classificação da imagem possam ser estratégias promissoras para mitigar esse problema.

Por fim, os resultados deste estudo reforçam a relevância da escolha cuidadosa do algoritmo em função das especificidades da tarefa de classificação. Estudos futuros poderiam beneficiar-se de experimentos com mais dados de treinamento e utilização de infraestrutura computacional de alta capacidade para aprimorar ainda mais o desempenho desses algoritmos.

## CONSIDERAÇÕES FINAIS

Neste estudo, foi realizada uma análise quali-quantitativa do desempenho dos algoritmos de aprendizado de máquina supervisionado, *Random Forest* (RF) e *Support Vector Machine* (SVM), para mapear cursos d'água em imagens RGB de alta resolução espacial. Ambos os algoritmos demonstraram alta eficácia, apesar da similaridade espectral significativa entre a feição de interesse (rios) e outras classes de uso do solo. O SVM apresentou um desempenho geral superior, refletido pelo F1-score mais elevado. No entanto, em cenários de maior complexidade, como a presença de áreas secas dentro das feições de rios, o RF foi superior ao SVM, gerando menos falsos-negativos.

Destaca-se a necessidade de estudos futuros voltados para o mapeamento de rios mais estreitos (por exemplo, até 10 m de largura) ou aqueles com indícios de assoreamento. A abordagem aqui apresentada tem potencial para ser aplicada em diferentes regiões geográficas, fornecendo informações geoespaciais em larga escala para apoiar projetos na área ambiental. Também é relevante investigar estratégias para mitigar a tendência do RF em gerar falsos-positivos, como o aumento das variáveis preditoras e a coleta de mais amostras durante o processo de classificação da imagem.

## AGRADECIMENTOS

À CAPES (Coordenação de Aperfeiçoamento de Pessoal de Nível Superior e Aperfeiçoamento de Pessoal de Nível Superior) pela bolsa de mestrado ao primeiro autor, e à Universidade do Oeste Paulista (UNOESTE).



## REFERÊNCIAS

- BAZZAN, T.; RENNÓ, C. D.; LIMA, D. L. C.; RECKZIEGEL, E. W. Integração da incerteza na amostragem e classificação Random Forest utilizando bandas e índices espectrais para o mapeamento de inundação. *Revista Geociências*, v. 41, n. 4, 2022. ISSN: 1980-900X. DOI: <https://doi.org/10.5016/geociencias.v41i04.16802>.
- BRASIL. **Lei nº 9.433, de 8 de janeiro, de 1997**. Institui a Política Nacional de Recursos Hídricos, cria o Sistema Nacional de Gerenciamento de Recursos Hídricos, regulamenta o inciso XIX do art. 21 da Constituição Federal, e altera o art. 1º da Lei nº 8.001, de 13 de março de 1990, que modificou a Lei nº 7.990, de 28 de dezembro de 1989.
- BREIMAN, L. Random forests. *Machine Learning*, v. 45, n. 1, p. 5-32, 2001. doi:10.1023/A:1010933404324.
- CELEBI, M. E.; KINGRAVI, H. A.; VELA, P. A. A comparative study of efficient initialization methods for the k-means clustering algorithm. *Expert Systems with Applications*, v. 40, n. 1, p. 200-210, 2013.
- COMANICIU, D.; MEER, P. Mean shift: a robust approach toward feature space analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, v. 24, n. 5, p. 603-619, 2002.
- CORTES, C.; VAPNIK, V. Support-vector networks. *Machine Learning*, v. 20, n. 3, p. 273-297, 1995.
- CRUZ, U. R. X.; OLIVEIRA, L. P. Comparativo entre os métodos de classificação MaxVer e Random Forest utilizando imagem Sentinel-2B. *Cadernos do Leste, Belo Horizonte*, v. 21, n. 21, p. [inserir as páginas], jan.-dez. 2021. ISSN 1679-5806.
- EMPRESA PAULISTA DE PLANEJAMENTO METROPOLITANO SA. Produtos cartográficos: projeto mapeia São Paulo, Ortofotos Digitais. São Paulo: Emplasa, 2014.
- GOODFELLOW, Ian; BENGIO, Yoshua; COURVILLE, Aaron. **Deep Learning**. (Adaptive Computation and Machine Learning). Cambridge: MIT Press, 2016. ISBN 978-0262035613.
- JAIN, P.; COOGAN, S. C. P.; SUBRAMANIAN, S. G.; CROWLEY, M.; TAYLOR, S.; FLANNIGAN, M. D. A review of machine learning applications in wildfire science and management. *Environmental Reviews*, v. 28, n. 4, p. 478-505, 2020. Disponível em: <https://doi.org/10.1139/er-2020-0019>.
- JENSEN, R. J. **Remote Sensing of the Environment: An Earth Resource Perspective**. 2. ed. Harlow: Pearson New International Edition, Pearson Education Limited Edinburgh Gate, 2014. 619 p.
- MARTINS, S.; BERNARDO, N.; OGASHAWARA, I. et al. Support Vector Machine algorithm optimal parameterization for change detection mapping in Funil Hydroelectric Reservoir (Rio de Janeiro State, Brazil). *Modeling Earth Systems and Environment*, v. 2, n. 138, 2016.
- NASCIMENTO, R. et al. An assessment of the Support Vector Machine for a CBERS-2 CCD image classification: a case study of a tropical reservoir in Brazil. *Revista Brasileira de Cartografia*, v. 65, n. 3, p. 431-439, 2013.
- OSCO, Lucas Prado et al. A review on deep learning in UAV remote sensing. *International Journal of Applied Earth Observation and Geoinformation*, v. 102, p. 102456, 2021.
- SILVA, L.; RAMBO, E.; BECKER, W. R.; COMINETI, H.; SILVA JÚNIOR, C. A.; JOHANN, J. Mineração de dados aplicada a imagens hiperespectrais e multiespectrais na classificação do uso e ocupação do solo. 2020. In: CONGRESSO BRASILEIRO DE ENGENHARIA AGRÍCOLA. Online: Editora, 2020. ISBN 978-65-87729-02-2.
- SOKOLOVA, M.; LAPALME, G. A systematic analysis of performance measures for classification tasks. *Information Processing & Management*, v. 45, n. 4, p. 427-437, 2009.
- TOMASONI, M.; PINTO, J.; SILVA, H. A questão dos recursos hídricos e as perspectivas para o Brasil. *GeoTextos*, Salvador, v. 5, n. 2, p. 107-127, dez. 2009.
- VINCENT, L.; SOILLE, P. Watersheds in digital spaces: an efficient algorithm based on immersion simulations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, v. 13, n. 6, p. 583-598, 1991.



JIANG, Wei et al. Multilayer Perceptron Neural Network for Surface Water Extraction in Landsat 8 OLI Satellite Images. *Remote Sensing*, [s.l.], v. 10, n. 5, p. 755, 2018.

### COMO CITAR ESTE TRABALHO

SOUZA, Mariany Kerriany Gonçalves de; PINHEIRO, Mayara Maezano Frita; FURUYA, Danielle Elis Garcia; OSO, Lucas Prado; JÚNIOR, José Marcato; GONÇALVES, Wesley Nunes; RAMOS, Ana Paula Marques. Mapeamento de rios em imagem RGB com aprendizagem de máquina supervisionada. *Revista Tamoios*, São Gonçalo, v. 20, n. 1, p. 87-101, 2024. Disponível em: <https://doi.org/10.12957/tamoios.2024.66545>. Acesso em: DD MM. AAAA.