ISSN 2675-004X
FATEC Rubens Lara
Santos/SP/BR
Fatecrl.edu.br/revistaconecta

Mais verde ou menos verde? Quando alta acurácia esconde muitos erros: uma análise da classificação de vegetação arbórea com SVM

More green or less green? When high overall accuracy hides significant errors: a critical analysis of SVM for arboreal vegetation classification

Daniel dos Santos Messa
Fatec Adamantina/BR

Paulo Roberto da Silva Ruiz
Fatec Adamantina/BR

Luiz Gustavo Teixeira
Fatec Adamantina/BR

Resumo

Este trabalho teve como objetivo desenvolver um método para classificação de cobertura arbórea no município de Adamantina - SP, utilizando imagens do satélite CBERS-4A e o algoritmo *Support Vector Machine* (SVM), objetivando subsidiar ações de gestão ambiental municipal. A metodologia compreendeu aquisição e pré-processamento de imagens, definição e obtenção de amostras de treinamento, extração de atributos espectrais, além da seleção e otimização de atributos. Os resultados demonstraram alta acurácia global (96,96%) e concordância substancial, com Índice *Kappa* de 0,749, com os índices NDVI e EVI emergindo como os atributos mais relevantes. Contudo, a análise detalhada revelou assimetria no desempenho do classificador: enquanto a classe "Outros" apresentou excelente precisão, a classe "Arbórea" registrou *recall* moderado, indicando sub identificação de aproximadamente um terço da vegetação arbórea. Conclui-se que o método se mostrou viável para aplicações de planejamento municipal em escala, porém com limitações significativas na detecção completa da cobertura arbórea. A pesquisa contribui com uma metodologia de baixo custo para monitoramento ambiental, destacando a importância da análise crítica de métricas da acurácia global em classificações automáticas de Sensoriamento Remoto.

Palavras-chave: aprendizado de máquina; CBERS 4A; índices de vegetação; sensoriamento remoto.

Abstract

This study aimed to develop a method for classifying arboreal coverage in the municipality of Adamantina, São Paulo State, Brazil, using images from the CBERS-4A satellite and the Support Vector Machine (SVM) algorithm, with the goal of supporting municipal environmental management actions. The methodology included image acquisition and preprocessing, definition and collection of training samples, extraction of spectral attributes, as well as attribute selection and optimization. The results demonstrated high overall accuracy (96.96%) and substantial agreement, with a Kappa Index of 0.749, while the NDVI and EVI

ISSN 2675-004X
FATEC Rubens Lara
Santos/SP/BR
Fatecrl.edu.br/revistaconecta

indices emerged as the most relevant attributes. However, detailed analysis revealed an asymmetry in classifier performance: while the "Other" class showed excellent precision, the "Arboreal" class recorded moderate recall, indicating the sub-identification of approximately one-third of the arboreal vegetation. It is concluded that the method proved feasible for large-scale municipal planning applications, but with significant limitations in the complete detection of arboreal coverage. The research contributes a low-cost methodology for environmental monitoring, emphasizing the importance of critical analysis of overall accuracy metrics in automated Remote Sensing classifications.

Keywords: CBERS-4A; machine learning; remote sensing; vegetation indices.

I. INTRODUÇÃO

O Sensoriamento Remoto (SR) tem-se consolidado como uma das principais ferramentas para aquisição de dados da superfície terrestre, possibilitando a análise, monitoramento e gestão dos recursos naturais em diferentes escalas espaciais e temporais. Através da captação de informações por meio de sensores a bordo de satélites, aviões ou drones, é possível observar áreas extensas e de difícil acesso, com elevada frequência e riqueza de dados espectrais (Novo, 2010).

Segundo Moreira (2011) e Anderson (1976), o uso de imagens de satélite, em particular, revolucionou o modo como pesquisadores, gestores públicos e empresas privadas lidam com informações ambientais e territoriais. Essas imagens permitem a obtenção de dados com maior precisão e atualização, os quais são fundamentais para o planejamento urbano, agricultura de precisão, gestão de bacias hidrográficas, análise de desmatamento e mudanças no uso e cobertura de terra, dentre outras aplicações.

A importância do SR também está relacionada à sua capacidade de integrar tecnologias de geoprocessamento, como os Sistemas de Informação Geográfica (SIGs), proporcionando análises espaciais mais completas e tomadas de decisão mais assertivas (Câmara; Souza; Freitas, 1996). Além disso, o uso de algoritmos de Aprendizado de Máquina (AM) proporciona a automatização na extração de informação, tornando essa etapa rápida e acessível aos tomadores de decisão em diferentes âmbitos institucionais.

O Governo do Estado de São Paulo, desde 2007, por meio da Secretaria de Estado do Meio Ambiente, atualmente Secretaria de Meio Ambiente, Infraestrutura e Logística, atua com o programa Município Verde Azul (PMVA) com o propósito de mediar e apoiar a eficiência da gestão ambiental, com a descentralização e valorização da agenda ambiental nos municípios, auxiliando os mesmos em estratégias para o desenvolvimento sustentável do Estado de São Paulo (Ambiente, 2025). O município de Adamantina recebeu o certificado em todas as edições

ISSN 2675-004X
FATEC Rubens Lara
Santos/SP/BR
Fatecrl.edu.br/revistaconecta

do programa, ficando de fora apenas na edição de 2016 (Adamantina, 2020). Além do certificado, essa preocupação com o meio ambiente no munícipio, contribui para realizar uma melhor gestão dos espaços, onde tem-se a noção do quanto e quais áreas verdes existem na cidade, podendo cadenciar suas manutenções e desenvolvimento, melhora da qualidade de vida dos moradores da cidade com a arborização urbana e diminuição da sensação térmica em épocas onde as temperaturas ficam mais elevadas.

Diante deste contexto, esta pesquisa tem como objetivo principal desenvolver e avaliar um método para classificação de cobertura arbórea no município de Adamantina-SP, utilizando imagens do satélite CBERS-4A e o algoritmo *Support Vector Machine* (SVM). Este trabalho se justifica pela necessidade de fornecer à gestão municipal ferramentas acessíveis para o planejamento urbano ambiental, em especial para o programa Município Verde Azul, considerando que o mapeamento detalhado da arborização urbana representa um grande desafio técnico. O trabalho adotou uma abordagem de aprendizado de máquina supervisionado, compreendendo etapas de aquisição e pré-processamento dos dados, extração e seleção de atributos espectrais, otimização de hiperparâmetros do algoritmos e validação dos resultados mediante métricas de acurácia, matriz de confusão e índice *Kappa*. Com isso, busca-se responder às seguintes questões: O SVM, quando alimentado com dados do CBERS-4A, é capaz de identificar a cobertura arbórea com a precisão necessária para aplicações de gestão municipal? Quais são os principais erros de classificação e suas possíveis causas?

Esta pesquisa torna-se importante ao demonstrar a viabilidade de uma metodologia de baixo custo para o monitoramento ambiental urbano, oferecendo subsídios tanto para a academia, no avanço das técnicas de SR, quanto para a sociedade, por meio de uma gestão pública mais eficiente na promoção da qualidade ambiental urbana.

2. FUNDAMENTAÇÃO TEÓRICA

2.1. Sensoriamento Remoto de alta resolução espacial

O SR de alta resolução espacial tem desempenhado um importante papel na análise detalhada da superfície terrestre. Segundo Jensen (2011), a resolução espacial representa o menor objeto detectável em uma imagem de satélite, sendo um fator determinante na qualidade da interpretação dos dados. Além disso, a resolução espectral define a capacidade do sensor de discriminar diferentes comprimentos de onda, possibilitando a identificação de materiais distintos com base em sua assinatura espectral. Já a resolução temporal refere-se à frequência

ISSN 2675-004X
FATEC Rubens Lara
Santos/SP/BR
Fatecrl.edu.br/revistaconecta

com que um sensor revisita a mesma área, o que é essencial para o monitoramento contínuo de mudanças ambientais e urbanas (Campbell; Wynne, 2011). Por fim, a resolução radiométrica está relacionada à sensibilidade do sensor em detectar variações na intensidade da radiação refletida, sendo um fator relevante para a precisão dos dados capturados (Schowengerdt, 2007).

Os avanços na tecnologia de SR permitiram o desenvolvimento de satélites de alta resolução espacial, como o CBERS 4A, proporcionando uma melhor definição de detalhes para aplicações urbanas e ambientais. Trata-se de um satélite brasileiro, parte do programa *China-Brazil Earth Resources Satellite* (CBERS), desenvolvido em colaboração com a China. Lançado em dezembro de 2017, o CBERS 4A é uma das mais recentes adições à série de satélites desse programa. Ele está equipado com sensores de alta performance para observação da Terra, com ênfase na coleta de dados sobre o uso da terra, cobertura vegetal, monitoramento de desmatamento e desenvolvimento urbano (INPE, 2020).

O satélite possui três câmeras que operam em diferentes bandas espectrais, incluindo a *Wide Field Panoramic Multiespectral* (WPM), a qual oferece uma resolução espacial de 2 metros, a *Multi-Spectral Scanner* (MUX) com uma resolução de 16,5 metros, além da *Wide Field Imager* (WFI), que possui 55 metros de resolução espacial. A WPM é útil para estudos urbanos, monitoramento agrícola e da vegetação, enquanto a MUX contribui para a análise de mudanças no uso do solo em áreas de grande extensão, já a WFI possui a ampla capacidade de revisita. Além disso, oferece uma cobertura de 250 km por imagem, o que é vantajoso para a captura de grandes áreas com boa definição (INPE, 2020).

Os dados do CBERS 4A são amplamente utilizados em projetos de monitoramento ambiental e gestão urbana, possibilitando a análise detalhada da dinâmica territorial, identificação de áreas de risco, planejamento de infraestrutura e preservação ambiental. A alta resolução das imagens coletadas permite uma análise precisa e detalhada da superfície terrestre, sendo um recurso essencial para estudos em diversas áreas, como o monitoramento de florestas, planejamento urbano, agricultura e gestão de desastres (Huang; Gong; Biging, 2020).

A aquisição de imagens orbitais de alta resolução espacial representa um desafio econômico significativo no planejamento de projetos de SR, restringindo o acesso a esses dados para projetos de grande escala. No caso de satélites comerciais como o *WorldView-3*, operado pela *Maxar Technologies*, uma única cena pode superar a faixa de milhares de dólares, dependendo da extensão da área solicitada e da necessidade de programação de coleta.

ISSN 2675-004X
FATEC Rubens Lara
Santos/SP/BR
Fatecrl.edu.br/revistaconecta

De acordo com Woodhouse (2006), o equilíbrio entre custo, resolução espacial e cobertura temporal constitui um dos principais desafios para o uso eficiente de dados de SR orbital. Assim, embora as imagens comerciais sejam tecnicamente superiores, seu alto custo pode limitar a sua adoção, especialmente em projetos com restrições orçamentárias, favorecendo o uso de alternativas gratuitas, como as imagens do CBERS 4A, em aplicações que toleram menor resolução espacial.

2.2. Sensoriamento Remoto urbano

O SR tem sido empregado para o estudo e planejamento das áreas urbanas, fornecendo dados para a gestão de cidades e análise da expansão urbana. O uso de imagens de satélite na análise urbana permite a identificação de padrões de ocupação do solo, contribuindo para um melhor entendimento das dinâmicas das cidades. Essa abordagem possibilita a identificação de edificações, áreas verdes, vias de circulação e outros elementos importantes para a organização do espaço urbano (Lin *et al*., 2021).

Além disso, o SR tem sido utilizado no planejamento de infraestrutura urbana, auxiliando na tomada de decisão em projetos de mobilidade, drenagem urbana e expansão habitacional (Tomasiello, 2016). Outros usos utilizando SR podem ser para análise da expansão urbana em metrópoles usando dados de alta resolução (Taubenböck *et al*., 2012) e uso de imagens aéreas e visão computacional para identificação de espécies de árvores (Branson *et al*., 2019).

2.3. Classificação de imagens

A classificação automática de imagens é um importante processo para a extração de informações, permitindo a segmentação de áreas conforme diferentes categorias de cobertura e uso do solo. Segundo Lillesand, Kiefer, Chipman (2015), os métodos de classificação podem ser supervisionados ou não supervisionados, dependendo da necessidade de treinamento do algoritmo.

A classificação supervisionada por regiões inicia-se com a segmentação, onde o algoritmo irá reunir pixels espectralmente semelhantes a partir de um limiar pré-definido. A seguir são definidas as classes de cobertura do solo e coletadas as amostras que serão utilizadas como base de conhecimento para o treinamento do algoritmo de classificação. A distribuição espacial das amostras é outro aspecto fundamental, Congalton e Green (2019) ressaltam que amostras de treinamento espacialmente dispersas ajudam a reduzir o viés espacial e garantem que o classificador aprenda padrões que são consistentes em toda a área de estudo. Dessa forma,

ISSN 2675-004X
FATEC Rubens Lara
Santos/SP/BR
Fatecrl.edu.br/revistaconecta

a coleta de dados não deve se concentrar apenas em áreas de fácil acesso ou visualmente distintas, mas deve abranger toda a variação espacial da paisagem.

Após a coleta, é prática comum dividir as amostras em dois subconjuntos: um para o treinamento do algoritmo e outro para a validação independente. Essa divisão permite avaliar a performance da classificação por meio de métricas como a acurácia global, o Índice *Kappa* e a matriz de confusão (Congalton; Green, 2019).

Um dos algoritmos utilizados no treinamento é o *Support Vector Machine* (SVM), em português Máquina de Vetor de Suporte, o qual busca encontrar um hiperplano ótimo para maximizar a separação entre diferentes classes no espaço de características (Vapnik, 1999). Ele trabalha identificando a fronteira de decisão que melhor separa os pontos de diferentes classes, utilizando um conceito chamado de margem máxima. Quando os dados não são linearmente separáveis, ele pode usar funções *kernel*, como o *radial basis function* (RBF) e o polinomial, para projetar os dados em um espaço de dimensão superior, onde a separação se torna viável. Esse método tem demonstrado alta precisão em classificações complexas, sendo aplicado em estudos de uso e cobertura do solo devido à sua capacidade de generalização e robustez em diferentes condições ambientais (Mountrakis; Im; Ogole, 2011).

Conforme Vapnik (1999), o SVM pode ser descrito considerando um conjunto de treinamento composto por D amostras rotuladas, denotadas por $\{x_i, y_i\}$, onde $i = 1, 2, ..., D$. Cada amostra é representada por um vetor de características $x_i \in \mathbb{R}^M$, enquanto seu rótulo correspondente y_i pertence ao conjunto discreto $\{-1, +1\}$. Assume-se que esses dados são provenientes de uma distribuição de probabilidade subjacente e desconhecida, $P(x, y)$, da qual as amostras de treinamento são extraídas de forma independente. O objetivo central do processo de aprendizado é, portanto, induzir um classificador capaz de aprender o mapeamento $x \rightarrow y$ a partir dos exemplos fornecidos. A máquina resultante deve generalizar adequadamente, sendo capaz de prever corretamente o rótulo de um novo par (x, y), não observado durante o treinamento, desde que este seja gerado pela mesma distribuição P.

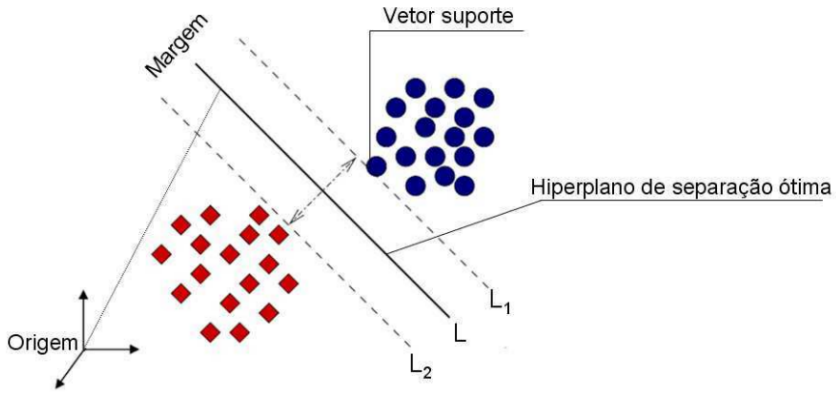
Uma característica fundamental que o distingue, prossegue o autor, é sua abordagem de minimização do risco esperado, representado por $\varepsilon(\varsigma)$, em uma tarefa de classificação. Esse risco é formalmente definido pela integral $\varepsilon(\varsigma) = \int (1/2) |y - f(x)| \, dP(x, y)$, que quantifica a expectativa de erro sob a distribuição verdadeira. Contudo, como $P(x, y)$ é desconhecida na prática, torna-se inviável calcular diretamente essa grandeza. Como alternativa, o SVM opera com base no risco empírico, $\varepsilon\_\psi(\varsigma)$, que corresponde à média da taxa de erro apurada

ISSN 2675-004X
FATEC Rubens Lara
Santos/SP/BR
Fatecrl.edu.br/revistaconecta

especificamente sobre o conjunto de treinamento disponível. Este risco empírico é calculado por $\varepsilon\_\psi(\varsigma) = (1/D) \sum\_\{i=1\}^D (1/2) |y_i - f(x_i)|$. Para um classificador hipotético ς e um conjunto de treinamento fixo, $\varepsilon\_\psi(\varsigma)$ assume um valor determinado.

Vapnik (1999) destaca que o critério de separação ótima de classes é materializado pela definição de um hiperplano de decisão (representado por *L* na Figura 1). A otimização deste hiperplano é orientada por dois princípios: (1) maximizar a margem de separação, que é a distância entre os hiperplanos de suporte de cada classe *(L_1 e L_2)*; e (2) assegurar que esse hiperplano ótimo seja equidistante dos pontos de dados mais próximos de cada classe, conhecidos como vetores de suporte. Essa estratégia objetiva encontrar a superfície de decisão que melhor se generaliza para novos dados.

Figura 1: Esquema de classificação por meio do SVM



Fonte: Adaptado de Huang, Davis, Townshend (2002)

2.4. Classificação de imagens

A verificação da avaliação da qualidade da classificação de imagens garante a confiabilidade dos resultados obtidos. Uma das ferramentas é a matriz de confusão, a qual permite analisar o desempenho em cada classe por meio dos erros de omissão e comissão. Esses erros são derivados da matriz de confusão e refletem as precisões do produtor e do usuário, respectivamente (Foody, 2002). Essa matriz permite a análise detalhada do desempenho do classificador, fornecendo diversos índices como o *Kappa* condicional, global e por classe (Congalton; Green, 2019).

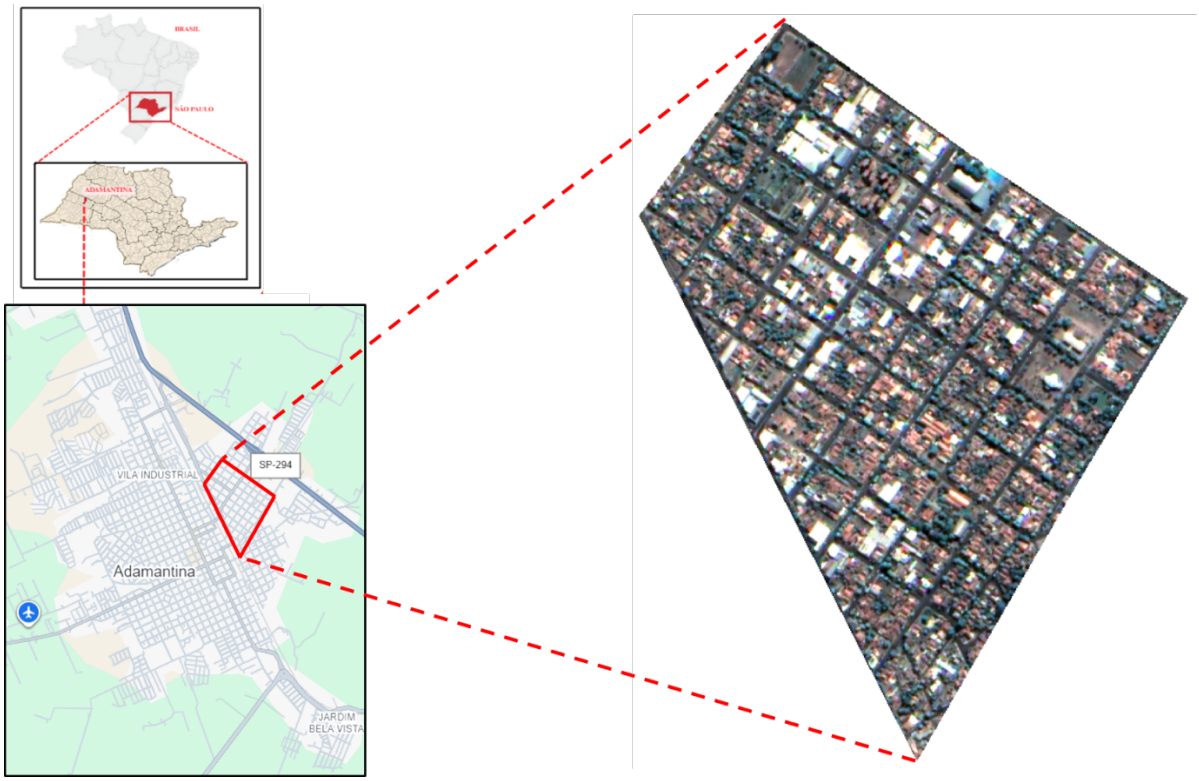
3. MÉTODO

3.1. Área de estudo

ISSN 2675-004X
FATEC Rubens Lara
Santos/SP/BR
Fatecrl.edu.br/revistaconecta

A área de estudo está localizada na cidade de Adamantina, região oeste do estado de São Paulo, possuindo as seguintes coordenadas centrais: 21º41'07"S e 51º04'21"O, com altitude de 453 m, área de 411,987 km². Segundo o Censo Demográfico realizado pelo IBGE em 2022, o município possui uma população estimada em 34.687 habitantes. Foi selecionada uma região da cidade contendo diferentes tipos de alvos urbanos e uma grande variedade arbórea. A Figura 2 apresenta a localização da área de estudo e a imagem em composição cor verdadeira.

Figura 2: Localização da área de estudo e imagem CBERS 4A



Fonte: Adaptado do Google (2025)

3.2. Dados de entrada

Para este trabalho, utilizou-se uma cena do satélite CBERS 4A, de seu sensor WPM, com cinco bandas espectrais, sendo a pancromática (450 a 900 nanometros - nm) com 2 m de resolução espacial, e as multiespectrais centradas no azul (450 a 520 nm), verde (520 a 590 nm), vermelho (630 a 690 nm) e infravermelho próximo (770 a 890 nm) com 8 m de resolução espacial (INPE, 2020). Trata-se de um transecto de uma região da cidade com a presença de diversas classes de cobertura do solo: diferentes tipos de telhados, pavimentação asfáltica, solo exposto, vegetação arbórea e rasteira. A imagem foi obtida em 06 de dezembro de 2024, com

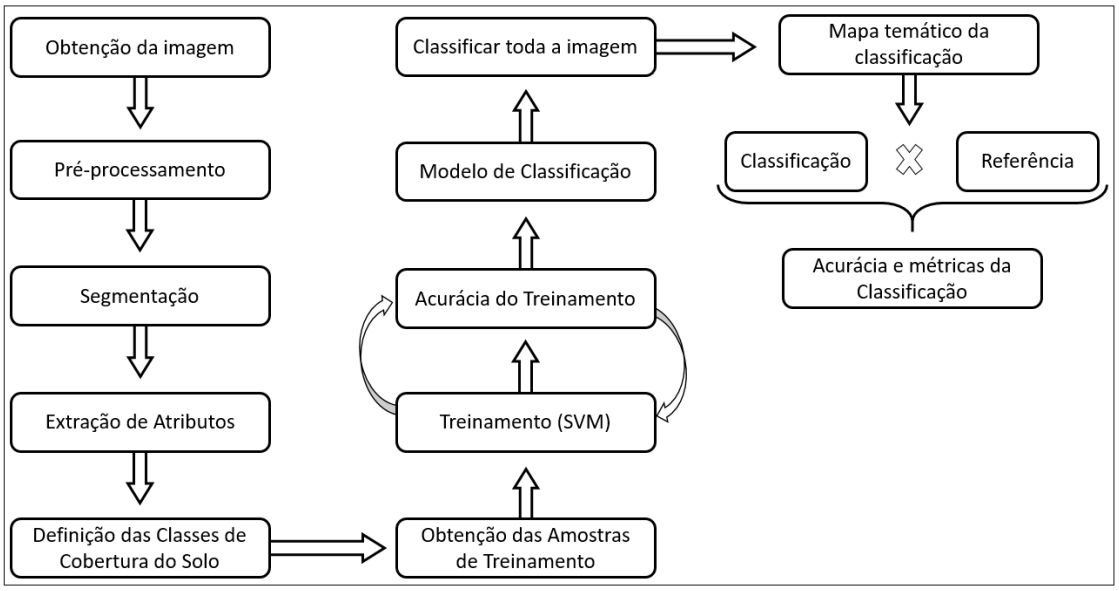
ISSN 2675-004X
FATEC Rubens Lara
Santos/SP/BR
Fatecrl.edu.br/revistaconecta

0% de nuvens. A alta gama de alvos urbanos e a diversidade de vegetação foram os motivos para a escolha deste local para realização deste trabalho.

3.3. Fluxograma metodológico

A Figura 3 apresenta o fluxograma metodológico, iniciando-se com a obtenção da imagem junto ao site do INPE. A seguir procede-se ao pré-processamento da imagem, seguido dos processos de: segmentação, definição das classes, obtenção de amostras de treinamento, extração de atributos, treinamento a partir de um algoritmo para construção do modelo de classificação e verificação da qualidade da classificação. Etapas descritas a seguir.

Figura 3: Fluxograma metodológico



Fonte: Autores (2025)

3.3.1. Pré-processamento da imagem

Uma etapa essencial em qualquer estudo é a preparação dos dados de entrada. Essas atividades ocorrem no pré-processamento. Nesta etapa, foram realizados os seguintes procedimentos: conversão dos dados de nível de cinza para radiância, onde transforma-se a incidência da radiação solar refletida pelos objetos para o espaço, captada pelos sensores do satélite. Seguido pela correção atmosférica, que considera a contaminação da atmosfera por gases e partículas que afetam os dados resultando na conversão dos dados em radiância para reflectância de superfície e, por fim, a fusão de bandas, combinando a melhor resolução espacial da banda pancromática com as bandas multiespectrais, sintetizando uma nova imagem multiespectral de melhor resolução espacial (Jensen, 2011).

ISSN 2675-004X
FATEC Rubens Lara
Santos/SP/BR
Fatecrl.edu.br/revistaconecta

3.3.2. Segmentação

Nesta etapa foram utilizados dois algoritmos disponíveis no *software* eCognition (Trimble, 2011): o *Multiresolution Segmentation* e o *Spectral Difference Segmentation*, cujas funções são, respectivamente, agrupar os pixels conforme a heterogeneidade, ajustando sua escala, forma e compacidade (Ruiz *et al*., 2025), além do refinamento de regiões segmentadas a partir de um limiar de diferenças espectrais permitidas dentro de uma mesma região (Baatz; Schape, 2018).

3.3.3. Definição de classes e amostras de treinamento

Foram definidas duas classes de cobertura do solo, sendo, vegetação arbórea e outros. A classe outros agrega todos os alvos não arbóreos, correspondendo a uma grande variedade espectral, já que agrega desde diferentes tipos de cobertura e pavimentação até solo exposto e corpos d´água. A partir daí, foram coletadas as amostras de treinamento para cada uma destas classes, sendo que houve uma preocupação em espalhá-las por toda a cena, além de escolher aquelas mais representativas para proporcionar a máxima diferenciação para a etapa subsequente de treinamento.

3.3.4. Extração de atributos

Cada atributo corresponde a um cálculo distinto de índices espectrais, sendo utilizado para extrair informações estatísticas e espectrais dos segmentos. Os atributos empregados incluíram: *Visible Atmospherically Resistant Index* (VARI), que quantifica a vegetação verde sob condições atmosféricas variáveis; o *Soil-Adjusted Vegetation Index* (SAVI), desenvolvido para corrigir a influência do solo; o *Normalized Difference Water Index* (NDWI), destinado à detecção de corpos d'água; o *Normalized Difference Vegetation Index* (NDVI), que estima a densidade de vegetação verde; o *Enhanced Vegetation Index* (EVI), atuando como complemento ao NDVI ao minimizar efeitos atmosféricos e do solo; e o *Green Chlorophyll Index* (GCI), focado na estimativa de clorofila. Adicionalmente, foram calculados o *Modified Soil-Adjusted Vegetation Index* (MSAVI), o *Green Normalized Difference Vegetation Index* (GNDVI), o *Optimized Soil-Adjusted Vegetation Index* (OSAVI), o *Renormalized Difference Vegetation Index* (RDVI) e o *Spectral Polygon Area* (SPI). Para caracterização espacial, utilizou-se a Diferença Máxima (*Max Diff*) e o Brilho entre valores espectrais, além da média e desvio padrão calculados para cada banda.

3.3.5. Treinamento e modelo de classificação

Esta etapa foi implementada em *Python* e executada no ambiente *Google Colab*. Inicialmente, os dados foram carregados a partir de um arquivo CSV contendo as amostras

ISSN 2675-004X
FATEC Rubens Lara
Santos/SP/BR
Fatecrl.edu.br/revistaconecta

normalizadas, onde os atributos (*features*) foram separados do vetor de rótulos (*target*). Para garantir a representatividade das classes durante a avaliação do modelo, adotou-se uma divisão dos dados, 70% para treinamento e 30% para teste (*test_size=0.3*), preservando a proporção original das classes em ambos os conjuntos (Figura 4).

Figura 4: Divisão dos dados em treinamento e teste

```
X_train, X_test, y_train, y_test = train_test_split(
    X_encoded, y_encoded, test_size=0.3, random_state=42, stratify=y_encoded)
```

Fonte: Autores (2025)

A análise exploratória dos dados começou com a visualização da distribuição das classes, criando um gráfico de barras para mostrar a quantidade de amostras e identificar possíveis desbalanceamentos no conjunto de dados (Figura 5, parte A). A seguir, foi implementada a análise de componentes principais (*Principal Component Analyses* - PCA), a fim de reduzir a dimensionalidade dos dados para dois componentes principais, agregando e reduzindo o conjunto de variáveis e informações das classes (Figura 5, parte B). Por fim, foi realizada a inspeção entre variáveis cria uma matriz de correlação entre todos os atributos a fim de identificar aqueles redundantes ou altamente correlacionados (Figura 5, parte C).

Figura 5: Análise exploratória dos dados

```
print("\n=== ANÁLISE EXPLORATÓRIA ===")
plt.figure(figsize=(15, 5))

plt.subplot(1, 3, 1)
y_series = pd.Series(y)
y_series.value_counts().plot(kind='bar')
plt.title('Distribuição das Classes')
plt.xticks(rotation=45)
```

A

```
pca_vis = PCA(n_components=2)
X_pca = pca_vis.fit_transform(X_encoded)
plt.subplot(1, 3, 2)
for class_label in np.unique(y_encoded):
    plt.scatter(X_pca[y_encoded == class_label, 0],
                X_pca[y_encoded == class_label, 1],
                label=le.inverse_transform([class_label])[0], alpha=0.7)
plt.title('PCA - Visualização dos Dados')
plt.legend()
```

B

```
plt.subplot(1, 3, 3)
correlation_matrix = X_encoded.corr()
sns.heatmap(correlation_matrix, cmap='coolwarm', center=0)
plt.title('Matriz de Correlação')
```

C

```
plt.tight_layout()
plt.show()
```

Fonte: Autores (2025)

Para a otimização do conjunto de entrada, foi empregado o método *Recursive Feature Elimination with Cross-Validation* (RFECV), o qual combina a eliminação recursiva de características com validação cruzada para identificar o subconjunto ótimo de variáveis

ISSN 2675-004X
FATEC Rubens Lara
Santos/SP/BR
Fatecrl.edu.br/revistaconecta

preditivas (Figura 6). O algoritmo foi inicializado com um modelo SVM linear como estimador base, utilizando validação cruzada de 5 *folds* e métrica de acurácia para avaliação. O método opera através de um processo iterativo que, em cada etapa, elimina os atributos menos importantes com base nos coeficientes do modelo, recalculando a performance via validação cruzada até determinar o número ideal de variáveis. Como resultado, o método selecionou automaticamente *X_train_selected* e *X_test_selected*, contendo apenas as características mais discriminativas, enquanto o gráfico de ranking visualiza a hierarquia de importância de todos os atributos originais, onde valores menores no ranking indicam variáveis mais relevantes para a classificação. Essa estratégia não apenas reduz a dimensionalidade e o custo computacional, mas também diminui o *overfitting* ao eliminar ruídos e redundâncias no conjunto de dados.

Figura 6: Otimização do conjunto de entrada

```
base_svm = SVC(kernel='linear', random_state=42)
rfecv = RFECV(estimator=base_svm, cv=5, scoring='accuracy', n_jobs=-1)
rfecv.fit(X_train, y_train)

X_train_selected = rfecv.transform(X_train)
X_test_selected = rfecv.transform(X_test)
```

Fonte: Autores (2025)

Na sequência, a otimização de hiperparâmetros foi realizada através do *BayesSearchCV*, uma abordagem de busca bayesiana que explora eficientemente o espaço de parâmetros definido para o SVM. Esse espaço incluiu o parâmetro de regularização C (variando de 10^{-3} a 10^{3}), o coeficiente *gamma* (de 10^{-4} a 10^{1}), o tipo de *kernel* (linear, RBF, polinomial ou sigmoidal) e o grau para *kernel* polinomial (de 2 a 5). O algoritmo executou 50 iterações de otimização utilizando validação cruzada de 5 *folds*, selecionando os próximos pontos a avaliar com base nos resultados anteriores, o que permitiu encontrar a combinação ótima de hiperparâmetros de forma mais eficiente que métodos de busca aleatória ou *grid search* tradicional (Figura 7).

Figura 7: Otimização de hiperparâmetros

```
param_space = {
    'C': Real(1e-3, 1e3, prior='log-uniform'),
    'gamma': Real(1e-4, 1e1, prior='log-uniform'),
    'kernel': Categorical(['rbf', 'linear', 'poly', 'sigmoid']),
    'degree': Integer(2, 5) }

bayes_search = BayesSearchCV(
    SVC(random_state=42),
    param_space,
    n_iter=50,
    cv=5,
    n_jobs=-1,
    random_state=42,
    scoring='accuracy' )
```

ISSN 2675-004X
FATEC Rubens Lara
Santos/SP/BR
Fatecrl.edu.br/revistaconecta

Fonte: Autores (2025)

A avaliação da etapa de treinamento foi realizada por meio de múltiplas métricas de desempenho, incluindo acurácia, validação cruzada estratificada de 5 *folds*, matriz de confusão e curva de aprendizado, permitindo analisar sua capacidade de generalização. Também foi realizada uma comparação entre diferentes *kernels* e uma análise dos padrões de erro, identificando instâncias problemáticas e fornecendo possíveis melhorias no processo de classificação.

3.3.6. Classificação e mapa temático

A partir do modelo de classificação criado na etapa de treinamento, toda a imagem foi classificada automaticamente e foi gerado um mapa temático no software QGIS.

3.3.7. Acurácia e métrica da classificação

Esta etapa foi conduzida por meio do cruzamento entre o mapa gerado pela classificação e o mapa de referência de solo. A partir de pontos aleatórios no mapa de classificação realizou-se o cruzamento com o mapa de referência a fim de verificar acertos e erros da classificação automática. Com isso, tornou-se possível construir a matriz de confusão e extrair diversos índices de qualidade.

A acurácia global representa a proporção de classificações corretas em relação ao total de amostras analisadas. No entanto, por não considerar os acertos ao acaso, é comum utilizar o índice *Kappa* como uma métrica complementar, medindo o grau de concordância entre a classificação automática e a referência, considerando a distribuição aleatória dos acertos (Foody, 2002). Ele é calculado a partir da matriz de confusão, utilizando a fórmula (Cohen, 1960):

$$kappa = \frac{P_0 - P_e}{1 - P_e},$$

onde: P_0 é a proporção de concordância observada, ou seja, a soma dos elementos da diagonal principal da matriz de confusão. P_e é a concordância esperada ao acaso.

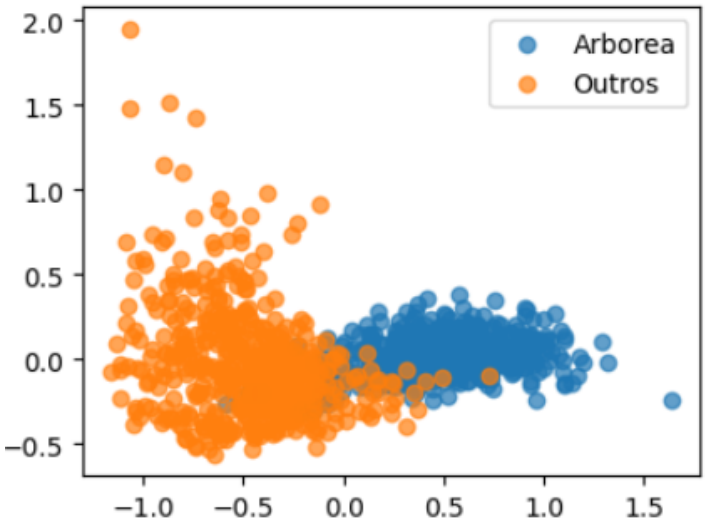
O valor de *Kappa* varia de -1 a 1, sendo que 1 indica uma perfeita concordância e 0 indica concordância aleatória, enquanto valores negativos são inferiores ao acaso (Congalton; Green, 2009). Apesar de útil para avaliar a classificação geral, o *Kappa* global pode ser influenciado por classes dominantes, o que pode mascarar falhas em classes minoritárias.

4. ANÁLISE E RESULTADOS

ISSN 2675-004X
FATEC Rubens Lara
Santos/SP/BR
Fatecrl.edu.br/revistaconecta

O conjunto de dados utilizado neste trabalho foi composto por 1.001 amostras, com 22 atributos espectrais, distribuídas entre duas classes: "Arbórea" (483 amostras) e "Outros" (518 amostras). Esta distribuição relativamente equilibrada entre as classes é vantajosa para o treinamento do modelo, minimizando vieses de classificação. A análise exploratória por PCA (Figura 8) revelou uma boa separabilidade entre as classes no espaço reduzido de duas componentes principais. A seleção de características via RFECV identificou que o NDVI e o EVI dentre todos os 22 atributos foram suficientes para atingir uma acurácia de 98,01% no conjunto de teste, indicando a presença de alta redundância espectral e a extrema relevância dos atributos selecionados. A junção do NDVI, que detecta o vigor da vegetação, com o EVI, que reduz interferências atmosféricas, foi suficiente para distinguir com precisão as classes analisadas. Por sua vez, a otimização de hiperparâmetros consolidou o desempenho do modelo ao identificar a configuração ótima: um *kernel* linear com alto valor de regularização (*C = 540,59*) e *gamma* de 2,28, resultando em um *score* de validação cruzada de 98,43%. Estes resultados, em conjunto, demonstram a eficácia do *pipeline* de pré-processamento e modelagem adotado.

Figura 8: Análise exploratória por PCA



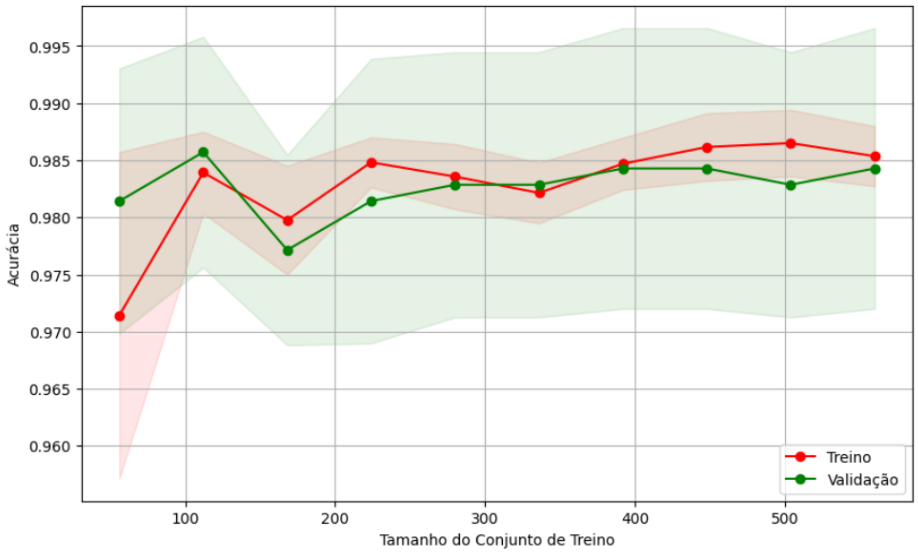
Fonte: Autores (2025)

O SVM demonstrou excelente desempenho na etapa de treinamento, com uma acurácia de 97,67% no conjunto de teste e 98,43% na validação cruzada. A Figura 9 apresenta a curva de aprendizado, confirmando a robustez do modelo ao mostrar que a acurácia se estabiliza em patamares elevados, conforme aumenta o tamanho do conjunto de treinamento, sem indícios de *overfitting*. Na comparação entre *kernels*, os modelos linear, RBF e polinomial apresentaram performances equivalentes, todos próximos de 98% de acurácia, enquanto o *kernel sigmoidal*

ISSN 2675-004X
FATEC Rubens Lara
Santos/SP/BR
Fatecrl.edu.br/revistaconecta

mostrou-se completamente inadequado para este problema, com apenas 4,86% de acurácia. Dessa forma, o modelo de classificação foi construído a partir do *Kernel* linear.

Figura 9: Curva de aprendizado do treinamento

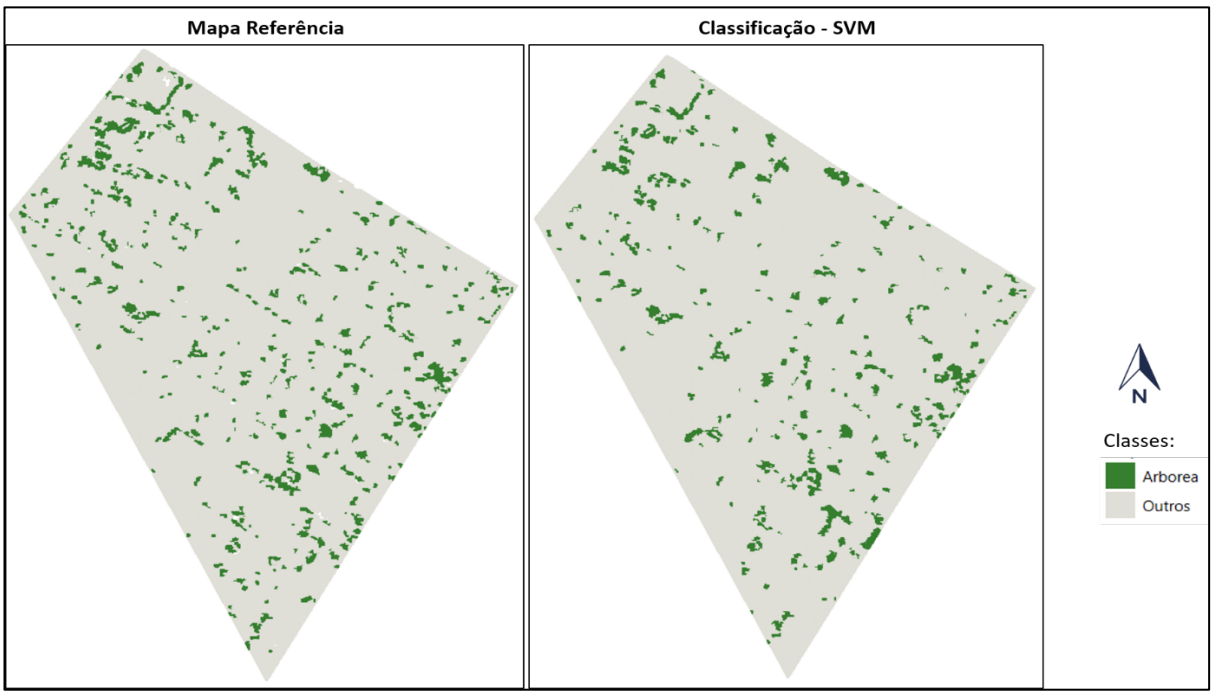


Fonte: Autores (2025)

A partir do modelo de classificação gerado, toda a imagem foi classificada (Figura 10). É possível verificar a distribuição da classe arbórea pelo transecto da cidade. A partir da comparação com o mapa de referência, verifica-se uma correspondência entre a localização das árvores, sobretudo quando elas estão espacialmente concentradas.

Figura 10: Classificação SVM comparada com o mapa de referência

ISSN 2675-004X
FATEC Rubens Lara
Santos/SP/BR
Fatecrl.edu.br/revistaconecta



Fonte: Autores (2025)

A análise das métricas da classificação demonstra um bom desempenho, porém com comportamentos distintos entre as classes. O índice de acerto global foi excelente, com 96,96%, abrangendo 2371 pontos aleatórios (amostras). Além disso, o Índice *Kappa* geral foi de 0,749, confirmando uma concordância substancial entre as classificações previstas e observadas, isso indica que o modelo superou a aleatoriedade na classificação.

Entretanto, uma análise detalhada por classe revela um contraste no desempenho do classificador (Tabela 1). A classe "Outros" apresentou um desempenho com Precisão de 0,974 e *Recall* de 0,994, resultando em um *F1-Score* de 0,984. Isto significa que das 2196 amostras desta classe, apenas 14 foram erroneamente classificados como "Arborea", demonstrando que o modelo é altamente confiável na identificação desta classe. Em contrapartida, a classe "Arborea" apresentou resultados mais moderados, com Precisão de 0,893, mas *Recall* de apenas 0,669, resultando em um *F1-Score* de 0,765. Esta assimetria indica que, embora quando o modelo classifica um ponto como "Arborea" ele esteja correto em 89,3% dos casos, ele deixa de identificar aproximadamente 33,1% dos pontos verdadeiros desta classe (58 das 175 amostras).

Tabela 1: Métricas da classificação SVM

ISSN 2675-004X
FATEC Rubens Lara
Santos/SP/BR
Fatecrl.edu.br/revistaconecta

Classe	Precisão	Recall	F1-Score	Suporte
Arborea	0.893	0.669	0.765	175
Outros	0.974	0.994	0.984	2196
GLOBAL	0.934	0.831	0,874	2371

Fonte: Autores (2025)

A matriz de confusão (Tabela 2) em valores absolutos revela que o principal desafio foi a sub identificação da classe "Arborea", com 58 pontos desta classe sendo erroneamente classificados como "Outros". Este padrão sugere que o modelo tende a um viés conservador na identificação de vegetação arbórea, possivelmente devido à similaridade espectral com outras coberturas ou à variabilidade interna da própria classe. A alta acurácia global é sustentada principalmente pelo domínio quantitativo da classe "Outros" no conjunto de dados, que representa 92,6% do total de amostras.

Tabela 2: Matriz de confusão da classificação SVM

		Arbórea	Outros
Classificação	Arbórea	117	58
	Outros	14	2182
		Referência	

Fonte: Autores (2025)

Os resultados deste trabalho indicam que, para aplicações onde a detecção de vegetação arbórea é crítica, estratégias específicas como ajuste de limiares ou amostragem seletiva podem ser necessárias para melhorar a sensibilidade do modelo para esta classe.

5. CONSIDERAÇÕES FINAIS

Este trabalho teve como objetivo desenvolver um método para classificação de cobertura arbórea em uma região selecionada no município de Adamantina-SP, utilizando imagens do satélite CBERS-4A e o algoritmo SVM, visando fornecer subsídios técnicos para a

ISSN 2675-004X
FATEC Rubens Lara
Santos/SP/BR
Fatecrl.edu.br/revistaconecta

gestão ambiental municipal. Os resultados obtidos permitem afirmar que o SVM, quando alimentado com dados do CBERS-4A, mostra-se capaz de identificar a cobertura arbórea com precisão satisfatória. No entanto, esta capacidade apresenta importantes ressalvas que devem ser consideradas em aplicações práticas.

Quanto aos principais erros de classificação, identificou-se que o modelo tende a um viés conservador na detecção de vegetação arbórea, resultando em significativa sub identificação desta classe. As possíveis causas para este comportamento incluem a similaridade espectral entre vegetação arbórea e outras coberturas vegetais, além da possível influência da resolução espacial e espectral do sensor na distinção de copas individuais. A análise da matriz de confusão revelou que o erro predominante foi a classificação errônea de árvores como "Outros", indicando que o modelo priorizou a precisão em detrimento da sensibilidade para a classe de interesse principal.

Entre as limitações deste trabalho, destaca-se a dependência de refinamento das amostras de treinamento e que essas sejam mais representativas das classes de cobertura. Para aplicações futuras, recomenda-se a incorporação de dados complementares, como informações de textura e geometria, além da validação em campo mais abrangente. A metodologia demonstrou ser viável para o programa Município Verde Azul, porém sua implementação operacional deve considerar as limitações identificadas, particularmente no que concerne à detecção completa da cobertura arbórea existente.

6. REFERÊNCIAS

ADAMANTINA. **Adamantina mais uma vez recebe o Certificado Município Verde Azul**. 2020. Disponível em: <https://www.adamantina.sp.gov.br/portal/noticias/0/3/4238/adamantina-mais-uma-vez-recebe-o-certificado-municipio-verdeazul>. Acesso em: 7 jun. 2025.

AMBIENTE. Secretaria do Meio Ambiente de São Paulo**. 21 Projetos Ambientais Estratégicos**. Disponível em < https://semil.sp.gov.br/mudancas-climaticas-e-sustentabilidade/ >. Acesso em 7 set. 2025.

ANDERSON, J. R. *et al*. A Land Use and Land Cover Classification System for Use with Remote Sensor Data. **U.S. Geological Survey**, Professional Paper 964, 1976. Disponível em: <https://pubs.usgs.gov/pp/0964/report.pdf>. Acesso em: 30 mar. 2025.

BRANSON, S. *et al*. From Google Maps to a fine-grained catalog of street trees. **ISPRS Journal of Photogrammetry and Remote Sensing**, v. 135, p. 13–30, jan. 2018. Disponível em: < https://arxiv.org/abs/1910.02675>. Acesso em: 07 jun. 2025.

ISSN 2675-004X
FATEC Rubens Lara
Santos/SP/BR
Fatecrl.edu.br/revistaconecta

BAATZ, M.; SCHAPE, A. Multiresolution segmentation: na optimization approach for high quality multi-scale image segmentation. In: **XII Angewandte Geographische Informationsverarbeitung**, 2000, Wichmann-Verlag, Heidelberg.

CÂMARA, G.; SOUZA, R. C. M.; FREITAS, U. M. de. Spring: Integrating remote sensing and GIS by object-oriented data modelling. **Computers & Graphics**, v. 20, n. 3, p. 395–403, 1996. Disponível em: <https://www.sciencedirect.com/science/article/pii/0097849396000088>. Acesso em: 30 mar. 2025.

CAMPBELL, J. B.; WYNNE, R. H. **Introduction to remote sensing**. Guilford press, 2011. Disponível em: <https://www.guilford.com/books/Introduction-to-Remote-Sensing/Campbell-Wynne/9781462549405>. Acesso em 30 mar. 2025.

COHEN, J. A coefficient of agreement for nominal scales. **Educational and psychological measurement**, v. 20, n. 1, p. 37-46, 1960. Disponível em: <https://journals.sagepub.com/doi/10.1177/001316446002000104>. Acesso em 30 mar. 2025.

CONGALTON, R. G.; GREEN, K. **Assessing the accuracy of remotely sensed data: principles and practices**. CRC press, 2019. Disponível em: <https://www.taylorfrancis.com/books/mono/10.1201/9780429052729/assessing-accuracy-remotely-sensed-data-russell-congalton-kass-green>. Acesso em: 30 mar. 2025.

CORTES, C.; VAPNIK, V. Support-vector networks. **Machine Learning**, v. 20, p. 273–297, 1995. Disponível em: <https://link.springer.com/article/10.1007/BF00994018>. Acesso em: 8 jun. 2025.

FOODY, G. M. Status of land cover classification accuracy assessment. **Remote sensing of environment**, v. 80, n. 1, p. 185-201, 2002. Disponível em: <https://www.sciencedirect.com/science/article/abs/pii/S0034425701002954>. Acesso em: 30 mar. 2025.

HUANG, C., GONG, P., BIGING, G. Satellite remote sensing for land use/land cover change: Advances and challenges. **International Journal of Remote Sensing**, 41(2), 658-708. 2020.

HUANG, C.; DAVIS, L. S.; TOWNSHEND, J. R. G. An assessment of support vector machines for land cover classification. **International Journal of Remote Sensing**, v. 23, n. 4, p. 725-749, 2002.

INPE - Instituto Nacional de Pesquisas Espaciais. **Satélite Sino-Brasileiro de Recursos Terrestres (CBERS)**. 2020. Disponível em: Acesso em 13 de maio de 2024.

ISSN 2675-004X
FATEC Rubens Lara
Santos/SP/BR
Fatecrl.edu.br/revistaconecta

JENSEN, J. R. **Sensoriamento remoto do ambiente: uma perspectiva em recursos terrestres**. São José dos Campos: Parêntese, 2011. Tradução José Carlos Neves Epiphanio.

LILLESAND, T.; KIEFER, R. W.; CHIPMAN, J. **Remote sensing and image interpretation**. John Wiley & Sons, 2015. Disponível em: <https://www.geokniga.org/bookfiles/geokniga-remote-sensing-and-image-interpretation.pdf>. Acesso em: 05 mai. 2025.

LIN, A. *et al*. Identifying Urban Building Function by Integrating Remote Sensing Imagery and POI Data. **IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing**, 2021. Disponível em: <10.1109/JSTARS.2021.3107543>. Acesso em: 17 mai. 2025.

MOREIRA, M. A. **Fundamentos do Sensoriamento Remoto e Metodologias de Aplicação**. 4. ed. Viçosa: UFV, 2011.

MOUNTRAKIS, G.; IM, J.; OGOLE, C.. Support vector machines in remote sensing: A review. **ISPRS journal of photogrammetry and remote sensing**, v. 66, n. 3, p. 247-259, 2011. Disponível em: <https://aboutgis.com/Publications/Mountrakis_SVM_review_in_remote_sensing_ISPRS2010.pdf>. Acesso em: 12 mai. 2025.

NOVO, E. M. L. M. **Sensoriamento Remoto: Princípios e aplicações**. 3. ed. São Paulo: Blucher, 2010. Disponível em: <https://api.pageplace.de/preview/DT0400.9788521216902_A47327924/preview-9788521216902_A47327924.pdf>. Acesso em: 05 mai. 2025.

RUIZ, P. R. S. **Classificação da cobertura do solo urbano usando árvores de decisão a partir de cenas WorldView-2 e WorldView-3 para diferentes níveis de legenda**. Dissertação (Mestrado em Sensoriamento Remoto) - Instituto Nacional de Pesquisas Espaciais (INPE), São José dos Campos, 2017, 181 p.

RUIZ, P. R. S. *et al*. Explorando a segmentação multirresolução em imagens urbanas: um estudo com CBERS 4A WPM. in: Anais do XXI Simpósio Brasileiro de Sensoriamento Remoto, 2025, Salvador. **Anais eletrônicos...**, Galoá, 2025. Disponível em: <https://proceedings.science/sbsr-2025/trabalhos/explorando-a-segmentacao-multirresolucao-em-imagens-urbanas-um-estudo-com-cbers?lang=pt-br> Acesso em: 17 Maio. 2025.

SCHOWENGERDT, R. A. **Remote sensing: models and methods for image processing**. Elsevier, 2006.

TAUBENBÖCK, H., WEGMANN, M., ROTH, A., MEHL, H., & DECH, S. (2012). Urbanization in India – Spatiotemporal analysis using remote sensing. **Computers, Environment and Urban Systems**, 36(3), 296–306. Disponível em: <https://www.sciencedirect.com/science/article/abs/pii/S0198971508000604>. Acesso em: 06 jun. 2025.

ISSN 2675-004X
FATEC Rubens Lara
Santos/SP/BR
Fatecrl.edu.br/revistaconecta

TOMASIELLO, D. B. **Modelos de rede transporte público e individual para estudos de acessibilidade em São Paulo**. 2016. Dissertação (Mestrado em Engenharia de Transporte) - Escola Politécnica, Universidade de São Paulo, São Paulo, 2016, 94 p. Disponível em: <https://www.teses.usp.br/teses/disponiveis/3/3138/tde-24062016-111306/en.php>. Acesso em: 13 agosto 2025.

TRIMBLE. **eCognition Developer 8.7 User Guide**. Munich, Germany: [s.n.], 2011. 258 p. Disponível em: < https://docs.ecognition.com/ Acesso em: 13 agosto 2025.

VAPNIK, Vladimir. **The nature of statistical learning theory**. Springer science & business media, 1999. Disponível em: <https://archive.org/details/natureofstatisti0000vapn>. Acesso em: 12 mai. 2025.

WOODHOUSE, Iain H. **Introduction to microwave remote sensing**. CRC press, 2017.