Composições de Imagens de Sensoriamento Remoto baseadas em Computação Evolutiva para Detecção de Desmatamento

Eduardo B. Neto, Paulo R. C. Pedro
Universidade Federal de São Paulo – São José dos Campos, Brasil,
Email: {ebneto, costa.paulo}@unifesp.br

***Resumo*—A conservação das florestas tropicais é um assunto atual de relevância social e ecológica, devido ao importante papel que elas desempenham no ecossistema global. Infelizmente, milhões de hectares são desmatados e degradados todo ano, sendo necessários programas – governamentais ou de iniciativas privadas – para monitoramento das florestas tropicais. Tais programas contam tanto com o auxílio de profissionais especializados quanto de sistemas computacionais para detecção de padrões. Para que um modelo computacional consiga classificar uma determinada área como desmatamento ou não, geralmente se utilizam imagens de satélite, que possuem a vantagem de ter resolução multiespectral de captura. Porém, a captura de uma imagem em vários comprimentos de onda leva à redundância da informação e à alta quantidade de bandas, causando a "maldição da dimensionalidade". Para contornar esse problema, este trabalho propõe o uso de algoritmos genéticos visando selecionar a combinação de bandas que apresenta melhor resultado na classificação quanto à presença de desmatamento e melhor resume as imagens coletadas pelos satélites.**

I. Introdução & Motivação

Florestas tropicais são florestas localizadas entre os Trópicos de Câncer e Capricórnio, próximas à linha do Equador. São encontradas na América do Sul e Central, na África e em regiões da Ásia e do Pacífico [1]. Possuem clima quente e úmido, com vegetação rica e exuberante. Até 2015 existiam 700 milhões de hectares de floresta primária (original) [2].

Embora as florestas tropicais cubram apenas 7% da superfície terrestre, estima-se que elas abrigam mais da metade das espécies do planeta [2]. Além da grande biodiversidade, as plantas e solo das florestas tropicais retêm de 460 a 575 bilhões de toneladas de carbono. Os processos de evaporação e evapotranspiração de suas plantas e árvores retornam grandes quantidades de água para a atmosfera local, promovendo a formação de nuvens e chuvas [3]. As florestas tropicais também são o lar de inúmeros povos indígenas.

Infelizmente, milhões de hectares de florestas tropicais tem sido perdidos a cada ano através de desmatamento e degradação [2], [4]. Desmatamento pode ser definido como a conversão da floresta primária em outras coberturas através de ações antropogênicas, sendo o desmatamento por corte raso a remoção completa da floresta em um curto período. Já a degradação é progressiva, que pode ocorrer durante anos, de maneira lenta, ao longo do processo de desmatamento, em que é feita contínua exploração de madeira e queimadas [5].

De acordo com um dos mais conhecidos e bem sucedidos programas de monitoramento, PRODES (Programa de Monitoramento da Floresta Amazônica Brasileira por Satélite), no período entre agosto/2018 e julho/2019, o desmatamento na Amazônia Legal Brasileira alcançou 10.129 km^2, correspondendo a um crescimento de 34% se comparado ao período anterior (agosto/2017 a julho/2018) [6].

O desmatamento das florestas tropicais acontece por diferentes e complexos motivos econômicos, como agricultura, pecuária, garimpo, extração madeireira e novas ameaças como exploração de gás e petróleo em reservas indígenas. Estima-se que quase metade da perda florestal no Brasil e 12% da perda florestal na Indonésia são ligadas a conversão de terra para a agro-indústria de larga escala [2].

Dependendo do tipo de atividade realizada, como agricultura intensiva, agricultura de sombra, pasto para gado, extração de madeira seletiva, a floresta pode demorar mais de 50 anos para regenerar, uma vez que quase todos os nutrientes são encontrados nas plantas e árvores, e não no solo das florestas tropicais, fazendo com que a supressão da vegetação e chuvas nesse solo exposto tornem-no cada vez menos fértil até chegar a seu esgotamento, em aproximadamente 3 anos, o que faz os fazendeiros abandonarem a área e procurarem um novo local [3].

O desmatamento pode trazer consequências irreversíveis e catastróficas. Muitas plantas e animais das florestas tropicais só podem ser encontrados em pequenas áreas, por necessitarem de um *habitat* especial para viverem, o que os torna muito vulneráveis a desmatamento. Estimava-se que 137 espécies desaparecem por dia, no mundo inteiro, por causa do desmatamento em florestas tropicais [3]. Outro fator que contribui para a perda da biodiversidade global é a fragmentação das florestas por rodovias e outros distúrbios [2].

Outras consequências que o desmatamento traz são o aumento da emissão dos gases do efeito estufa, impactando na mudança do clima, desertificação, escassez de água potável, aumento de doenças e até surgimento de pandemias [2], [7]–[10].

Recentemente, dois pesquisadores apontaram que a Floresta Amazônica está chegando a um ponto de não-retorno, em que o desmatamento e as queimadas causariam consequências no ciclo hidrológico que incapacitariam o ecossistema tropical, transformando partes da Floresta Amazônia em savana. Eles

destacam que as secas mais severas dos anos de 2005, 2010 e de 2015 a 2016 podem ser indicadores que o sistema da floresta está oscilando e que devem ser tomadas atitudes para reduzir o desmatamento para menos de 20% [11].

Como a conservação das florestas tropicais é urgente e necessária, programas de monitoramento foram criados por agências governamentais e instituições sem fins lucrativos. Esses programas utilizam imagens de sensoriamento remoto, processamento de imagens, técnicas de Aprendizado de Máquina e fotointerpretação para analisar, identificar e quantificar mudanças na cobertura florestal [12].

Algoritmos de aprendizado supervisionado tem sido utilizados para tarefas de classificação, e seus sucessos dependem da representatividade do conjunto de treinamento. Geralmente uma grande quantidade de dados é necessária para treinar um classificador. Esses dados são classificados por especialistas e extensa análise manual, tornando o processo dispendioso. Assim, é desejável usar conjuntos de treinamento pequenos para obter alta acurácia na classificação. Uma abordagem para alcançar esse objetivo é de construir o conjunto de treinamento iterativamente, adicionando apenas amostras que poderiam trazer melhor representatividade ao treinamento. Essa técnica de amostragem é conhecida na literatura de Aprendizado de Máquina como Aprendizado Ativo [13].

Ainda assim, a classificação de especialistas ainda é necessária. A escassez de mão-de-obra especializada e/ou a grande quantidade de dados a serem analisados, faz com que o processo possa ser custoso tanto financeiramente quanto em relação ao tempo, o que constitui um grande desafio para as tecnologias da informação e comunicação (TIC) [14]. Uma possível solução a esse problema é utilizar Ciência Cidadã (*Citizen Science*, em inglês), onde voluntários não-especializados coletam, analisam e classificam dados para resolverem vários problemas técnicos e científicos [15], [16].

Inúmeros projetos de Ciência Cidadã, desenvolvidos em diferentes partes do mundo, vêm atraindo a atenção de importantes revistas científicas, como *Nature* [17], [18] e *Science* [19]. Recentemente, a União Europeia[1] e os governos dos EUA[2] e da Austrália[3], lançaram programas oficiais para catalogar e apoiar projetos de Ciência Cidadã [20].

Muitos projetos de Ciência Cidadã propõem envolver os cidadãos no monitoramento ambiental. Nesse contexto, em 2012, foi lançado o projeto *ForestWatchers*, com o objetivo de utilizar voluntários no monitoramento do desmatamento em florestas tropicais. Ele era composto de 3 aplicações. Em uma delas, chamada de *Correct Classification*, imagens de sensoriamento remoto eram analisadas por um algoritmo classificador, rotulando os *pixels* das imagens em floresta ou não-floresta e computando uma taxa de confiança para a rotulagem. Finalmente, as regiões com uma taxa de confiança abaixo de um certo limite eram enviadas aos voluntários, que as analisavam, mantinham ou alteravam a classificação em floresta ou não-floresta [12].

As classificações dos voluntários nas regiões de baixa confiança, somadas ao restante das rotulagens pelo algoritmo, podem gerar um mapa do desmatamento na floresta. No entanto, essas classificações vindas dos voluntários poderiam ser utilizadas como treinamento de uma técnica de Aprendizado de Máquina. Assim, foi concebido o projeto *ForestEyes* [21].

O projeto *ForestEyes* foi lançado em abril/2019 e é hospedado na conhecida plataforma de Ciência Cidadã, *Zooniverse* [22], [23]. O projeto tem como objetivo aliar Ciência Cidadã e Aprendizado de Máquina no monitoramento de florestas, utilizando voluntários para classificarem segmentos de sensoriamento remoto e usando essas classificações para construir um pequeno, mas eficiente conjunto de treinamento. Esse conjunto será utilizado para treinar técnicas de aprendizado supervisionado ou semi-supervisionado (como Aprendizado Ativo) para classificar novas imagens de sensoriamento remoto.

O projeto foi testado em áreas da Amazônia Legal Brasileira, especificamente no estado de Rondônia, uma vez que esse local possui um programa de monitoramento bem conceituado, o PRODES, do Instituto Nacional de Pesquisas Espaciais (INPE), que serviu como validação das classificações obtidas. A ambição é que, com o aprimoramento do projeto, os dados gerados possam servir de fonte complementar aos programas já existentes, podendo ainda ser utilizado em regiões florestais onde não existam sistemas de monitoramento.

Um fator diferencial de imagens de satélites, mesmo ópticos, é a resolução espectral de captura. A maioria dos satélites modernos disponibiliza imagens multiespectrais, ou seja, realizam a captura da imagem em diversos comprimentos de onda. A alta quantidade de bandas, bem como a redundância de informação, pode acarretar na ”maldição da dimensioalidade” [24]. Isso representa um problema para sistemas de classificação de imagens de sensoriamento remoto, podendo acarretar perda de desempenho. Por isso, a tarefa de *feature selection* ou, particularmente, seleção de bandas, é de grande importância para tais sistemas.

Nesse contexto, o uso de algoritmos genéticos para a busca da combinação ideal mostra-se uma alternativa interessante, por ser mais eficiente em tempo e custo computacional do que métodos tradicionais, como uma *grid search* convencional [25]. Na literatura, diversos autores fizeram uso dessa estratégia para resolver o problema. Assim, propomos uma abordagem semelhante para determinar a melhor combinação de bandas para a detecção de desmatamento. A Seção II contém alguns conceitos fundamentais que norteiam este trabalho. Na Seção III, são destacados trabalhos relacionados ao nosso objeto de estudo. A metodologia experimental é descrita na Seção V. Finalmente, definimos o que esperamos entregar na **??**

[1]https://www.ecsite.eu/activities-and-services/projects/eu-citizenscience, acessado em 18-07-2020

[2]https://www.citizenscience.gov/, acessado em 18-07-2020

[3]https://citizenscience.org.au/, acessado em 18-07-2020

II. Conceitos Fundamentais

A. Aprendizado Supervisionado

No contexto do Aprendizado de Máquina, o Aprendizado Supervisionado contempla tarefas em que se busca aprender uma função com base em pares (amostras) de exemplos de suas entradas e saídas – estas, também chamadas de verdades ou *ground truths*) [26].

B. Análise de Componentes Principais

Dentro da Mineração de Dados, a análise de componentes principais é uma técnica de pré-processamento dos dados que visa reduzir a dimensionalidade dos vetores de entrada, por meio da remoção de variáveis que são combinações lineares de outras que permanecerão como dado de entrada, podendo ser usadas para tarefas de classificação, regressão, agrupamento e afins. [27], [28].

C. Algoritmo MaskSLIC

Trata-se de um algoritmo de segmentação de imagens que aprimora o algoritmo SLIC ao permitir a produção de *superpixels* em máscaras. Para tal conquista, três passos são feitos: no primeiro, utiliza-se uma transformada de distância Euclidiana para distribuir espacialmente os pontos de semente dentro da máscara, de modo que tais pontos fiquem o mais distante possível entre si; no segundo, aplica-se o SLIC nos pontos de sementes, a fim de posicioná-los uniformemente na máscara;no terceiro, por fim, aplica-se de novo o SLIc, a fim de obter superpixels mais consistentes para a região de interesse [29].

D. SVM

Na área do Aprendizado de Máquina, as Máquinas de Vetores Suporte são uma das técnicas mais utilizadas para tarefa de categorização de amostras de dados em classes e apresentam um resultado bem satisfatório em problemas não lineares, superando, por vezes, as Redes Neurais Artificiais. Seu funcionamento baseia-se nos teoremas de Cover e de Mercer e no conceito de Kernel Trick [30].

E. Descritores de Haralick

No contexto dos descritores de imagens, os descritores de Haralick são aqueles que descrevem a textura de uma imagem, por meio de cálculos feitos com matrizes de co-ocorrência, os quais permitem produzir uma matriz de probabilidades de ocorrência de combinações entre níveis de cinza. Com essas probabilidades, obtém-se o valor dos atributos de textura [31].

F. UMDA

No contexto dos algoritmos genéticos, o *Univariate Marginal Distribution Algorithm* (UMDA) é um algoritmo de estimação de distribuição, que utiliza modelos probabilísticos das melhores soluções para conduzir a busca pela solução do problema. Para gerar uma nova população, o UMDA estima a distribuição marginal univariada de cada variável, a partir dos melhores indivíduos da geração anterior. Portanto, a ideia desse algoritmo é utilizar análises estatísticas sobre as melhores soluções encontradas até o momento para orientar a procura por soluções melhores, permitindo uma busca mais eficiente. [32], [33]

G. Taxa de Homogeneidade (HoR)

Para quantificar a qualidade dos segmentos produzidos por um segmentador como o MaskSLIC [29], que não atribui classes aos segmentos produzidos, [34] propõe o uso da Taxa de Homogeneidade (ou *HoR*, na sigla em inglês para *Homogenity Ratio*). A *HoR* é calculada a partir da verdade sobre a região segmentada. O valor da taxa é definido como a quantidade de *pixels* da classe majoritária (segundo a verdade) dividido pela quantidade total de *pixels* no segmento. A Equação 1 formaliza essa definição aplicada ao contexto do presente trabalho, em que trabalha-se com uma segmentação binária (classes de ”floresta”e ”não floresta”). Nela, ”NFP”refere-se ao número de *pixels* da classe ”floresta”enquanto ”NNP”refere-se à quantidade de *pixels* da classe ”não-floresta”.

$$HoR = \frac{max(NFP, NNP)}{NFP + NNP} \tag{1}$$

H. Combinações de Bandas

Dentro do processamento de imagens, a combinação de bandas é um processo que envolve a fusão ou combinação de diferentes bandas espectrais de uma imagem para obter uma imagem mais informativa. A Figura 1 ilustra algumas combinações de bandas comumente utilizadas para o satélite Landsat-8. Na Tabela I, estão detalhadas as bandas captadas pelo sensor.

Tabela I
Bandas suportadas pelo Landsat-8 e seus respectivos comprimentos de onda.

Banda	**Nome**	**Faixa espectral**
1	*Coastal Aerosol*	0,430 - 0,450 µm
2	*Blue*	0,450 - 0,510 µm
3	*Green*	0,530 - 0,590 µm
4	*Red*	0,640 - 0,670 µm
5	*Near-Infrared*	0,850 - 0,880 µm
6	*Shortwave Infrared 1*	1,570 - 1,650 µm
7	*Shortwave Infrared 2*	2,110 - 2,290 µm

III. Trabalhos Relacionados

Em [25], abordou-se o problema da seleção de bandas de imagens hiperespectrais (compostas por centenas de bandas) para redução de dimensionalidade de imagens de sensoriamento remoto. Por meio de algoritmos genéticos, buscou-se, dentre as 124 bandas de uma imagem hiperespectral, a melhor combinação de três bandas para a tarefa de classificação de tipos de vegetação me imagens de sensoriamento remoto.

Zhang et al.,2009 [35] também propõe o uso de algoritmos genéticos na tarefa de seleção de bandas. Assim como em [25], buscou-se uma combinação de uma quantidade fixa de bandas. A abordagem ao problema é interessante e se assemelha à nossa, particularmente em relação à codificação dos genes.

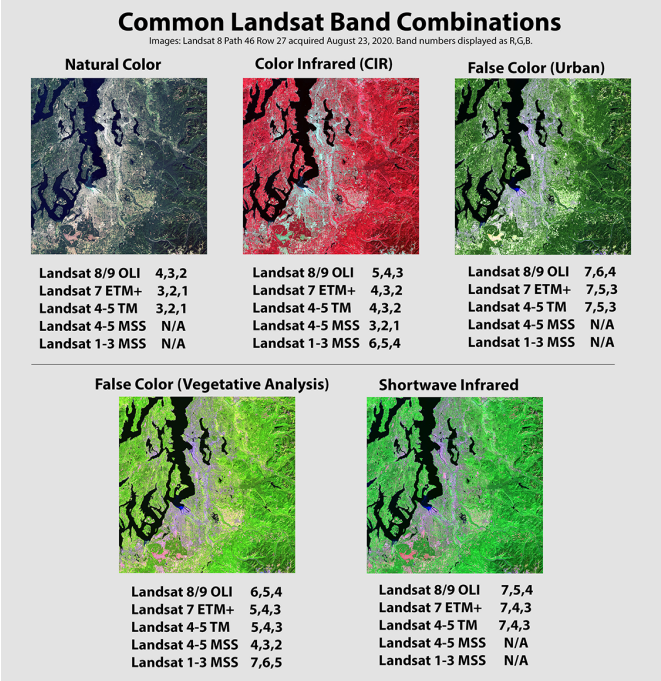

Figura 1. Combinações de bandas comuns para o Landsat-8. Fonte: USGS.

Isto é, na forma de representação dos indivíduos da população: um vetor com uma quantidade fixa de dimensões em que cada posição é referente a uma banda espectral. Para cada banda, o valor ”1”indica que ela é incluída na combinação, enquanto ”0”representa sua exclusão.

O uso de algoritmos genéticos para melhorar o desempenho em classificações de imagens de sensoriamento remoto pode contemplar outros escopos além da seleção de bandas. Lin et al., 2020 [36] faz uso dessa técnica para determinar os hiperparâmetros de penalidade e da função kernel em um classificador SVM, obtendo resultados superiores aos de métodos tradicionais de seleção, como o de máxima verossimilhança.

Em Yao & Tian, 2003 [37], algoritmos genéticos são propostos para reduzir a dimensionalidade de imagens de sensoriamento remoto de forma conjunta com a técnica PCA. O método proposto é baseado na técnica de Análise de Componentes Principais, porém há uma seleção primária de bandas para a realização da PCA. Essa selecão é feita utilizando algoritmos genéticos.

Na área da saúde, [38] utilizou de algoritmos genéticos para a seleção de características em imagens de mamografia, na tarefa de detecção de tumores. Ao contrário do visto em imagens de sensoriamento remoto, as características a serem selecionadas não são bandas espectrais: neste trabalho, toma-se como características os vetores produzidos pela técnica PCA e, também, os descritores de Haralick. A implementacão da função *fitness* é baseada em um classificador de rede neural. Ou seja, são selecionadas as combinações de descritores que promovem as melhores métricas de desempenho no classificador. Os resultados indicam que o uso da seleção de características promoveu um aumento de 4% na métrica de desempenho utilizada.

Também no campo do sensoriamento remoto, [39] propõe o uso da estratégia Wrapper unida ao classificador SVM, que faz uso de informação espectral e espacial, para seleção de bandas multiobjetivo, que visa aprimorar a classificação de pixels e reduzir a quantidade de bandas de imagens hiperespectrais, comuns em diversas aplicações, como na agricultura. Esse, portanto, trabalho visa elaborar um classificador que apresente um bom resultado, mas que não use todas as bandas, melhorando o custo computacional da tarefa de categorização, por meio da redução da dimensionalidade.

A seleção de bandas via algorimtos genéticos não fica limitada apenas à área de sensoriamento remoto. No campo da agronomia, [40] realiza um estudo de seleção de bandas em imagens produzidas por câmeras hiperespectrais para a identificação de pragas em plantações de soja. A partir de um algoritmo genético, determinou-se a melhor combinação espectral – neste caso, de 6 bandas – para a identificação da doença a partir de imagens hiperespectrais dos talos da plantacão. A função *fitness* (ou seja, a função que norteia o algoritmo genético) é baseada no desempenho de classificação das imagens em um classificador SVM. Os resultados com a combinação de bandas obtida pelo algoritmo genético apontam uma acurácia de classificação 26,1% melhor em relação a imagens RGB convencionais.

Há, ainda, abordagens ao problema de *feature selection* que combinam outros métodos a algoritmos genéticos. Li et al. [41] propõe uma abordagem com agrupamento prévio de bandas, em que as bandas de uma imagem hiperespectral são divididas de forma disjunta em grupos por meio de sua correlação: isto é, bandas com maior correlação entre si concentram-se nos mesmos grupos. A partir do agrupamento, é feita uma redução do espaço de busca: toma-se uma banda de cada grupo. Aplica-se um algoritmo genético de busca sobre essas bandas de forma a determinar-se a melhor combinação de bandas. Um detalhe importante de impelmentação deste trabalho é a definição dos operadores genéticos. Adotou-se o sistema numérico dos inteiros em base 10 ao invés do sistema de codificação binária utilizado na maioria dos trabalhos relacionados. Os autores fizeram essa escolha para que, na operação de mutação, as alterações fiquem restritas aos seus próprios grupos. Isto é, caso um indivíduo sofra mutação em uma banda pertencente a um grupo, há a certeza de que a nova banda será pertencente ao mesmo grupo que a banda anterior, o que garante que todos os indivíduos da população sejam compostos por apenas uma banda de cada grupo. Finalmente, para reduzir novamente a dimenisionalidade das imagens, é utilizado um algoritmo de otimização para remover bandas redundantes dentre as restantes.

Um exemplo de seleção de bandas em imagens de sensoriamento remoto sem a utilização de algoritmos genéticos pode ser encontrado em Dallaqua, 2020 [34]. No trabalho, busca-se determinar, dentre as 7 bandas de imagens capturadas pelo sensor óptico Landsat-8 e as 3 bandas produzidas pelo PCA dessas respectivas imagens, a melhor combinação para a detecção de desmatamento. Um total de 135 combinações

(ou seja, todas as combinações possíveis) de bandas foram estudadas e rankeadas a partir de seu desempenho em um classificador SVM.

Além das abordagens vistas nos trabalhos relatados acima, [42] mostra que o uso de aprendizado por reforço profundo também pode ser um grande aliado do sensoriamento remoto. Para selecionar as bandas que melhor refletem as imagens hiperespectrais, este trabalho realizou o treino de um modelo que defronta esse problema como um processo de decisão de Markov, em que as escolhas devem ser sequencialmente feitas e a consequência de uma tomada de decisão não é clara para o agente (modelo) [43], que a parametriza. Após o treino (parametrização), o modelo explora completamente as imagens hiperespectrais e escolhe as bandas que melhor a representam. Além disso, duas medidas de recompensa foram definidas para simular um aprendizado por reforço nesse modelo de aprendizado profundo: a entropia da informação, que é utilizada para medir a riqueza das informações que as bandas selecionadas trazem e o coeficiente de correlação, que quantifica a força da correlação intra-bandas e entre as bandas selecionadas. No caso desse trabalho, a medida de correlação utilizada foi a de Pearson.

IV. Objetivo

Neste trabalho, objetiva-se, fazendo uso de algoritmos genéticos, determinar a combinação de bandas espectrais do sensor Landsat-8 que promove o melhor resultado na tarefa de detecção de desmatamento por meio de modelos de classificação.

V. Metodologia Experimental

A. Geração de segmentos via algoritmo MaskSLIC

Primeiramente, serão criados os segmentos a partir do algoritmo MaskSLIC. Esta etapa é necessária pois, subsequentemente, no algoritmo genético, esses segmentos serão classificados por uma SVM. Seguindo [34], os segmentos serão produzidos pelo algoritmo MaskSLIC sobre as imagens da base de dados na combinação PCA.

B. Algoritmo Genético

Para cada segmento, serão calculados seus respectivos descritores de Haralick. Isso será feito tratando cada banda de forma individual, ou seja: cada segmento terá 7 vetores de características, um referente a cada banda espectral capturada pelo Landsat-8. Então, no Algoritmo Genético, cada indivíduo da população será representado por um vetor de 7 dimensões, em que se adota a codificação estocástica. Dessa forma, cada posição do vetor contém um valor contínuo no intervalo $[0; 1]$, que representa sua probabilidade de ocorrência na população. Em um indivíduo i, para cada banda b, se seu valor correspondente no vetor estiver contido em um intervalo entre dois *tresholds* (limites superiores e inferiores), é realizada a sua inclusão na composição de bandas de i. Não há nenhuma restrição de tamanho para as combinações: um indivíduo pode utilizar desde 1 até todas as 7 bandas em sua composição. Como função *fitness*, que desempenha o papel de agente de seleção natural, será adotada a acurácia balanceada de um classificador SVM (Subseção II-D). Além da evolução estocástica mencionada anteriormente, o algoritmo implementado contém as operações de mutação na geração de novos indivíduos e de *crossover* entre os indivíduos selecionados para reprodução. Parâmetros como a tamanho da população, limites superior e inferior, quantidade de individuos selecionados para reprodução e número de gerações, pertinentes ao algoritmo genético, foram obtidos a partir da testagem manual. No escopo do classificador SVM, utilizado como agente seletor, foram testadas as funções *kernel* de "base radial"(*RBF*, na sigla em inglês) e sigmóide. Além disso, foram feitos experimentos com o valor "C"ou parâmetro de regularização. Foram testadas duas configurações: pesos iguais para ambas classes e pesos "balanceados", que fazem com que classes representadas em menor proporção no *dataset* (no caso, "não floresta") possuam margens mais estreitas (menor tolerância a erros). Ambas configurações são disponibilizadas nativamente para o classificador SVM da biblioteca *scikit-learn*. Os parâmetros referentes ao algortimo UMDA que produziram os melhores resultados (Seção VI) estão detalhados na Tabela II. Para o classificador SVM, a função *kernel* escolhida foi a *RBF* e a configuração dos parâmetros de regularização para cada classe foi a "balanceada". Uma visão geral da metodologia experimental adotada neste trabalho está exposta na Figura 2.

Tabela II
Valores utilizados para os parâmetros do classificador e do AG que foram ajustados

Parâmetro	Valor adotado
População inicial	80
Número de gerações	30
Número de indivíduos selecionados	10
Número de descentes gerados	20
Limite inferior	0.25
Limite superior	0.85

C. Base de Dados

A base de dados é composta por 11898 segmentos: 11471 da classe "floresta"e 427 da classe "não floresta", provenientes de dez regiões diferentes, sendo que a colaboração de cada região nesses números está exposta na Tabela III. Os segmentos foram criados, seguindo [34], pelo algoritmo MaskSLIC [29] sobre uma imagem em falsa-cor produzida pela aplicação da técnica PCA nas 7 bandas de uma cena capturada pelo sensor Landsat-8. A partir de dois critérios, foi feita uma filtragem dos segmentos disponíveis: selecionou-se os segmentos com Taxa de Homogeneidade (*HoR*) maior ou igual a 0.70 e área maior ou igual a 70 *pixels*, conforme feito em [34]. Os segmentos foram, então, salvos no formato de 7 bandas. Em seguida, foram calculados os descritores de Haralick [44] para cada um dos segmentos, utilizando a biblioteca *"mahotas"* [45]. Então, para cada segmento, calculou-se os descritores de cada banda separadamente. Além disso, por padrão de implementação da biblioteca, apenas os 13 primeiros descritores são imple-

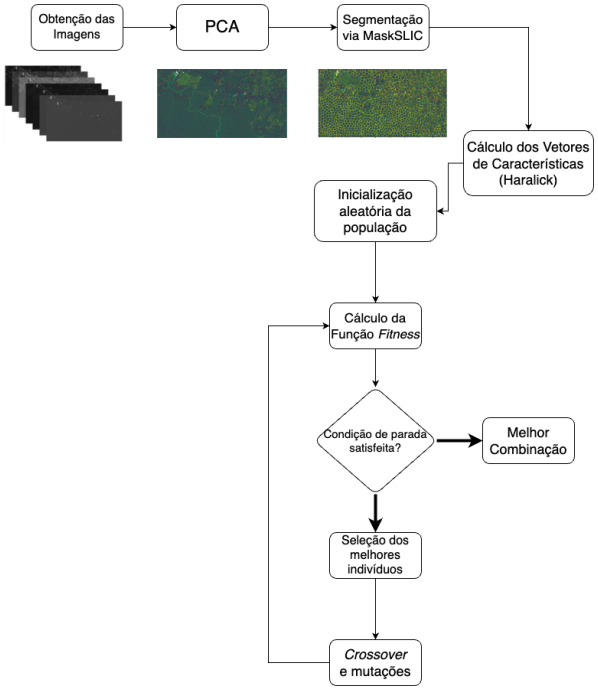

Figura 2. *Workflow* da abordagem proposta.

mentados. A Figura 3 detalha a distribuição das taxas de homogeneidade ao longo de toda a base de dados (isto é, todas as 10 regiões) por classe.

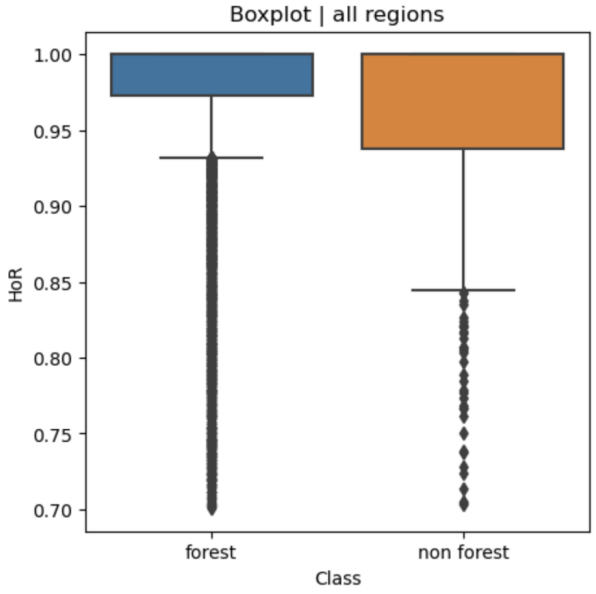

Figura 3. *Boxplot* das taxas de homogeneidade para a base de dados.

D. Protocolo de Validação

Após a criação da base de dados, foi realizada a sua divisão em treino, validação e teste. A separação dos 3 conjuntos foi feita de forma aleatória, tal que os conjuntos de treino, validação e teste fossem compostos, respectivamente, por 70%, 15% e 15% da base de dados total. A escolha destas proporções foi refeita tendo em vista a alta quantidade de imagens disponíveis no *dataset* decorrente das mudanças adotadas nesta entrega (Subseção V-C).

Tabela III
QUANTIDADE DE SEGMENTOS DISPONÍVEIS POR CLASSE E NO TOTAL EM CADA CONJUNTO DA BASE DE DADOS.

Região	Floresta	Não-floresta	Total
x 01	1585 segmentos	107 segmentos	1692 segmentos
x 02	1182 segmentos	134 segmentos	1316 segmentos
x 03	1167 segmentos	25 segmentos	1192 segmentos
x 04	475 segmentos	9 segmentos	484 segmentos
x 05	612 segmentos	36 segmentos	648 segmentos
x 06	1104 segmentos	17 segmentos	1121 segmentos
x 07	2016 segmentos	9 segmentos	2025 segmentos
x 08	928 segmentos	36 segmentos	964 segmentos
x 09	1582 segmentos	19 segmentos	1601 segmentos
x 10	820 segmentos	35 segmentos	855 segmentos

VI. Resultados & Discussões

Neste trabalho, implementou-se, com sucesso, o algoritmo UMDA para a tarefa de *feature selection* das bandas do sensor Landsat-8 para detecção de desmatamento. Os resultados obtidos com o UMDA superaram o modelo evolucional simples utilizado nas etapas anteriores do projeto em mais de $4,0\%$, graças à obtenção da combinação das bandas *[1, 5 e 6]* (Tabela I) do sensor Landsat-8, que possibilitou uma acurácia de teste de $93,5\%$ no classificador SVM. A Tabela IV expõe as três melhores combinações de bandas obtidas, bem como suas respectivas acurácias de validação e de teste. Estes resultados foram obtidos a partir da testagem manual de alguns parâmetros do algoritmo evolucional e do classificador SVM (utilizado como função *fitness*).

Tabela IV
MELHORES INDIVÍDUOS E OS VALORES DE ACURÁCIA BALANCEADA APRESENTADOS POR ELES NA TAREFA DE CLASSIFICAÇÃO

Bandas selecionadas	Acurácia validação	Acurácia teste
1, 5 e 6	$88,5\%$	$93,5\%$
1, 4, 5 e 6	$88,5\%$	$93,4\%$
1, 4 e 6	$84,7\%$	$91,9\%$

VII. Conclusões & Trabalhos Futuros

Neste artigo, buscou-se determinar se o uso de algoritmos evolucionais – em particular, o de Algoritmo de Distribuição Marginal Univariada (*UMDA*, na sigla em inglês) – é uma abordagem viável ao problema de determinação das melhores combinações de bandas para a detecção automátizada de desmatamento em imagens de sensoriamento remoto. A metodologia adotada foi largamente inspirada no trabalho de Dallaqua, 2020 [34] nas etapas de construção da base de dados e no uso do classificador SVM. Nossos resultados (Tabela IV) são congruentes com os encontrados por Dallaqua, que realizou uma busca extensiva entre todas as combinações possíveis

de bandas do Satélite Landsat-8 e alcançou a combinação ótima das bandas 4 e 6. Destacamos que, das três melhores bandas, todas continham a banda 6 (infravermelho de ondas curtas), além de que a banda 4 estava presente em duas das três melhores. Além disso, este resultado aponta que tais bandas são efetivas na detecção de desmatamento em regiões diversas, uma vez que nossas regiões de interesse, embora contidas no mesmo bioma, são diferentes das estudadas por Dallaqua. Como trabalhos futuros, os autores propõem, no âmbito da base de dados, a adoção de um protocolo de validação por separação de regiões, conferindo mais robustez aos resultados. Além disso, a realização de uma busca extensiva (*grid-search*) para os parâmetros do classificador SVM pode ser benéfica ao experimento, garantindo uma função *fitness* mais criteriosa. No mais, mostramos que algoritmos genéticos são uma abordagem válida para a tarefa de *feature selection* de bandas em imagens de satélite multiespectrais alcançando, por meio do algoritmo UMDA, resulados próximos aos obtidos por uma busca extensiva.

REFERÊNCIAS

[1] B. Dupuy, H. Maitre, and I. Amsallem, "Tropical forest management techniques: a review of the sustainability of forest management practices in tropical countries. Working paper: FAO/FPIRS/04 prepared for the World Bank Forest Policy Implementation Review and Strategy," 1999.

[2] C. Martin, *On the Edge: The State and Fate of the World's Tropical Rainforests*. Greystone Books Ltd, 2015.

[3] G. Urquhart, W. Chomentowski, D. Skole, and C. Barber, "Tropical deforestation," 1998.

[4] M. C. Hansen, P. V. Potapov, R. Moore, M. Hancher, S. Turubanova, A. Tyukavina, D. Thau, S. Stehman, S. Goetz, T. R. Loveland *et al.*, "High-resolution global maps of 21st-century forest cover change," *science*, vol. 342, no. 6160, pp. 850–853, 2013.

[5] A. Souza, A. M. Vieira Monteiro, C. Daleles Rennó, C. A. Almeida, D. de Morisson Valeriano, F. Morelli, L. Vinhas, L. E. P. Maurano, M. Adami, M. I. Sobral Escada, M. da Motta, and S. Amaral, "Metodologia Utilizada nos Projetos PRODES e DETER," *São José dos Campos: INPE*, 2019.

[6] INPE, "A estimativa da taxa de desmatamento por corte raso para a Amazônia Legal em 2019 é de 9.762 km^2," http://www.inpe.br/noticias/noticia.php?Cod_Noticia=5294, 2019, accessed: 2020-02-13.

[7] D. Campbell-Lendrum, J.-P. Dujardin, E. Martinez, M. D. Feliciangeli, J. E. Perez, L. N. M. P. d. Silans, and P. Desjeux, "Domestic and peridomestic transmission of American cutaneous leishmaniasis: changing epidemiological patterns present new control opportunities," *Memórias do Instituto Oswaldo Cruz*, vol. 96, no. 2, pp. 159–162, 2001.

[8] J. Walsh, D. Molyneux, and M. Birley, "Deforestation: effects on vector-borne disease," *Parasitology*, vol. 106, no. S1, pp. S55–S75, 1993.

[9] J. R. Coura and J. Borges-Pereira, "Chagas disease: 100 years after its discovery. A systemic review," *Acta tropica*, vol. 115, no. 1-2, pp. 5–13, 2010.

[10] A. Afelt, R. Frutos, and C. Devaux, "Bats, coronaviruses, and deforestation: Toward the emergence of novel infectious diseases?" *Frontiers in microbiology*, vol. 9, p. 702, 2018.

[11] T. E. Lovejoy and C. Nobre, "Amazon Tipping Point," *Science Advances*, vol. 4, no. 2, 2018. [Online]. Available: https://advances.sciencemag.org/content/4/2/eaat2340

[12] E. F. Luz *et al.*, "The ForestWatchers: A Citizen Cyberscience Project for Deforestation Monitoring in the Tropics," *Human Computation*, vol. 1, pp. 137–145, 2014.

[13] D. Tuia, M. Volpi, L. Copa, M. Kanevski, and J. Munoz-Mari, "A survey of active learning algorithms for supervised remote sensing image classification," *IEEE Journal of Selected Topics in Signal Processing*, vol. 5, no. 3, pp. 606–617, 2011.

[14] M. D. Soares, R. Santos, N. Vijaykumar, and L. Dutra, "Citizen science-based labeling of imprecisely segmented images: Case study and preliminary results," in *Collaborative Systems-Simposio Brasileiro de Sistemas Colaborativos (SBSC), 2010 Brazilian Symposium of. IEEE*, 2010, pp. 87 – 94.

[15] F. Grey, "Viewpoint: The age of citizen cyberscience," *Cern Courier*, vol. 29, 2009.

[16] J. Silvertown, "A new dawn for citizen science," *Trends in Ecology & Evolution*, vol. 24, no. 9, pp. 467–471, 2009.

[17] A. Irwin, "No PhDs needed: how citizen science is transforming research," *Nature*, vol. 562, no. 7728, p. 480—482, October 2018. [Online]. Available: https://doi.org/10.1038/d41586-018-07106-5

[18] T. Gura, "Citizen science: amateur experts," *Nature*, vol. 496, no. 7444, pp. 259–261, 2013.

[19] C. J. Guerrini, M. A. Majumder, M. J. Lewellyn, and A. L. McGuire, "Citizen science, public policy," *Science*, vol. 361, no. 6398, pp. 134–136, 2018.

[20] R. Bonney, C. Cooper, and H. Ballard, "The theory and practice of citizen science: Launching a new journal," *Citizen Science: Theory and Practice*, vol. 1, no. 1, 2016.

[21] F. Dallaqua, A. Fazenda, and F. Faria, "ForestEyes Project: Can Citizen Scientists Help Rainforests?" in *IEEE 15th International Conference on eScience*. IEEE, 9 2019, pp. 18–27.

[22] C. Buongiorno, L. Grove, and A. Salata, *Into the Zooniverse*. Zooniverse, 2019.

[23] A. M. Smith, S. Lynn, and C. J. Lintott, "An introduction to the zooniverse," in *First AAAI conference on human computation and crowdsourcing*, 2013.

[24] P. S. Singh and S. Karthikeyan, "Enhanced classification of remotely sensed hyperspectral images through efficient band selection using autoencoders and genetic algorithm," *Neural Computing and Applications*, vol. 34, no. 24, pp. 21 539–21 550, 2022.

[25] J.-P. Ma, Z.-B. Zheng, Q.-X. Tong, and L.-F. Zheng, "An application of genetic algorithms on band selection for hyperspectral image classification," in *Proceedings of the 2003 international conference on machine learning and cybernetics (IEEE Cat. No. 03EX693)*, vol. 5. IEEE, 2003, pp. 2810–2813.

[26] P. Norvig and S. Russell, *Inteligência Artificial*. ELSEVIER EDITORA, 2013, ch. 17. [Online]. Available: https://books.google.com.br/books?id=KhUQvgAACAAJ

[27] M. R. Heinen and F. S. Osório, "Autenticaçao de assinaturas utilizando análise de componentes principais e redes neurais artificiais," in *1st Workshop on Computational Intelligence (WCI 2006)*, 2006, pp. 1–6.

[28] J. Schmitt *et al.*, "Pré-processamento para a mineração de dados: uso da análise de componentes principais com escalonamento ótimo," 2005.

[29] B. Irving, "maskSLIC: regional superpixel generation with application to local pathology characterisation in medical images," *arXiv preprint arXiv:1606.09518*, 2016.

[30] A. C. Lorena and A. C. De Carvalho, "Uma introdução às support vector machines," *Revista de Informática Teórica e Aplicada*, vol. 14, no. 2, pp. 43–67, 2007.

[31] R. M. Haralick, K. Shanmugam, and I. Dinstein, "Textural features for image classification," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. SMC-3, no. 6, pp. 610–621, 1973.

[32] A. R. Gonçalves and F. J. Von Zuben, "Aprendizado probabilístico em algoritmos evolutivos para otimização em ambientes dinâmicos."

[33] J. Branke, C. Lode, and J. L. Shapiro, "Addressing sampling errors and diversity loss in umda," in *Proceedings of the 9th annual conference on Genetic and evolutionary computation*, 2007, pp. 508–515.

[34] F. Beatriz Rojas Dallaqua, "Projeto foresteyes - ciÊncia cidadÃ e aprendizado de mÁquina na detecÇÃo de Áreas desmatadas em florestas tropicais," Ph.D. dissertation, 2020.

[35] X. Zhang, Q. Sun, and J. Li, "Optimal band selection for high dimensional remote sensing data using genetic algorithm," in *Second International Conference on Earth Observation for Global Changes*, vol. 7471. SPIE, 2009, pp. 522–528.

[36] Z. Lin and G. Zhang, "Genetic algorithm-based parameter optimization for eo-1 hyperion remote sensing image classification," *European Journal of Remote Sensing*, vol. 53, no. 1, pp. 124–131, 2020. [Online]. Available: https://doi.org/10.1080/22797254.2020.1747949

[37] H. Yao and L. Tian, "A genetic-algorithm-based selective principal component analysis (ga-spca) method for high-dimensional data feature extraction," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 41, no. 6, pp. 1469–1478, 2003.

[38] S. H. Amroabadi, M. R. Ahmadzadeh, and A. Hekmatnia, "Mass detection in mammograms using ga based pca and haralick features selection," in *2011 19th Iranian Conference on Electrical Engineering*, 2011, pp. 1–4.

[39] D. Saqui, "Um novo método wrapper multiobjetivo para seleção de bandas de imagens hiperspectrais," 2020.

[40] K. Nagasubramanian, S. Jones, S. Sarkar, A. K. Singh, A. Singh, and B. Ganapathysubramanian, "Hyperspectral band selection using genetic algorithm and support vector machines for early identification of charcoal rot disease in soybean stems," *Plant methods*, vol. 14, pp. 1–13, 2018.

[41] S. Li, H. Wu, D. Wan, and J. Zhu, "An effective feature selection method for hyperspectral image classification based on genetic algorithm and support vector machine," *Knowledge-Based Systems*, vol. 24, no. 1, pp. 40–48, 2011.

[42] L. Mou, S. Saha, Y. Hua, F. Bovolo, L. Bruzzone, and X. X. Zhu, "Deep reinforcement learning for band selection in hyperspectral image classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–14, 2022.

[43] J. Pellegrini and J. Wainer, "Processos de decisão de markov: um tutorial," *Revista de Informática Teórica e Aplicada*, vol. 14, no. 2, pp. 133–179, 2007.

[44] R. M. Haralick, K. Shanmugam *et al.*, "Textural features for image classification," *IEEE Transactions on systems, man, and cybernetics*, no. 6, pp. 610–621, 1973.

[45] L. P. Coelho, "Mahotas: Open source software for scriptable computer vision," *Journal of Open Research Software*, vol. 1, 2013. [Online]. Available: http://dx.doi.org/10.5334/jors.ac