sid.inpe.br/mtc-m21d/2021/06.24.13.18.26-TDI

AVALIAÇÃO DE MÉTRICAS EXTRAÍDAS DE SÉRIES TEMPORAIS DE IMAGENS DE SATÉLITE EM APLICAÇÕES DE APRENDIZADO DE MÁQUINA

Felipe Carvalho de Souza

Dissertação de Mestrado do Curso de Pós-Graduação em Computação Aplicada, orientada pelos Drs. Karine Reis Ferreira Gomes, e Rafael Duarte Coelho dos Santos, aprovada em 11 de junho de 2021.

URL do documento original:
<http://urlib.net/8JMKD3MGP3W34T/44TSDCA>

INPE
São José dos Campos
2021

PUBLICADO POR:

Instituto Nacional de Pesquisas Espaciais - INPE
Coordenação de Ensino, Pesquisa e Extensão (COEPE)
Divisão de Biblioteca (DIBIB)
CEP 12.227-010
São José dos Campos - SP - Brasil
Tel.:(012) 3208-6923/7348
E-mail: pubtc@inpe.br

sid.inpe.br/mtc-m21d/2021/06.24.13.18.26-TDI

AVALIAÇÃO DE MÉTRICAS EXTRAÍDAS DE SÉRIES TEMPORAIS DE IMAGENS DE SATÉLITE EM APLICAÇÕES DE APRENDIZADO DE MÁQUINA

Felipe Carvalho de Souza

Dissertação de Mestrado do Curso de Pós-Graduação em Computação Aplicada, orientada pelos Drs. Karine Reis Ferreira Gomes, e Rafael Duarte Coelho dos Santos, aprovada em 11 de junho de 2021.

URL do documento original:
<http://urlib.net/8JMKD3MGP3W34T/44TSDCA>

INPE
São José dos Campos
2021

Dados Internacionais de Catalogação na Publicação (CIP)

Souza, Felipe Carvalho de.

So94a Avaliação de métricas extraídas de séries temporais de imagens de satélite em aplicações de aprendizado de máquina / Felipe Carvalho de Souza. – São José dos Campos : INPE, 2021.

xx + 93 p. ; (sid.inpe.br/mtc-m21d/2021/06.24.13.18.26-TDI)

Dissertação (Mestrado em Computação Aplicada) – Instituto Nacional de Pesquisas Espaciais, São José dos Campos, 2021.

Orientadores : Drs. Karine Reis Ferreira Gomes, e Rafael Duarte Coelho dos Santos.

1. Séries temporais de imagens de satélites. 2. Cubos de dados de observações da Terra. 3. Extração de métricas temporais. 4. Agrupamento de séries temporais. 5. Classificação de séries temporais. I.Título.

CDU 519.246.8:528.8

MINISTÉRIO DA CIÊNCIA, TECNOLOGIA E INOVAÇÕES

INSTITUTO NACIONAL DE PESQUISAS ESPACIAIS

DEFESA FINAL DE DISSERTAÇÃO DE FELIPE CARVALHO DE SOUZA
BANCA Nº 154/2021, REG 866864/2019

No dia 11 de junho de 2021, às 14h00min, por Vídeo Conferência, o(a) aluno(a) mencionado(a) acima defendeu seu trabalho final (apresentação oral seguida de arguição) perante uma Banca Examinadora, cujos membros estão listados abaixo. O(A) aluno(a) foi APROVADO(A) pela Banca Examinadora por unanimidade, em cumprimento ao requisito exigido para obtenção do Título de Mestre em Computação Aplicada . O trabalho precisa da incorporação das correções sugeridas pela Banca Examinadora e revisão final pelo(s) orientador(es).

Título: “Avaliação de métricas extraídas de séries temporais de imagens de satélite em aplicações de aprendizado de máquina”

Eu, Thales Sehn Korting, Presidente da Banca Examinadora, assino esta ATA, em nome de todos os membros, com o consentimento dos mesmos.

Membros da Banca

Dr. Thales Sehn Korting - Presidente - INPE
Dr. Karine Reis Ferreira Gomes - Orientador - INPE
Dr. Rafael Duarte Coelho dos Santos - Orientador - INPE
Dr. Michel Eustáquio Dantas Chaves - Membro Interno - INPE
Dr. Diego Furtado Silva - Membro Externo - UFSCAR

Documento assinado eletronicamente por **Thales Sehn Korting**, **Pesquisador**, em 15/06/2021, às 11:03 (horário oficial de Brasília), com fundamento no art. 6º do Decreto nº 8.539, de 8 de outubro de 2015.

A autenticidade deste documento pode ser conferida no site http://sei.mctic.gov.br/verifica.html, informando o código verificador **7602864** e o código CRC **B541F6AA**.

AGRADECIMENTOS

Primeiramente, gostaria de agradecer à minha Santa Teresinha, que me trouxe conforto para que eu pudesse finalizar este trabalho.

Aos meus pais, Ana e Luciano, que sempre me apoiaram e incentivaram em todos os momentos.

Agradeço aos meus orientadores, Karine Reis e Rafael Santos, pelas ideias, discussões, revisões e por todo apoio do início ao fim deste trabalho.

Aos colaboradores deste trabalho, Rolf Simões, Lorena Santos, Alber Sanchez, Felipe Carlos e Amita Muralikrishna, que me ajudaram com ideias, revisões e sugestões. Muito obrigado, pessoal!

Aos meus amigos do ap 406, Adriano, Gabriel, Guilherme, Helvécio, Marujo, muito obrigado pela companhia e conversas descontraídas.

Aos amigos(as) da faculdade, Gustavo (gus), Gabriel (god), Ítalo (impera) e Daniela (Dani), que sempre me apoiaram nessa jornada acadêmica.

Um agradecimento especial ao meu amigo Felipe Carlos, que sempre me incentivou em todas as empreitadas e puxou minha orelha nos momentos certos. Muito obrigado, Minino!

Aos colaboradores do INPE, em especial à Jéssica e à Glória, que sempre me auxiliaram com as questões administrativas ou quando eu perdia a chave do laboratório.

À todas as pessoas que me ajudaram de alguma forma durante essa jornada do mestrado. Muito obrigado!

À Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) pelo apoio financeiro.

RESUMO

As atividades humanas estão impactando ecossistemas e paisagens em todo o Planeta. Pesquisadores procuram entender como esses impactos podem interferir na dinâmica dos sistemas terrestres, tais como regime de chuvas e mudanças climáticas. Uma das formas de quantificar parte desses impactos é através de mapas de uso e cobertura da terra gerados a partir de imagens de sensoriamento remoto. Com os avanços tecnológicos que tornam as observações da superfície terrestre cada vez mais precisas e consistentes temporalmente, pesquisadores organizam esse grande volume de imagens de satélite como cubos de dados de observação da Terra, integrando a dimensão temporal desses dados. Uma das principais aplicações de cubos de dados de observação da Terra é a análise de séries temporais de imagens de satélites, um importante tema de pesquisa no campo de uso e cobertura da Terra. Séries temporais de imagens de satélites têm demonstrado potencial em detectar, identificar e monitorar alterações em mudanças no uso e cobertura da terra. Com a facilidade da coleta de séries temporais, provindas de diferentes sensores, através de cubos de dados de observação da Terra, torna-se cada vez mais necessária a extração de informações dessa grande massa de dados. Os avanços na área de inteligência artificial permitem que grandes volumes de dados sejam processados de forma automática a partir de modelos de aprendizado de máquina, o que permite que cientistas usem cada vez mais dados para entender que mudanças estão ocorrendo em nossos ecossistemas. Várias abordagens que usam aprendizado de máquina para a geração de mapas de uso e cobertura da terra através de séries temporais foram desenvolvidas nos últimos anos. Duas delas são objeto de investigação nesta dissertação: a abordagem que usa séries temporais completas e a abordagem que usa métricas derivadas de séries temporais. Nesta pesquisa, são conduzidos três estudos de casos com diferentes aplicações para comparar essas abordagens. No primeiro objetivou-se avaliar as amostras de uso e cobertura da terra usando o método de agrupamento de Mapas Auto-Organizáveis. No segundo estudo de caso, foi realizada uma classificação de corpos d'água em que se objetivou avaliar os desempenhos entre as abordagens de séries temporais completas e de métricas. Por fim, no último estudo de caso, foi realizada uma classificação de uso e cobertura da terra a partir das duas abordagens e objetivou-se avaliar os respectivos desempenhos computacionais e de acurácia na classificação. Os resultados mostram que a abordagem de métricas de séries temporais produzem mapas de uso e cobertura com qualidade igual ou superior aos gerados usando séries temporais completas. Além de melhores resultados, o tempo de processamento é um fator determinante para o uso da abordagem de métricas de séries temporais.

Palavras-chave: Séries Temporais de Imagens de Satélites. Cubos de Dados de Observações da Terra. Extração de Métricas Temporais. Agrupamento de Séries Temporais. Classificação de Séries Temporais. Seleção de Atributos.

EVALUATION OF METRICS EXTRACTED FROM SATELLITE IMAGE TIME SERIES IN MACHINE LEARNING APPLICATIONS

ABSTRACT

Human activities are impacting ecosystems and landscapes all over the planet. Researchers seek to understand how these impacts can interfere with the dynamics of earth systems, such as rainfall regime and climate change. One way to quantify part of these impacts is through land use and land cover maps generated from remote sensing images. Through technological advances that make observations of the Earth's surface increasingly accurate and temporally consistent, researchers are organizing this large volume of satellite imagery as Earth observation data cubes, integrating the temporal dimension of this data. One of the main applications of Earth observation data cubes is the analysis of satellite image time series, an important research topic in the field of land use and land cover. Time series of satellite images have shown potential in detecting, identifying, and monitoring changes in land use and land cover. Given the ease of collecting time series from different sensors through Earth observation data cubes, it is becoming increasingly necessary to extract information from this large mass of data. Advances in artificial intelligence allow large volumes of data to be processed automatically from machine learning models, allowing scientists to use increasingly more data to understand what changes are taking place in our ecosystems. Several approaches that use machine learning for the generation of land use and land cover maps through time series have been developed in recent years. Two of these are investigated in this work: the approach that uses complete time series and the approach that uses metrics derived from time series. In this research, the authors conducted three case studies with different applications to compare these approaches. In the first one, it was aimed to evaluate the land use and land cover samples using the Self-Organizing Maps clustering method. For the second case study, the aim was to evaluate the performance of the complete time series and metric approaches in classifying water bodies. Finally, in the last case study, the aim was to evaluate the respective computational performance and accuracy in land use and land cover classification from the two approaches. The results show that the time series metrics approach produces maps of land use and cover with equal or higher quality than those generated using complete time series. Besides better results, processing time is a determining factor for using the time series metrics approach.

Keywords: Satellite Image Time Series. Earth Observation Data Cubes. Temporal Metrics Extraction. Time Series Clustering. Time Series Classification. Feature Selection.

LISTA DE FIGURAS

Pág.

LISTA DE TABELAS

Pág.

LISTA DE ABREVIATURAS E SIGLAS

API	–	*Application Programming Interface*
BDC	–	*Brazil Data Cube*
EO	–	*Earth Observation*
EVI	–	*Enhanced Vegetation Index*
GEMI	–	*Global Environmental Monitoring Index*
GNDVI	–	*Green Normalized Difference Vegetation Index*
GRRF	–	*Guided Regularized Random Forest*
INPE	–	Instituto Nacional de Pesquisas Espaciais
IBGE	–	Instituto Brasileiro de Geografia e Estatística
ML	–	*Machine Learning*
MNDWI	–	*Modified Normalized Difference Water Index*
NDVI	–	*Normalized Diference Vegetation Index*
NDWI	–	*Normalized Difference Water Index*
RF	–	*Random Forests*
SOM	–	*Self Organizing-Maps*
SVM	–	*Support Vector Machines*
SITS	–	*Satellite Image Time Series*

LISTA DE SÍMBOLOS

min	–	minutos
m	–	metros
ha	–	hectare
GB	–	Gigabyte

SUMÁRIO

Pág.

1 INTRODUÇÃO

A pressão por recursos naturais tem produzido um impacto ambiental sem precedentes ao planeta Terra. O monitoramento da mudança de uso e cobertura da terra tornou-se uma ferramenta fundamental para quantificar esses impactos e subsidiar políticas públicas que promovam a preservação do meio ambiente e o gerenciamento de recursos naturais. A expansão da agricultura, das pastagens e de áreas urbanas, são alguns dos principais usos da terra no Brasil. Quando não manejadas corretamente, essas mudanças podem provocar perdas na biodiversidade, diminuir a capacidade dos ecossistemas de manter os recursos de água doce e florestais, regular o clima e a qualidade do ar (FOLEY et al., 2005), além de contribuir para a emissão de gases de efeito estufa. A conciliação entre produção e preservação é um dos principais desafios da humanidade.

O Brasil é um dos principais produtores e exportadores de *commodities* do mundo, e produtos como a soja, açúcar, café, laranja, aves, carne bovina e etanol são exemplos de matérias-primas exportadas pelo país (MARTINELLI et al., 2010). Conforme a FOOD AND AGRICULTURAL ORGANIZATION - FAO (2021), entre os anos de 1994 até 2019, a cana-de-açúcar foi a *commodity* agrícola mais produzida no Brasil. Isso deveu-se ao acesso do produto aos mercados globais, os elevados preços mundiais do açúcar e as políticas de energia renovável brasileiras e de outros países (SPERA et al., 2017).

Além de promover a qualidade do meio ambiente e subsidiar políticas públicas, o monitoramento de uso e cobertura da terra é de suma importância para se medir a produtividade agrícola. Imagens de satélites são a principal fonte de informações para a geração de mapas devido à ampla cobertura geográfica e a observação periódica da superfície do planeta (GÓMEZ et al., 2016). Somente em 2019, estima-se que, pelo menos cinco petabytes de imagens foram produzidas pelos satélites Landsat-7, Landsat-8, MODIS (unidades Terra e Aqua) e Sentinel-1/2/3 (SOILLE et al., 2018). Devido às políticas de dados abertos adotadas por várias agências espaciais, ampliou-se a facilidade de aquisição de imagens de satélite em um volume e frequência jamais vistas. Atualmente, grandes volumes de imagens de observação da Terra (do inglês, *Earth Observation*, EO) são gerados e disponibilizados para o público de forma aberta.

Isso possibilita o uso de técnicas baseadas em dados multi-temporais e multi-espectrais para monitorar as alterações dos ecossistemas (COPPIN et al., 2004). Uma ampla gama de informações pode ser derivada de dados de satélites de EO (PETTO-

RELLI et al., 2005), fornecendo indicadores espacialmente explícitos com atualizações constantes (HÜTTICH et al., 2009). Por exemplo, Pekel et al. (2016) produziram uma máscara global de corpos d'água, entre os anos de 1985 até 2020, através de imagens da família de satélites Landsat.

Para suportar a análise de séries temporais, esse grande volume de imagens têm sido organizadas como cubos de dados de EO (NATIVI et al., 2017). Cubos de dados de EO são estruturas de dados multidimensionais, com três ou mais dimensões, incluindo espaço e tempo. São usadas para tornar grandes coleções de imagens de satélite prontas para análise e facilmente acessíveis (APPEL; PEBESMA, 2019).

Atualmente, existem diversas iniciativas para se gerar cubos de dados em diversos países e continentes, por exemplo, *Australian Data Cube* (LEWIS et al., 2017), *Swiss Data Cube* (GIULIANI et al., 2017) e *Africa Regional Data Cube* (KILLOUGH, 2019). A principal iniciativa no Brasil é o projeto *Brazil Data Cube* (BDC), conduzido pelo Instituto Nacional de Pesquisas Espaciais (INPE) desde 2019 (FERREIRA et al., 2020).

O projeto BDC visa produzir dados prontos para análise (do inglês, *Analysis-Ready Data*, ARD) e cubos de dados multidimensionais para todo o território nacional a partir de grandes volumes de imagens de sensoriamento remoto de média resolução dos satélites CBERS-4 e 4A, Sentinel-2 e Landsat-8. Além de desenvolver tecnologias para acessar e visualizar grandes volumes de imagens de sensoriamento remoto, o projeto desenvolve algoritmos de aprendizado de máquina (do inglês, *machine learning*, ML) e análise de séries temporais de imagens de satélite visando extrair informações de uso e cobertura da terra a partir dos cubos de dados gerados pelo projeto (SIMOES et al., 2020; PICOLI et al., 2020; SANTOS et al., 2021a; SANTOS et al., 2021b).

A análise de séries temporais têm sido amplamente aplicada na produção de informações sobre uso e cobertura da terra (EPIPHANIO et al., 2010; GRIFFITHS et al., 2014; FRANKLIN et al., 2015; GÓMEZ et al., 2016; PICOLI et al., 2018; SIMOES et al., 2020), tais como geração de mapas (SIMOES et al., 2020) e avaliação de amostras de uso e cobertura da terra (SANTOS et al., 2021c). Essa abordagem é chamada em Aghabozorgi et al. (2015), Gómez et al. (2016) de séries temporais completas e usa todos os dados disponíveis das séries temporais para derivar modelos de ML para fins de agrupamento ou classificação.

Devido ao grande volume de dados envolvidos na geração de mapas nessa abordagem, é comum o uso de métricas derivadas de séries temporais para reduzir o espaço dimensional de atributos e, consequentemente, a quantidade de dados a serem manipulados durante esse processo (PARENTE et al., 2019; POTAPOV et al., 2020). Essa técnica de reduzir a dimensionalidade das séries temporais é chamada de abordagem baseada em métricas (AGHABOZORGI et al., 2015), na qual os atributos são derivados de séries temporais para obter valores que as caracterizam usando um menor espaço dimensional. Por exemplo, em Potapov et al. (2020) são utilizadas métricas estatísticas para a redução de dimensionalidade e extração de informações de séries temporais para gerar um mapa global de uso e cobertura da terra. A abordagem de métricas de séries temporais é baseada em técnicas de extração de atributos que fazem a sumarização das séries temporais para cada banda ou índice espectral. Nesta dissertação, delimitou-se ao uso do conjunto de métricas propostas pelos autores (KÖRTING et al., 2013). Essas métricas foram usadas em diversos trabalhos de sensoriamento remoto para a classificação de uso e cobertura da terra (NEVES et al., 2016; UEHARA et al., 2020; RODRIGUES et al., 2020; SOARES et al., 2020).

Considerando as abordagens de séries temporais completas e métricas de séries temporais, uma questão que emerge é: *a abordagem baseada em métricas de séries temporais possui a mesma capacidade de generalização que a abordagem baseada em séries temporais completas no contexto de modelos de ML?*

Em uma meta-análise de artigos publicados desde 2001 em periódicos de alto impacto em sensoriamento remoto (KHATAMI et al., 2016) constatou-se que o uso de abordagens multi-temporais, tais como séries temporais completas, aumentam a acurácia global em 6.9% quando comparadas acurácia global média nos diferentes métodos empregados nos trabalhos analisados. Entretanto, a meta-análise não considerou trabalhos que fizeram uso de abordagem baseada em métricas de séries temporais. Na literatura de sensoriamento remoto consultada, não foi encontrado nenhum trabalho que comparasse os resultados obtidos pelas abordagens de métricas de séries temporais e de séries temporais completas.

Para responder à questão de pesquisa, neste trabalho foram realizados três estudos de caso comparando as abordagens de métricas de séries temporais e séries temporais completas em aplicações de uso e cobertura da terra usando métodos de ML. O primeiro estudo de caso realizou uma avaliação de um conjunto de amostras de uso e cobertura da terra proposta em Santos et al. (2019) usando as duas abordagens. O segundo e o terceiro estudos de caso realizam uma classificação de uso e cobertura

da terra para diferentes classes e áreas de estudo. A hipótese de trabalho foi que a abordagem de métricas extraídas de séries temporais possui, pelo menos, a mesma capacidade de generalização em modelos de ML com menor custo computacional. Os objetivos que conduziram os estudos de caso são apresentados a seguir.

1.1 Objetivos

O objetivo geral deste trabalho é avaliar o uso de métricas extraídas de séries temporais de imagens de satélites em diferentes aplicações de uso e cobertura da terra usando técnicas de ML e comparar as mesmas aplicações com o uso de séries temporais completas.

Os objetivos específicos deste trabalho são:

a) Comparar as abordagens de séries temporais completas e de métricas de séries temporais no contexto de avaliação de amostras de uso e cobertura da terra.

b) Comparar as abordagens de séries temporais completas e de métricas de séries temporais no contexto de classificação de uso e cobertura da terra.

c) Avaliar o custo computacional das diferentes abordagens para a geração de mapas de uso e cobertura da terra.

d) Avaliar as métricas propostas por Körting et al. (2013) que são mais relevantes para as diferentes aplicações de uso e cobertura da terra.

1.2 Contribuições

As contribuições deste trabalho são dividas em duas partes: acadêmicas e tecnológicas. Nas contribuições acadêmicas, as comparações das abordagens de séries temporais completas e métricas de séries temporais foram realizadas em diferentes aplicações, o que aumenta o alcance das conclusões. Com o estudo em diferentes regiões utilizando distintos conjuntos de amostras e cubos de dados, procurou-se reduzir o viés espacial dos resultados.

Os resultados obtidos neste trabalho permitem fazer recomendações sobre as principais métricas que podem ser geradas no âmbito do projeto BDC para aprimorar a análise e geração de produtos de uso e cobertura da terra.

As contribuições decorrentes do desenvolvimento deste trabalho consistem em publicações de artigos em congressos nacionais e internacionais, pacotes de *software* e implementação de funcionalidades em pacotes de *softwares* já existentes. A seguir é apresenta uma lista das principais contribuições:

a) Publicação do artigo: "ggsom: ferramenta de visualização baseada em mapas auto-organizáveis" (SOUZA et al., 2019), no congresso de Encontro Nacional de Modelagem Computacional XXII.

b) Co-autoria do artigo: "rstac: an R package to access spatiotemporal asset catalog satellite imagery" (aceito para publicação na conferência internacional IGARSS 2021 - *IEEE International Geoscience and Remote Sensing Symposium*, Julho, 2021).

c) Pacote **ggsom** para visualização de séries temporais na linguagem R. Link: `github.com/OldLipe/ggsom`.

d) Pacote cliente em R **rstac** para o acesso aos metadados de cubo de dados padronizados pelo *SpatioTemporal Asset Catalog* (STAC). Link: `github.com/brazil-data-cube/rstac`

e) Pacote **sitsdraft** para a organização e padronização de classificações de uso e cobertura da terra de acordo com as premissas de pesquisa reprodutível. Link: `github.com/OldLipe/sitsdraft/`.

f) Pacote **sitsfeats** para a extração de métricas básicas e polares usando **RcppArmadillo**. Link: `github.com/OldLipe/sitsfeats/`.

g) Integração do pacote **sits** com os cubos de dados do *Brazil Data Cube*, Sentinel-AWS e *Digital Earth Africa*. Link: `e-sensing.github.io/sitsbook`.

h) Integração do pacote **sits** para a geração de cubo de dados a partir da biblioteca **gdalcubes**. Link: `e-sensing.github.io/sitsbook`.

i) Implementação do submódulo de geração de cubos de métricas no pacote **sits**. Link: `github.com/OldLipe/sits/tree/experimentos-v0.11.0`.

1.3 Organização

Esta dissertação está organizada da seguinte forma: o Capítulo 2, apresenta uma revisão dos principais conceitos usados neste trabalho. O Capítulo 3 descreve a

metodologia comum relacionada a extração e seleção de atributos que foi aplicada nos estudos de caso.

Em seguida, nos três capítulos seguintes, são apresentados os estudos de caso. O Capítulo 4 descreve o primeiro estudo de caso, em que é conduzida uma avaliação de amostras de uso e cobertura da terra. O Capítulo 5 apresenta o segundo estudo de caso, em que é realizada uma classificação de corpos d'água avaliando as métricas mais importantes para a detecção dessa classe de cobertura. O Capítulo 6 descreve o terceiro estudo de caso que compara a geração de mapas de uso e cobertura da terra utilizando as abordagens de séries temporais completas e de métricas de séries temporais. Alguns detalhes específicos dos dados e métodos utilizados, bem como dos resultados obtidos, são descritos nesses capítulos. Por fim, as considerações finais são apresentadas no Capítulo 7.

2 REFERENCIAL TEÓRICO

Neste Capítulo, são apresentados os conceitos necessários para o entendimento deste trabalho. Apresenta-se, na Seção 2.1, o conceito de cubos de dados de observação de Terra. Na Seção 2.2, são apresentados os conceitos de séries temporais de imagens de satélites de observação da Terra. Apresenta-se, na Seção 2.3, o conceito de aprendizado de máquina e suas aplicações na área de sensoriamento remoto. Na Seção 2.4 apresentam-se as técnicas de classificação de séries temporais. Na Seção 2.5, apresenta-se a técnica de agrupamento de Mapas Auto-Organizáveis. Por fim, na Seção 2.6, aborda-se o conceito de redução de dimensionalidade em séries temporais.

2.1 Cubos de dados de observação da Terra

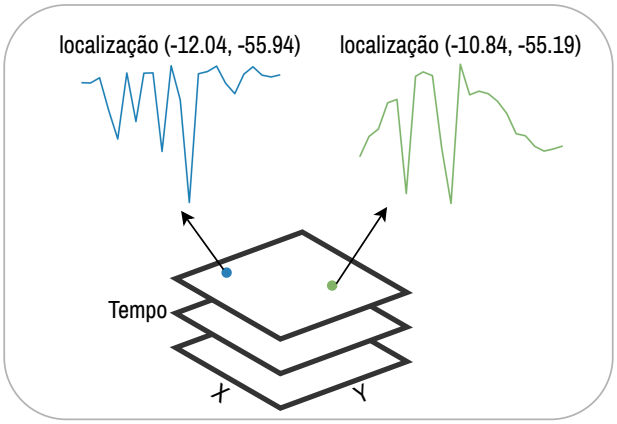
Uma forma de organizar o grande volume de imagens de satélite disponibilizadas por diferentes provedores de dados é através de cubos de dados de observação da Terra (GIULIANI et al., 2017; NATIVI et al., 2017). Cubos de dados de observação da Terra (ou simplesmente chamado aqui de Cubo de dados) são estruturas de dados multidimensionais, com três ou mais dimensões, que incluem espaço, tempo e propriedades espectrais. Os cubos de dados são usados para tornar grandes coleções de imagens de satélite prontas para análise facilmente acessíveis (APPEL; PEBESMA, 2019).

Cubos de dados podem ser definidos como um conjunto de séries temporais associadas à *pixels* alinhados espacialmente. Lu et al. (2018) definem cubo de dados como uma matriz de quatro dimensões: x (longitude); y (latitude); t (tempo) e as bandas espectrais. Um cubo pode ser visto como uma função que associa valores de observações espectrais a uma localização no cubo (x, y, t). Com o uso de cubos de dados, operações como extração de séries temporais tornam-se mais simples (APPEL; PEBESMA, 2019). Em um cubo de dados, uma série temporal pode ser obtida fornecendo-se uma localização espacial e um intervalo de tempo. Uma imagem pode ser recuperada a partir de um retângulo envolvente (intervalo espacial em duas dimensões) e uma data específica. A Figura 2.1 apresenta um exemplo de extração de séries temporais a partir de duas localizações espaciais.

A dimensão espacial é discretizada de acordo com uma resolução espacial, geralmente a mesma resolução das imagens usadas para gerar o cubo. Similarmente, a dimensão temporal é discretizada de acordo com um período pré-definido. Uma vez estabelecidos uma região espacial e um intervalo temporal para os quais o cubo será válido, o cubo de dados é gerado selecionando as imagens de satélite obtidas em intervalos re-

gulares. Quando mais de um valor estiver disponível para a mesma posição no cubo (x, y, t), geralmente, um método de escolha do melhor *pixel* é usado para decidir quais imagens serão selecionadas. Nessa definição, toda localização espacial possuirá as mesmas ocorrências temporais e, de modo inverso, cada ocorrência temporal possuirá valores para o mesmo domínio espacial, garantindo que toda localização estará associada a um valor observado (FERREIRA et al., 2020).

Figura 2.1 - Exemplo de extração de séries temporais em um cubo de dados.



Fonte: Próprio Autor.

Atualmente existem diversas iniciativas de estruturar cubos de dados em nível nacional, tais como o cubo de dados da Suíça (GIULIANI et al., 2017) e da Austrália (LEWIS et al., 2017). No Brasil, a principal iniciativa é realizada pelo Instituto Nacional de Pesquisas Espaciais (INPE) com o projeto BDC. O projeto BDC visa criar cubos de dados multidimensionais prontos para análise para todo o território brasileiro. Esse projeto tem quatro objetivos principais (FERREIRA et al., 2020):

- Criar dados prontos para análise (do inglês *Analysis-Ready Data*, ARD), a partir de imagens de sensoriamento remoto de média resolução espacial (20 a 60 metros) dos satélites Landsat-8, CBERS-4 e Sentinel-2, para o Brasil;
- Modelar dados de observação da Terra como cubos multidimensionais incluindo dimensões espaciais, temporais e de atributos;
- Propor e desenvolver novos métodos e tecnologias de *big data* para armazenar e processar esse grande volume de dados de observação da Terra e

para analisar e extrair informações de uso e cobertura da terra a partir desses dados usando técnicas de análise de séries temporais, aprendizado de máquina e procedimentos de processamento de imagens;

- Gerar informações sobre mudanças de uso e cobertura da terra utilizando os cubos de dados e métodos desenvolvidos neste projeto.

Neste cenário, existem algumas ferramentas que suportam o uso de cubos de dados e que permitem a análise de séries temporais. Por exemplo, o *Open Data Cube* (GOMES et al., 2020) que provê diversas funcionalidades para se trabalhar com cubos de dados de vários satélites. Outro exemplo é o pacote em código aberto desenvolvido na linguagem R, *sits* (SIMOES et al., 2021), que fornece uma interface de programação (do inglês, *Application Programming Interface*, API) simples e intuitiva que facilita o uso de cubo de dados para a geração de mapas de uso e cobertura da terra usando modelos de aprendizado de máquina e séries temporais. O pacote possui a capacidade de acessar cubos de dados em diferentes plataformas de computação em nuvem.

2.2 Séries temporais de imagens de satélites de observação da Terra

Séries temporais podem ser definidas como um conjunto de valores obtidos a partir de medições sequenciais feitas temporalmente (ESLING; AGON, 2012). Dados de séries temporais são aplicáveis em diversos problemas reais, situados em vários campos de pesquisa, por exemplo, previsão econômica (SONG; LI, 2008), detecção de intrusão em redes de computadores (ZHONG et al., 2007), classificação de expressão genética (LIN et al., 2008) e monitoramento médico (BURKOM et al., 2007).

No contexto de observação da Terra, aplicações que envolvem o monitoramento de uso da terra destacam-se pelo uso de séries temporais, pois possibilitam uma análise ao longo do tempo do alvo de estudo. Estudos relacionados à detecção de fenologia em culturas (ZHENG et al., 2016), análise e classificação de mudanças no uso e cobertura da terra (TOURE et al., 2018; SHAO et al., 2016), são exemplos do uso de séries temporais de observação da Terra.

As séries temporais extraídas de imagens de satélites podem ser compostas por diferentes atributos como bandas e índices espectrais. Cada banda espectral exibe certas características de seus alvos (KUENZER et al., 2015). Por exemplo, bandas espectrais no comprimento de onda da luz visível medem a reflectância dos pigmentos das superfícies. Bandas em outros comprimentos de onda podem capturar característi-

cas físicas como a temperatura da superfície terrestre, ou do topo da atmosfera (do inglês, *Top of Atmosphere*, TOA) (KUENZER et al., 2015).

Os índices espectrais são amplamente utilizados em trabalhos que envolvem a caracterização do uso e da cobertura da terra (ZENG et al., 2020). Eles são obtidos através de transformações realizadas nas bandas espectrais observadas. Vários fatores físicos associados ao tipo de cobertura influenciam os níveis de reflectância nas diferentes bandas do espectro. Os índices facilitam a detecção de diferentes tipos de cobertura pela acentuação ou atenuação de comportamentos espectrais do alvo (ZENG et al., 2020; REED et al., 1994).

As vantagens do uso de índices incluem: a minimização do efeito de reflectância do solo e outros efeitos de fundo, e a redução do espaço de atributos (REED et al., 1994). O uso de índices de vegetação, como o Índice de Vegetação por Diferença Normalizada (do inglês, *Normalized Diference Vegetation Index*, NDVI), em séries temporais, podem fornecer informações do comportamento temporal sobre um determinado alvo. Por exemplo, na Figura 2.2 é possível observar um padrão fenológico correspondente a um ciclo anual de uma vegetação. Os valores do NDVI crescem até atingir o pico no período que a cobertura apresenta alta atividade fotossintética e depois gradativamente diminuem indicando a senescência da vegetação (PETTORELLI et al., 2005).

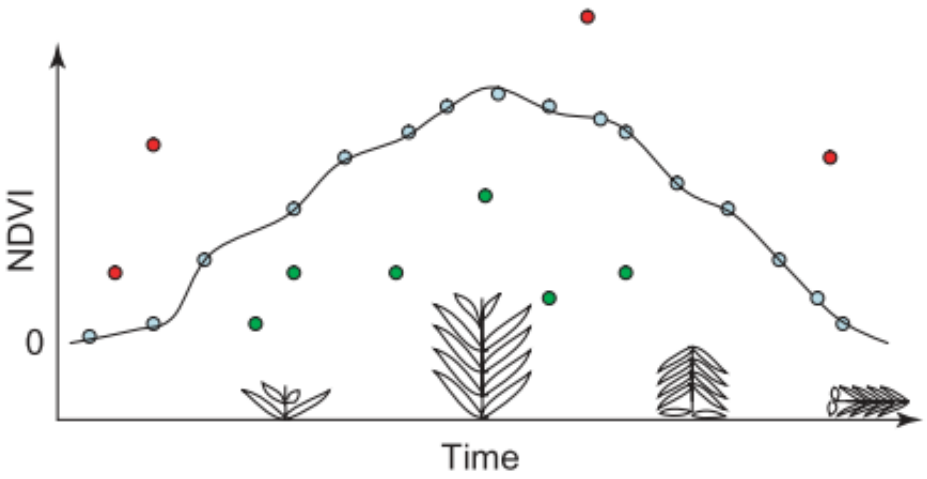
Nos últimos anos, a análise de séries temporais de imagens de satélites de observação da Terra tem sido amplamente empregada em algoritmos de aprendizado de máquina para a detecção de mudanças de uso e cobertura da terra (EPIPHANIO et al., 2010; GRIFFITHS et al., 2014; FRANKLIN et al., 2015; GÓMEZ et al., 2016; PICOLI et al., 2018; SIMOES et al., 2020).

2.3 Aprendizado de máquina

De acordo com Michalski et al. (2013), o aprendizado de máquina pode ser definido como um modelo estatístico que aprende com base em experiências passadas na execução de alguma tarefa. Na área de aprendizado de máquina, entende-se como aprendizado o processo de descoberta de padrões e associações em um conjunto de dados. Os dados podem ter diversas origens e formatos, como imagens, dados estáticos (sem variação no tempo) e séries temporais (HAN et al., 2011).

Técnicas de aprendizado de máquina têm sido amplamente usadas por quase duas décadas em dados não geográficos. No entanto, no sensoriamento remoto, o uso

Figura 2.2 - Exemplo de um perfil espectro-temporal de um ciclo de vegetação. Os círculos representados por diferentes cores, correspondem um cenário no qual os dados podem ser coletados: dados coletados durante um dia nublado (círculos verdes); em um dia sem ruídos (círculos azuis) e valores com *outliers* (círculos vermelhos).



Fonte: Pettorelli et al. (2005).

desses métodos são mais recentes e limitados devido ao grande volume de dados gerados diariamente (LARY et al., 2016). Com os avanços computacionais, tanto em *hardware* quanto em *software*, os métodos de ML vêm sendo aplicados cada vez mais nas diversas áreas de sensoriamento remoto.

Por exemplo, no trabalho de Sanchez et al. (2019) foram utilizadas técnicas de ML para identificar pontos de desmatamento e corte seletivo em regiões da Amazônia brasileira. Nguyen et al. (2020) usaram aprendizado de máquina para a caracterização do uso e cobertura da terra no estado da Dakota do Sul, nos EUA. Ma et al. (2020) identificaram construções civis que foram afetadas por desastres naturais, em vários locais do Planeta.

Os métodos de ML podem ser usados para diferentes objetivos. Seus principais usos são:

- **Métodos de Classificação:** Usados para encontrar um modelo ou função que descreve e distingue as classes a partir dos dados.

- **Métodos de Regressão:** Usados para a previsão de valores em dados numéricos ausentes ou indisponíveis, ao invés de classes.

- **Métodos de Agrupamento:** Usados para a criação de grupos em um conjunto de dados. Grupos são formados por objetos que possuem alguma similaridade entre si, e possuem dissimilaridade com objetos de outros grupos.

Os métodos de classificação e regressão pertencem à categoria de aprendizado supervisionado. A partir desses métodos criam-se modelos usando conhecimento prévio de classes associadas à cada amostra de um conjunto de dados. Por outro lado, métodos de agrupamento, que não usam classes de amostras em seus algoritmos, pertencem à categoria de aprendizado não-supervisionado. Nesta dissertação, são abordados os métodos de classificação e agrupamento. Mais informações sobre esses métodos podem ser vistas em Han et al. (2011).

2.4 Classificação de séries temporais

Em aprendizado de máquina, tarefas que envolvem a predição de dados categóricos a partir de um modelo estatístico são aplicadas em diversos contextos, como na caracterização de uma transição bancária ou na identificação de *spam* em *emails*. O processo de criação de um modelo classificador pode ser dividido em duas etapas: treinamento e classificação. Na etapa de treinamento, o modelo classificador aprende com uma entrada X e um vetor saída y, denominado conjunto de treinamento. Após o treinamento, a etapa seguinte consiste na classificação, em que parte dos dados que não foram utilizados no treinamento do modelo são usados como conjunto de teste para avaliar o desempenho do modelo treinado (HAN et al., 2011).

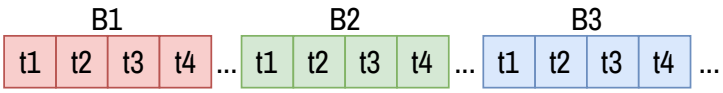
O processo de classificação de séries temporais pode ser realizado de modo similar. Entre a diversidade de métodos de classificação de séries temporais aplicados no contexto de sensoriamento remoto, mais especificamente, na classificação de uso e cobertura da terra, verificou-se que os classificadores *ensembles* e baseados em *kernel* obtiveram melhores resultados em uma ampla gama de aplicações (CHAVES et al., 2020). Definem-se como modelos *ensemble*, algoritmos que combinam múltiplos classificadores em sua arquitetura com o objetivo de produzir conjuntos de modelos mais robustos, que após o treinamento, são selecionados através de algum algoritmo de *ranking*. O modelo *ensemble* mais utilizado em aplicações de uso e cobertura é o *Random Forests* (RF), sendo aplicado em diferentes contextos, desde classificação

de áreas urbanas (ANJOS et al., 2017) até detecção de desmatamento (SAHA et al., 2020), o que ilustra a versatilidade do classificador em diferentes aplicações.

Já os modelos baseados em *kernels* utilizam funções de *kernel* para determinar o melhor encaixe em um espaço multidimensional através de um hiperplano, que divide esse espaço em regiões lineares. De acordo com Maxwell et al. (2018), o método *Support Vector Machine* (SVM) é o mais utilizado em aplicações de classificação de uso e cobertura terra. Por exemplo, Simoes et al. (2020) utilizaram o método SVM com a função de *kernel* radial para realizar a classificação do estado do Mato Grosso.

A organização dos dados de séries temporais para fins de treinamento e classificação segue o formato de organização conhecido como *wide dataset*. Nesse formato, as instâncias temporais são organizadas em colunas, e a cada nova instância de tempo uma nova coluna é adicionada. A Figura 2.3 apresenta um exemplo dessa organização com três bandas espectrais e quatro instâncias de tempo.

Figura 2.3 - Organização das séries temporais em formato *wide*.



Fonte: Próprio autor.

2.4.1 *Random Forests*

Os modelos baseados em *Random Forests* (RF) foram introduzidos por Breiman (2001), os quais usam de um conjunto de árvores de decisões para comporem suas classificações. Árvores de decisão usam da estrutura de árvores para representar a divisão do conjunto de dados de acordo com regras pré-estabelecidas em cada ramo da árvore, em que os nós não-folha representam uma regra associada a um determinado atributo e os nós-folha representam as classes do conjunto de dados.

Os resultados expressivos do RF devem-se à estratégia de reamostragem denominada *bagging* (CHAN; PAELINCKX, 2008), que desenvolve uma população de árvores de decisão pouco correlacionadas combinadas para formar um modelo de previsão conduzido por voto majoritário. Este procedimento de agregação melhora substancialmente a capacidade preditiva de árvores individuais.

Durante a fase de treinamento, a criação das árvores ocorre através de *bootstrap*, em que uma reamostragem aleatória com substituição a partir da qual, a cada passo, uma árvore de decisão T é produzida considerando apenas um subconjunto das observações originais. Com isso, cada árvore divide seus ramos por uma seleção aleatória de atributos. Em cada ramo, são selecionados m atributos aleatoriamente a partir do conjunto completo de p atributos dos dados de entrada. De acordo com um critério de pureza, o melhor atributo $\hat{m} \in \{1, \ldots, p\}$ é selecionado entre os m candidatos ($1 < m \leq p$), que então é usado para dividir aquele ramo em algum valor S do atributo escolhido que minimiza o índice de impureza.

O processo de divisão continua até que as observações em uma determinada folha possua a mesma classe, quando o índice de impureza atinge zero. Geralmente, pode-se usar qualquer critério de impureza, tais como o índice de GINI ou *cross-entropy* para decidir qual atributo deve ser usado para dividir o nó de cada árvore.

A seleção aleatória de atributos ajuda a diminuir as correlações entre as árvores de decisão produzidas pelo algoritmo do RF. Este procedimento gera um conjunto de b árvores de decisão formando o modelo RF final. A predição é realizada por um esquema de consenso simples: aplicando cada árvore gerada a um dado de entrada e computando a classe resultante como um voto. Assim, escolhe-se a classe mais votada entre b árvores de decisão como a classe de previsão do modelo RF.

Os principais parâmetros do algoritmo de RF são: o número de atributos amostrados divididos por nó ($mtry$); o número de árvores de decisão (b); o tamanho mínimo do nó (N_{min}) e a fração de amostra a ser sorteada a cada iteração (λ).

2.4.2 *Support Vector Machine*

As máquinas de vetores de suporte (do inglês, *support vector machines*, SVM) são uma generalização do classificador de separação simples por hiperplanos (HASTIE et al., 2009). Esse combina as noções de hiperplanos separadores ideais, a suavização das margens separadoras e a ampliação do espaço de atributos de entrada através de funções *kernel*. Em um modelo simples de separação por hiperplano, o classificador funciona apenas em conjuntos de amostras linearmente separáveis, em que o algoritmo procura por um separador linear no espaço de atributos de entrada que divida o espaço dimensional em duas partes, de forma a maximizar a margem desse separador aos dados de treinamento.

Caso as amostras sejam linearmente separáveis, o hiperplano separador realizará uma classificação ótima, na qual não possuirá confusão entre as amostras. Entretanto, a condição sob a qual um classificador por separação simples de hiperplanos tem solução é restritiva, pois é necessário um conjunto de treinamento linearmente separável. Isso dificilmente ocorre em uma situação real. Para contornar esta restrição, o SVM introduz um termo suavizador no problema de otimização do hiperplano separador. Além disso, permite que algumas amostras violem o seu rótulo de classe por se situarem na parte errada do subespaço particionado. O conjunto de amostras que violam a separação das classes ou que se situam próximas ao hiperplano, compõem os vetores de suporte. A solução dos coeficientes do hiperplano depende apenas dessas amostras. Essa modificação deixa o algoritmo SVM menos suscetível à *outliers* (CORTES; VAPNIK, 1995; HASTIE et al., 2009; JAMES et al., 2013).

Os hiperplanos são separadores lineares e separam o espaço de atributos em dois subespaços. O SVM realiza classificações não lineares ampliando o espaço de atributos das amostras de treinamento. Visto que a ampliação do espaço de atributos pode ser computacionalmente custosa, o algoritmo SVM usa funções *kernel* para contornar essa limitação. As funções *kernel* permitem ampliar o espaço de atributos com um custo computacional muito baixo (CORTES; VAPNIK, 1995). Por exemplo, a função *kernel* de base radial (RBF)

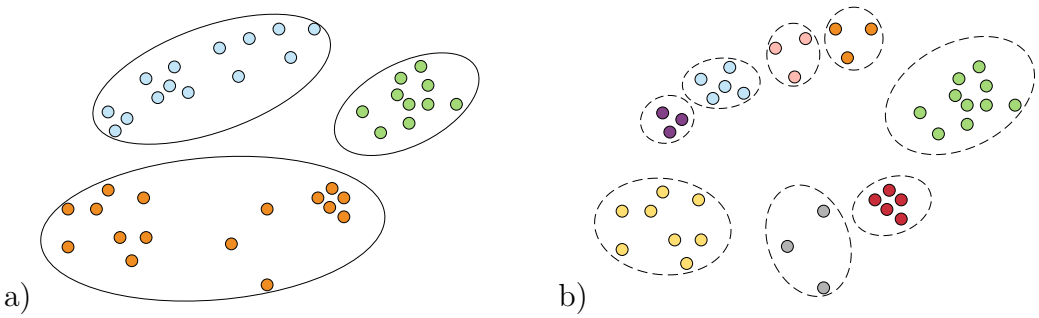
$$K(\mathrm{u}, \mathrm{v}) = \exp(-\gamma \|\mathrm{u} - \mathrm{v}\|^2), \tag{2.1}$$

em que u e v são vetores do espaço de atributos, mapeia o espaço de entrada para um espaço de dimensões infinitas que, de outro modo, seria obtido por uma expansão da série de potências. O uso de *kernels* é uma estratégia computacional eficiente para produzir separadores não lineares no espaço de atributos de entrada, geralmente conseguindo uma melhor separação entre classes de treinamento (HASTIE et al., 2009).

2.5 Agrupamento de séries temporais

As técnicas de agrupamento são ferramentas amplamente difundidas na área de mineração de dados, dado que identificam padrões em conjuntos de dados não rotulados através da organização desses em grupos. Nessa abordagem, o objetivo é formar grupos a partir de elementos ou objetos que possuem a máxima semelhança entre si e a mínima semelhança com elementos de outros grupos (AGHABOZORGI et al., 2015). A Figura 2.4 apresenta dois exemplos de agrupamento, (a) $C = 3$ e (b) $C = 8$, em que C representa a quantidade de grupos.

Figura 2.4 - Exemplo de dois agrupamentos baseados no mesmo conjunto de dados. Em (a) observam-se três grupos, $C = 3$, e em (b) oito grupos, $C = 8$.



Fonte: Adaptado de Esling e Agon (2012).

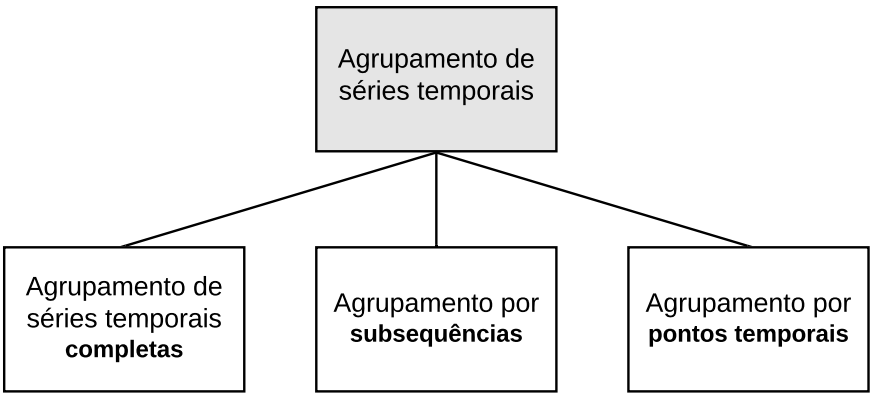
Segundo Han et al. (2011), além de aplicações convencionais como descoberta de padrões, detecção de *outliers* e análise de distribuições intrínsecas nos dados; as técnicas de agrupamento podem ser usadas em etapas de pré-processamento para outros algoritmos, por exemplo, na caracterização, seleção de subconjuntos de atributos e classificação. Ainda, de acordo com Han et al. (2011), pesquisas em técnicas de agrupamentos concentram-se na escalabilidade dos métodos de agrupamento, na eficácia dos métodos para agrupar formas complexas (por exemplo, não-convexas) e nos tipos de dados (por exemplo, texto, gráficos e imagens), nas técnicas de agrupamento de alta dimensão (por exemplo, objetos de agrupamento com centenas de atributos) e em métodos para agrupar dados numéricos e nominais em grandes bases de dados.

Pesquisas em agrupamento de séries temporais envolvem uma série de desafios, como os descritos pelos autores Aghabozorgi et al. (2015): o primeiro desafio refere-se à forma de armazenamento das séries temporais, que devido ao seu tamanho, são armazenadas em discos rígidos, tal operação diminui exponencialmente o acesso aos dados por conta das operações de I/O; outro desafio diz respeito à alta dimensionalidade contida em dados de séries temporais, o que dificulta a manipulação dos dados no uso de técnicas de agrupamento e aumenta o tempo de processamento. Por fim, os autores abordam sobre as medidas de similaridades utilizadas nos agrupamentos, pois algumas desconsideram as características das séries temporais.

De acordo com Aghabozorgi et al. (2015), o agrupamento de séries temporais pode ser dividido em três categorias: agrupamento por séries temporais `completas`, agrupamento por `subsequências` e agrupamento por `pontos temporais`. Tais categorias

formam o que é conhecido como a taxonomia do agrupamento de séries temporais (AGHABOZORGI et al., 2015), ilustrada na Figura 2.5.

Figura 2.5 - A taxonomia do agrupamento de séries temporais.



Fonte: Adaptado de Aghabozorgi et al. (2015).

O agrupamento por séries temporais completas é realizado da mesma forma que em dados não temporais. Dado um conjunto de séries temporais individuais, o objetivo é criar grupos de séries temporais semelhantes. Por outro lado, no agrupamento por subsequências criam-se grupos de segmentos de uma série temporal, os segmentos são extraídos por uma janela deslizante em uma única série temporal. No agrupamento por pontos temporais, semelhante à abordagem de agrupamento por subsequência, os pontos são agrupados pela combinação da proximidade temporal e pela similaridade dos valores correspondentes (AGHABOZORGI et al., 2015).

Como mencionado anteriormente, técnicas baseadas nas categorias de agrupamento por subsequência e de pontos temporais não produzem grupos de séries temporais, e sim, grupos de determinados pontos extraídos da série temporal. Keogh e Lin (2005) provam que o agrupamento por subsequência produz grupos aleatórios, pois os algoritmos desta categoria são totalmente independentes dos dados de entrada. Considerando as informações mencionadas e com o objetivo de criar grupos a partir da série temporal como um todo, neste trabalho, optou-se por utilizar técnicas de agrupamento baseadas em séries completas.

Aghabozorgi et al. (2015) definem quatro componentes essenciais à categoria de agrupamento por séries temporais completas, sendo: **redução de dimensionalidade**, **medidas de distância**, **técnicas de agrupamento** e **agrupamento de**

protótipos. A redução de dimensionalidade tem por objetivo transformar uma série temporal em um vetor de atributos com baixa dimensão; as medidas de distâncias são métodos de comparação entre duas séries temporais; as técnicas de agrupamento são modelos baseados em diferentes características para agrupar séries temporais e, por fim, no agrupamento de protótipo busca-se encontrar protótipos significativos para melhorar a qualidade dos grupos. Nesta dissertação, o componente de agrupamento de protótipo não é abordado, caso o leitor tenha interesse, os autores recomendam o artigo de Ratanamahatana e Keogh (2005).

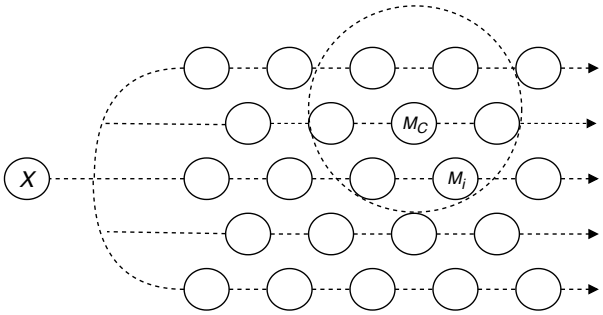
Neste trabalho, os atributos extraídos com as métricas temporais são avaliados com base em técnicas de agrupamento e classificação. Diante da extensa lista de métodos de agrupamento propostos na literatura, é difícil uma categorização definitiva desses, pois alguns métodos compartilham características semelhantes pertencendo a diferentes categorias. Em geral, com base nos trabalhos de Jain et al. (2000) e Han et al. (2011), métodos com as seguintes características são considerados fundamentais: métodos baseados em partição; hierarquia; densidade e na auto-organização através de redes neurais.

Em especial, nos métodos baseados em auto-organização, a técnica de agrupamento de Mapas Auto-Organizáveis (do inglês, *Self-Organizing Maps*, SOM) vem ganhando destaque nos últimos anos na área de uso e cobertura da terra. Por exemplo, em Santos et al. (2021a), o SOM é utilizado na identificação de padrões espaço-temporais em amostras de uso e cobertura da terra. Já no trabalho de Picoli et al. (2020), usou-se o SOM para a limpeza de amostras que possuem confusões espectro-temporais ou que foram rotuladas de forma errônea. Desta forma, diante da literatura consultada, optou-se por utilizar o método SOM. A seguir descreve-se o método SOM e as medidas de avaliação de agrupamento que são utilizadas nesta dissertação.

2.5.1 Mapas Auto-Organizáveis

O SOM (KOHONEN, 1982) é uma rede neural não supervisionada, que aplica o procedimento de aprendizado competitivo para mapear os vetores de entrada multidimensionais em uma grade bidimensional retangular ou hexagonal de baixa dimensão. Um mapa auto-organizado é, portanto, caracterizado pela formação de um mapa topográfico dos padrões de entrada, em que as localizações espaciais dos neurônios na grade são indicativas das características estatísticas intrínsecas contidas nos padrões de entrada, originando o nome "mapa auto-organizado" (HAYKIN, 2010).

Figura 2.6 - Arquitetura da rede SOM com a topologia hexagonal.



Fonte: Adaptado de Kohonen (2013).

Em resumo, na rede SOM os nós de saída competem entre si pelos vetores de entrada, e ao final de cada iteração é determinado o nó vencedor (BMU, do inglês *best match unit*), aquele que possui a menor distância, comumente euclidiana, com o vetor de entrada. Após a escolha do BMU, todos os nós vizinhos em um determinado raio atualizam seus valores, para se aproximarem do padrão escolhido no BMU (KOHONEN, 2013). A Figura 2.6 apresenta um exemplo de arquitetura do SOM com topologia hexagonal, na qual é possível observar o vetor de entrada X, neurônio vencedor M_c, o raio de vizinhança (círculo) e o neurônio atingido pela taxa de atualização M_i.

2.5.2 Medidas de avaliação de agrupamento

Após a aplicação de técnicas de agrupamento é importante avaliar a qualidade dos grupos criados. Uma série de medidas podem ser usadas. Por exemplo, alguns métodos medem o quão bem os grupos se ajustam ao conjunto de dados, enquanto outros medem o quão bem os grupos correspondem às classes. As classes podem ser consideradas como supervisão na forma de "rótulos dos grupos" (HAN et al., 2011).

Os métodos que medem a qualidade dos agrupamentos podem ser separados em dois grupos, conforme a disponibilidade ou não de classes. Caso as classes estejam disponíveis, faz-se uso de índices externos, os quais são usados para medir a semelhança de grupos formados com as classes fornecidas externamente, sendo considerado o método de avaliação mais popular na literatura (HALKIDI et al., 2001; AGHABOZORGI et al., 2015). Caso as classes dos grupos não sejam fornecidas, faz-se uso de índices internos, que avaliam a qualidade de um agrupamento considerando a semelhança

dos elementos presentes em cada grupo (HAN et al., 2011). Neste trabalho, são utilizados os índices de avaliação externos, uma vez que as classes das amostras de uso e cobertura da terra são disponibilizadas juntos aos dados.

A principal tarefa dos métodos externos é atribuir uma pontuação $Q(C, C_g)$, dado um agrupamento C, com as classes C_g. Assim, de acordo com Han et al. (2011), uma medida Q é considerada eficaz caso atenda as quatro categorias:

a) **Homogeneidade de Grupos:** Exige-se que, quanto mais puros os grupos em um agrupamento, melhor seja o agrupamento. Em detalhes, sejam dois agrupamentos C_1 e C_2. Considere um conjunto de dados D, que possui classes de $L_1, ..., L_n$. Considere também, que o agrupamento C_1 possui apenas um grupo g_1, tal grupo contém dois objetos de diferentes rótulos L_i, L_j $(1 \leq i < j \leq n)$. Considere que o agrupamento C_2 dispõe de dois grupos g_1 e g_2, nos quais cada um possui objetos da mesma classe, L_i e L_j, respectivamente. Assim, uma medida de qualidade Q, com base no critério de homogeneidade de grupos, deve atribuir uma pontuação mais alta ao agrupamento C_2, pelo fato do mesmo possuir grupos de objetos com classes únicas.

b) **Integridade de Grupos:** Exige-se que um agrupamento de objetos com a mesma classe (de acordo com as verdades de campo) devem pertencer ao mesmo grupo. Em detalhes, seja um agrupamento C_1, com dois grupos g_1 e g_2, nos quais os objetos contidos em g_1 e g_2 tenham a mesma classe de acordo com a verdade de campo. Considere outro agrupamento C_2, idêntico ao C_1, exceto que g_1 e g_2 são mesclados em apenas um grupo no C_2. Assim, uma medida de qualidade Q, com base no critério de integridade de grupos, deve atribuir uma pontuação mais alta ao agrupamento $C2$, pelo fato do mesmo possuir um único grupo com objetos da mesma classe.

c) ***Rag Bag*:** Em diversos cenários reais, existe uma categoria que contém objetos que não podem ser misturados com outros, tal categoria, por vezes, contém nomes como "diversos" ou "outros", assim denominados como *rag bag*. No critério de *rag bag* exige-se que, objetos heterogêneos colocados em grupos puros devem ser mais penalizados do que colocá-lo em um grupo de *rag bag*. Em detalhes, seja um agrupamento C_1 com apenas um grupo g_1. Todos os objetos contidos em g_1, com exceção de um, denotado por o, pertencem à mesma classe de acordo com a verdade de campo. Considere também um agrupamento C_2, idêntico ao C_1, porém o objeto o é associado

a outro grupo g_2. Tal grupo g_2 do agrupamento C_2 possui objetos de diferentes classes, assim é considerado um grupo *rag bag*. Desta forma, uma medida de qualidade Q, com base no critério *rag bag*, deve atribuir uma pontuação mais alta ao agrupamento C_2, pelo fato do mesmo possuir um grupo com objetos de diferentes classes.

d) **Preservação de pequenos grupos:** O critério de preservação de pequenos grupos afirma que dividir uma classe com poucos objetos em pedaços é mais prejudicial do que dividir uma classe com maior número de objetos. Por exemplo, seja um conjunto de dados D com dez objetos, desses objetos, oito pertencem à classe $X = \{x_1, x_2, ..., x_7, x_8\}$, e dois pertencem à classe $Y = \{y_1, y_2\}$. Suponha que o agrupamento C_1 tenha três grupos, $g_1 = \{x_1, ..., x_8\}$, $g_2 = \{y_1\}$ e $g_3 = \{y_2\}$. Suponha também, que o agrupamento C_2 tenha três grupos, $g_1 = \{x_1, ..., x_7\}$, $g_2 = \{x_8\}$ e $g_3 = \{y_1, y_2\}$. Uma medida de qualidade Q, com base no critério de preservação de pequenos grupos, deve atribuir uma pontuação mais alta ao agrupamento C_2.

Seguindo as diretrizes mencionadas acima, as medidas de avaliação externa mais conhecidas são: *Rand Index* (RI) (RAND, 1971), *Adjusted Rand Index* (ARI) (HUBERT; ARABIE, 1985) e o Coeficiente Jaccard (J) (JAIN; DUBES, 1988; VENDRAMIN et al., 2010). Neste trabalho, além dos índices mencionados, foi utilizada a entropia como medida de pureza entre os grupos. As medidas de avaliação RI, ARI e J são baseadas em pares de objetos entre cada grupo e seus valores são medidos de acordo com as seguintes regras:

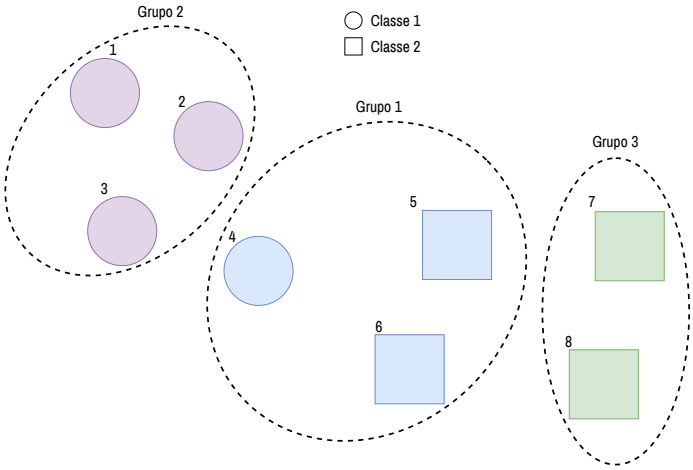
a) Número de pares de objetos que pertencem ao mesmo grupo e à mesma classe;

b) Número de pares de objetos que pertencem às mesmas classes, mas são de grupos diferentes;

c) Número de pares de objetos que são de diferentes classes e estão no mesmo grupo;

d) Número de pares de objetos que pertencem às diferentes classes e diferentes grupos.

Figura 2.7 - Exemplo de um agrupamento com três grupos e duas classes.



Fonte: Adaptado de Vendramin et al. (2010).

Para exemplificar o uso destas medidas, considere a Figura 2.7, em que os objetos geométricos representam as classes, as cores representam os grupos (cada grupo está envolvido em uma elipse) e os números acima de cada observação representam os identificadores únicos de cada objeto. A Classe 1 (círculo) é composta pelos objetos 1, 2, 3 e 4, enquanto a classe 2 (quadrado) é composta pelos objetos 5, 6, 7 e 8. Em relação aos grupos, o grupo 1 (em azul) possui os objetos 4, 5 e 6; o grupo 2 (em roxo) possui os objetos 1, 2 e 3, e o grupo 3 (em azul) possui os objetos 7 e 8. Logo, conforme a lógica especificada acima, podemos definir da seguinte forma:

a) Os pares de objetos que pertencem ao mesmo grupo e à mesma classe são: (1,2), (1,3), (2,3), (5,6), e (7,8). $a = 5$.

b) Os pares de objetos que pertencem à mesma classe, mas são de grupos diferentes são: (3,4), (2,4), (1,4), (5,7), (5,8), (6,8) e (6,7). $b = 7$.

c) Os pares de objetos que pertencem a diferentes classes e estão no mesmo grupo são: (4,5) e (4,6). $c = 2$.

d) Os pares de objetos que pertencem às diferentes classes e diferentes grupos são: (1,5), (1,6), (1,7), (1,8), (2,5), (2,6), (2,7), (2,8), (3,5), (3,6), (3,7), (3,8), (4,7), (4,8). $d = 14$.

A medida de avaliação RI é definida da seguinte forma:

$$RI = \frac{a+d}{a+b+c+d} \tag{2.2}$$

Nota-se que os termos a e d medem a consistência dos grupos, ou seja, o número de pares de objetos da mesma classe que estão no mesmo grupo ou em grupos únicos. Por outro lado, os termos b e c medem a inconsistência dos grupos, ou seja, o número de pares de objetos de classes diferentes que estão no mesmo grupo. Assim, o RI varia entre 0 a 1, em caso de 0 corresponde a grupos inconsistentes e a 1 a grupos totalmente homogêneos. Resolvendo a Equação 2.2, temos $RI \approx 0.68$ (VENDRAMIN et al., 2010).

No entanto, após críticas ao RI (FACELI et al., 2005), pelo fato do índice não produzir valor igual a 0 quando um conjunto aleatório é avaliado ou 1 quando todos os objetos do grupo são da mesma classe, criou-se o ARI para ajustar esse detalhe. O índice ARI é dado por:

$$ARI = \frac{a - \frac{(a+c)(a+b)}{M}}{\frac{(a+c)+(a+b)}{2} - \frac{(a+c)(a+b)}{M}} \tag{2.3}$$

Em que $M = a + b + c$. Por outro lado, o Coeficiente de Jaccard remove o termo d da formulação do índice RI 2.2, pois o índice RI fornece o mesmo peso para os termos a e d (VENDRAMIN et al., 2010). Logo, temos:

$$J = \frac{a+d}{a+b+c} \tag{2.4}$$

Por fim, a entropia avalia a pureza entre cada grupo, como neste trabalho é usado o método de agrupamento SOM, considera-se cada neurônio como um grupo e podemos defini-la da seguinte forma:

$$e_j = \sum_{i=1}^{L} = p_{ij} log_2 p_{ij} \tag{2.5}$$

em que $p_{ij} = \frac{m_{ij}}{m_j}$, sendo m_j a quantidade de objetos do grupo j e m_{ij} a quantidade de objetos de classe i correspondentes ao neurônio j.

De acordo com Faceli et al. (2005), o índice RI possui a tendência de apresentar valores mais elevados para agrupamentos com grandes números de grupos. Já o índice J apresenta valores mais elevados para agrupamentos com menor número de grupos. O índice ARI não apresenta essas características. Os índices podem ser interpretados da seguinte forma: valores altos nos índices RI, ARI e J indicam forte concordância, dado que seus intervalos vão de 0 a 1. Já para entropia, valores menores indicam maior concordância entre os grupos (FACELI et al., 2005).

2.5.3 Análise de agrupamento para avaliação de amostras

O uso de técnicas de agrupamento para análise de amostras de uso e cobertura da terra é um tema recorrente de pesquisa em sensoriamento remoto. Por exemplo, os autores Scrivani et al. (2014) utilizaram o método de agrupamento K-*means* para a identificação de áreas inundadas, florestas, culturas e pastagens usando séries temporais do sensor MODIS. Já os autores Souza et al. (2019) usaram o SOM para analisar a separabilidade das amostras utilizando séries temporais do produto harmonizado de imagens Landsat e Sentinel-2 (do inglês, *Harmonized Landsat and Sentinel-2*, HLS).

Na avaliação de amostras, o interesse não é apenas avaliar a separabilidade das classes, mas identificar potenciais erros de rotulagem de classe no conjunto de amostras. Esse tipo de erro ocorre na atribuição de classes às amostras, em que a classe rotulada difere daquela da verdade de campo (PELLETIER et al., 2017).

O processo de avaliação da qualidade das amostras é uma etapa crucial para a obtenção de bons resultados de classificação. Santos et al. (2021b) propõem um método para redução de ruído de classes em amostras de séries temporais. O método é baseado na técnica de agrupamento SOM. Após a realização do agrupamento, os neurônios são rotulados de acordo com a classe majoritária das amostras contidas nos neurônios. Caso um neurônio não possua amostras associadas, ele será considerado um neurônio sem classe, que recebe o rótulo de "NoClass". Se houver empate, será considerada a vizinhança definida pela topologia da rede SOM.

Para avaliar a confiabilidade e a qualidade das amostras são computadas duas medidas de probabilidades. A primeira, corresponde a frequência de amostras por classe em cada neurônio. Essa é considerada a probabilidade *a priori* de cada amostra pertencer a uma classe. A segunda é obtida usando a estrutura de vizinhança do SOM a partir da qual é realizada uma atualização da probabilidade *a priori* para se obter uma probabilidade *a posteriori* através da inferência Bayesiana. A última

etapa consiste em definir limites de probabilidade para identificar aquelas amostras com rótulos errados, ou que precisam ser re-avaliadas e sinalizadas para investigação posterior feita por um especialista. Em detalhes, seja os limiares das probabilidades *a priori* (τ_p) e *a posteriori* (τ_c), são consideradas as seguintes regras:

a) Caso a probabilidade *a priori* seja $< \tau_c$, então as amostras são descartadas;

b) Caso as probabilidades *a priori* seja $\geq \tau_c$ e *a posteriori* $\geq \tau_p$, então as amostras são mantidas;

c) Caso as probabilidades *a priori* seja $\geq \tau_c$ e *a posteriori* seja $< \tau_p$, então as amostras são marcadas para uma análise mais detalhada.

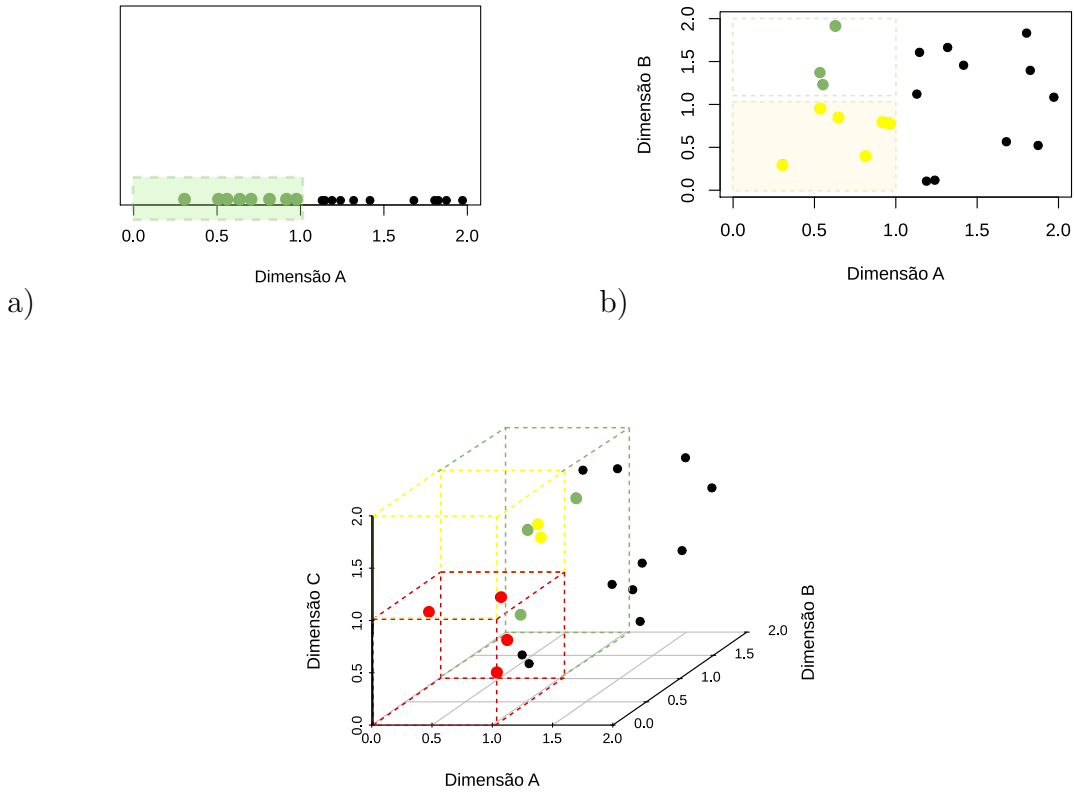
Em Santos et al. (2021b) mostram como o procedimento de remoção de amostras ruidosas aumentam a acurácia global de uma validação cruzada de 94% para 98% usando um conjunto de mais de 50 mil amostras de treinamento com séries temporais extraídas do sensor MODIS.

2.6 Redução de dimensionalidade em séries temporais

Na área de aprendizado de máquina, a dimensionalidade refere-se ao número de atributos ou características de um conjunto de dados. A dificuldade de se trabalhar com dados de alta dimensão no contexto de agrupamento e classificação é exaltada por diversos autores. Por exemplo, Aghabozorgi et al. (2015) destacam os problemas relacionados à eficiência computacional, dado que o cálculo da distância entre duas séries temporais com alta dimensão é computacionalmente caro. Já os autores Parsons et al. (2004) mencionam o efeito da "maldição da dimensionalidade" (KORN et al., 2001), em que à medida que o número de dimensões de um conjunto de dados aumenta, as medidas de distância tornam-se cada vez mais sem sentido, pois em um espaço de dimensão muito elevado os pontos ficam equidistantes entre si.

A Figura 2.8, baseada no artigo de Parsons et al. (2004), ilustra o conceito de "maldição da dimensionalidade". O conjunto de dados apresentado nesta Figura possui 20 observações geradas em uma distribuição uniforme. Observe que, com a adição das dimensões B e C, a quantidade de pontos que estão a uma unidade de distância diminui. Na Figura 2.8(b), é possível observar seis pontos no retângulo amarelo, visto que os outros três se espalharam com a adição da dimensão B. Por fim, com a adição da dimensão C sobram apenas 4 pontos, apresentados no retângulo vermelho. Desta forma, com a adição de mais dimensões, os pontos continuarão a distribuir-se até estarem todos igualmente distantes (PARSONS et al., 2004).

Figura 2.8 - Exemplo do efeito da "maldição da dimensionalidade". De acordo com a adição de novas dimensões, os pontos a uma unidade de distância diminuem. Em a) há 9 pontos a uma unidade de distância; em b) 6 e em c) 4.

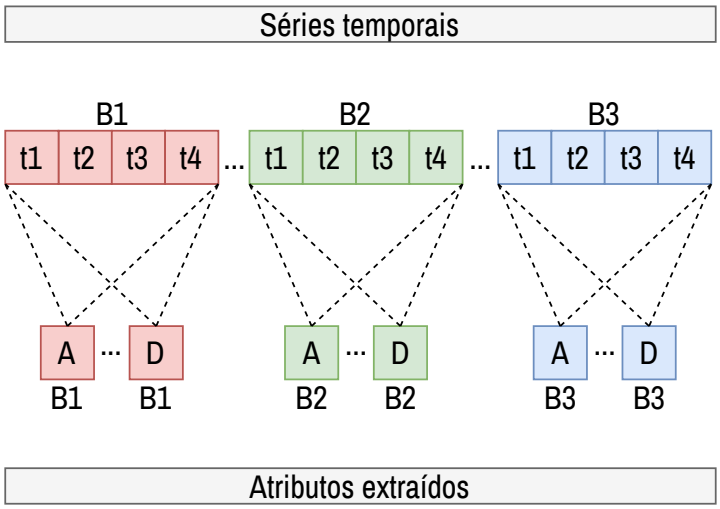


Fonte: Adaptado de Parsons et al. (2004).

Séries temporais completas e "cruas", por natureza, dispõem de alta dimensionalidade, além de inerentes ruídos, ocasionados pela forma de obtenção desses dados, os quais podem interferir no processo de agrupamento e classificação. Desta forma, a redução da dimensionalidade deste tipo de dado pode contribuir com maiores acurácias, tempo de processamento e entendimento dos dados (MAXWELL et al., 2018). A redução de dimensionalidade pode ser dividida em duas categorias: extração e seleção de atributos. As técnicas baseadas em extração de atributos criam novos atributos a partir dos originais. Nesta dissertação, são usadas técnicas de extração de atributos baseadas em sumarização. Nessa categoria, os elementos espectro-temporais de uma única banda são sumarizados e representados através de um valor único. A Figura 2.9 ilustra um exemplo da técnica de sumarização de atributos aplicada em séries

temporais, os atributos são extraídos com base em todas as instâncias temporais de cada banda espectral, por exemplo, os atributos A e D foram extraídos das séries temporais da banda *B1*, *B2* e *B3*.

Figura 2.9 - Exemplo de extração de atributos a partir de séries temporais, em que B1, B2 e B3 representam as bandas no tempo t_n extraídas pelas métricas A e D.



Fonte: Próprio autor.

Na seleção de atributos, no entanto, busca-se determinar um subconjunto ótimo a partir dos atributos originais do conjunto de dados. Diversos classificadores realizam o ranqueamento dos atributos internamente, no entanto, eles não realizam a seleção deles.

2.6.1 Extração de atributos

As técnicas de extração de atributos avaliadas neste trabalho fazem a sumarização das séries temporais (Figura 2.9). Nesta dissertação, para seguir de acordo com a literatura (KÖRTING et al., 2013; SOARES et al., 2020), as técnicas de extração de atributos são chamadas de métricas. As métricas usadas neste trabalho, segundo Soares et al. (2020), são o atual estado-da-arte de métricas para séries temporais de uso e cobertura da terra.

As métricas usadas neste trabalho são divididas em dois grupos: métricas básicas e polares. As métricas básicas fazem a extração de atributos baseados no espaço eucli-

diano. Seus métodos são compostos por estatísticas básicas, como média, mediana, desvio padrão; estatísticas baseadas em histograma, por exemplo, valores mínimos, máximos, primeiro, segundo e terceiro quartil e diferença interquartil; métodos baseados em análises séries temporais, como primeira inclinação, amplitude e média da energia espectral.

As métricas polares, abordagem proposta por Körting et al. (2013), baseiam-se na representação polar para descrever eventos cíclicos, cujos eventos são comuns em aplicações agrícolas. Entende-se ciclo como eventos que possuem recorrência. Para permitir a visualização de cada ciclo, os autores adaptaram a técnica de visualização proposta por Edsall et al. (1997), projetando os valores em ângulos no intervalo $[0, 2\pi]$.

Em detalhes, seja uma função $f(x, y, T)$, em que (x, y) corresponde a posição espacial de um ponto em um intervalo de tempo T em $t_1, ..., t_N$, N corresponde ao número de observações de um ciclo. Para criar uma representação polar de um conjunto de valores $v_i \in V$, define-se uma função $g(V) \implies \{A, O\}$, em que A representa o eixo das abscissas e O o eixo das ordenadas em um plano cartesiano. Assim, para cada ponto em A e O, têm-se as Equações 2.6 e 2.7, respectivamente (KÖRTING et al., 2013).

$$a_i = v_i \cos \frac{2\pi i}{N} \in A, i = 1, ..., N \tag{2.6}$$

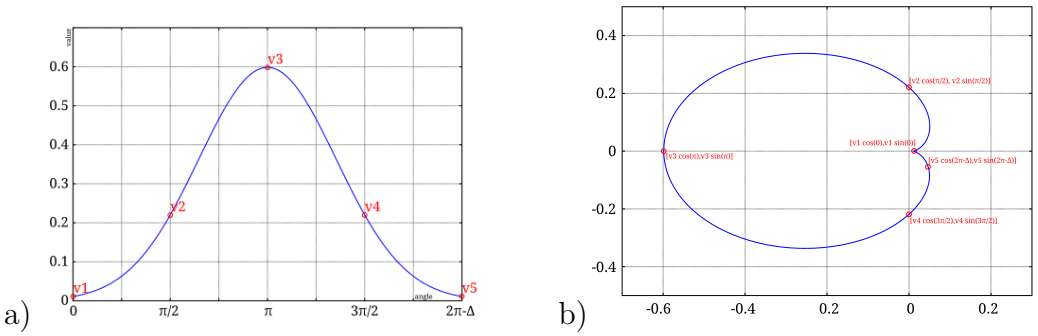
$$o_i = v_i \sin \frac{2\pi i}{N} \in O, i = 1, ..., N \tag{2.7}$$

O uso da transformação polar é necessário para criar uma forma fechada, para qual se tem $a_{N+1} = a_1$ e $o_{N+1} = o_1$, como ilustrado na Figura 2.10(b). Assim, um ciclo com valores constantes resulta em um círculo, de modo que ciclos diferentes desenham formas diferentes de acordo com suas propriedades. Com a forma fechada, uma série de métricas podem ser extraídas, tais como área, perímetro, direção principal, elipse delimitadora, excentricidade e raio (KÖRTING et al., 2013).

2.6.2 Seleção de atributos

Atualmente, com o acesso gratuito a grandes bases de imagens de observação da Terra e o uso de cubo de dados para representá-las, o número de atributos em aplicações de sensoriamento remoto aumentou. Assim, tornando necessário o desen-

Figura 2.10 - Em (a) valores de um ciclo associados a determinados ângulos, em (b) figura com formato fechado criado através da transformação polar.



Fonte: Körting et al. (2013).

volvimento de novas abordagens para a redução e identificação de características mais relevantes em tarefas de classificação, regressão e agrupamento em aplicações em observação da Terra.

Diante da extensa lista de métodos de seleção de atributos contidos na literatura, baseados em ranqueamento ou correlação, os métodos de seleção de atributos *wrapper*, aqueles em que os atributos são selecionados em tempo de treinamento, detém os melhores resultados em diversas aplicações. O método RF se tornou o mais popular seletor de atributos em aplicações de sensoriamento remoto. No entanto, a dificuldade no uso do RF está relacionada à quantidade de atributos que serão selecionados após a classificação. Caso o usuário não seja um especialista, a escolha do número de atributos, por vezes, pode ser complexa. Por exemplo, dados que possuem alta dimensionalidade, geralmente, dispõem de atributos com alta correlação e isso pode causar efeitos negativos no classificador. Além disso, é necessário fornecer um limiar para a seleção de atributos, tal complexidade pode retornar atributos com altas correlações, mas com pouca validade no domínio de aplicação (IZQUIERDO-VERDIGUIER; ZURITA-MILLA, 2020).

Ainda que o RF forneça a importância de cada atributo, informação que auxilia no entendimento das contribuições, com o crescimento exponencial da quantidade de atributos ainda é um desafio investigar as pontuações de importância de um grande número de atributos retornados (DENG, 2013). Desta forma, os autores Deng e Runger (2013) desenvolveram o método de seleção de atributos *Guided Regularized Random Forest* (GRRF).

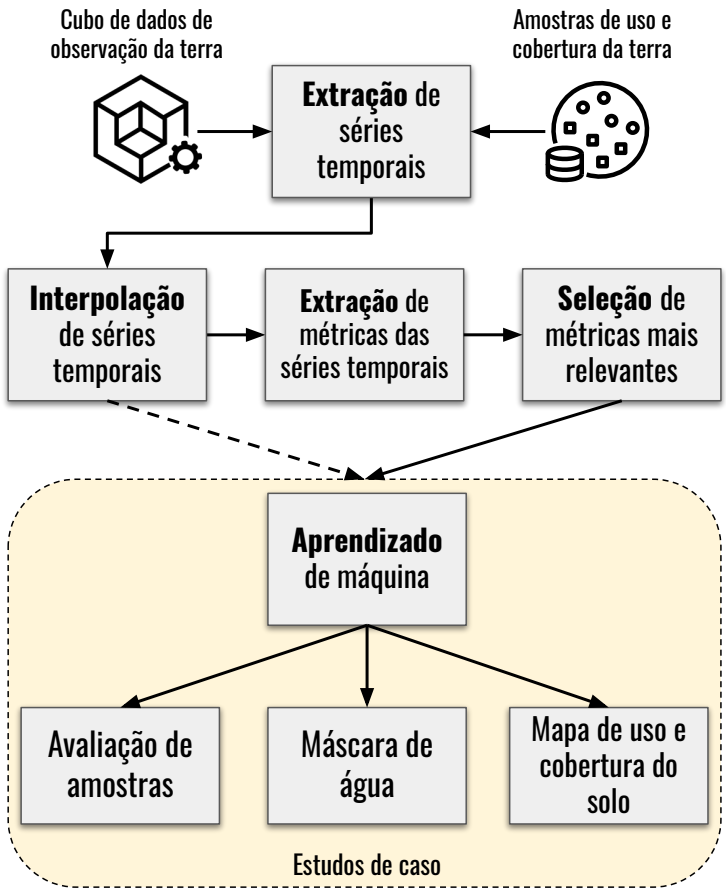
O GRRF usa as medidas de importância dos atributos retornados por um RF. Com isso, essa medida é escalada com os objetivos de encontrar a medida de importância para todas as árvores do RF, ao final forçando os atributos mais relevantes a serem testados em todas as RF avaliadas. Para realizar a seleção, o GRRF atribui um fator de peso (γ) pré-definido pelo usuário para filtrar os atributos mais relevantes na divisão de cada nó de um RF (WUNDERVALD et al., 2020; IZQUIERDO-VERDIGUIER; ZURITA-MILLA, 2020).

A avaliação realizada por Izquierdo-Verdiguier e Zurita-Milla (2020) mostrou que em diferentes cenários o uso do GRRF em aplicações de sensoriamento remoto se sobressaiu se comparado com os atributos selecionados por um RF normal. Desta forma, neste trabalho, é utilizado o GRRF para fazer a seleção de métricas extraídas das séries temporais de imagens de satélites, visto que essa extração gera um grande número de atributos.

3 METODOLOGIA

Para comparar as abordagens de métricas e séries temporais completas, foram realizados três estudos de caso cuja metodologia é descrita neste Capítulo. A Figura 3.1 ilustra de forma geral, os passos realizados em cada um dos estudos de caso que estão indicados na Figura.

Figura 3.1 - Metodologia adotada neste trabalho.



Fonte: Próprio Autor.

Duas fontes de dados são usadas nos estudos de caso: cubos de dados de observação da Terra e amostras de uso e cobertura da terra. O primeiro passo é a extração das séries temporais das amostras a partir dos cubos de dados (etapa extração de séries temporais). As observações identificadas como nuvem ou sombra de nuvem são substituídas por valores interpolados a partir das observações válidas da vizinhança

temporal (etapa de interpolação de séries temporais). As métricas usadas nos estudos de caso são computadas para cada banda ou índice espectral das séries temporais (etapa de extração de métricas das séries temporais). A etapa seguinte consiste na identificação das métricas mais importantes na discriminação das classes (etapa de seleção de métricas mais relevantes). Em seguida, os modelos de aprendizado de máquina são treinados de acordo com os subconjuntos de métricas previamente selecionadas (etapa de aprendizado de máquina). A linha tracejada que liga a etapa de interpolação de séries temporais à etapa de aprendizado de máquina refere-se aos experimentos dos estudos de caso que usam séries temporais completas a título de comparação com a abordagem de métricas. As Seções seguintes detalham cada uma dessas etapas. No final do Capítulo, é apresentado um resumo com as principais informações sobre cada estudo de caso (Tabela 3.4).

3.1 Cubos de dados de observação da Terra

Os cubos de dados de observação da Terra usados como fonte de séries temporais foram produzidos e gerados no âmbito do projeto BDC. Foram utilizados os cubos de dados dos satélites Sentinel-2, com 10 metros de resolução espacial; Landsat-8, com 30 metros de resolução espacial; e CBERS-4, com 64 metros de resolução espacial. Todos com a mesma periodicidade de 16 dias de resolução temporal. As bandas espectrais usadas para extração de séries temporais estão listadas na Tabela 3.1.

Além das bandas, também foram utilizados os seguintes índices espectrais: índice de vegetação melhorado (do inglês, *enhanced vegetation index*, EVI), índice de vegetação de diferença normalizada (do inglês, *normalized difference vegetation index*, NDVI), índice de monitoramento ambiental global (do inglês, *global environmental monitoring index*, GEMI), índice de vegetação de diferença normalizada do verde (do inglês, *green normalized difference vegetation index*, GNDVI), índice de água de diferença normalizada (do inglês, *normalized difference water index*, NDWI), índice de água da diferença normalizada modificada (do inglês, *modified normalized difference water index*, MNDWI) e índice de produtividade, versatilidade e resiliência (do inglês, *productivity, versatility, and resiliency*, PVR), todos descritos na Tabela 3.2. Os índices EVI, NDVI, GEMI e GNDVI são tipicamente sensíveis a coberturas vegetais (ROUSE et al., 1973; PINTY; VERSTRAETE, 1992; GITELSON et al., 1996; HUETE et al., 1999). O EVI usa a banda do azul que é sensível à atmosfera para corrigir a banda do vermelho quanto à influência de aerossol. É conhecido que o NDVI satura rapidamente para vegetações densas. Os índices MNDWI e NDWI são sensíveis a coberturas de corpos d'água (GAO, 1996; XU, 2006). Já o índice PVR é

Tabela 3.1 - Cubos de dados do projeto BDC usados nos estudos de caso. As bandas listadas referem-se àquelas usadas nos estudos de caso.

Estudo de caso	**Cubo de dados BDC**	**Satélite Sensor**	**Res. Temporal (dias)**	**Res. Espacial (m)**	**Bandas**
Avaliação de amostras	S2_10_16D_STK-1	Sentinel-2 (A&B) MSI	16	10	coastal blue green red nir swir16 swir22 quality
Máscara de água	LC8_30_16D_STK-1	Landsat-8 OLI	16	30	coastal blue green red nir swir16 swir22 quality
Mapa de uso e cobertura da terra	CB4_64_16D_STK-1	CBERS-4 AWFI	16	64	blue green red nir quality

As bandas blue, green e red representam as faixas do azul, verde e vermelho, respectivamente. As bandas nir, swir16 e swir22 representam as faixas do espectro do infravermelho próximo, infravermelho de ondas curtas de $1.6\mu m$ e de $2.2\mu m$, respectivamente. A banda quality representa a camada de qualidade dos *pixels* a partir da qual são detectadas as coberturas de nuvem e sombra de nuvem.

Fonte: Próprio Autor.

usado para o monitoramento de áreas agrícolas e indica maior aptidão para a produção de culturas intensivas e de longo prazo, especialmente para culturas alimentícias (METTERNICHT, 2003). Os índices NDVI e EVI são fornecidos nos cubos de dados do BDC. Os demais índices utilizados nos estudos de caso foram computados de acordo com as fórmulas apresentadas na Tabela 3.2.

Tabela 3.2 - Índices espectrais usados nos estudos de caso.

Disponibilidade	**Índice**	**Fórmula**	**Referência**
Gerados pelo BDC	EVI	$2.5\frac{\text{nir}-\text{red}}{\text{nir}+6\text{red}-7.5\text{blue}+1}$	Huete et al. (1999)
	NDVI	$\frac{\text{nir}-\text{red}}{\text{nir}+\text{red}}$	Rouse et al. (1973)
Computados localmente	GEMI	$\eta(1-0.25\eta)-\frac{(\text{red}-0.125)}{(1-\text{red})}$, $\eta=\frac{(2(\text{nir}^2-\text{red}^2)+1.5\text{nir}+0.5\text{red}}{(\text{nir}+\text{red}+0.5)}$	Pinty e Verstraete (1992)
	GNDVI	$\frac{\text{nir}-\text{green}}{\text{nir}+\text{green}}$	Gitelson et al. (1996)
	MNDWI	$\frac{\text{green}-\text{swir}}{\text{green}+\text{swir}}$	Xu (2006)
	NDWI	$\frac{\text{nir}-\text{swir}}{\text{nir}+\text{swir}}$	Gao (1996)
	PVR	$\frac{\text{green}-\text{red}}{\text{green}+\text{red}}$	Metternicht (2003)

As bandas red, green, nir e swir representam as bandas do vermelho, verde, infravermelho-próximo e infravermelho de ondas curtas, respectivamente, no cubo de dados correspondente ao estudo de caso (ver Tabela 3.1).

Fonte: Próprio Autor.

3.2 Extração e interpolação de séries de temporais

As séries temporais foram extraídas usando conjuntos de amostras consideradas verdades de campo. Uma amostra é constituída pelas seguintes informações: uma localização geográfica (latitude e longitude), um intervalo de tempo (data de início e data de fim) e uma classe de uso e cobertura da terra para a qual a amostra é válida.

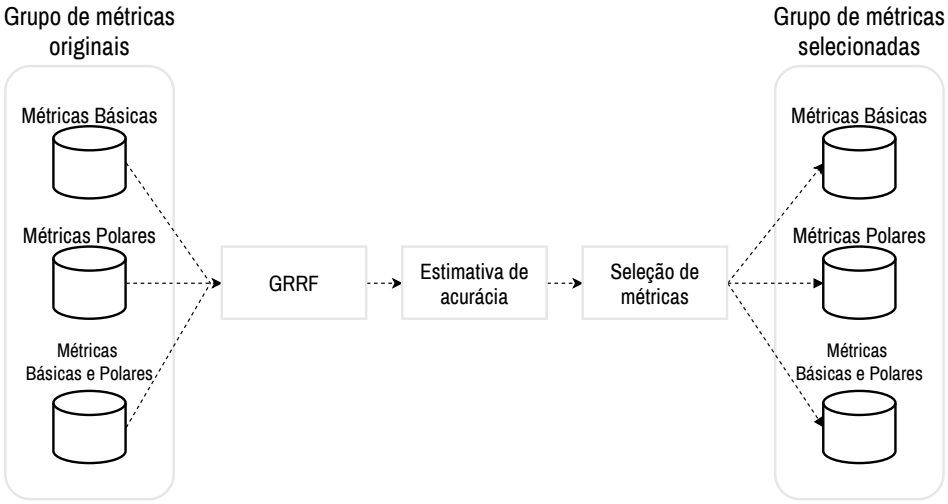
Todas as amostras usadas nos três estudos de caso são referentes ao mesmo intervalo de um ano, que vai de 1 setembro de 2018 a 31 de agosto de 2019, o que corresponde a 23 observações de acordo com a periodicidade de 16 dias dos cubos de dados utilizados. Utilizando informações de qualidade do *pixel*, os valores de cobertura de nuvens e de sombra de nuvem são removidos e interpolados linearmente usando os valores válidos da vizinhança temporal. Esse intervalo de um ano coincide com o calendário agrícola praticado nas localidades estudadas. Para extrair e interpolar as séries temporais, foi utilizado o pacote de código aberto `sits` que consegue acessar os cubos de dados do projeto BDC.

3.3 Extração e seleção de métricas

Nesta etapa, foram definidas 25 métricas a serem extraídas de cada atributo das séries temporais (i.e. bandas e índices espectrais). Essas métricas fornecem um valor que captura uma certa característica das séries temporais. Elas estão divididas em dois grupos, as básicas, com 15 métricas, e as polares, com 10 métricas. A Tabela 3.3 fornece a lista das métricas utilizadas e uma breve descrição de cada uma delas. O nome entre parênteses refere-se a como a métrica será denominada neste trabalho.

Após a extração das métricas de cada banda e índice espectral das séries temporais, procedeu-se à seleção das métricas mais relevantes na discriminação das classes de uso e cobertura da terra. O algoritmo utilizado para identificar a combinação de métricas mais relevantes foi o GRRF, fornecido pelo pacote RRF[1]. Para esta etapa, foram consideradas as métricas básicas e polares isoladamente, e uma terceira seleção foi realizada considerando todas as métricas juntas, criando assim três grupos de métricas. Para cada grupo de métricas, foram testados diferentes valores para os parâmetros: *mtry* e γ. Nesses foram variados a quantidade de atributos amostrados aleatoriamente como candidatos em cada divisão, e a importância de cada atributo dada pelo peso na divisão de cada nó, respectivamente.

Figura 3.2 - Etapas efetuadas para a seleção de atributos em cada grupo de métricas.



Fonte: Próprio Autor.

[1] https://CRAN.R-project.org/package=RRF

Tabela 3.3 - Métricas de sumarização temporal avaliadas neste trabalho.

Métrica	**Descrição**	***Range***
Amplitude (amplitude)	A diferença entre os valores máximo e mínimo do ciclo. Uma pequena amplitude significa um ciclo estável.	$[0, 1]$
Ângulo (angle)	O ângulo principal da forma fechada criada pela visualização polar. Um ângulo pequeno define uma forma possivelmente estável ao longo das estações, enquanto ângulos diferentes apontam para picos de EVI em uma estação específica.	$[0, \pi]$
Max (max)	Retorna o valor máximo da série.	$[0, 1]$
Min (min)	Retorna o valor mínimo da série	$[0, 1]$
Média (mean)	Retorna a média da série.	$[0, 1]$
Mediana (median)	Retorna o valor médio da série.	$[0, 1]$
Somatório (sum)	Retorna a soma de todos os pontos da série.	≥ 0
Desvio padrão (std)	Retorna o desvio padrão da série.	$[0, 1]$
Somatório absoluto (abs_sum)	Retorna a soma absoluta dos pontos da série.	≥ 0
Derivada média absoluta (amd)	Retorna a média absoluta da diferença entre cada ponto da série.	$[0, 1]$
Média da energia espectral (mse)	A densidade de energia espectral média calcula a energia da série temporal que é distribuída pela frequência.	≥ 0
Primeiro quartil (fqr)	Retorna o valor do primeiro quartil da série (0,25)	$[0, 1]$
Segundo quartil (sqr)	Retorna o valor do segundo quartil da série (0,50)	$[0, 1]$
Terceiro quartil (tqr)	Retorna o valor do terceiro quartil da série (0,75)	$[0, 1]$

Tabela 3.3 - Conclusão.

Métrica	**Descrição**	***Range***
Faixa interquartil (iqr)	Retorna a faixa interquartil (diferença entre o terceiro e o primeiro quartil).	$[0, 1]$
Excentricidade (ecc_metric)	Retorna valores próximos a 0 se a forma for um círculo e 1 se a forma for semelhante a uma linha.	$[0, 1]$
Primeira inclinação (first_ slope)	Valor máximo da primeira inclinação do ciclo. Indica quando o ciclo apresenta alguma mudança abrupta na curva. A inclinação entre dois valores relaciona a rapidez das fases de crescimento ou senescência.	$[-1, 1]$
Área (area)	Área da forma fechada. Um valor mais alto indica um ciclo com altos valores de VI.	≥ 0
Área por estação (area_q1, area_q2, area_q3, area_q4)	Área parcial da forma fechada, proporcional a um quadrante específico da representação polar. Alto valor na temporada de verão pode estar relacionado ao desenvolvimento fenológico de uma terra cultivada.	≥ 0
CSI (csi)	Medida quantitativa sem dimensão da morfologia,que caracterizam o desvio padrão de um objeto de um círculo.	≥ 0
Gyradius (gyradius_radius)	Igual à distância média entre cada ponto dentro da forma e o centroide da forma. Quanto mais semelhante a um círculo for a forma, maior a probabilidade do centroide estar dentro dele e, portanto, esse recurso estará mais próximo de 0.	≥ 0
Balanço polar (polar_balance)	O desvio padrão das áreas por estação, considerando as 4 estações. Um valor pequeno pode apontar para um ciclo constante, como o EVI da água (com um pequeno valor de Área) ou floresta (com um valor médio de Área).	≥ 0

Fonte: Adaptado de Körting et al. (2013)

Em detalhes, de modo a garantir que diferentes valores de parâmetros fossem testados na seleção de atributos, no parâmetro *mtry* variou-se a quantidade de atributos amostrados indo de $\sqrt{p}$ até $0.95p$ em cinco passos igualmente espaçados (WUNDERVALD et al., 2020), em que p é a quantidade total de atributos das séries temporais. Para o parâmetro de peso (γ), os valores variaram de 0.1 a 1, com incrementos de 0.1. A variação de parâmetros resultou em um total de 50 testes para cada um dos três grupos de métricas. Em cada teste, um conjunto de métricas foi retornado pelo algoritmo GRRF. Em seguida, foi computada uma estimativa de acurácia global para cada subconjunto de métricas selecionadas usando o algoritmo RF. Para essa estimativa separou-se 70% das amostras para treinamento e 30% para teste. Para garantir que nenhum viés fosse introduzido durante o treinamento de cada conjunto de métricas selecionadas de cada grupo, foi utilizado a técnica de reamostragem, em que cada conjunto de métricas foi computado 30 vezes. Assim, estabelecendo a média e o desvio padrão da acurácia global de cada conjunto selecionado.

Após estimada a acurácia resultante, selecionou-se o conjunto de métricas mais representativo para cada grupo (básicas, polares e métricas combinadas) utilizando-se do método de ótimo de Pareto (IZQUIERDO-VERDIGUIER; ZURITA-MILLA, 2020; BOX; MEYER, 1986), em que objetivou-se encontrar os subconjuntos mais representativos com a maior média de acurácia global e a menor quantidade de atributos selecionados. Essa otimização avaliou a capacidade discriminatória das métricas para distinguir as classes de uso e cobertura da terra usando um espaço de atributos reduzido. A Figura 3.2 apresenta as etapas realizadas para a seleção das métricas.

A extração das métricas foi feita pelo pacote R *sitsfeats*[2], desenvolvido neste trabalho. As funções de processamento das séries temporais do pacote, utilizadas para o cálculo das métricas, foram implementadas em C++ com o auxílio da biblioteca *RcppArmadillo*. Os cálculos são realizados de forma matricial, o que otimiza o desempenho de processamento. As funcionalidades do pacote desenvolvido neste trabalho é análoga às disponibilizadas na biblioteca *stmetrics*[3], do ecossistema Python.

3.4 Estudos de casos

Para comparar as abordagens de séries temporais completas e métricas de séries temporais, foram realizados três estudos de casos com diferentes aplicações. No primeiro objetivou-se avaliar as amostras de uso e cobertura da terra usando o método proposto por Santos et al. (2019) que usa o método de agrupamento SOM.

[2]https://github.com/oldlipe/sitsfeats
[3]https://github.com/brazil-data-cube/stmetrics

No segundo estudo de caso, foi realizada uma classificação de corpos d'água em que se objetivou avaliar os desempenhos entre as abordagens de séries temporais completas e de métricas. Por fim, no último estudo de caso, foi realizada uma classificação de uso e cobertura da terra a partir das duas abordagens e objetivou-se avaliar os respectivos desempenhos computacionais e de acurácia na classificação. Esses estudos de caso são descritos, respectivamente, nos Capítulos 4, 5 e 6.

Tabela 3.4 - Resumo dos estudos de caso realizados.

Item	**Estudo de caso 1** **Análise de amostras**	**Estudo de caso 2** **Máscaras de água**	**Estudo de caso 3** **Mapas de uso e cobertura**
Objetivo	Avaliar a separabilidade das classes de uso e cobertura das amostras usando métricas de séries temporais	Avaliar a geração de máscaras de água usando métricas de séries temporais	Avaliar a geração de mapas de uso e cobertura da terra usando métricas de séries temporais
Área de estudo	Região compreendida entre os estados de MT, MS e GO.	Região localizada no centro-oeste do estado de Minas Gerais.	Mesorregião do extremo oeste da Bahia.
Número de amostras	Amostras de avaliação: 852	Treinamento: 200 Validação: 36733	Treinamento: 922 Validação: 1453
Classes de uso e cobertura	Agricultura (256), Cana-de-açúcar (134), Floresta (245) e Pastagem (217)	Água (100) e Não-Água (100)	Agricultura (242), Vegetação Natural (422) e Pastagem (258)
Cubo de dados	BDC Sentinel-2 Bandas: 7 Índices: 6	BDC Landsat-8 Bandas: 9 Índices: 2	BDC CBERS-4 Bandas: 4 Índices: 2
Série temporal	Período: 1 ano (23 observações) De: 2018-09-01 Até: 2019-08-31	*	*
Métricas	Métricas básicas: 15 Métricas polares: 10 Método de seleção: GRRF Estimação de acurácia: validação cruzada	*	*
Aprendizado de máquina	*Self-Organized Maps*	*Random Forests* *Support Vector Machine*	*Random Forests* *Support Vector Machine*
Tipo de aprendizado	Não-supervisionado	Supervisionado	Supervisionado

Os pontos * correspondem ao mesmo texto em toda linha.

Fonte: Próprio Autor.

4 ESTUDO DE CASO 1: AVALIAÇÃO DE AMOSTRAS

4.1 Contextualização

Um dos grandes desafios de gerar modelos de aprendizado de máquina supervisionados é conseguir um conjunto de amostras representativas da área de interesse e com boa qualidade. Amostras ruidosas e não representativas podem causar um efeito negativo no desempenho da classificação (FRÉNAY; VERLEYSEN, 2013; SANTOS et al., 2021c). A avaliação de amostras é uma importante etapa para se obter bons resultados.

No contexto de séries temporais de imagens de satélites e métricas derivadas dessas séries temporais, duas fontes de erros podem afetar a qualidade das amostras de uso de cobertura da terra (PELLETIER et al., 2017). A primeira está relacionada ao processo de aquisição do dado de sensoriamento remoto e as condições nas quais essa aquisição é realizada. Ruídos provocados por coberturas de nuvens ou inconsistências causadas na calibração da imagem são exemplos dessa fonte de erros. A segunda fonte de erros está relacionada com o levantamento das amostras. Ruídos de classe devido a erros de interpretação do dado ou rotulagens errôneas são exemplos dessa fonte de erro.

Uma das formas de melhorar a qualidade de um conjunto de amostras é realizando um pré-processamento que identifique aquelas amostras potencialmente ruidosas no conjunto dos dados avaliados. Amostras ruidosas tendem a exibir comportamentos anômalos que podem ser detectados usando técnicas de análise de agrupamento (SANTOS et al., 2019).

Neste estudo de caso, é realizado uma comparação entre as abordagens de séries temporais completas e de métricas de séries temporais para a análise de amostras de uso e cobertura da terra usando o método proposto em Santos et al. (2019).

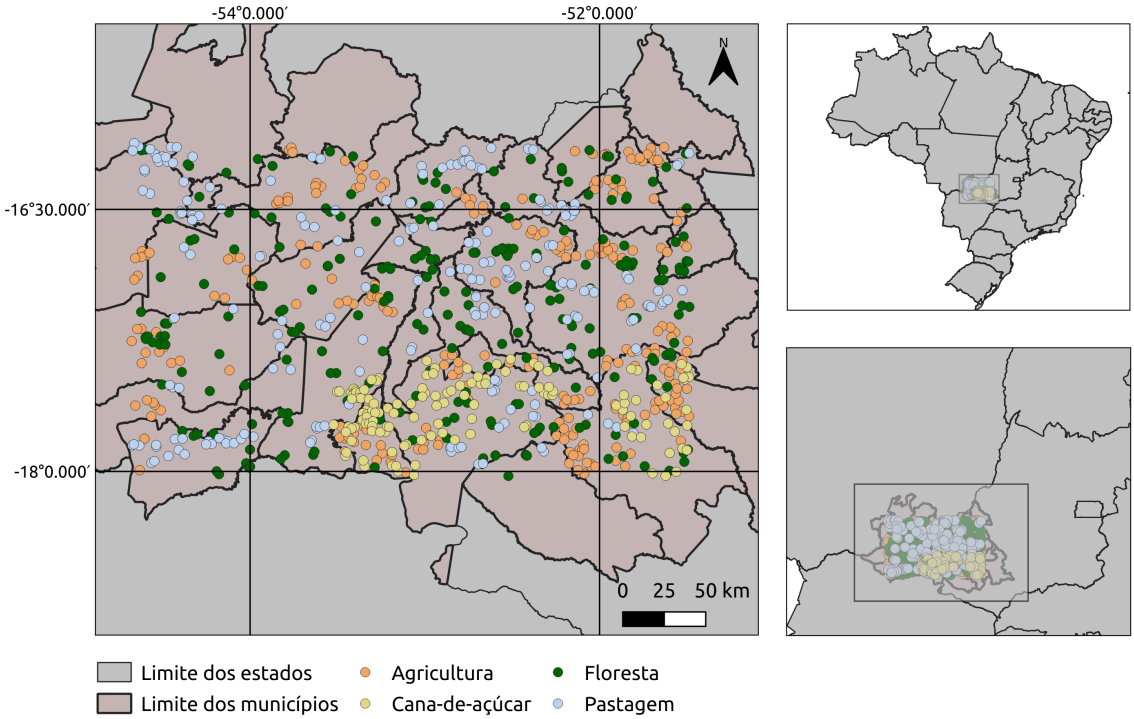
4.2 Materiais e métodos

4.2.1 Área de estudo

A área de estudo compreende os estados do Mato Grosso, Mato Grosso do Sul e Goiás, conforme apresentado na Figura 4.1. A região de estudo foi escolhida com base no conjunto de amostras coletadas em Picoli et al. (2020) e que foram utilizadas neste estudo de caso. A região, compreendida pelo bioma Cerrado, é caracterizada pela

presença de variadas fitofisionomias vegetais e de áreas agrícolas com predominância de agriculturas temporárias, como cana-de-açúcar e soja, e pastagem.

Figura 4.1 - Localização da área de estudo avaliada no estudo de caso de avaliação de amostras.



Fonte: Próprio Autor.

4.2.2 Dados de entrada

O conjunto de 852 amostras usado neste estudo de caso foi coletado por interpretação visual de imagens de alta resolução com auxílio de séries temporais (PICOLI et al., 2020). O conjunto possui as seguintes classes e quantidades de amostras: Agricultura (256 amostras), Cana-de-açúcar (134 amostras), Floresta (245 amostras) e Pastagem (217 amostras).

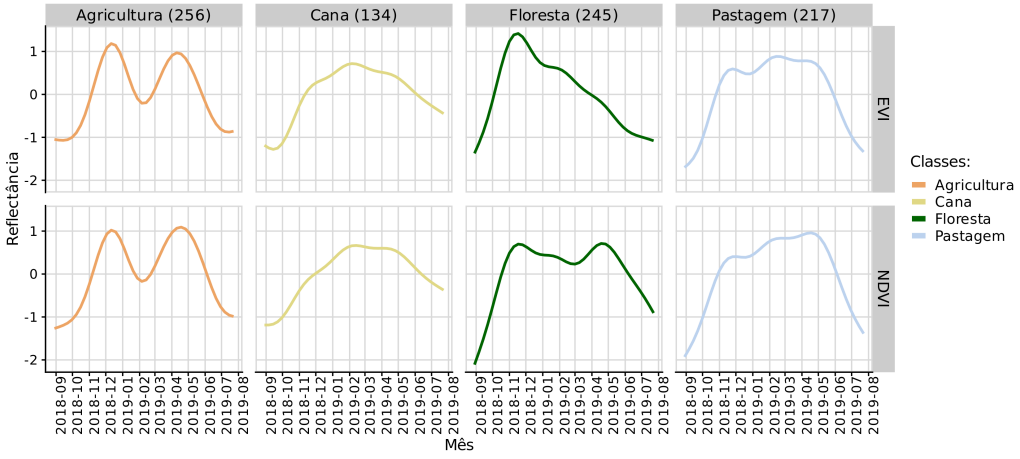
As amostras compreendem o intervalo de um ano agrícola que vai de setembro de 2018 até agosto de 2019. A classe de Agricultura refere-se a culturas temporárias cujo ciclo vegetativo tem duração inferior a um ano e, após a colheita, o terreno está disponível para um novo plantio. A classe Cana-de-açúcar é considerada uma

cultura semi-perene em que o cultivo pode ter uma duração entre um e dois anos (INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA - IBGE, 2013).

As séries temporais das amostras foram extraídas do cubo de dados Sentinel-2 (cubo de dados do BDC S2_10_16D_STK-1), através das bandas do espectro visível (blue, green e red), primeira banda do *red-edge*, infravermelho próximo (nir), as bandas do infravermelho curto (swir16 e swir22), e os índices de vegetação EVI, NDVI, GNDVI, PVR, NDWI e GEMI. A escolha das bandas e dos índices foi baseada na revisão sistemática publicada por Chaves et al. (2020).

A Figura 4.2 apresenta os padrões espectro-temporais de cada classe para os índices EVI e NDVI obtidos por modelo aditivo generalizado. É possível observar que o padrão da classe de Agricultura possui dois ciclos de cultivo, onde o primeiro ciclo compreende o período de setembro de 2018 até fevereiro de 2019, e o segundo ciclo vai de março de 2019 até agosto do mesmo ano. Outro detalhe importante é a similaridade do perfil espectro-temporal das classes Cana-de-açúcar e Pastagem, o que sugere uma potencial fonte de confusão. Para uma visualização mais detalhada das séries temporais, veja no Anexo A.1.

Figura 4.2 - Padrões espectro-temporais do EVI e NDVI obtidos por modelo aditivo generalizado das amostras utilizadas neste experimento.



Fonte: Próprio Autor.

4.3 Resultados e discussões

4.3.1 Atributos selecionados

Após a extração das séries temporais do cubo de dados, procedeu-se à extração e seleção das métricas conforme descrito no Capítulo 3, Seção 3.3. A Figura 4.3 apresenta os resultados desses procedimentos para cada grupo de métricas. As acurácias globais médias (eixo y dos gráficos) foram obtidas por reamostragem com 30 realizações de RF onde 70% das amostras de cada classe foram usadas para treinamento e 30% para teste. As barras de erro correspondem ao desvio padrão obtido. Para cada grupo de métricas aplicou-se o método ótimo de Pareto para selecionar as métricas mais representativas. O método selecionou o subconjunto que minimiza a quantidade de atributos e maximiza a acurácia global média.

Tabela 4.1 - Quantidade de atributos extraídos e selecionados para cada grupo de métricas. Os valores em parênteses correspondem ao desvio padrão da acurácia global.

Grupo de Métricas	**Atributos extraídos**	**Atributos selecionados**	**Acurácia global média**
Métricas Básicas	195	22	96.7% (± 1.2)
Métricas Polares	130	25	97.0% (± 0.9)
Métricas Básicas e Polares	325	42	97.5% (± 0.9)

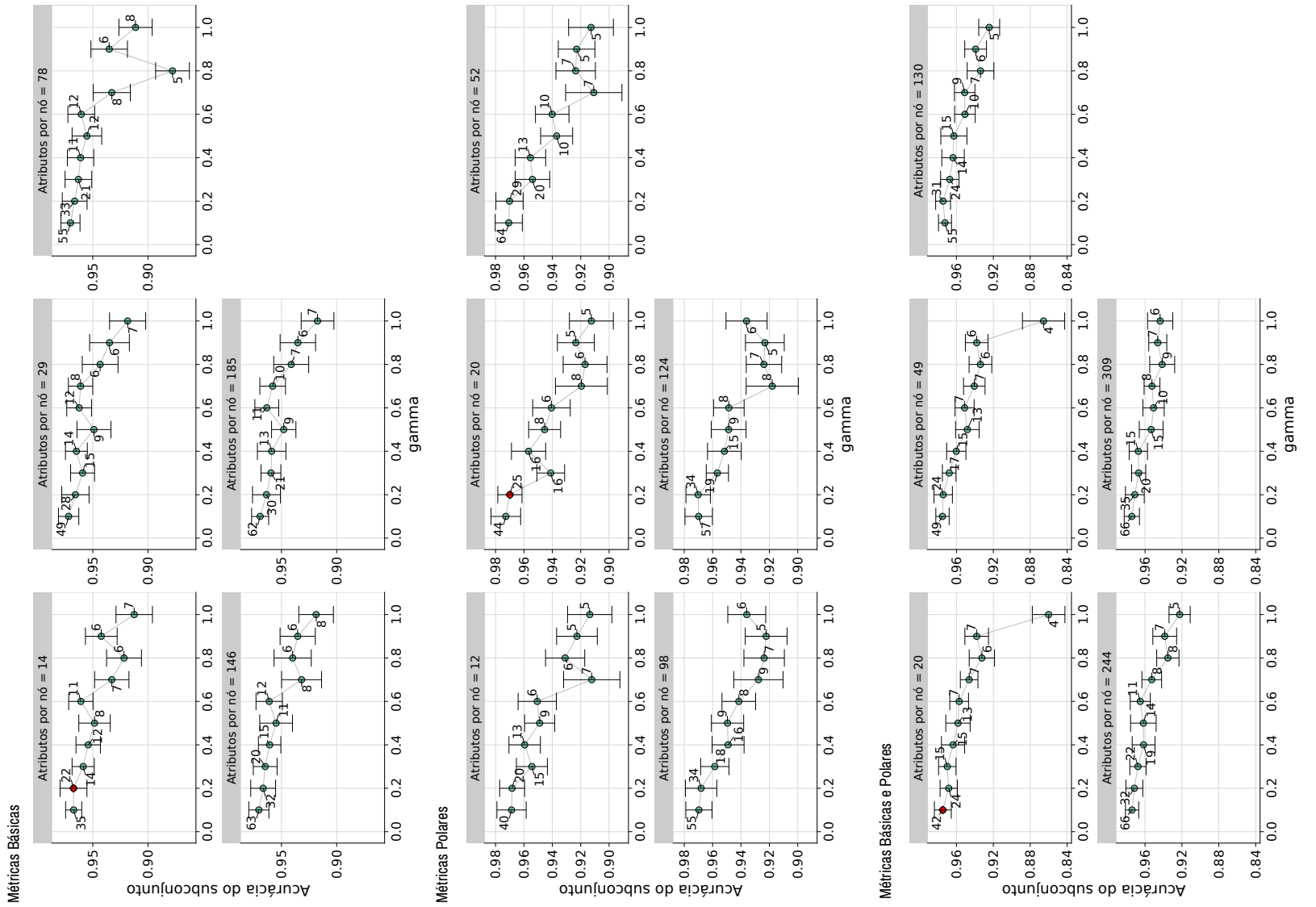
Fonte: Próprio Autor.

Nota-se na Figura 4.3 que conforme aumentam os valores de γ diminuem a quantidade de subconjuntos selecionados. Os subconjuntos selecionados pelo critério de Pareto para cada grupo de métrica tiveram o parâmetro $\gamma \leq 0.2$. Não se observou uma relação significativa entre o parâmetro *mtry* e a quantidade de atributos nos subconjuntos obtidos. A Tabela 4.1 apresenta um resumo dessa etapa de extração e seleção indicando a quantidade de atributos extraídos e selecionados.

4.3.2 Modelos selecionados

A análise das amostras seguiu os procedimentos propostos em Santos et al. (2019). Para determinar os parâmetros do modelo de agrupamento SOM, foram realizados 20 experimentos variando-se o parâmetro do tamanho da grade de neurônios para cada grupo de métricas e para as séries temporais completas. Para cada um dos experimentos foram computados os índices de avaliação externa ARI, RI, J e entropia.

Figura 4.3 - Acurácia de cada subconjunto de métricas selecionadas a partir do modelo GRRF. Os valores indicados nos círculos em azul mostram a quantidade de atributos selecionados, e o losango em vermelho o subconjunto selecionado de métricas.



Fonte: Próprio Autor.

Tabela 4.2 - Resultados dos agrupamentos gerados neste estudo de caso. Os valores em negrito representam os índices de avaliação externa que apresentaram melhores resultados.

Dados	Grade	ARI	RI	Jaccard	Entropia
Séries Temporais	8x8	0.834	0.936	0.781	0.247
	10x10	0.879	0.953	0.836	0.179
	12x12	0.915	0.967	0.881	**0.118**
	15x15	0.909	0.965	0.874	0.115
	18x18	0.926	0.971	0.896	0.087
Métricas Básicas	8x8	0.876	0.952	0.834	0.191
	10x10	0.897	0.960	0.858	0.170
	12x12	0.909	0.965	0.874	0.140
	15x15	0.916	0.968	0.883	0.099
	18x18	0.936	0.975	0.910	0.076
Métricas Polares	8x8	0.903	0.962	0.866	0.156
	10x10	**0.939**	**0.976**	**0.914**	**0.087**
	12x12	0.929	0.972	0.900	0.096
	15x15	0.937	0.976	0.911	0.074
	18x18	0.962	0.985	0.945	0.052
Métricas Básicas e Polares	8x8	**0.937**	**0.976**	**0.911**	**0.11**
	10x10	0.936	0.975	0.909	0.104
	12x12	**0.946**	**0.979**	**0.924**	**0.074**
	15x15	**0.962**	**0.985**	**0.945**	**0.051**
	18x18	**0.953**	**0.982**	**0.933**	**0.056**

Fonte: Próprio Autor.

Os índices de avaliação dos agrupamentos obtidos são apresentados na Tabela 4.2. Em geral, em todos os cenários avaliados, o uso de métricas apresentaram os melhores resultados. Nota-se que a combinação entre métricas básicas e polares produziram agrupamentos mais homogêneos para as grades de 8x8, 12x12, 15x15 e 18x18, e as métricas polares para as grades 10x10. Em relação aos índices de validação externa, conforme o crescimento da grade do SOM, os neurônios se tornam mais especializados, e o índice RI foi o que mais aumentou devido ao seu termo d (veja em 2.5.2), como apresentado por Faceli et al. (2005) e Vendramin et al. (2010). Outro ponto importante é a concordância entre os índices de avaliação externa e a entropia, que demonstram a homogeneidade dos agrupamentos gerados.

Considerando apenas os experimentos usando métricas, observou-se que o subconjunto selecionado de métricas básicas apresentou os piores resultados se comparados com os demais. Isso mostra a dificuldade de se obter um conjunto de dados represen-

tativo o suficiente para separar as amostras de Cana-de-Açúcar e Pasto. Por outro lado, o subconjunto de métricas básicas e polares combinadas apresentou os melhores resultados em grande parte dos agrupamentos, o que mostra que a junção das métricas pode resultar em um conjunto de atributos mais representativos de cada grupo.

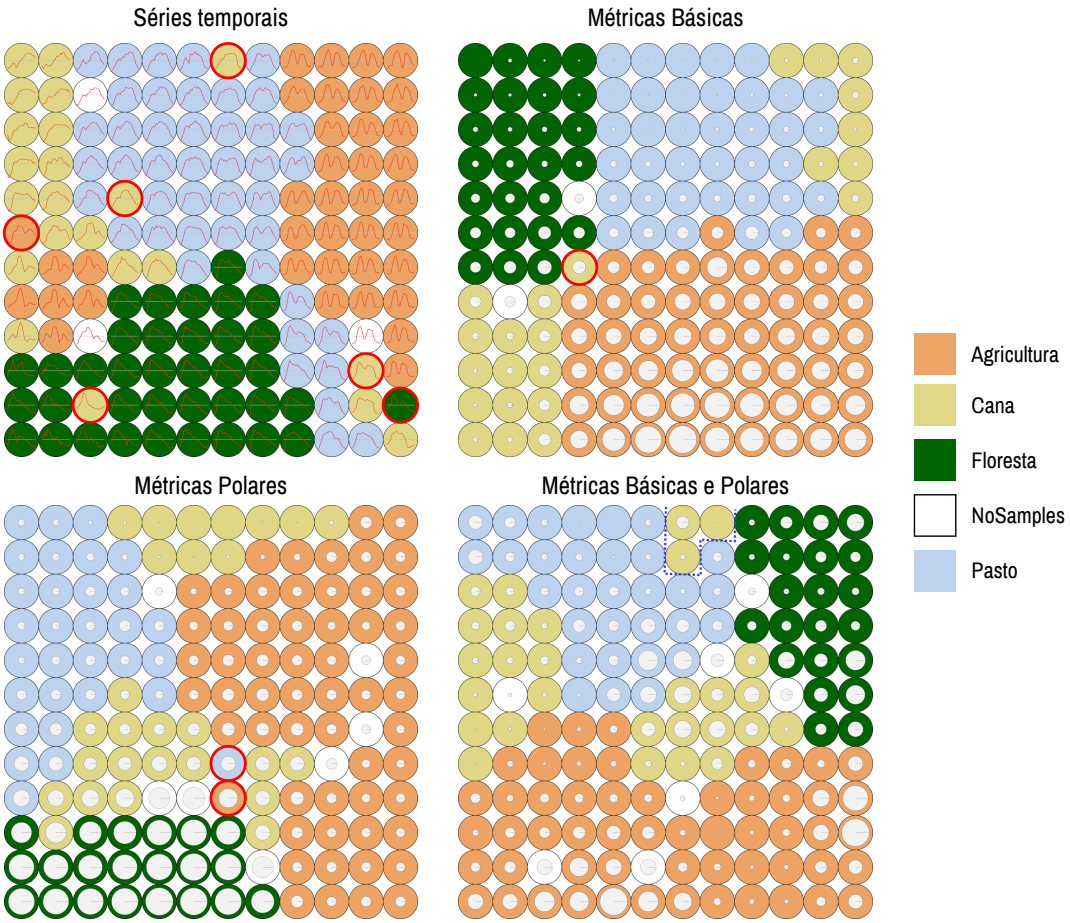
Com base nos resultados apresentados na Tabela 4.2 e em uma análise visual dos agrupamentos gerados, observou-se que os agrupamentos com grade maiores que 12x12 produziram muitos neurônios vazios. Assim, foram escolhidos os agrupamentos com grade de 12x12 neurônios para realizar a avaliação da separabilidade das amostras. Tal configuração de grade corrobora com a heurística apresentada por Vesanto e Alhoniemi (2000), na qual indica que a quantidade ideal de neurônios no SOM corresponde a $5\sqrt{P}$, em que P é a quantidade total de observações.

4.3.3 Avaliação das amostras

Os resultados do método de avaliação de amostras proposto por Santos et al. (2021c) são apresentados na Figura 4.4. Nos agrupamentos gerados por séries temporais, métricas básicas e métricas polares nota-se a presença de neurônios *outliers*, destacados com círculos vermelhos, cuja classe majoritária difere das classes majoritárias de suas vizinhanças. Esses neurônios não são necessariamente resultantes de amostras rotuladas erroneamente, eles podem representar amostras que possuem diferentes padrões de classes de uso e cobertura da terra no espaço ou tempo, ou amostras que não são separáveis utilizando séries temporais ou métricas extraídas a partir delas.

Visualmente é possível observar que o agrupamento baseado em séries temporais apresenta uma vizinhança mais heterogênea, principalmente referente aos neurônios rotulados como Cana. Além das maiores confusões entre as classes de Pasto e Cana, no agrupamento baseado em séries temporais observa-se também que as amostras de Floresta podem apresentar confusões quando comparadas com outras classes. Por outro lado, os agrupamentos baseados em métricas possuem grupos com vizinhanças mais homogêneas, gerando menos confusões entre as classes. Essa homogeneidade vai de acordo com os resultados apresentados na Tabela 4.2, em que os erros baseados em entropia são menores principalmente para as métricas polares e a combinação das métricas básicas com as polares.

Figura 4.4 - Resultados dos agrupamentos gerados através de séries temporais, métricas básicas, métricas polares e métricas básicas e polares em uma grade de 12x12. Os círculos vermelhos representam possíveis neurônios *outliers*.

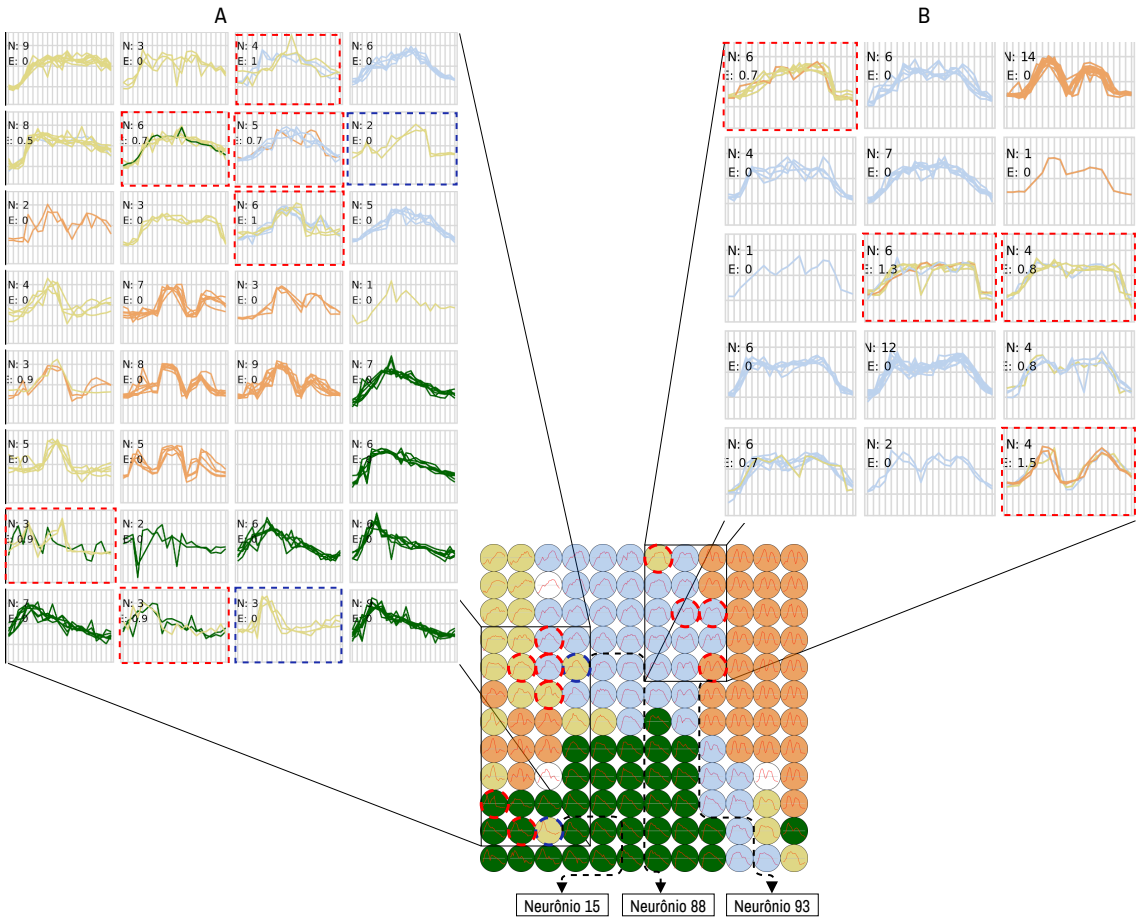


Fonte: Próprio Autor.

Para uma análise mais profunda do método de avaliação de amostras proposto por Santos et al. (2021a), foi utilizada a ferramenta criada por Souza et al. (2019), apresentada na Figura 4.5 para visualizar ampliadamente as informações correspondentes às amostras agrupadas em cada neurônio, entre elas as séries temporais, a quantidade de amostras (N), e a pureza baseada na entropia (E). Na Figura 4.5 no lado (A), dois neurônios *outliers* rotulados como Cana foram destacados em linha tracejada azul, neurônios 15 e 88. Nota-se que esses neurônios são homogêneos, porém ambos apresentam vizinhança onde a maioria contém amostras das classes de Floresta e Pasto. Considerando que as amostras utilizadas neste estudo de caso

correspondem ao calendário agrícola de um ano, nota-se que as séries temporais EVI agrupadas no neurônio 15 possuem padrões de monocultura, porém apresentam ciclos que duram em torno 90 dias, com início em novembro e finalizando no início de fevereiro. Já no caso do neurônio 88, as séries temporais EVI apresentam um tipo de monocultura com ciclo de aproximadamente 120 dias, iniciando em janeiro e indo até abril. Por conta dos perfis temporais apresentados pelas amostras agrupadas nestes neurônios, provavelmente são neurônios *outliers* que indicam amostras que foram rotuladas erroneamente para o ano agrícola de 2018/19.

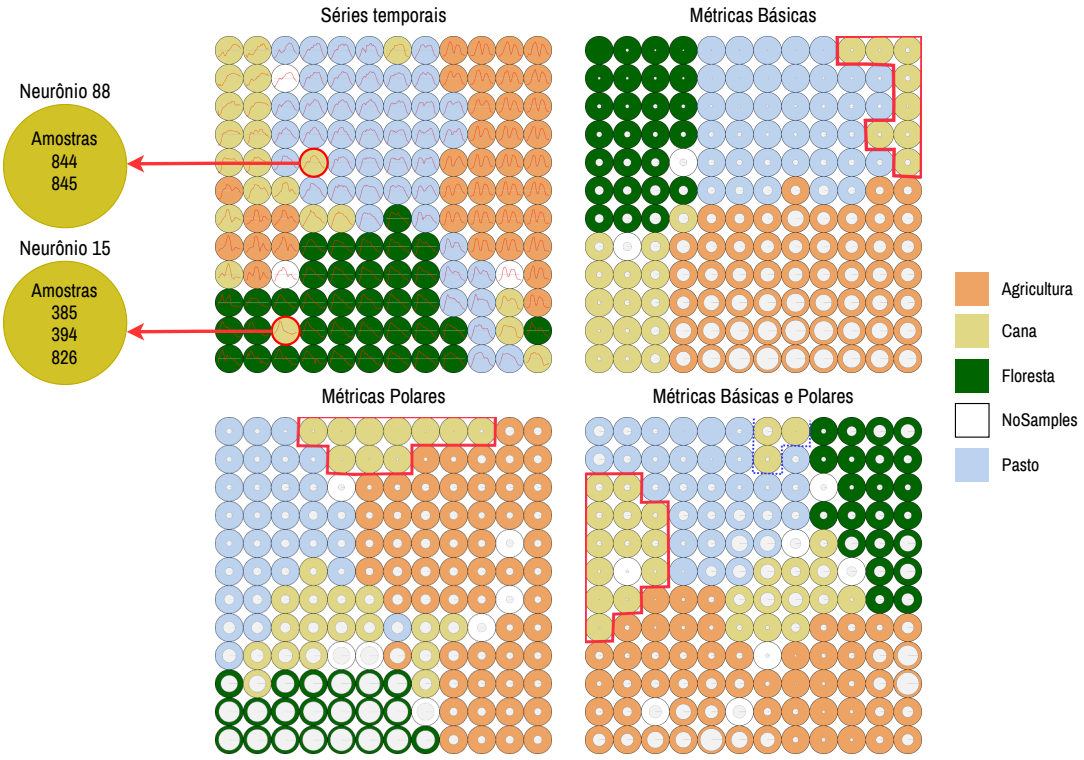
Figura 4.5 - Confusão entre neurônios no agrupamento de séries temporais.



Fonte: Próprio Autor.

É importante destacar que as amostras que foram agrupadas nos neurônios 15 e 88 deixaram de ser mapeadas em neurônios *outliers* em vizinhanças de Floresta e Pasto quando os agrupamentos por métricas foram aplicados. A Figura 4.6 ilustra os subgrupos em que as seis amostras foram mapeadas em cada um dos agrupamentos baseado em métricas. Neste caso, os agrupamentos baseados em métricas criaram uma nova vizinhança com os neurônios de Cana que possuem o padrão de monocultura com o período mais curto do que as demais amostras de Cana.

Figura 4.6 - Identificação das amostras atribuídas a neurônios *outliers* nos agrupamentos baseados em séries temporais nos agrupamentos baseado em métricas.



Fonte: Próprio Autor.

Em relação aos neurônios apresentados no lado (B) da Figura 4.5, observam-se confusões entre Pasto e Cana, tal comportamento é esperado devido as suas semelhanças espectro-temporais (XAVIER et al., 2006; GUSSO et al., 2009). É importante destacar que o neurônio 93 apresentou maior entropia em comparação com os demais, 1.5. Isto se deve ao fato da heterogeneidade causada pela diversidade de classes agrupada

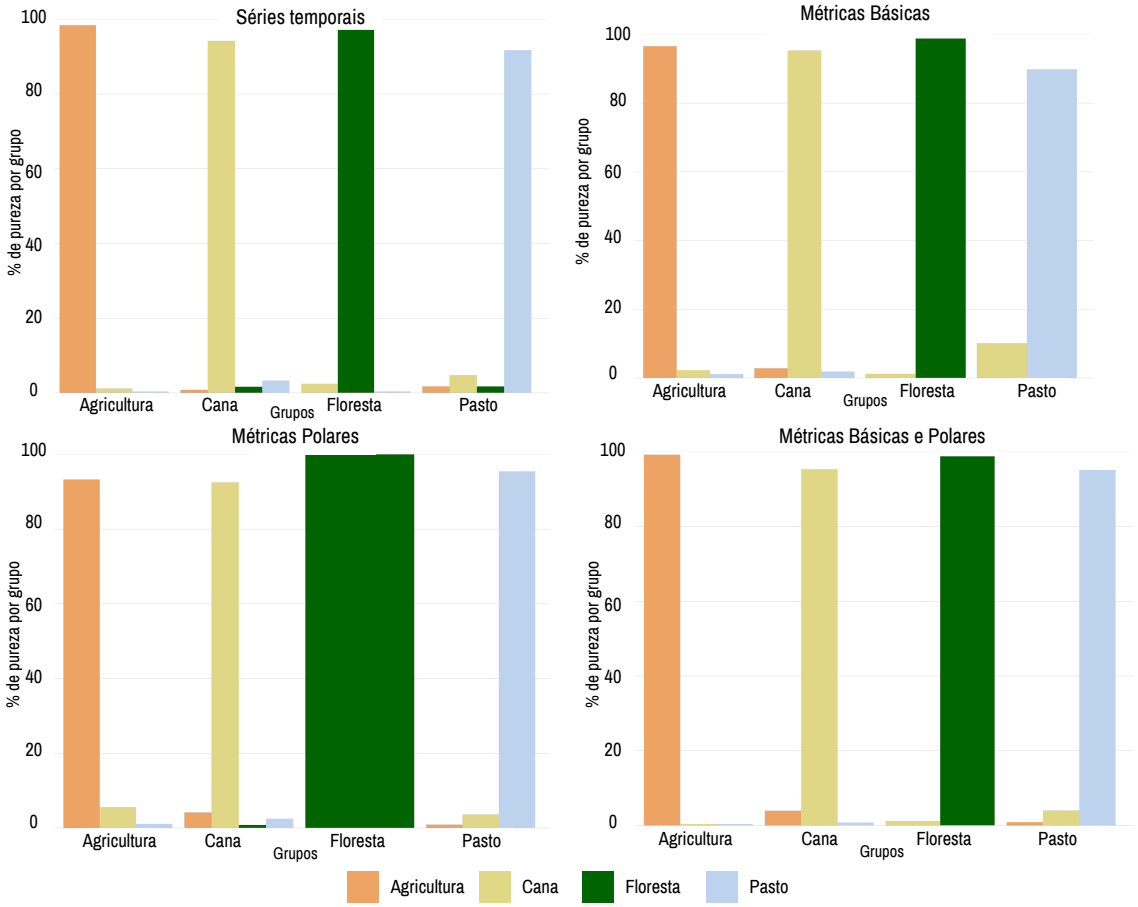
nele, entre elas duas amostras de Agricultura, uma de Pasto e Cana. Além disso, nota-se que as amostras de Cana e Pasto apresentam padrões de amostras com o perfil de Agricultura de cultivo duplo. Neste caso, o método proposto por Santos et al. (2021c) sinaliza a remoção dessas amostras de Cana e Pasto do conjunto de dados de forma automática, sem a necessidade da análise de um especialista da área, por conta dos baixos valores de probabilidades que essas amostras indicam.

Para ter uma visão geral da qualidade dos agrupamentos gerados por séries temporais e métricas, a Figura 4.7 apresenta as confusões entre classes que foram geradas para cada grupo. Neste caso, considera-se um grupo o conjunto de neurônios que foram rotulados com a mesma classe, como proposto em Santos et al. (2021c). No geral, para todos os agrupamentos, tiveram-se grupos de boa qualidade, com o grau de pureza acima de 89%. No entanto, no agrupamento baseado na combinação de métricas básicas e polares obtiveram-se grupos com a melhor separabilidade.

No conjunto de dados utilizado, as amostras de Floresta são as que apresentam a melhor separabilidade quando comparadas com as demais classes da área de estudo selecionada. Além disso, houve uma melhora na separabilidade entre as classes de Floresta com Cana e Pasto quando os agrupamentos baseados em métricas foram aplicados. Esse comportamento é visto principalmente no agrupamento com métricas polares, que o grupo de Floresta foi 100% puro.

Já nos grupos de Agricultura houve maiores confusões com as classes de Cana e Pasto. Como mencionado anteriormente, as confusões entre as amostras Pasto e Agricultura podem ter ocorrido por conta dos padrões de amostras rotuladas como Cana que possuem o perfil espectro-temporal de monocultura. Enquanto que a confusão entre Agricultura e Pasto pode ter ocorrido por conta de séries temporais ruidosas de amostras de Pasto, causadas por ocorrência de nuvem, que geram assinaturas temporais semelhantes à Agricultura de duplo cultivo.

Figura 4.7 - Porcentagem de confusão entre os grupos do agrupamento SOM com grande 12x12.



Fonte: Próprio Autor.

Após a etapa de agrupamento e avaliação das amostras, foram aplicadas as regras propostas por Santos et al. (2021c) (veja em 2.5.3) para identificar amostras que possuem boa qualidade e separabilidade através de séries temporais, amostras que sinalizam que devem ser analisadas por especialistas, e por fim amostras ruidosas que são sinalizadas para serem removidas do conjunto de dados. Nesse método é proposto um limiar para identificar o grau de confiabilidade das amostras através dos valores das probabilidades *a priori* e *a posteriori*. Sendo assim, foi usado o limiar de 60% para ambas as probabilidades mencionadas. A Tabela 4.3 apresenta os resultados da análise de ruídos de cada classe para cada agrupamento. As classes filtradas são aquelas consideradas boas, e as marcadas são consideradas duvidosas e requerem a avaliação de especialistas. Observa-se que as métricas combinadas sinalizaram a remoção de menos amostras de Cana, Floresta e Pasto do que as séries temporais.

Com isso, para avaliar o efeito de remoção das amostras ruidosas, treinou-se um modelo RF com 1000 árvores utilizando a técnica *k-fold* ($k = 5$), para avaliar a qualidade das amostras antes e depois da remoção. Na limpeza foram apenas consideradas as amostras confiáveis de cada classe. A Tabela 4.4 mostra as acurácias do produtor, usuário e global para as amostras originais e filtradas. É possível observar que o uso de métricas atingiu melhores resultados em ambos os cenários. Outro detalhe é a melhora da acurácia do produtor para Cana, que após a remoção dos ruídos obtiveram-se maiores acurácias.

Tabela 4.3 - Resultado da limpeza das amostras utilizando séries temporais e os três grupos de métricas.

Dados	**Classes**	**Mantidas**	**Marcadas**	**Removidas**
Séries Temporais	Agricultura	96.5%	0.781%	2.73%
	Cana-de-Açúcar	79.1%	10.4%	10.4%
	Floresta	96.3%	0.816%	2.86%
	Pasto	88.0%	6.91%	5.07%
Métricas Básicas	Agricultura	95.7%	2.34%	1.95%
	Cana-de-Açúcar	65.7%	15.7%	18.7%
	Floresta	100%	-	-
	Pasto	92.2%	3.69%	4.15%
Métricas Polares	Agricultura	100%	-	-
	Cana-de-Açúcar	67.9%	24.6%	7.46%
	Floresta	100%	-	-
	Pasto	94.0%	1.38%	4.61%
Métricas Básicas e Polares	Agricultura	100%	-	-
	Cana-de-Açúcar	74.6%	17.9%	7.46%
	Floresta	98.4%	0.816%	0.816%
	Pasto	96.3%	1.38%	2.30%

Fonte: Próprio Autor.

Neste estudo de caso foi apresentado o uso de métricas temporais para analisar a separabilidade de amostras de uso e cobertura da terra. Observou-se que o uso de métricas resultou em agrupamentos mais homogêneos, auxiliando na separabilidade de amostras complexas. Tais resultados elucidam que o uso de métricas podem representar melhores resultados em agrupamentos SOM aplicados neste trabalho.

Tabela 4.4 - Resultado do treinamento dos modelo RF com as amostras antes e após a filtragem das observações ruidosas.

Dados	**Classes**	**AP Ori.**	**AP Filt.**	**AU Ori.**	**AU Filt.**	**AG Ori.**	**AG Filt.**
Séries Temporais	Agricultura	96.88%	98.79%	98.02%	98.79%	95.66%	97.69%
	Cana-de-Açúcar	89.55%	92.45%	92.31%	94.23%		
	Floresta	96.73%	97.88%	97.53%	98.72%		
	Pasto	96.77%	98.95%	92.92%	96.92%		
Métricas Básicas	Agricultura	**98.83**%	**100**%	97.68%	99.59%	96.95%	**99.10%**
	Cana-de-Açúcar	89.55%	**95.45**%	93.02%	96.55%		
	Floresta	**100**%	**100**%	98.79%	99.59%		
	Pasto	95.85%	98.50%	96.30%	98.99%		
Métricas Polares	Agricultura	96.88%	99.17%	97.64%	98.35%	97.07%	98.59%
	Cana-de-Açúcar	91.79%	94.51%	91.11%	94.51%		
	Floresta	99.59%	100%	99.19%	100%		
	Pasto	**97.70**%	98.04%	97.70%	99.01%		
Métricas Básicas e Polares	Agricultura	98.44%	99.60%	98.44%	98.80%	**97.65**%	99%
	Cana-de-Açúcar	**92.54**%	95.00%	93.94%	97.94%		
	Floresta	**100%**	**100%**	99.59%	100%		
	Pasto	97.24%	**99.04%**	96.79%	98.57%		

Em que **Ori** refere-se as amostras originais; **Filt** amostras filtrada; **AP** refere-se a acurácia do produtor; **AU** acurácia do usuário, e **AG** Acurácia Global.

Fonte: Próprio Autor.

5 ESTUDO DE CASO 2: MÁSCARA DE ÁGUA

5.1 Contextualização

Os corpos de água da superfície terrestre, sendo eles lagos, rios e reservatórios são essenciais para os ecossistemas terrestres e para a civilização humana (FENG et al., 2016). Visto que a água é essencial para a sobrevivência de todo tipo de organismo, logo, garantir que todos possam ter acesso a este recurso é um requisito mínimo para a sobrevivência dos seres vivos na Terra (OKI; KANAE, 2006).

Devido a ações antropogênicas, esse recurso vem sofrendo impactos que ameaçam grande parte da biodiversidade. Nas últimas três décadas, mais de $162.000km^2$ dos corpos d'água continentais apresentaram não ser permanentes, quase $90.000km^2$ desapareceram por completo e mais de $72.000km^2$ tiveram transição do estado permanente para o estado sazonal (PEKEL et al., 2016).

O mapeamento de corpos d'água é essencial para o monitoramento dessas mudanças. Uma forma de obter essas informações é através de classificação de imagens de satélite de observação da Terra. Um dos trabalhos mais proeminentes nesta área é de Pekel et al. (2016), que mapearam, globalmente, as mudanças mensais que ocorrem nas águas da superfície terrestre usando imagens dos satélites Landsat 5 (sensor TM), Landsat 7 (sensor ETM+) e Landsat 8 (sensor OLI) de 1985 até 2016.

Os mapas de corpos d'água também podem ser úteis como máscara de pós-processamento em aplicações de classificação de uso e cobertura da terra. Por exemplo, o mapa gerado por Pekel et al. (2016) foi utilizado como máscara de água por Simoes et al. (2020) em que foi classificado o estado do Mato Grosso através de imagens do sensor MODIS. Esse tipo de procedimento é útil, pois permite isolar as classes de interesse na geração de mapas de uso e cobertura da terra.

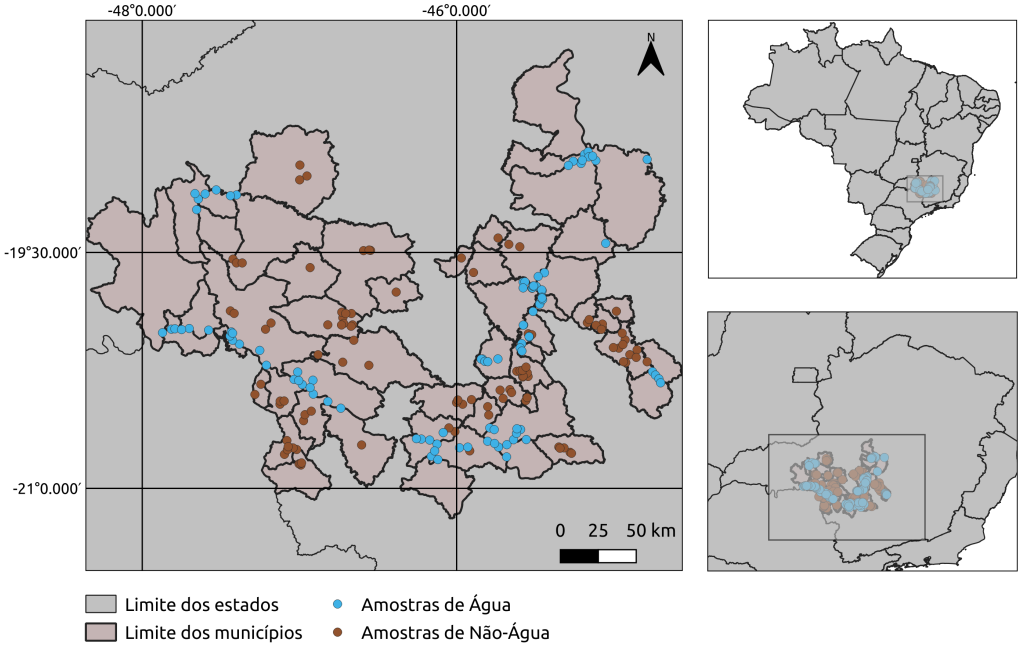
Neste estudo de caso, é realizada uma comparação entre as abordagens de séries temporais completas e de métricas de séries temporais para a geração de máscara de água. Foram gerados mapas de uso e cobertura da terra usando algoritmos RF e SVM treinados nas diferentes abordagens.

5.2 Materiais e métodos

5.2.1 Área de estudo

A área de estudo compreende a região centro-oeste do estado de Minas Gerais, conforme apresentado na Figura 5.1. As amostras foram coletadas pelo próprio autor, sendo elas: Água (100 amostras) e Não-água (100 amostras). Para as amostras de Água buscou-se coletar pontos de diferentes rios, uma vez que cada rio pode abranger diferentes composições, como a presença de sedimentos e de variadas bactérias, o que pode alterar o perfil espectro-temporal das amostras. Durante a coleta das amostras de água foram considerados os seguintes rios da região de estudo: Rio Grande, Rio São Francisco e o Rio Paraopeba. Em relação às amostras de não-água foram incluídas amostras de Área Urbana, Pastagem, Vegetação Natural e de regiões Agrícolas. A etapa de coleta de amostras foi guiada pelo mapa temático do Projeto **TerraClass 2013**[1] e uma imagem alta de resolução espacial.

Figura 5.1 - Área de estudo considerada neste experimento juntamente com as amostras de uso e cobertura da terra.



Fonte: Próprio Autor.

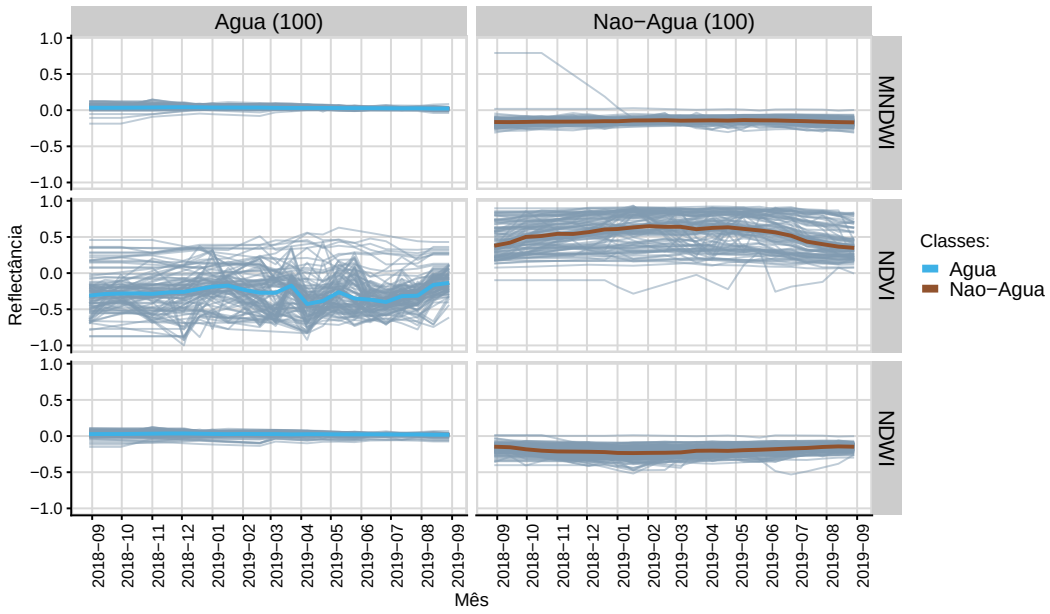
[1] http://www.dpi.inpe.br/tccerrado

5.2.2 Dados de entrada

Com base nos trabalhos mencionados anteriormente, foram utilizadas séries temporais extraídas dos cubos de dados do satélite Landsat 8 (Tabela 3.1), com as seguintes bandas: Banda *Coastal* (banda 1), bandas do espectro visível (azul, verde e vermelho), infravermelho próximo (NIR), as bandas do infravermelho curto (SWIR - banda 6 e banda 7), índices de vegetação EVI e NDVI. Além disso, foram gerados dois índices de água: NDWI e MNDWI (Tabela 3.2).

O período de estudo corresponde ao mês de setembro de 2018 até agosto de 2019. A Figura 5.2 apresenta as séries temporais das amostras coletadas neste estudo de caso, em que as linhas com maiores espessuras representam as medianas. Em detalhes, para os índices específicos de água, NDWI e MNDWI, valores maiores que 0 correspondem a alvos de água, e menor que zero não-água (GAO, 1996). No índice MNDWI observa-se uma série temporal com valor alto nas amostras de Não-Água, tal amostra pode ser de áreas inundadas em curto período de cheia ou região de fronteira em rio que se teve cheia. Já no índice NDVI essa variabilidade no comportamento espectro-temporal das amostras de Água pode ter ocorrido pelo fato das regiões arbóreas que ficam às margens dos rios.

Figura 5.2 - Séries temporais extraídas das amostras de uso e cobertura da terra utilizadas neste experimento.



Fonte: Próprio Autor.

5.3 Resultados e discussões

5.3.1 Atributos selecionados

Após a extração das séries temporais, procedeu-se à seleção dos atributos em cada grupo de métricas, a Figura 5.3 apresenta os resultados da seleção de atributos dos subconjuntos selecionados em cada grupo de métricas (Tabela C.1). A quantidade de atributos selecionados é destacada na Tabela 5.1.

Assim como no estudo de caso anterior, não houve indícios de que a quantidade de atributos por nó influenciaram em uma melhor acurácia global. O motivo da alta acurácia nos atributos selecionados, se deve ao fato do conjunto de dados possuir duas classes que são linearmente separáveis (Figura 5.2). Tal característica resultou em poucos atributos em cada subconjunto selecionado pelo GRRF.

Tabela 5.1 - Quantidade de atributos extraídos e selecionados em cada grupo de métricas. Os valores em parênteses correspondem ao desvio padrão da acurácia global.

Grupo de métricas	**Atributos extraídos**	**Atributos selecionados**	**Acurácia global média**
Métricas Básicas	165	2	99.3% (± 0.8)
Métricas Polares	110	1	98.9% (± 1.0)
Métricas Básicas e Polares	275	2	99.6% (± 0.7)

Fonte: Próprio Autor.

Em relação aos atributos selecionados, as métricas polares tiveram um padrão em todos os cenários retornados pelo método ótimo de Pareto, em que foi selecionada a métrica de área do quarto quartil para o índice NDWI; e para as métricas básicas, os dois atributos de água foram derivados do índice MNDWI. Por fim, para as métricas combinadas, foram selecionadas duas métricas básicas, sendo a média e o desvio padrão, da banda do infravermelho curto (banda 6).

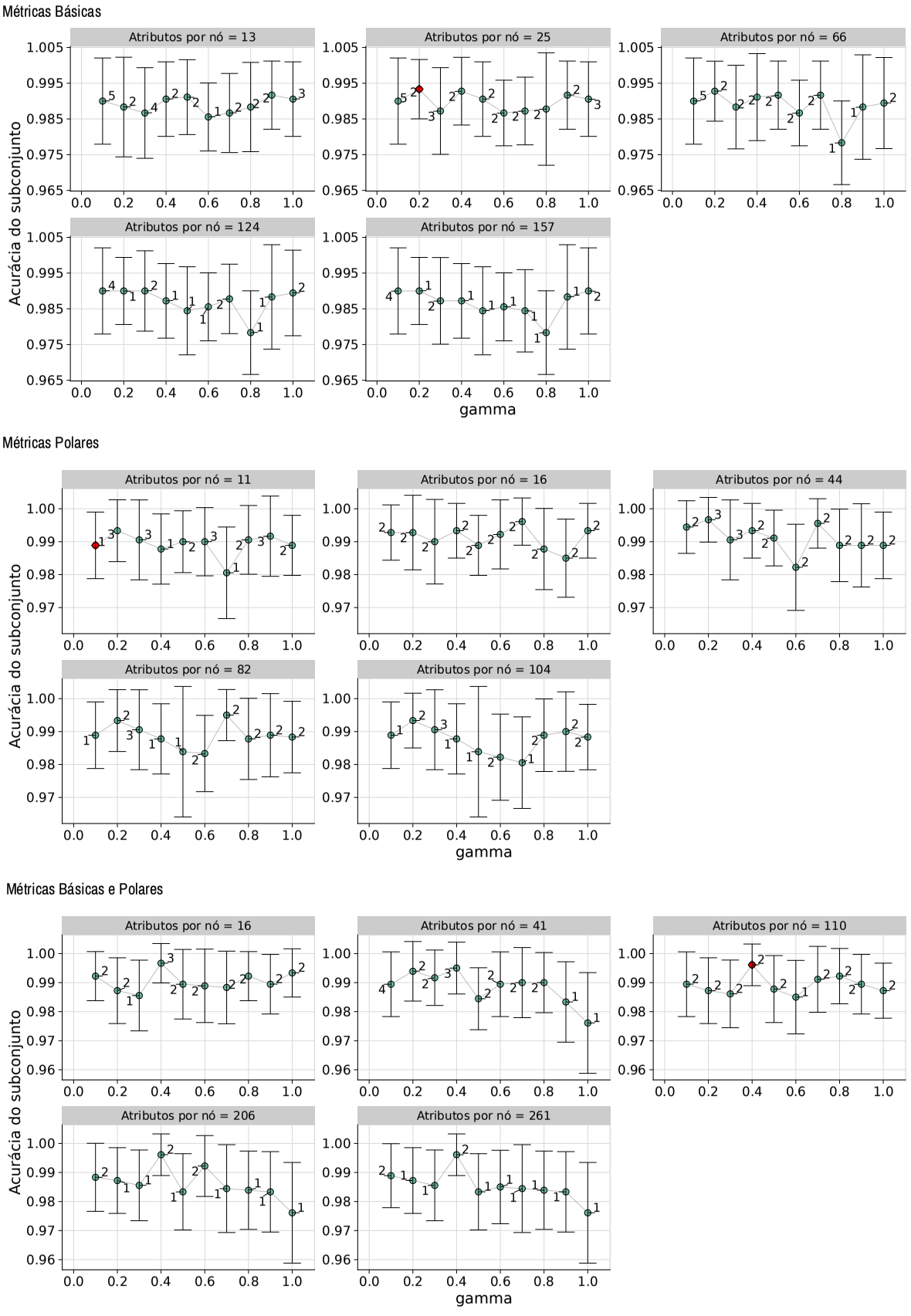


Figura 5.3 - Subconjuntos de atributos selecionados a partir do algoritmo GRRF com variações da quantidade de atributos por nó e taxa de penalização (γ).

Fonte: Próprio Autor

5.3.2 Modelos selecionados

Para a seleção dos algoritmos de classificação utilizou-se a técnica de validação cruzada *k-fold* (k = 5) a partir das séries temporais com o intuito de comparar o uso dos mesmos parâmetros com a classificação das métricas. Para o modelo RF foram diversificadas a quantidade de árvores entre 500, 1000 e 2000. Já no modelo SVM foram diversificados os seguintes *kernels*: radial, linear, polinomial e de sigmóide e o parâmetro de regularização, começando de 0.1 indo até 1, com o passo de 0.1.

Assim, pelo fato de ser uma aplicação binária, para o RF foi selecionado o modelo 500 árvores com 99% de acurácia global. Para o SVM, foi selecionado o *kernel* linear com o valor de regularização igual a 0.4 com 99% de acurácia global.

5.3.3 Desempenho e acurácia das classificações

A Tabela 5.2 apresenta os tempos de execução para cada classificação realizada neste estudo de caso, é possível observar que a classificação dos cubos de métricas reduzem o tempo de execução em ≈ 97%. O tempo de geração para os cubos de métricas básicas é de ≈ 6 *min* e polares de ≈ 36 *min*. Os experimentos foram executados em um servidor Linux Ubuntu 20.04, com 20 GB de memória e 10 núcleos.

Tabela 5.2 - Tempo da classificação em minutos das séries temporais e métricas avaliadas no estudo de caso da região de Minas Gerais. Os experimentos foram executados em um servidor Linux Ubuntu 20.04, com 20 GB de memória e 10 núcleos.

Modelo	**Séries Temporais**	**Métricas básicas**	**Métricas polares**	**Métricas básicas e polares**
RF	50.5 *min*	1.7 *min*	1.2 *min*	1.7 *min*
SVM	43.2 *min*	0.8 *min*	0.6 *min*	0.6 *min*

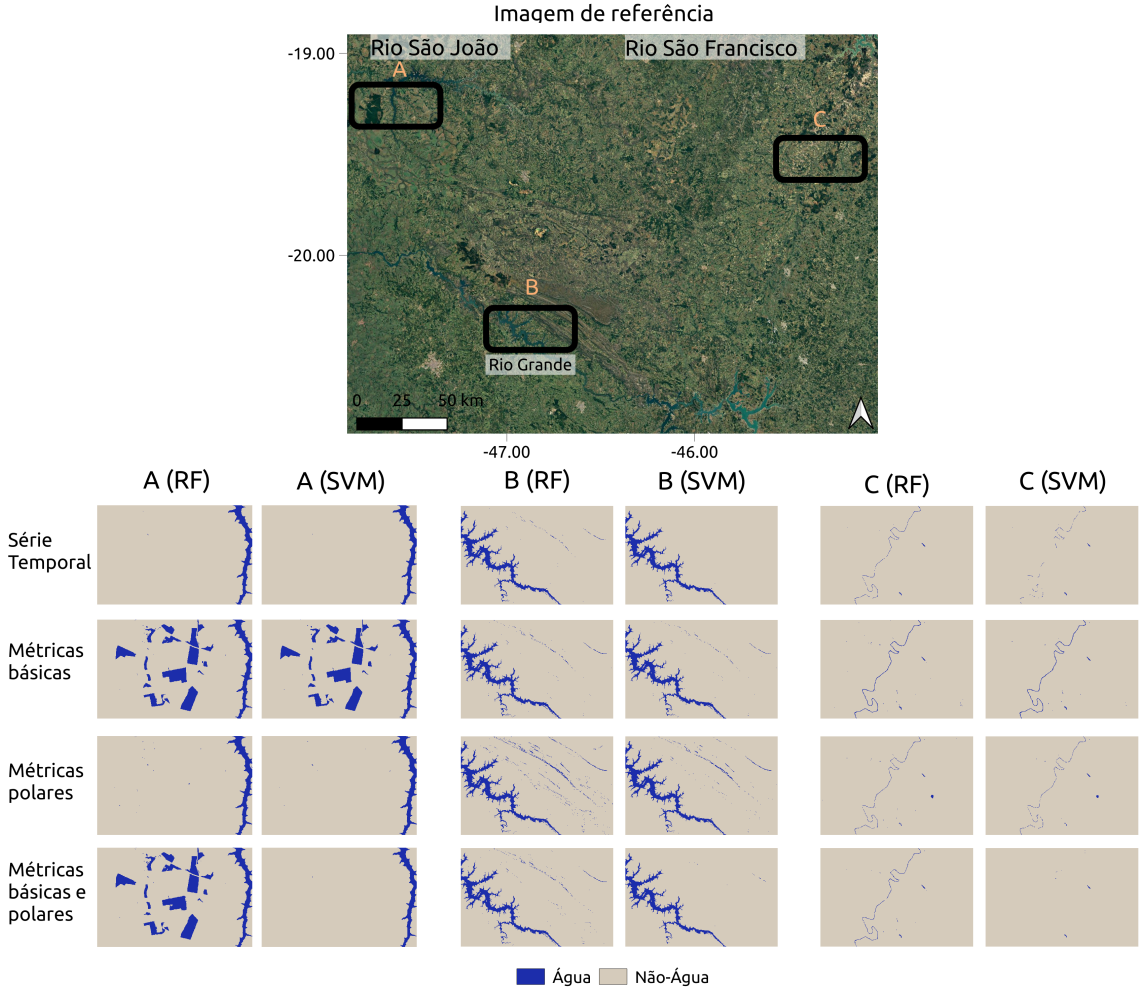
Fonte: Próprio Autor.

A Figura 5.4 apresenta os mapas de uso e cobertura da terra classificados com os modelos RF e SVM neste estudo de caso. Em geral, todos os classificadores detectaram os rios mais abundantes da região, entre eles: Rio São João, Rio Grande e o Rio São Francisco. No entanto, alguns mapas tiveram confusões em diversas regiões. Por exemplo, na região A, os mapas baseados em métricas básicas e na combinação de métricas básicas e polares classificaram a região de vegetação natural

como água. Tal confusão pode ter ocorrido por diversos motivos, por exemplo, caso seja uma vegetação úmida ou uma região em que houve uma inundação em determina data do ano. Ambos mapas que apresentaram essa confusão foram classificados a partir da métrica do segundo quartil, o que pode indicar um padrão na detecção de regiões de vegetação úmida ou inundada.

Em relação aos classificadores, observa-se que os mapas gerados a partir do SVM geraram classificações mais limpas, ou seja, sem ruídos ou efeitos *salt pepper*, nas regiões B e C. Porém, os mapas gerados pelo RF detectaram as curvas dos corpos d'água com mais precisão, o que pode ser facilmente observado no mapa de métricas básicas na região C.

Figura 5.4 - Máscaras d'água geradas neste estudo de caso.



Fonte: Próprio Autor.

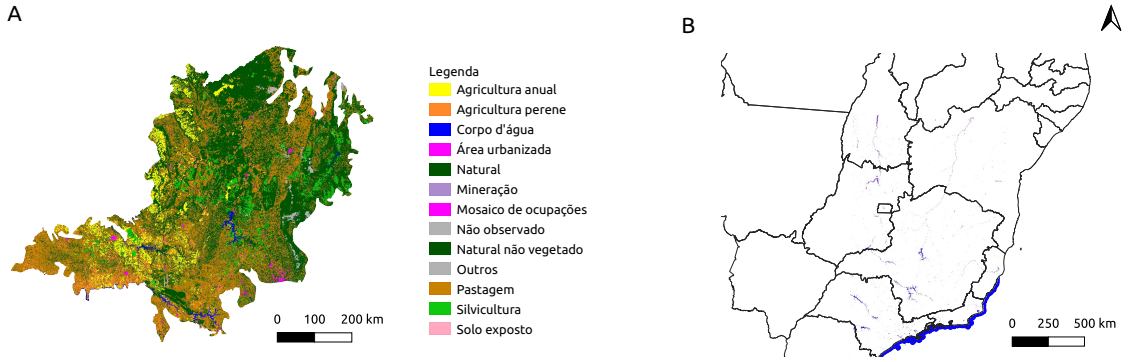
Outra região que apresentou intensa confusão espectral entre seus elementos encontra-se na região B da Figura 5.4, mais especificamente, entre o Rio Grande e o Parque Serra da Canastra, em que parte do relevo rochoso foi classificado como corpos d'água. A confusão pode ter ocorrido por três razões: a primeira se deve ao fato da região possuir uma alta altitude, visto que o pico mais elevado corresponde ao ponto dos chapadões, com altitude de 1.300 m, desta forma, o classificador pode ter confundido com as sombras formadas por essa região. A segunda baseia-se na possibilidade do classificador ter confundido com as cachoeiras da região, sendo elas: Cachoeira Zé Carlinho e a Cachoeira da Maria Concebida. Por fim, após uma análise mais detalhada com imagens de alta resolução, a confusão parece ocorrer devido à presença de sombras das regiões rochosas próximas à cachoeira da Posse.

Para validar os resultados gerados neste estudo de caso, utilizou-se como verdade de campo os mapas gerados pelo projeto TerraClass Cerrado[2] 2013 e a máscara d'água global gerada pelos autores Pekel et al. (2016). O projeto TerraClass Cerrado, criado pelo Instituto Nacional de Pesquisas Espaciais, visa produzir mapas de uso e cobertura da terra para o bioma do Cerrado, a partir de imagens do satélite Landsat-8. Já o trabalho de Pekel et al. (2016) foi criada uma máscara de água global, dos anos de 1985 até 2020, com diferentes níveis de ocorrência de água, variando entre 1% a 100%, em que 1% é a probabilidade mínima de ocorrer água no *pixel* analisado e 100% é a probabilidade máxima.

A Figura 5.5 apresenta os dois mapas supracitados: mapa do Projeto TerraClass Cerrado 2013 para o bioma do Cerrado (Figura 5.5A) e a máscara de água produzida pelos autores Pekel et al. (2016). Para a validação baseada no mapa do Cerrado foram coletados 18.887 pontos para cada classe, sorteando os pontos de forma aleatória no mapa. Em relação ao mapa de Pekel et al. (2016), foram coletados 17.846 pontos para cada classe, Água ou Não-água, também de forma aleatória. Para o mapa de Pekel et al. (2016) foi considerado o limiar de 50% de chance de ocorrer água em determinado ponto.

[2]http://www.dpi.inpe.br/tccerrado/

Figura 5.5 - Mapas de referência usados para validar as classificações geradas neste experimento. Em **A** Mapa TerraClass Cerrado 2013 e **B** Mapa global de ocorrência de corpos da água de Pekel et al. (2016).



Fonte: Próprio Autor.

A Tabela 5.3 apresenta os resultados de validação baseados nos mapas do projeto TerraClass Cerrado 2013 e de Pekel et al. (2016). Em geral, as métricas básicas e polares obtiveram resultados semelhantes ou melhores do que as séries temporais. Observa-se que as métricas polares atingiram as maiores acurácias em ambos os mapas validados, o que mostra a eficiência dessas métricas na identificação de corpos d'água. Em detalhes, as métricas polares atingiram 83% (RF) e 88% (RF) na acurácia do produtor para Água, com base nos mapas TerraClass e Pekel et al. (2016), respectivamente. No entanto, por ser uma classificação binária, é interessante analisar a acurácia do usuário para a classe oposta, na qual obteve 85% (RF) e 89% (RF) para Não-Água. O que mostra que não houve extrapolação de *pixels* para a classificação dos corpos d'água.

Em relação aos modelos de classificação, observa-se que o RF obteve as melhores acurácias do produtor para Água e do usuário para Não-Água na classificação com métricas. Por outro lado, as séries temporais obtiveram melhores resultados em todos os cenários com SVM, o que mostra que a seleção dos parâmetros foi eficaz para esse modelo.

Neste estudo de caso foi apresentado o uso de métricas temporais para a sumarização de séries temporais aplicadas na classificação de corpos d'água. Observou-se que as métricas polares apresentaram os melhores resultados em todos os cenários avaliados, o que responde à hipótese deste trabalho para aplicações que envolvem corpos d'água.

Tabela 5.3 - Matriz com as acurácias respectivas aos mapas classificados em relação aos mapas de referência para a área de estudo na região centro-oeste de MG.

			AP		**AU**		**AG**	
Ref	**Dados**	**Classes**	**RF**	**SVM**	**RF**	**SVM**	**RF**	**SVM**
Terra Class Cerrado	Séries Temporais	Água	82%	85%	99%	99%	88%	89%
		Não-Água	99%	99%	85%	87%		
	Métricas Básicas	Água	84%	84%	99%	99%	89%	89%
		Não-Água	99%	99%	86%	86%		
	Métricas Polares	Água	**88**%	**88**%	99%	99%	**91**%	**91**%
		Não-Água	99%	99%	**89**%	**89**%		
	Métricas Básicas e Polares	Água	81%	79%	99%	99%	87%	86%
		Não-Água	99%	99%	84%	82%		
Mapa Pekel	Séries Temporais	Água	76%	78%	99%	99%	91%	92%
		Não-Água	99%	99%	80%	82%		
	Métricas Básicas	Água	79%	78%	99%	99%	92%	92%
		Não-Água	99%	99%	82%	82%		
	Métricas Polares	Água	**83**%	**82**%	99%	99%	**94**%	**94**%
		Não-Água	99%	99%	**85**%	**84**%		
	Métricas Básicas e Polares	Água	76%	72%	99%	99%	90%	89%
		Não-Água	99%	99%	80%	78%		

Em que **AP** refere-se a acurácia do produtor; **AU** acurácia do usuário, e **AG** Acurácia Global.

Fonte: Próprio Autor.

6 ESTUDO DE CASO 3: MAPAS DE USO E COBERTURA

6.1 Contextualização

Informações sobre uso e cobertura da terra são essenciais para o monitoramento de biomas e para estimar emissões de gases de efeito estufa devido a conversões de uso da terra (GÓMEZ et al., 2016). Com a disponibilidade de grandes bases de imagens de satélites que recobrem periodicamente todo o Planeta de forma consistente, é possível gerar mapas de uso e cobertura da terra utilizando séries temporais.

Diversos trabalhos utilizam séries temporais de imagens de satélite para a classificação de mapas de uso e cobertura da terra. Por exemplo, em Simoes et al. (2020) foram gerados mapas de uso e cobertura da terra para o estado de Mato Grosso para os anos de 2001 até 2017 usando séries temporais MODIS. Em Picoli et al. (2020), séries temporais do satélite sino-brasileiro CBERS-4 foram usadas para gerar classificação de uso e cobertura da terra de áreas agrícolas no estado do Mato Grosso. Em Ferreira et al. (2020) foram geradas três classificações usando cubos de dados de diferentes satélites.

Neste estudo de caso, é realizado uma comparação entre as abordagens de séries temporais completas e métricas de séries temporais para gerar mapas de uso e cobertura da terra. Para isso, foram gerados mapas usando os algoritmos RF e SVM treinados usando as duas abordagens.

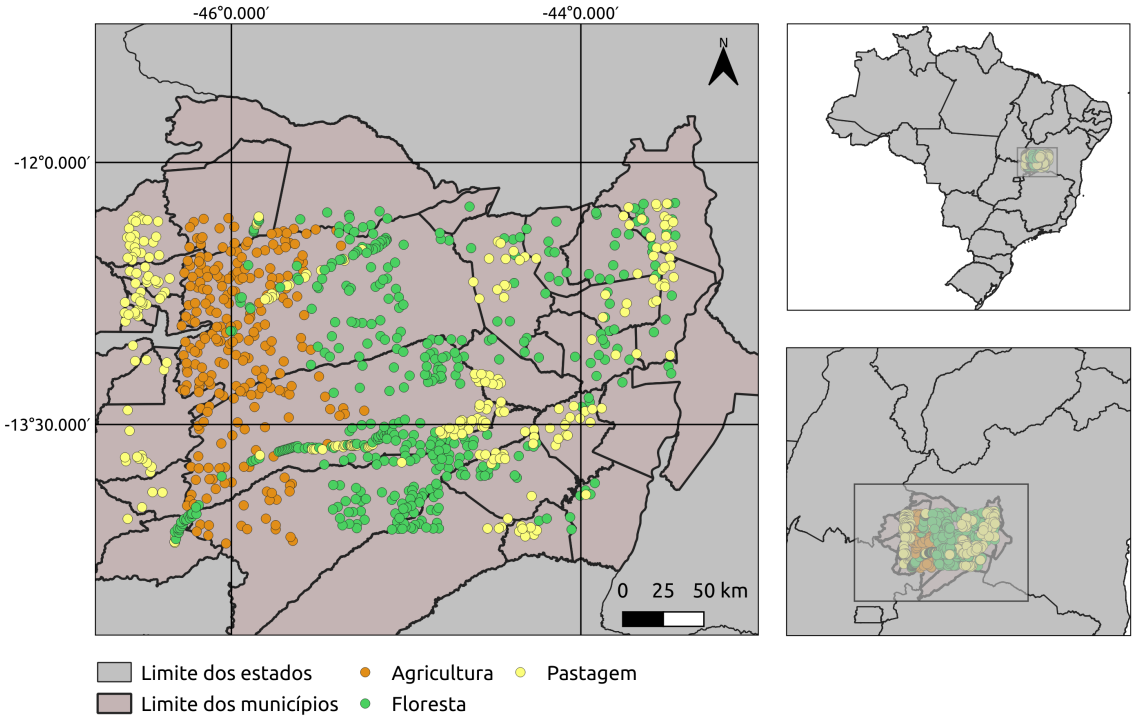
6.2 Materiais e métodos

6.2.1 Área de estudo

A área de estudo compreende a mesorregião do extremo oeste do estado da Bahia, conforme apresentado na Figura 6.1. A região oeste da Bahia é marcada por fortes atividades agrícolas, com uma intensa produção de grãos, dentre eles milho e soja, além de extensas áreas de pastagens para a criação de gado (SANO et al., 2011; BORGES; SANO, 2014).

A região abrange os biomas Cerrado e Caatinga apresentando períodos bem definidos de seca e precipitação. Cerca de 90% da precipitação é concentrada entre os meses de outubro a abril (FERREIRA et al., 2003). A estação de seca é marcada por uma deficiência hídrica (MACENA et al., 2008; GIROLAMO, 2018).

Figura 6.1 - Localização da área de estudo avaliada no estudo de caso do Oeste da Bahia.



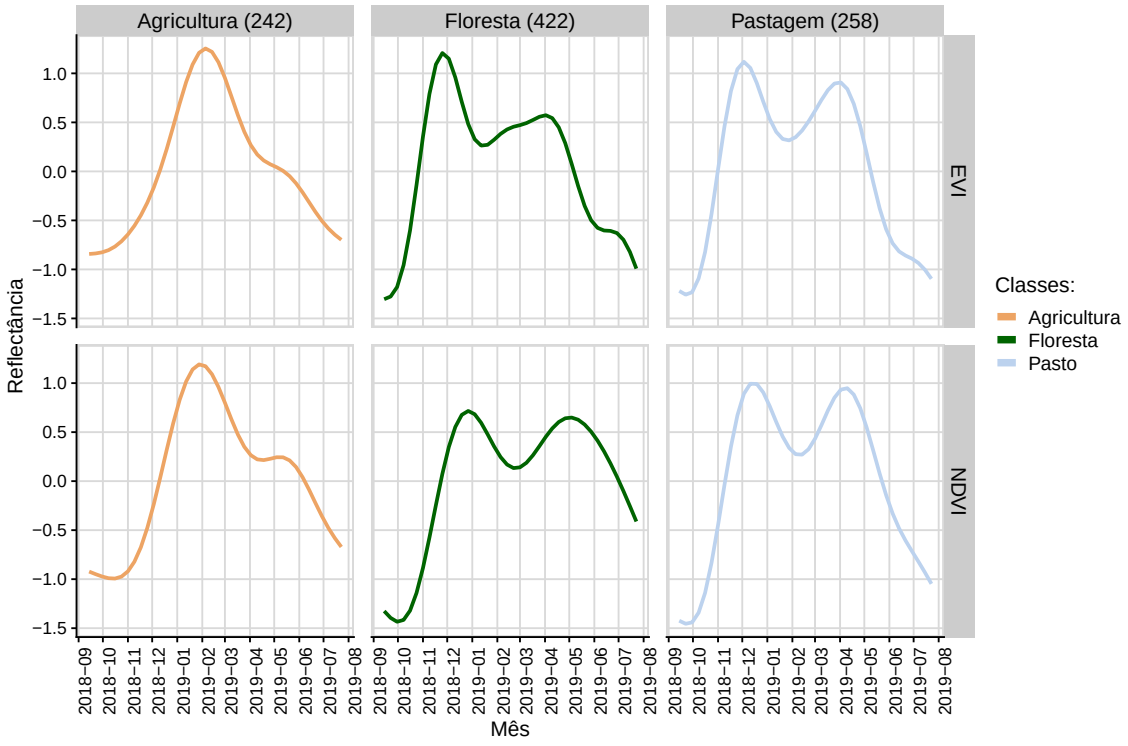
Fonte: Próprio Autor.

6.2.2 Dados de entrada

Para este estudo de caso, foi usado o cubo de dados CBERS-4 (BDC: CB4_64_-16D_STK-1) para gerar os mapas de uso e cobertura da terra. As séries temporais das bandas do espectro visível (blue, green e red), do infravermelho próximo (nir) e dos índices de vegetação EVI e NDVI foram extraídas do cubo de dados a partir das mesmas amostras utilizadas em Ferreira et al. (2020), que possuem as seguintes classes e quantidades: Agricultura (242 amostras), Floresta (422 amostras) e Pastagem (258 amostras).

A Figura 6.2 ilustra os padrões temporais das amostras usadas. Nota-se que o tipo de Agricultura é de único cultivo, diferente do que foi apresentado no estudo de caso do Capítulo 4. As quedas nos padrões temporais do NDVI das classes de Floresta e Pastagem estão associadas aos períodos de estiagem para essa época do ano. As séries temporais completas podem ser visualizadas na Figura C.1 dos apêndices.

Figura 6.2 - Padrões temporais das amostras utilizadas neste experimento localizadas no Oeste da Bahia.



Fonte: Próprio Autor.

Conforme o sistema de classificação multinível de cobertura e uso da terra descrito pelo INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA - IBGE (2013), as classes Agricultura, Cana-de-açúcar e Pastagem pertencem à categoria de áreas antrópicas agrícolas, e a classe Floresta à categoria de áreas de vegetação natural.

A categoria de áreas antrópicas agrícolas compreende áreas de uso para a produção de alimentos, fibras ou outras matérias-primas, que podem ser utilizadas no setor industrial, por exemplo, na produção de etanol através da cana-de-açúcar, ou na pecuária através do cultivo de pastagem, destinadas ao pastoreio de gado. Por outro lado, as áreas de vegetação natural abrangem florestas, campos originais, florestas secundárias, arbustivas, herbáceas e/ou gramíneo-lenhosas. É considerado floresta as formações arbóreas com porte superior a 5 m. De outro modo, é considerado como áreas campestres regiões que se caracterizam por um estrato predominantemente arbustivo (INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA - IBGE, 2013).

Para validar os mapas gerados foram utilizadas 1.258 amostras coletadas em Ferreira et al. (2020). As amostras de validação foram extraídas do produto PRODES Cerrado, ano 2018, que mapeia áreas antropisadas como máscara de incrementos, isto é, uma vez identificada uma área antropisada, ela entra na máscara do PRODES. As amostras extraídas a partir do mapa do PRODES foram: Antrópico (663 amostras) e Vegetação Natural (595 amostras). Para realizar a validação, as classes de Agricultura e Pastagem dos mapas gerados foram convertidas para Antrópico e a classe de Floresta foi convertida para Vegetação Natural.

6.3 Resultados e discussões

6.3.1 Atributos selecionados

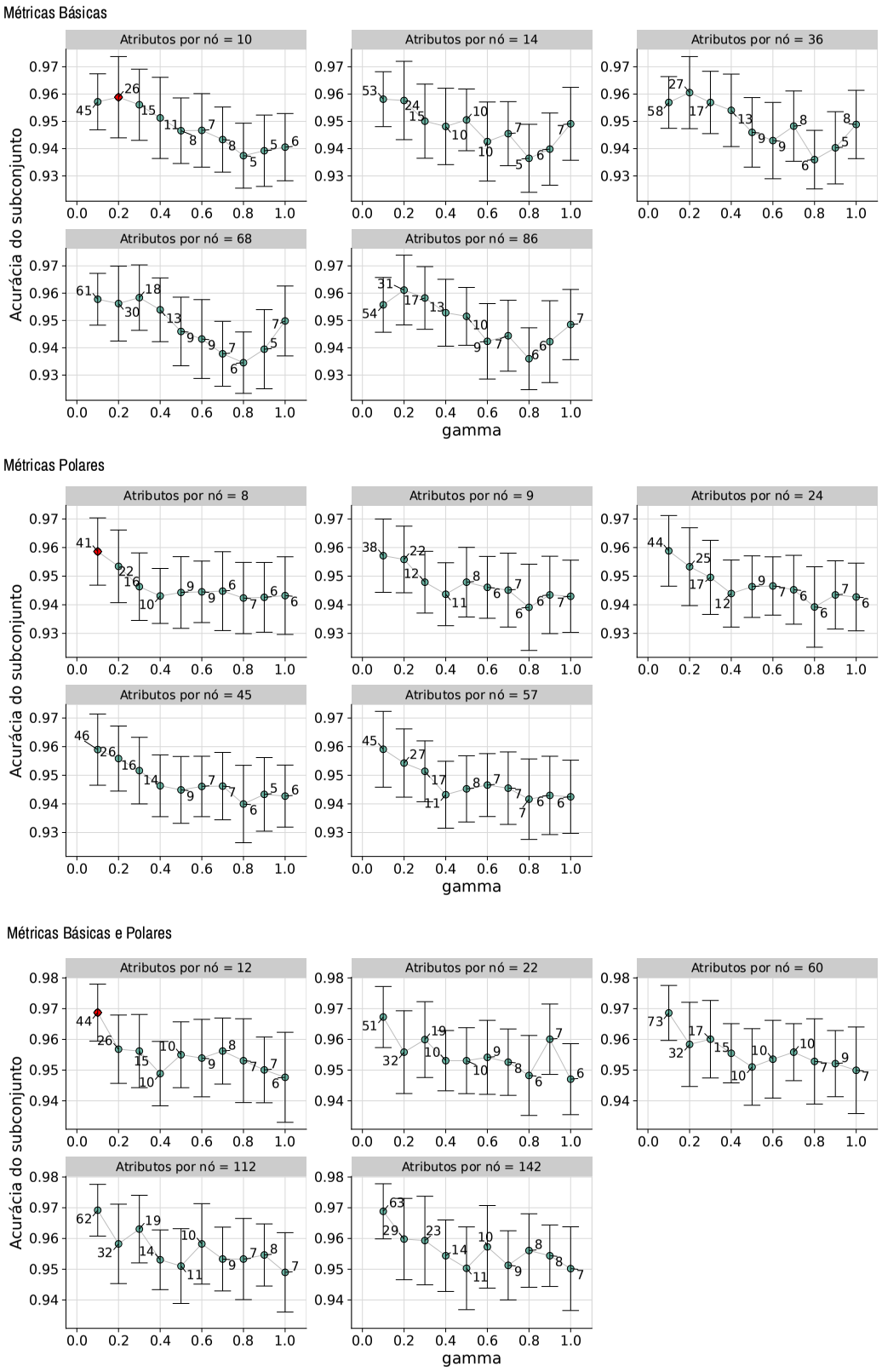
A geração e seleção de atributos foi realizada de acordo com os procedimentos descritos no Capítulo 3. A Figura 6.3 apresenta os resultados da seleção de atributos dos subconjuntos. De cada grupo de métricas (Tabela C.1) selecionou-se o melhor subconjunto de atributos usando o método ótimo de Pareto que minimiza a quantidade de atributos e maximiza a acurácia média (IZQUIERDO-VERDIGUIER; ZURITA-MILLA, 2020). Assim como nos outros estudos de caso, o parâmetro γ domina o parâmetro *mtry* cujo efeito na acurácia não é percebida. As quantidades de atributos selecionados para cada grupo de métricas são listadas na Tabela 6.1.

Tabela 6.1 - Quantidade de atributos extraídos e selecionados para cada grupo de métricas. Os valores em parênteses correspondem ao desvio padrão da acurácia global.

Grupo de métricas	**Atributos extraídos**	**Atributos selecionados**	**Acurácia Global Média**
Métricas Básicas	90	26	95.9% (± 1.4)
Métricas Polares	60	41	96.9% (± 1.1)
Métricas Básicas e Polares	150	44	99.6% (± 0.9)

Fonte: Próprio Autor.

Observa-se na Figura 6.3 que um maior valor do parâmetro γ tende a diminuir a quantidade de atributos nos subconjuntos selecionados e, consequentemente, reduz a acurácia média dos experimentos. Os subconjuntos selecionados pelo ótimo de Pareto tiveram o parâmetro $\gamma \leq 0.2$. Isso evidencia que os ganhos do índice GINI calculados no crescimento da árvore de decisão no método GRRF foram satisfatórios e as reduções das quantidades de atributos selecionados que apresentaram alta acurácia média se deu para valores abaixo de γ.



Fonte: Próprio Autor.

Figura 6.3 - Subconjuntos de atributos selecionados a partir do algoritmo GRRF com variações da quantidade de atributos divididos por nó e taxa de penalização (γ) para a região de estudo do Oeste da Bahia.

6.3.2 Modelos selecionados

Para determinar os parâmetros dos modelos de aprendizado de máquina foi definido um espaço de parâmetros para os algoritmos RF e SVM. Para o RF, variou-se a quantidade de árvores em 500, 1000 e 2000. Para o SVM, foram avaliados os *kernels* radial, linear, polinomial e sigmóide, bem como o parâmetro de regularização, iniciando em 0.1 indo até 1, com o passo de 0.1. Para determinar os parâmetros ótimos dos dois algoritmos, procedeu-se uma validação cruzada *k-fold* ($k = 5$) onde, em cinco rodadas, uma parte em cinco das amostras são usadas para teste e o restante para treinamento.

Desta forma, foram selecionados os modelos RF com 2000 árvores (acurácia global de 97.4%) e SVM com *kernel* radial e o valor de regularização de 0.9 (acurácia global 96.2%). Utilizando os mesmos parâmetros, foram treinados dois modelos (RF e SVM) para cada subconjunto de métricas selecionadas e para as séries temporais completas, totalizando assim 8 classificadores que foram usados para gerar mapas de uso e cobertura da terra.

6.3.3 Desempenho e acurácia das classificações

Os tempos de processamento dos mapas gerados pelos modelos treinados estão apresentados na Tabela 6.2. Observa-se que os tempos de execução das métricas foram consideravelmente menores, com uma redução de $\approx$ 80% no tempo de processamento, no modelo SVM treinado com métricas básicas. O tempo de geração para os cubos de métricas básicas é de ≈ 6 *min* e polares de ≈ 36 *min*. Os experimentos foram executados em um servidor Linux Ubuntu 20.04, com 40 GB de memória e 20 núcleos.

Percebe-se que os tempos de processamento são proporcionais ao número de atributos necessários para a classificação. Aqui, as séries temporais apresentaram o maior tempo por conter o maior espaço de atributos dentre os quatro conjuntos de amostras.

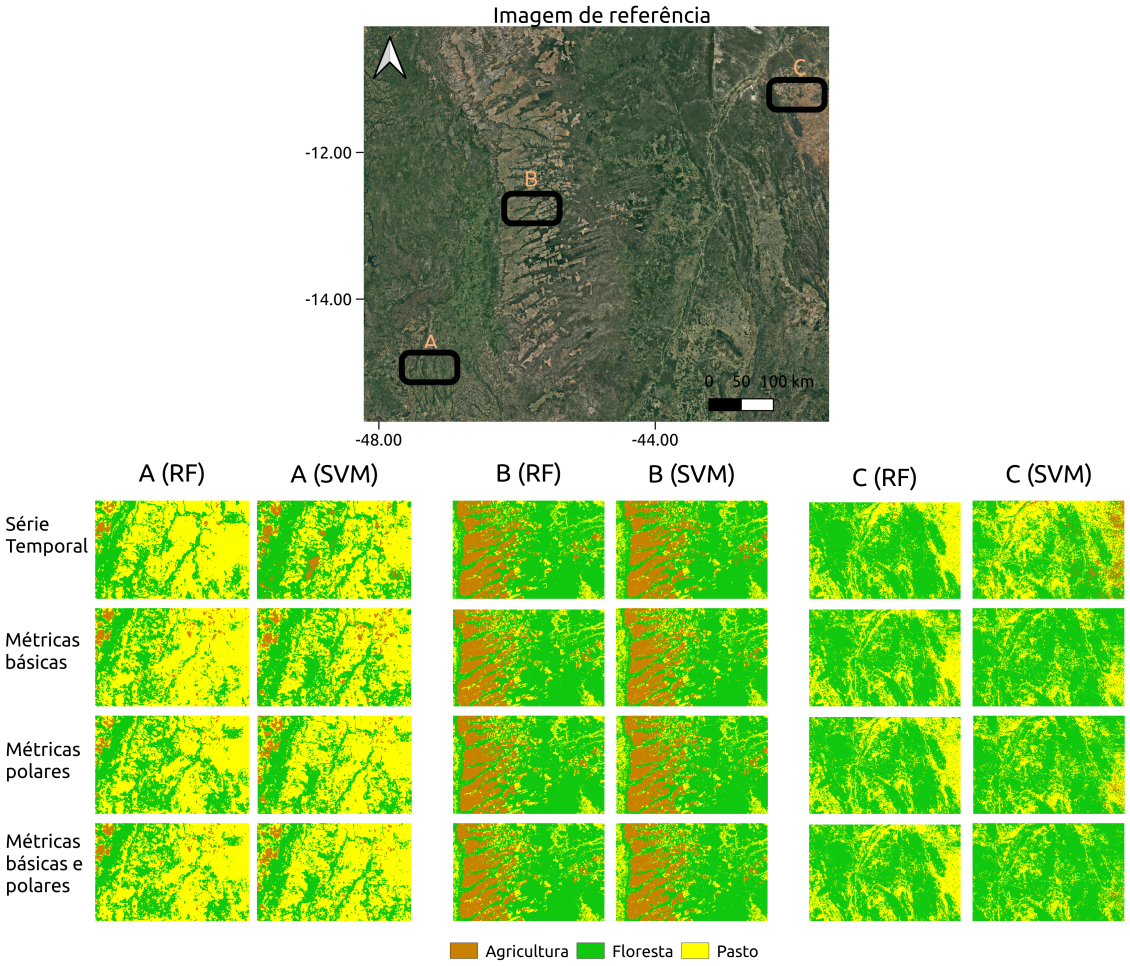
A Figura 6.4 apresenta os mapas de uso e cobertura da terra gerados pelos modelos treinados com as séries temporais completas e com os subconjuntos de métricas. Observa-se que, no geral, os classificadores delimitam de forma consistente os talhões agrícolas, localizados ao centro da imagem (detalhe B). Os erros de classificação notáveis são apresentados no detalhe A (SVM).

Tabela 6.2 - Tempo de classificação em minutos das séries temporais e métricas avaliadas no estudo de caso da região de Bahia. Os experimentos foram executados em um servidor Linux Ubuntu 20.04, com 40 GB de memória e 20 núcleos.

Modelo	Séries Temporais	Métricas básicas	Métricas polares	Métricas básicas e polares
RF	13.6 *min*	3.5 *min*	5.2 *min*	6.5 *min*
SVM	13.2 *min*	2.6 *min*	3.7 *min*	4.8 *min*

Fonte: Próprio Autor.

Figura 6.4 - Mapas de uso e cobertura da terra classificados neste experimento.



Fonte: Próprio Autor.

Os mapas classificados por métricas produziram menos regiões de pastagens do que os mapas classificados com séries temporais, tal característica pode ser observada na região C.

A Tabela 6.3 apresenta os resultados obtidos a partir das amostras de validação dos mapas de uso e cobertura da terra. Nota-se que a acurácia global das classificações ficaram semelhantes, o que pôde ser notado através de uma análise visual. No entanto, a maior acurácia global (AG) obtida foi através da combinação das métricas básicas e polares com 88.2%, seguida da classificação pelo modelo SVM no mapa de séries temporais, com 88.0%.

Tabela 6.3 - Matriz com as acurácias respectivas aos mapas classificados na área de estudo no Oeste da Bahia.

			AP		**AU**		**AG**	
Ref	**Dados**	**Classes**	**RF**	**SVM**	**RF**	**SVM**	**RF**	**SVM**
PRODES Cerrado	Séries Temporais	Antrópico	82%	83.7%	92.2%	**92.9**%	86.9%	**88**%
		Vegetação Natural	92.2%	**92.9**%	82.35%	83.6%		
	Métricas Básicas	Antrópico	85.8%	83.2%	90%	91.2%	87.6%	86.9%
		Vegetação Natural	89.7%	91%	85%	83%		
	Métricas Polares	Antrópico	84.9%	83.7%	91.5%	90%	**87.9**%	86.7%
		Vegetação Natural	91.2%	90%	84.4%	83.2%		
	Métricas Básicas e Polares	Antrópico	**86.1**%	83.1%	91%	89%	**88.2**%	85.6%
		Vegetação Natural	90.5%	88.5%	**85.4**%	82.4%		

Em que **AP** refere-se a acurácia do produtor; **AU** acurácia do usuário, e **AG** Acurácia Global.

Fonte: Próprio Autor.

Em relação às amostras, atingiu-se 86.1% de acurácia do produtor (AP) para as amostras de Antrópico através da combinação de métricas, e 92.9 para as amostras de Vegetação Natural com o uso de séries temporais. Na acurácia do usuário atingiu-se 85.4% para as amostras de Vegetação Natural na classificação de métricas combinadas, o que apresenta mais especificidade para essa classe. Por fim, para Antrópico atingiu-se 92.6% de acurácia do usuário para a classificação com séries temporais, o que reflete na maior generalização de áreas de pastagens apresentadas na Figura 6.4.

7 CONSIDERAÇÕES FINAIS

Neste trabalho, apresentou-se uma avaliação do uso de métricas extraídas de séries temporais de uso e cobertura da terra em aplicações de aprendizado de máquina. A metodologia aplicada para avaliar as métricas é composta pelas seguintes etapas: extração e interpolação de séries temporais, seleção de métricas temporais e aplicação de modelos de aprendizado de máquina. Com isso, foram efetuados três estudos de casos, sendo eles: avaliação de amostras de uso e cobertura da terra a partir de técnicas de agrupamento, criação de máscara de água e geração de mapas de uso e cobertura da terra. Nos três estudos de casos foi aplicado o método de seleção de atributos GRRF para diminuir o espaço dimensional. Desta forma, em todos os estudos de caso foram comparados os resultados obtidos pelas séries temporais e pelas métricas extraídas das séries temporais. Por fim, para evitar qualquer viés espacial nos estudos de casos, foram realizados três estudos de casos em diferentes regiões e cubos de dados.

No primeiro estudo de caso, observou-se que em todos os agrupamentos gerados por métricas obteve-se melhores resultados se comparados com séries temporais. Além de produzir grupos mais homogêneos, as métricas produziram menos neurônios com amostras *outliers*, o que evidenciou que a abordagem de métricas de séries temporais foi capaz de promover uma melhor separabilidade de amostras de uso e cobertura de terra avaliadas neste estudo de caso.

No segundo estudo de caso, as métricas polares produziram os melhores resultados em todos os cenários avaliados, nos quais se obteve 91% de acurácia global para a validação baseada no mapa do projeto TerraClass Cerrado e 94% na validação baseada no mapa de Pekel et al. (2016). É interessante ressaltar que a classificação baseada em métricas polares possui apenas uma métrica, sendo ela a área do quarto quartil para o índice de água NDWI (Tabela B.1). Tal informação pode ser útil para futuras aplicações no contexto do projeto do BDC. Além de produzir os melhores resultados, o tempo de processamento é um fator determinante para o uso desta abordagem, visto que, o tempo de processamento para o modelo SVM foi 54× mais rápido em comparação com o tempo das séries temporais e 30× mais rápido no RF.

Por fim, no último experimento, observou-se que as métricas atingiram resultados semelhantes ou um pouco melhores que as séries temporais para o modelo RF. As métricas básicas apresentaram melhores resultados para o modelo RF com o tempo de processamento quase 4× mais eficaz. Outro detalhe interessante são as métricas

selecionadas pelo GRRF para o grupo de métricas básicas e polares, que forneceram 86% de acurácia do produtor para o modelo RF.

Desta forma, com base nos resultados desta dissertação, conclui-se que o uso da abordagem de métricas de séries temporais é igual ou superior à alternativa de usar todas as instâncias temporais por conta do desempenho computacional e da redução de dimensionalidade espacial. Além disso, o método de seleção de atributos GRRF foi satisfatório, pois permitiu a redução do conjunto de métricas consideradas para o treinamento dos modelos de ML, o que certamente contribuiu para obter mapas de uso e cobertura mais precisos.

Pretende-se utilizar as informações de métricas selecionadas neste trabalho para a produção de cubos de métricas para diferentes regiões do Brasil, tendo em vista que no projeto BDC, estão sendo geradas classificações para todo o território nacional e manipulando um grande volume de dados de observação da Terra. Portanto, o uso de métricas se apresenta uma excelente alternativa para diminuir o custo computacional dessas classificações.

7.1 Trabalhos futuros

Como trabalhos futuros espera-se adicionar mais métricas na metodologia adotada neste trabalho, visto que diferentes métricas podem resultar em diferentes tipos de informações que podem ser extraídas nos alvos de uso e cobertura da terra. Além disso, pretende-se avaliar outros tipos de modelos de classificação, principalmente, modelo de Redes Neurais Profundas e Redes Neurais Convolucionais, pois ambas conseguem extrair mais detalhes na classificação.

Em relação aos cubos de métricas, pretende-se melhorar o tempo de geração dos cubos de métricas para otimizar a geração de classificações em escalas regionais. Com isso, pretende-se avaliar outras técnicas de seleção de atributos para obter mais informações sobre diferentes classes em diferentes aplicações.

REFERÊNCIAS BIBLIOGRÁFICAS

AGHABOZORGI, S.; SHIRKHORSHIDI, A. S.; WAH, T. Y. Time-series clustering–a decade review. **Information Systems**, v. 53, p. 16–38, 2015. 2, 3, 15, 16, 17, 19, 25

ANJOS, C. S. dos; LACERDA, M. G.; ANDRADE, L. do L.; SALLES, R. N. Classification of urban environments using feature extraction and random forest. In: INTERNATIONAL GEOSCIENCE AND REMOTE SENSING SYMPOSIUM (IGARSS). **Proceedings...** USA: IEEE, 2017. p. 1205–1208. 13

APPEL, M.; PEBESMA, E. On-demand processing of data cubes from satellite image collections with the gdalcubes library. **Data**, v. 4, n. 3, p. 92, 2019. 2, 7

BORGES, E. F.; SANO, E. E. Caracterização fenológica da cobertura vegetal do oeste da bahia a partir de séries temporais de evi do sensor modis. **Revista Brasileira de Cartografia**, v. 66, n. 6, 2014. 65

BOX, G. E.; MEYER, R. D. An analysis for unreplicated fractional factorials. **Technometrics**, v. 28, n. 1, p. 11–18, 1986. 38

BREIMAN, L. Random forests. **Machine Learning**, v. 45, n. 1, p. 5–32, 2001. 13

BURKOM, H. S.; MURPHY, S. P.; SHMUELI, G. Automated time series forecasting for biosurveillance. **Statistics in Medicine**, v. 26, n. 22, p. 4202–4218, 2007. 9

CHAN, J. C.-W.; PAELINCKX, D. Evaluation of random forest and adaboost tree-based ensemble classification and spectral band selection for ecotope mapping using airborne hyperspectral imagery. **Remote Sensing of Environment**, v. 112, n. 6, p. 2999–3011, 2008. 13

CHAVES, M. E.; PICOLI, M. C.; SANCHES, I. D. Recent applications of landsat 8/oli and sentinel-2/msi for land use and land cover mapping: a systematic review. **Remote Sensing**, v. 12, n. 18, p. 3062, 2020. 12, 43

COPPIN, P.; JONCKHEERE, I.; NACKAERTS, K.; MUYS, B.; LAMBIN, E. Review articledigital change detection methods in ecosystem monitoring: a review. **International Journal of Remote Sensing**, v. 25, n. 9, p. 1565–1596, 2004. Disponível em: <https://doi.org/10.1080/0143116031000101675>. 1

CORTES, C.; VAPNIK, V. Support-vector networks. **Machine Learning**, v. 20, n. 3, p. 273–297, 1995. 15

DENG, H. Guided random forest in the rrf package. **arXiv preprint arXiv:1306.0237**, 2013. 29

DENG, H.; RUNGER, G. Gene selection with guided regularized random forest. **Pattern Recognition**, v. 46, n. 12, p. 3483–3489, 2013. 29

EDSALL, R. M.; KRAAK, M.-J.; MACEACHREN, A. M.; PEUQUET, D. J. Assessing the effectiveness of temporal legends in environmental visualization. In: GIS/LIS ANNUAL MEETING, 1997. **Proceedings...** Ohio, USA, 1997. v. 97, p. 28–30. 28

EPIPHANIO, R. D. V.; FORMAGGIO, A. R.; RUDORFF, B. F. T.; MAEDA, E. E.; LUIZ, A. J. B. Estimating soybean crop areas using spectral-temporal surfaces derived from modis images in Mato Grosso, Brazil. **Pesquisa Agropecuária Brasileira**, v. 45, n. 1, p. 72–80, 2010. 2, 10

ESLING, P.; AGON, C. Time-series data mining. **ACM Computing Surveys (CSUR)**, v. 45, n. 1, p. 1–34, 2012. 9, 16

FACELI, K. et al. **Validação de algoritmos de agrupamento.** São Carlos, SP, Brasil: USP, 2005. 23, 24, 46

FENG, M.; SEXTON, J. O.; CHANNAN, S.; TOWNSHEND, J. R. A global, high-resolution (30-m) inland water body dataset for 2000: first results of a topographic–spectral classification algorithm. **International Journal of Digital Earth**, v. 9, n. 2, p. 113–133, 2016. 55

FERREIRA, K. R. et al. Earth observation data cubes for Brazil: requirements, methodology and products. **Remote Sensing**, v. 12, n. 24, p. 4033, 2020. 2, 8, 65, 66, 68

FERREIRA, L.; YOSHIOKA, H.; HUETE, A.; SANO, E. Seasonal landscape and spectral vegetation index dynamics in the brazilian cerrado: an analysis within the large-scale biosphere–atmosphere experiment in Amazônia (lba). **Remote Sensing of Environment**, v. 87, n. 4, p. 534–550, 2003. 65

FOLEY, J. A. et al. Global consequences of land use. **Science**, v. 309, n. 5734, p. 570–574, 2005. 1

FOOD AND AGRICULTURAL ORGANIZATION - FAO. **FAOSTAT**. 2021. Disponível em: <http://www.fao.org/faostat/en/#data/QC/visualize>. 1

FRANKLIN, S. E.; AHMED, O. S.; WULDER, M. A.; WHITE, J. C.; HERMOSILLA, T.; COOPS, N. C. Large area mapping of annual land cover dynamics using multitemporal change detection and classification of landsat time series data. **Canadian Journal of Remote Sensing**, v. 41, n. 4, p. 293–314, 2015. 2, 10

FRÉNAY, B.; VERLEYSEN, M. Classification in the presence of label noise: a survey. **IEEE Transactions on Neural Networks and Learning Systems**, v. 25, n. 5, p. 845–869, 2013. 41

GAO, B.-C. Ndwi—a normalized difference water index for remote sensing of vegetation liquid water from space. **Remote Sensing of Environment**, v. 58, n. 3, p. 257–266, 1996. 32, 34, 57

GIROLAMO, C. D. N. **Identificação de fitofisionomias de cerrado no Parque Nacional de Brasília utilizando Random Forest aplicado a imagens de alta e média resoluções espaciais.** Tese (Doutorado em Sensoriamento Remoto) — Instituto Nacional de Pesquisas Espaciais (INPE), São José dos Campos, 2018. 65

GITELSON, A. A.; MERZLYAK, M. N.; LICHTENTHALER, H. K. Detection of red edge position and chlorophyll content by reflectance measurements near 700 nm. **Journal of Plant Physiology**, v. 148, n. 3-4, p. 501–508, 1996. 32, 34

GIULIANI, G.; CHATENOUX, B.; BONO, A. D.; RODILA, D.; RICHARD, J.-P.; ALLENBACH, K.; DAO, H.; PEDUZZI, P. Building an earth observations data cube: lessons learned from the swiss data cube (sdc) on generating analysis ready data (ard). **Big Earth Data**, v. 1, n. 1-2, p. 100–117, 2017. 2, 7, 8

GOMES, V. C.; QUEIROZ, G. R.; FERREIRA, K. R. An overview of platforms for big earth observation data management and analysis. **Remote Sensing**, v. 12, n. 8, p. 1253, 2020. 9

GÓMEZ, C.; WHITE, J. C.; WULDER, M. A. Optical remotely sensed time series data for land cover classification: a review. **ISPRS Journal of Photogrammetry and Remote Sensing**, v. 116, p. 55–72, 2016. 1, 2, 10, 65

GRIFFITHS, P.; KUEMMERLE, T.; BAUMANN, M.; RADELOFF, V. C.; ABRUDAN, I. V.; LIESKOVSKY, J.; MUNTEANU, C.; OSTAPOWICZ, K.;

HOSTERT, P. Forest disturbances, forest recovery, and changes in forest types across the carpathian ecoregion from 1985 to 2010 based on landsat image composites. **Remote Sensing of Environment**, v. 151, p. 72–88, 2014. 2, 10

GUSSO, A.; ADAMI, M.; SILVA, W. F. da; AGUIAR, D. A. de; RUDORFF, B. F. T. Aplicação de séries temporais evi/modis na identificação do uso e ocupação do solo anterior ao cultivo da cana-de-açúcar. In: SIMPÓSIO BRASILEIRO DE SENSORIAMENTO REMOTO, 14., 2009, NATAL - RN. **Anais...** São José dos Campos: INPE, 2009. p. 5851–5856. 50

HALKIDI, M.; BATISTAKIS, Y.; VAZIRGIANNIS, M. On clustering validation techniques. **Journal of Intelligent Information Systems**, v. 17, n. 2-3, p. 107–145, 2001. 19

HAN, J.; PEI, J.; KAMBER, M. **Data mining: concepts and techniques**. [S.l.]: Elsevier, 2011. 10, 12, 16, 18, 19, 20

HASTIE, T.; TIBSHIRANI, R.; J., F. **The elements of statistical learning. Data mining, inference, and prediction**. New York: Springer, 2009. 14, 15

HAYKIN, S. **Neural networks and learning machines, 3/E**. [S.l.]: Pearson Education India, 2010. 18

HUBERT, L.; ARABIE, P. Comparing partitions. **Journal of Classification**, v. 2, n. 1, p. 193–218, 1985. 21

HUETE, A.; JUSTICE, C.; LEEUWEN, W. V. **Modis vegetation index (MOD13): algorithm theoretical basis document.** Greenbelt: NASA Goddard Space Flight Center, 1999. 32, 34

HÜTTICH, C.; GESSNER, U.; HEROLD, M.; STROHBACH, B. J.; SCHMIDT, M.; KEIL, M.; DECH, S. On the suitability of modis time series metrics to map vegetation types in dry savanna ecosystems: a case study in the Kalahari of NE Namibia. **Remote Sensing**, v. 1, n. 4, p. 620–643, 2009. 2

INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA - IBGE. **Manual técnico de uso da terra**. [S.l.]: IBGE, 2013. 43, 67

IZQUIERDO-VERDIGUIER, E.; ZURITA-MILLA, R. An evaluation of guided regularized random forest for classification and regression tasks in remote sensing. **International Journal of Applied Earth Observation and Geoinformation**, v. 88, p. 102051, 2020. 29, 30, 38, 68

JAIN, A. K.; DUBES, R. C. **Algorithms for clustering data**. [S.l.]: Prentice-Hall, 1988. 21

JAIN, A. K.; DUIN, R. P. W.; MAO, J. Statistical pattern recognition: a review. **IEEE Transactions on Pattern Analysis and Machine Intelligence**, v. 22, n. 1, p. 4–37, 2000. 18

JAMES, G.; WITTEN, D.; HASTIE, T.; TIBSHIRANI, R. **An introduction to statistical learning: with applications in R**. New York, EUA: Springer, 2013. 441 p. p. 15

KEOGH, E.; LIN, J. Clustering of time-series subsequences is meaningless: implications for previous and future research. **Knowledge and Information Systems**, v. 8, n. 2, p. 154–177, 2005. 17

KHATAMI, R.; MOUNTRAKIS, G.; STEHMAN, S. V. A meta-analysis of remote sensing research on supervised pixel-based land-cover image classification processes: general guidelines for practitioners and future research. **Remote Sensing of Environment**, v. 177, p. 89–100, 2016. 3

KILLOUGH, B. The impact of analysis ready data in the Africa Regional Data Cube. In: INTERNATIONAL GEOSCIENCE AND REMOTE SENSING SYMPOSIUM, 2019, Yokohama, Japan. **Proceedings...** [S.l.]: IEEE, 2019. p. 5646–5649. ISBN 978-1-5386-9154-0. 2

KOHONEN, T. Self-organized formation of topologically correct feature maps. **Biological Cybernetics**, v. 43, n. 1, p. 59–69, 1982. 18

_____. Essentials of the self-organizing map. **Neural Networks**, v. 37, p. 52–65, 2013. 19

KORN, F.; PAGEL, B.-U.; FALOUTSOS, C. On the"dimensionality curse"and the"self-similarity blessing". **IEEE Transactions on Knowledge and Data Engineering**, v. 13, n. 1, p. 96–111, 2001. 25

KÖRTING, T. S.; FONSECA, L. M. G.; CÂMARA, G. Geodma—geographic data mining analyst. **Computers & Geosciences**, v. 57, p. 133–145, 2013. 3, 4, 27, 28, 29, 37

KUENZER, C.; DECH, S.; WAGNER, W. **Remote sensing time series - revealing land surface dynamics**. [S.l.: s.n.], 2015. ISBN 978-3-319-15966-9. 9, 10

LARY, D. J.; ALAVI, A. H.; GANDOMI, A. H.; WALKER, A. L. Machine learning in geosciences and remote sensing. **Geoscience Frontiers**, v. 7, n. 1, p. 3–10, 2016. 11

LEWIS, A. et al. The australian geoscience data cube—foundations and lessons learned. **Remote Sensing of Environment**, v. 202, p. 276–292, 2017. 2, 8

LIN, T.-h.; KAMINSKI, N.; BAR-JOSEPH, Z. Alignment and classification of time series gene expression in clinical studies. **Bioinformatics**, v. 24, n. 13, p. i147–i155, 2008. 9

LU, M.; APPEL, M.; PEBESMA, E. Multidimensional arrays for analysing geoscientific data. **ISPRS International Journal of Geo-Information**, v. 7, n. 8, p. 313, 2018. 7

MA, H.; LIU, Y.; REN, Y.; WANG, D.; YU, L.; YU, J. Improved cnn classification method for groups of buildings damaged by earthquake, based on high resolution remote sensing images. **Remote Sensing**, v. 12, n. 2, p. 260, 2020. 11

MACENA, F.; ASSAD, E.; EVANGELISTA, B. Caracterização climática do bioma cerrado. **Cerrado: Ecologia E Flora**, p. 69–88, 01 2008. 65

MARTINELLI, L. A.; NAYLOR, R.; VITOUSEK, P. M.; MOUTINHO, P. Agriculture in Brazil: impacts, costs, and opportunities for a sustainable future. **Current Opinion in Environmental Sustainability**, v. 2, n. 5-6, p. 431–438, 2010. 1

MAXWELL, A. E.; WARNER, T. A.; FANG, F. Implementation of machine-learning classification in remote sensing: an applied review. **International Journal of Remote Sensing**, v. 39, n. 9, p. 2784–2817, 2018. 13, 26

METTERNICHT, G. Vegetation indices derived from high-resolution airborne videography for precision crop management. **International Journal of Remote Sensing**, v. 24, n. 14, p. 2855–2877, 2003. 34

MICHALSKI, R. S.; CARBONELL, J. G.; MITCHELL, T. M. **Machine learning: an artificial intelligence approach**. [S.l.]: Springer Science & Business Media, 2013. 10

NATIVI, S.; MAZZETTI, P.; CRAGLIA, M. A view-based model of data-cube to support big earth data systems interoperability. **Big Earth Data**, v. 1, n. 1-2, p. 75–99, dec 2017. ISSN 2096-4471, 2574-5417. 2, 7

NEVES, A. K.; BENDINI, H. do N.; KORTING, T. S.; FONSECA, L. M. G. Combining time series features and data mining to detect land cover patterns: a case study in northern Mato Grosso state, Brazil. **Revista Brasileira de Cartografia**, v. 68, n. 6, 2016. 3

NGUYEN, L. H.; JOSHI, D. R.; CLAY, D. E.; HENEBRY, G. M. Characterizing land cover/land use from multiple years of landsat and modis time series: a novel approach using land surface phenology modeling and random forest classifier. **Remote Sensing of Environment**, v. 238, p. 111017, 2020. 11

OKI, T.; KANAE, S. Global hydrological cycles and world water resources. **Science**, v. 313, n. 5790, p. 1068–1072, 2006. 55

PARENTE, L.; MESQUITA, V.; MIZIARA, F.; BAUMANN, L.; FERREIRA, L. Assessing the pasturelands and livestock dynamics in Brazil, from 1985 to 2017: a novel approach based on high spatial resolution imagery and Google Earth Engine cloud computing. **Remote Sensing of Environment**, v. 232, p. 111301, out. 2019. ISSN 0034-4257. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0034425719303207>. 3

PARSONS, L.; HAQUE, E.; LIU, H. Subspace clustering for high dimensional data: a review. **Acm Sigkdd Explorations Newsletter**, v. 6, n. 1, p. 90–105, 2004. 25, 26

PEKEL, J.-F.; COTTAM, A.; GORELICK, N.; BELWARD, A. S. High-resolution mapping of global surface water and its long-term changes. **Nature**, v. 540, n. 7633, p. 418–422, 2016. xii, 2, 55, 62, 63, 73

PELLETIER, C.; VALERO, S.; INGLADA, J.; CHAMPION, N.; SICRE, C. M.; DEDIEU, G. Effect of training class label noise on classification performances for land cover mapping with satellite image time series. **Remote Sensing**, v. 9, n. 2, p. 173, 2017. 24, 41

PETTORELLI, N.; VIK, J. O.; MYSTERUD, A.; GAILLARD, J.-M.; TUCKER, C. J.; STENSETH, N. C. Using the satellite-derived ndvi to assess ecological responses to environmental change. **Trends in Ecology & Evolution**, v. 20, n. 9, p. 503–510, 2005. 2, 10, 11

PICOLI, M. C.; SIMOES, R.; CHAVES, M.; SANTOS, L. A.; SANCHEZ, A.; SOARES, A.; SANCHES, I. D.; FERREIRA, K. R.; QUEIROZ, G. R. Cbers data cube: a powerful technology for mapping and monitoring brazilian biomes. **ISPRS**

Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences, v. 3, p. 533–539, 2020. 2, 18, 41, 42, 65

PICOLI, M. C. A. et al. Big earth observation time series analysis for monitoring brazilian agriculture. **ISPRS Journal of Photogrammetry and Remote Sensing**, v. 145, p. 328–339, 2018. 2, 10

PINTY, B.; VERSTRAETE, M. Gemi: a non-linear index to monitor global vegetation from satellites. **Vegetatio**, v. 101, n. 1, p. 15–20, 1992. 32, 34

POTAPOV, P.; HANSEN, M. C.; KOMMAREDDY, I.; KOMMAREDDY, A.; TURUBANOVA, S.; PICKENS, A.; ADUSEI, B.; TYUKAVINA, A.; YING, Q. Landsat analysis ready data for global land cover and land cover change mapping. **Remote Sensing**, v. 12, n. 3, p. 426, jan. 2020. Disponível em: <https://www.mdpi.com/2072-4292/12/3/426>. 3

RAND, W. M. Objective criteria for the evaluation of clustering methods. **Journal of the American Statistical Association**, v. 66, n. 336, p. 846–850, 1971. 21

RATANAMAHATANA, C. A.; KEOGH, E. Multimedia retrieval using time series representation and relevance feedback. In: INTERNATIONAL CONFERENCE ON ASIAN DIGITAL LIBRARIES, 2005. **Proceedings...** [S.l.]: Springer, 2005. p. 400–405. 18

REED, B. C.; BROWN, J. F.; VANDERZEE, D.; LOVELAND, T. R.; MERCHANT, J. W.; OHLEN, D. O. Measuring phenological variability from satellite imagery. **Journal of Vegetation Science**, v. 5, n. 5, p. 703–714, 1994. 10

RODRIGUES, M.; BENDINI, H.; SOARES, A.; KÖRTING, T.; FONSECA, L. Remote sensing image time series metrics for distinction between pasture and croplands using the random forest classifier. In: LATIN AMERICAN GRSS & ISPRS REMOTE SENSING CONFERENCE, 2020. **Proceedings...** [S.l.]: IEEE, 2020. p. 149–154. 3

ROUSE, J.; HAAS, R.; SCHELL, J.; DEERING, D. Monitoring vegetation systems in the great plains with erts. In: NASA GODDARD SPACE FLIGHT CENTER SYMPOSIUM, 3. **Proceedings...** [S.l.]: NASA, 1973. v. 3, p. 309–317. 32, 34

SAHA, S.; SAHA, M.; MUKHERJEE, K.; ARABAMERI, A.; NGO, P. T. T.; PAUL, G. C. Predicting the deforestation probability using the binary logistic regression, random forest, ensemble rotational forest, reptree: a case study at the

gumani river basin, India. **Science of The Total Environment**, v. 730, p. 139197, 2020. 13

SANCHEZ, A.; PICOLI, M. C. A.; ANDRADE, P. R. de; SIMÕES, R. E. O.; SANTOS, L. A.; CHAVES, M.; BEGOTTI, R. A.; CAMARA, G. Land cover classifications of clear-cut deforestation using deep learning. In: BRAZILIAN SYMPOSIUM ON GEOINFORMATICS (GEOINFO), 20., 2019, São José dos Campos, SP. **Proceedings...** São José dos Campos: INPE, 2019. p. 48–56. 11

SANO, E. E.; SANTOS, C. C. M. dos; SILVA, E. M. da; CHAVES, J. M. Fronteira agrícola do oeste baiano: considerações sobre os aspectos temporais e ambientais. **Geociências (São Paulo)**, v. 30, n. 3, p. 479–489, 2011. 65

SANTOS, L.; FERREIRA, K. R.; PICOLI, M.; CAMARA, G. Self-organizing maps in earth observation data cubes analysis. In: INTERNATIONAL WORKSHOP ON SELF-ORGANIZING MAPS, 2019. **Proceedings...** [S.l.]: Springer, 2019. p. 70–79. 3, 38, 41, 44

SANTOS, L. A.; FERREIRA, K.; PICOLI, M.; CAMARA, G.; ZURITA-MILLA, R.; AUGUSTIJN, E.-W. Identifying spatiotemporal patterns in land use and cover samples from satellite image time series. **Remote Sensing**, v. 13, n. 5, p. 974, 2021. 2, 18, 48

SANTOS, L. A.; FERREIRA, K. R.; CAMARA, G.; PICOLI, M. C.; SIMOES, R. E. Quality control and class noise reduction of satellite image time series. **ISPRS Journal of Photogrammetry and Remote Sensing**, v. 177, p. 75–88, 2021. ISSN 0924-2716. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0924271621001155>. 2, 24, 25

_____. _____. **ISPRS Journal of Photogrammetry and Remote Sensing**, v. 177, p. 75–88, 2021. 2, 41, 47, 51, 52

SCRIVANI, R.; AMARAL, B. F. do; GONÇALVES, R. d. V.; SOUSA, E. P. M. de; JUNIOR, J. Z.; ROMANI, L. A. S. Identificação da mudança de uso da terra usando técnicas de agrupamento de séries temporais de imagens de satélite. In: SIMPÓSIO DE GEOTECNOLOGIAS NO PANTANAL, 5., 2014, CAMPO GRANDE, MS. **Anais...** [S.l.], 2014. 24

SHAO, Y.; LUNETTA, R. S.; WHEELER, B.; IIAMES, J. S.; CAMPBELL, J. B. An evaluation of time-series smoothing algorithms for land-cover classifications

using modis-ndvi multi-temporal data. **Remote Sensing of Environment**, v. 174, p. 258–265, 2016. 9

SIMOES, R.; CAMARA, G.; SOUZA, F.; ANDRADE, P.; SANTOS, L.; FERREIRA, K.; QUEIROZ, G.; de Carvalho, A. Y.; MAUS, V. **sits: data analysis and machine learning using satellite image time series**. Sao Jose dos Campos, Brazil, 2021. Disponível em: <https://github.com/e-sensing/sits>. 9

SIMOES, R.; PICOLI, M. C.; CAMARA, G.; MACIEL, A.; SANTOS, L.; ANDRADE, P. R.; SÁNCHEZ, A.; FERREIRA, K.; CARVALHO, A. Land use and cover maps for Mato Grosso state in Brazil from 2001 to 2017. **Scientific Data**, v. 7, n. 1, p. 1–10, 2020. 2, 10, 13, 55, 65

SOARES, A. R.; BENDINI, H. N.; VAZ, D. V.; UEHARA, T. D.; NEVES, A. K.; LECHLER, S.; KÖRTING, T. S.; FONSECA, L. M. Stmetrics: a python package for satellite image time-series feature extraction. In: IEEE INTERNATIONAL GEOSCIENCE AND REMOTE SENSING SYMPOSIUM (IGARSS), 2020. **Proceedings...** [S.l.]: IEEE, 2020. p. 2061–2064. 3, 27

SOILLE, P.; BURGER, A.; DE MARCHI, D.; KEMPENEERS, P.; RODRIGUEZ, D.; SYRRIS, V.; VASILEV, V. A versatile data-intensive computing platform for information retrieval from big geospatial data. **Future Generation Computer Systems**, v. 81, p. 30–40, 2018. ISSN 0167739X. Disponível em: <https://doi.org/10.1016/j.future.2017.11.007>. 1

SONG, H.; LI, G. Tourism demand modelling and forecasting—a review of recent research. **Tourism Management**, v. 29, n. 2, p. 203–220, 2008. 9

SOUZA, F.; SANTOS, R.; FERREIRA, K. R. ggsom: ferramenta de visualização baseada em mapas auto-organizáveis. In: ENCONTRO NACIONAL DE MODELAGEM COMPUTACIONAL, 22.; ENCONTRO DE CIÊNCIA E TECNOLOGIA DE MATERIAIS, 10., 2019. **Anais...** [S.l.], 2019. 5, 24, 48

SPERA, S.; VANWEY, L.; MUSTARD, J. The drivers of sugarcane expansion in Goiás, Brazil. **Land Use Policy**, v. 66, p. 111–119, 2017. 1

TOURE, S. I.; STOW, D. A.; SHIH, H.-c.; WEEKS, J.; LOPEZ-CARR, D. Land cover and land use change analysis using multi-spatial resolution data and object-based image analysis. **Remote Sensing of Environment**, v. 210, p. 259–268, 2018. 9

UEHARA, T. D. T.; SOARES, A. R.; QUEVEDO, R. P.; KöRTING, T. S.; FONSECA, L. M. G.; ADAMI, M. Land cover classification of an area susceptible to landslides using random forest and ndvi time series data. In: INTERNATIONAL GEOSCIENCE AND REMOTE SENSING SYMPOSIUM, 2020. **Proceedings...** [S.l.]: IEEE, 2020. p. 1345–1348. 3

VENDRAMIN, L.; CAMPELLO, R. J.; HRUSCHKA, E. R. Relative clustering validity criteria: a comparative overview. **Statistical Analysis and Data Mining: the ASA Data Science Journal**, v. 3, n. 4, p. 209–235, 2010. 21, 22, 23, 46

VESANTO, J.; ALHONIEMI, E. Clustering of the self-organizing map. **IEEE Transactions on Neural Networks**, v. 11, n. 3, p. 586–600, 2000. 47

WUNDERVALD, B.; PARNELL, A. C.; DOMIJAN, K. Generalizing gain penalization for feature selection in tree-based models. **IEEE Access**, v. 8, p. 190231–190239, 2020. 30, 38

XAVIER, A. C.; RUDORFF, B. F.; SHIMABUKURO, Y. E.; BERKA, L. M. S.; MOREIRA, M. A. Multi-temporal analysis of modis data to classify sugarcane crop. **International Journal of Remote Sensing**, v. 27, n. 4, p. 755–768, 2006. 50

XU, H. Modification of normalised difference water index (ndwi) to enhance open water features in remotely sensed imagery. **International Journal of Remote Sensing**, v. 27, n. 14, p. 3025–3033, 2006. 32, 34

ZENG, L.; WARDLOW, B. D.; XIANG, D.; HU, S.; LI, D. A review of vegetation phenological metrics extraction using time-series, multispectral satellite data. **Remote Sensing of Environment**, v. 237, p. 111511, 2020. 10

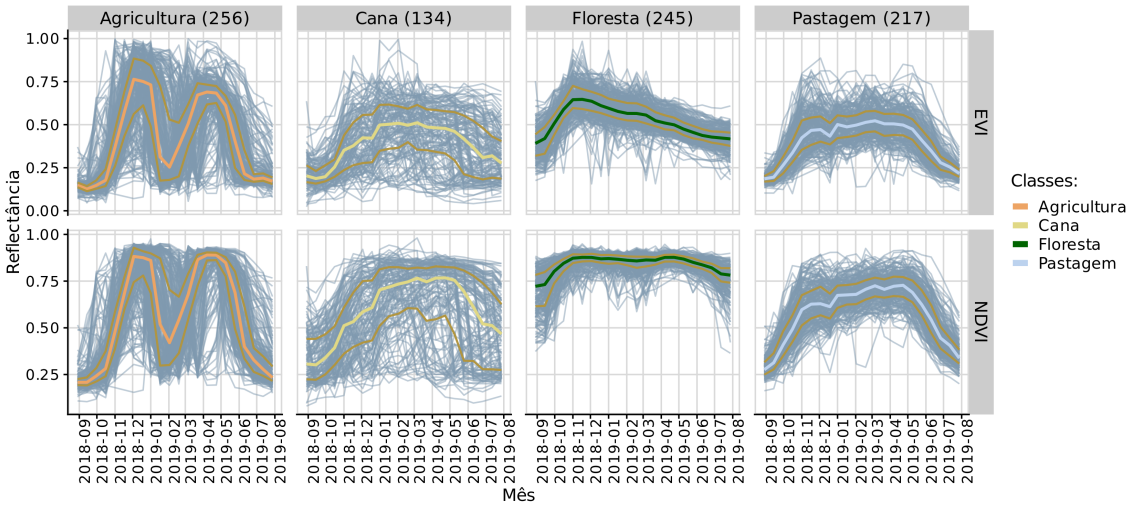
ZHENG, Y.; WU, B.; ZHANG, M.; ZENG, H. Crop phenology detection using high spatio-temporal resolution data fused from spot5 and modis products. **Sensors**, v. 16, n. 12, p. 2099, 2016. 9

ZHONG, S.; KHOSHGOFTAAR, T. M.; SELIYA, N. Clustering-based network intrusion detection. **International Journal of reliability, Quality and safety Engineering**, v. 14, n. 02, p. 169–187, 2007. 9

APÊNDICE A - INFORMAÇÕES ADICIONAIS DO CAPÍTULO 4

A.1 Figura

Figura A.1 - Séries temporais extraídas de cubo de dados Sentinel-2 com resolução temporal de 16 dias e resolução espacial de 10 m.



Os contornos mais fortes mostram a mediana de cada classe. As linhas que estão acima e abaixo da mediana mostram o primeiro e terceiro quartil.

Fonte: Próprio Autor.

A.2 Tabela

Tabela A.1 - Atributos selecionados no estudo de caso da região do Mato Grosso.

Banda	Métricas básicas	Métricas polares	Métricas básicas e polares
Banda 2	std	area_q4	min
	fslope	angle	-
	fqr	-	-
Banda 3	median	csi	area_q1
	-	-	area_q4
	-	-	csi
	-	-	sum
	-	-	fqr
Banda 4	mse	area_q3	fslope
	-	gyration_radius	tqr
Banda 5	mean	area_q1	sum
	-	area_q3	median
	-	area_q4	std
	-	area_ts	fqr
Banda 8	std	area_q3	area_q2
	fqr	-	-
Banda 11	amplitude	area_q4	area_q4
	-	polar_balance	angle
	-	area_ts	max
	-	csi	-
Banda 12	fslope	polar_balance	area_q3
	-	csi	area_ts
	-	-	std
	-	-	amd
EVI	max	csi	area_q2
	sum	-	area_q3
	amplitude	-	std
	amd	-	amd
	amd	-	tqr
GEMI	min	-	ecc_metric
	amd	-	amplitude
	-	-	fslope
GNDVI	amd	csi	csi
	iqr	-	amd
	iqr	-	
PVR	iqr	area_q3	area_q3
	-	angle	area_q4
	-	-	csi
	-	-	abs_sum
	-	-	amd
	-	-	mse
NDVI	amplitude	area_q4	ecc_metric
	abs_sum	area_ts	csi
	amd	gyration_radius	std
	-	csi	amd
	-	-	mse
	-	-	fqr
NDWI		polar_balance	

Fonte: Próprio Autor.

APÊNDICE B - INFORMAÇÕES ADICIONAIS DO CAPÍTULO 5

B.1 Tabela

Tabela B.1 - Atributos selecionados no estudo de caso da região de Minas Gerais.

Banda	Métricas básicas	Métricas polares	Métricas básicas e polares
Banda 6	-	-	mean
	-	-	sqr
NDWI	-	area_q4	-
MNDWI	abs_sum	-	-
	sqr	-	-

Fonte: Próprio Autor.

APÊNDICE C - INFORMAÇÕES ADICIONAIS DO CAPÍTULO 6

C.1 Tabela

Tabela C.1 - Atributos selecionados no estudo de caso da região Oeste da Bahia.

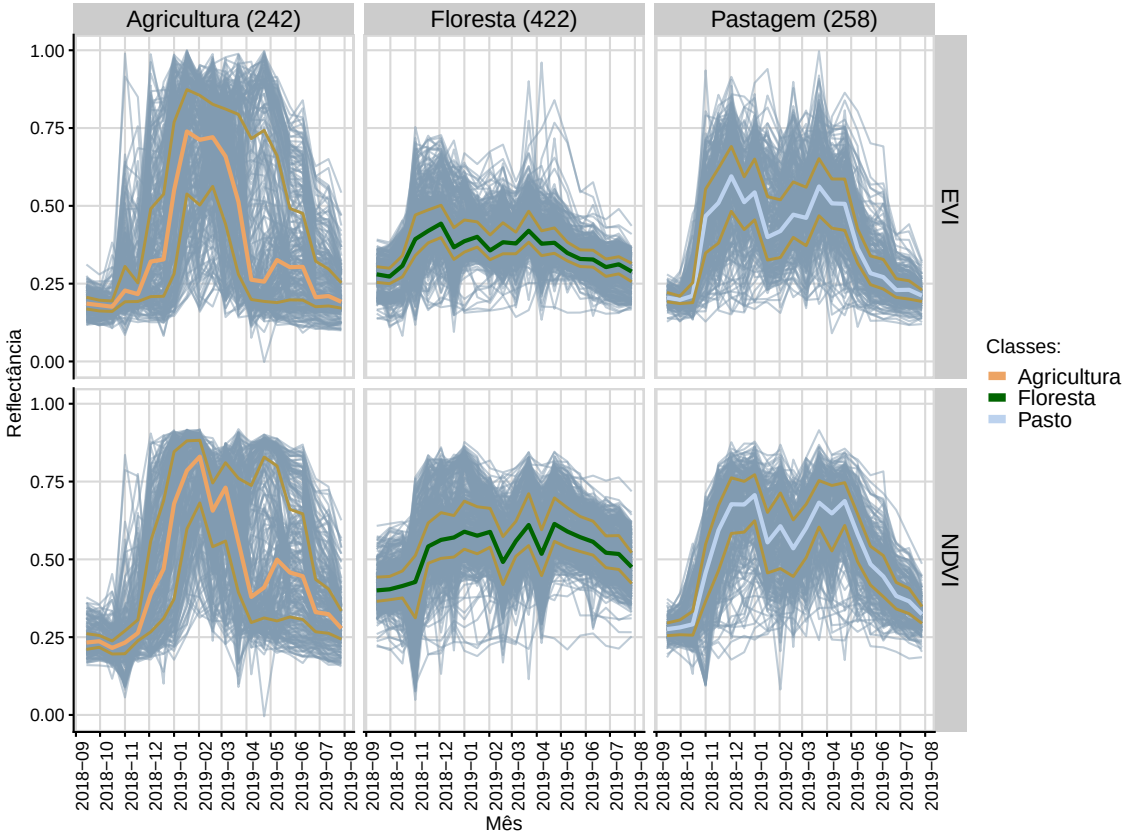
Banda	Métricas básicas	Métricas polares	Métricas básicas e polares
Banda 13	min	area_q1	area_q3
	sum	area_q3	area_ts
	median	area_q4	min
	std	area_ts	amplitude
	fslope	ecc_metric	sqr
	mse	csi	tqr
	tqr	-	-
	iqr	-	-
Banda 14	max	area_q3	area_q1
	min	area_q4	polar_balance
	sum	angle	angle
	amd	ecc_metric	min
	mse	gyration_radius	mse
	fqr	-	tqr
	sqr	-	-
	tqr	-	-
	iqr	-	-
Banda 15	max	area_q2	area_q2
	min	area_q4	area_q4
	sum	polar_balance	angle
	std	angle	ecc_metric
	amplitude	area_ts	std
	fslope	ecc_metric	amplitude
	amd	csi	tqr
	fqr	-	iqr
	sqr	-	-
	tqr	-	-
	iqr	-	-
Banda 16	max	area_q1	area_q1
	min	area_q3	area_q2
	median	area_q4	area_q3

	std	polar_balance	area_q4
	fslope	angle	polar_balance
	abs_sum	area_ts	angle
	amd	ecc_metric	gyration_radius
	mse	gyration_radius	abs_sum
	tqr	csi	amd
	iqr	-	tqr
EVI	max	area_q1	area_q1
	min	area_q2	area_q2
	median	area_q3	area_q4
	std	area_q4	polar_balance
	amplitude	polar_balance	gyration_radius
	fslope	ecc_metric	max
	amd	gyration_radius	min
	fqr	csi	std
	tqr	-	amplitude
	iqr	-	fslope
	-	-	fqr
NDVI	max	area_q2	area_q2
	min	area_q3	area_q4
	sum	area_q4	max
	median	polar_balance	min
	std	ecc_metric	median
	amplitude	csi	std
	fslope	-	amplitude
	amd	-	abs_sum
	fqr	-	amd
	tqr	-	fqr

Fonte: Próprio Autor.

C.2 Figura

Figura C.1 - Séries temporais extraídas de cubo de dados CBERS-4 com resolução temporal de 16 dias e resolução espacial de 64 m.



Os contornos mais fortes mostram a mediana de cada classe. As linhas que estão acima e abaixo da mediana mostram o primeiro e terceiro quartil.

Fonte: Próprio Autor.