

**UNIVERSIDADE ESTADUAL PAULISTA "JÚLIO DE MESQUITA FILHO"**  
FACULDADE DE CIÊNCIAS - CAMPUS BAURU  
DEPARTAMENTO DE COMPUTAÇÃO  
BACHARELADO EM CIÊNCIA DA COMPUTAÇÃO

ALEXANDRE DE TOMY SILVA

**APLICAÇÃO DE MÉTODOS DE APRENDIZADO DE MÁQUINA  
PARA CLASSIFICAÇÃO DE USO E COBERTURA DA TERRA EM  
IMAGENS DE SENSORIAMENTO REMOTO**

BAURU  
2019

ALEXANDRE DE TOMY SILVA

**APLICAÇÃO DE MÉTODOS DE APRENDIZADO DE MÁQUINA  
PARA CLASSIFICAÇÃO DE USO E COBERTURA DA TERRA EM  
IMAGENS DE SENSORIAMENTO REMOTO**

Trabalho de Conclusão de Curso do Curso de Bacharelado em Ciência da Computação da Universidade Estadual Paulista “Júlio de Mesquita Filho”, Faculdade de Ciências, Campus Bauru.

Orientador: Prof. Dr. Kelton Augusto Pontara da Costa

ALEXANDRE DE TOMY SILVA    APLICAÇÃO DE MÉTODOS DE APREN-  
DIZADO DE MÁQUINA PARA CLASSIFICAÇÃO DE USO E COBERTURA  
DA TERRA EM IMAGENS DE SENSORIAMENTO REMOTO/ ALEXANDRE  
DE TOMY SILVA. – Bauru, 2019-    37 p. : il. ; 30 cm.  
Orientador: Prof. Dr. Kelton Augusto Pontara da Costa  
Trabalho de Conclusão de Curso – Universidade Estadual Paulista “Júlio de  
Mesquita Filho”  
Faculdade de Ciências  
Bacharelado em Ciência da Computação, 2019.  
1. Sensoriamento Remoto 2. Aprendizado de Máquina 3. R 4. Observação da  
Terra

ALEXANDRE DE TOMY SILVA

# **APLICAÇÃO DE MÉTODOS DE APRENDIZADO DE MÁQUINA PARA CLASSIFICAÇÃO DE USO E COBERTURA DA TERRA EM IMAGENS DE SENSORIAMENTO REMOTO**

Trabalho de Conclusão de Curso do Curso de Bacharelado em Ciência da Computação da Universidade Estadual Paulista "Júlio de Mesquita Filho", Faculdade de Ciências, Campus Bauru.

Banca Examinadora

---

**Prof. Dr. Kelton Augusto Pontara da  
Costa**

Orientador

Universidade Estadual Paulista "Júlio de  
Mesquita Filho"

Faculdade de Ciências

Departamento de Computação

---

**Profª. Drª Simone Domingues Prado**

Universidade Estadual Paulista "Júlio de  
Mesquita Filho"

Faculdade de Ciências

Departamento de Computação

---

**Prof. Dr. Aparecido Nilceu Marana**

Universidade Estadual Paulista "Júlio de  
Mesquita Filho"

Faculdade de Ciências

Departamento de Computação

Bauru, \_\_\_\_\_ de \_\_\_\_\_ de \_\_\_\_\_.  
\_\_\_\_\_  
\_\_\_\_\_  
\_\_\_\_\_

# Resumo

Imagens geradas a partir de satélites possuem grande relevância no contexto de observação da terra para monitoramento dos recursos terrestres. Com a grande disponibilidade desses dados atualmente, torna-se relevante o desenvolvimento, aplicação e disseminação de metodologias de análise e inferência. Em sensoriamento remoto, algoritmos preditivos focam em classificações de uso e cobertura da terra, possibilitando a diferenciação de diferentes padrões de classes. Com isso em vista, o objetivo deste trabalho é o desenvolvimento e estudo de um processo de classificação a partir de dados gerados pela iniciativa MapBiomas bem como imagens do sensor MSI/Sentinel-2, tendo como área de interesse a região do município de Bauru - SP. O ambiente de desenvolvimento utilizado foi o Rstudio e o Google Earth Engine; o resultado está disponível no GitHub no formato html, juntamente com o código, dados e informações necessárias para reprodução.

**Palavras-chave:** Sensoriamento remoto, aprendizado de máquina, R, observação da terra

# Listas de figuras

Figura 1 – Fluxograma do processo de desenvolvimento em etapas . . . . .	20
Figura 2 – Composição com as bandas vermelha, verde e azul do Sentinel-2/MSI para a região da região do município de Bauru . . . . .	22
Figura 3 – Amostras aleatórias de pontos contidos na região de cerrado do Estado de São Paulo . . . . .	24
Figura 4 – Representação do cálculo do índice NDVI . . . . .	26
Figura 5 – Parâmetros selecionados após treinamento com o algoritmo maquina de vetores suporte com núcleo radial . . . . .	27
Figura 6 – Parâmetros selecionados após treinamento com o algoritmo maquina de vetores suporte linear com regularização . . . . .	28
Figura 7 – Parâmetros selecionados após treinamento com o algoritmo florestas aleatórias	29
Figura 8 – Mapa com classes de uso e cobertura da terra obtido a partir da predição do modelo utilizando MVS com núcleo radial . . . . .	30
Figura 9 – Mapa com classes de uso e cobertura da terra obtido a partir da predição do modelo utilizando MVS linear com regularização . . . . .	31
Figura 10 – Mapa com classes de uso e cobertura da terra obtido a partir da predição do modelo utilizando FA's . . . . .	32
Figura 11 – Mapa com classes de uso e cobertura da terra obtido a partir da classificação realizada pela iniciativa MapBiomas . . . . .	33

# **Lista de tabelas**

Tabela 1 – Características das bandas do sensor multiespectral MSI do Satélite Sentinel-2	21
Tabela 2 – Classes de uso e cobertura da terra da coleção 3.1 do MapBiomass para o bioma Cerrado . . . . .	23
Tabela 3 – Quantidade de exemplos de treinamento para cada classe . . . . .	25
Tabela 4 – Tabela de avaliação de precisão das previsões realizadas . . . . .	33

# Sumário

<b>1</b>	<b>INTRODUÇÃO</b>	<b>9</b>
<b>1.1</b>	<b>Objetivos</b>	<b>10</b>
1.1.1	Objetivos Gerais	10
1.1.2	Objetivos Específicos	10
<b>2</b>	<b>FUNDAMENTAÇÃO TEÓRICA</b>	<b>11</b>
<b>2.1</b>	<b>Geotecnologias</b>	<b>11</b>
2.1.1	Dados geoespaciais	11
<b>2.2</b>	<b>Sensoriamento Remoto</b>	<b>12</b>
2.2.1	Energia	12
2.2.1.1	Interferências Atmosféricas	12
2.2.1.2	Espectro Eletromagnético	13
2.2.2	Sensores	13
2.2.2.1	Resolução dos sensores	13
<b>2.3</b>	<b>Aprendizado de Máquina</b>	<b>14</b>
2.3.1	Aprendizado supervisionado	14
2.3.1.1	Alto e baixo viés	14
2.3.1.2	Validação Cruzada	15
2.3.2	Aprendizado de Máquina e Sensoriamento remoto	15
2.3.3	Algoritmos	15
2.3.3.1	Máquina de Vetores Suporte	15
2.3.3.2	Florestas Aleatórias	16
2.3.4	Avaliação de Desempenho	16
<b>3</b>	<b>METODOLOGIA</b>	<b>17</b>
<b>3.1</b>	<b>Ferramentas e Dados</b>	<b>17</b>
3.1.1	Linguagem R	17
3.1.1.1	Ambiente R	17
3.1.1.2	Rmarkdown	17
3.1.1.3	Bibliotecas	18
3.1.2	Google Earth Engine	18
3.1.3	Iniciativas	18
3.1.3.1	Mapbiomas	18
3.1.4	Satélites	19
<b>3.2</b>	<b>Etapas</b>	<b>19</b>

<b>4</b>	<b>DESENVOLVIMENTO</b>	<b>21</b>
<b>4.1</b>	<b>Etapas</b>	<b>21</b>
4.1.1	Aquisição dos dados	21
4.1.1.1	Área de Estudo	21
4.1.1.2	Imagens de Satélite	21
4.1.1.3	Seleção de Amostras	22
4.1.2	Extração de Características	24
4.1.2.1	Extração dos Valores das Amostras	24
4.1.2.2	Índices de Vegetação	25
4.1.2.3	Seleção de Variáveis	26
4.1.3	Modelagem	26
4.1.3.1	Reamostragem	27
4.1.3.2	Treinamento	27
<b>4.2</b>	<b>Resultados</b>	<b>29</b>
4.2.1	Predição dos Dados	29
4.2.1.1	Avaliação de Precisão	32
4.2.1.2	Considerações	33
<b>5</b>	<b>CONCLUSÃO</b>	<b>35</b>
<b>5.1</b>	<b>Trabalhos Futuros</b>	<b>35</b>
	<b>REFERÊNCIAS</b>	<b>36</b>

# 1 Introdução

Dados gerados a partir da Observação da Terra são ricas fontes para descobrir como a Terra está mudando. Imagens obtidas a partir de satélites que orbitam o globo possibilitam uma visão de conjunto multitemporal da superfície terrestre, o que possibilita o estudo e monitoramento dos impactos causados por fenômenos naturais e antrópicos (FLORENZANO, 2002). Portanto, o desenvolvimento de técnicas e abordagens que viabilizam a análise desses dados é essencial.

As mudanças na ocupação do solo afetam o clima global, logo torna-se relevante o devido monitoramento do uso e cobertura da terra em escala global (WULDER; COOPS, 2014). Isso é possível através da **coleta, distribuição e análise** de dados obtidos por Sensoriamento Remoto.

Sensoriamento remoto é uma técnica de obtenção de imagens dos objetos da superfície terrestre sem que haja um contato físico de qualquer espécie entre o sensor e o objeto.(MENESES; ALMEIDA, 2012, p. 3).

**Cobertura da terra** pode ser definido como a cobertura biofísica observada na superfície terrestre. Já o termo **uso da terra** abrange os arranjos, atividades e insumos empreendidos pelas pessoas em um tipo de cobertura da terra para produzir, alterar ou manter. (GREGORIO, 2016)

Pesquisas na área de Sensoriamento Remoto envolvendo métodos digitais de classificação de imagens chamam a atenção porque seus resultados são a base para muitas aplicações ambientais e socioeconômicas (LU; WENG, 2007). Além disso, metodologias de aprendizado de máquina têm bons resultados em aplicações reais, especialmente para tarefas de classificação e regressão, uma vez que não é necessário um conhecimento a priori sobre o modelo de distribuição dos dados disponíveis nem o relacionamento entre as variáveis independentes precisam ser assumidos. Essas são propriedades desejáveis para o sucesso desses métodos para a análise de imagens de sensoriamento remoto. (WASKE, 2009)

Hoje em dia, há uma alta demanda para aplicações de alta performance em geoprocessamento, num contexto onde muitos dados georreferenciados são produzidos a todo instante, a exemplo dos *smartphones* que possuem GPS e são amplamente utilizados. Além disso, **Sensoriamento Remoto (SR)** para monitoramento é uma tarefa que exige um muito processamento e espaço em disco disponível (LOVELACE; NOWOSAD; MUENCHOW, 2019).

A solução atualmente é a computação e armazenamento em aplicações baseadas em nuvem, ou seja, serviços disponibilizados por estruturas capazes de lidar com uma grande quantidade de dados de maneira eficiente. Com isso, a coleta e distribuição de dados gerados

por sensoriamento remoto se torna muito mais viável, possibilitando o monitoramento em escala global. ([MATHIEU; AUBRECHT, 2018](#))

## 1.1 Objetivos

### 1.1.1 Objetivos Gerais

Levando em consideração a importância da aplicação de métodos analíticos para monitoramento da cobertura terrestre, bem como os meios que tornam essa tarefa viável, este trabalho teve como objetivo a aplicação e estudo de um processo de construção e comparação de modelos preditivos, a fim de realizar a classificação de novos dados.

O desenvolvimento foi realizado a partir da coleta de dados da iniciativa MapBiomass, e também imagens do sensor MSI do satélite Sentinel-2, ambos disponíveis gratuitamente. Os ambientes utilizados foram a IDE RStudio da linguagem R, bem como plataforma do Google Earth Engine. A área de estudo foi a região do município de Bauru - SP. Os métodos de classificação de imagens aplicados foram: Florestas Aleatórias (ou *Random Forests*) e Máquina de Vetores Suporte (ou *Support Vector Machines*) linear com regularização e também com função de núcleo não linear.

### 1.1.2 Objetivos Específicos

- Revisão teórica: organizar um estudo dos conceitos chave acerca dos temas que envolvem o trabalho, ou seja, sensoriamento remoto, aprendizado de máquina, classificação de imagens.
- Análise e comparação: discussão dos métodos e dados a serem coletados, bem como apresentação das ferramentas utilizadas.
- Análise das classes de uso e cobertura da terra para o bioma Cerrado, a partir da metodologia adotada pela iniciativa MapBiomass
- Implementação: desenvolvimento do modelo e aplicação.
- Teste e validação dos resultados: realizar uma avaliação de precisão de classificação, a fim de comparar os modelos aplicados e discussão de resultados

## 2 Fundamentação teórica

### 2.1 Geotecnologias

"As geotecnologias são o conjunto de tecnologias para coleta, processamento, análise e oferta de informação com referência geográfica."(ROSA, 2005)

Podem ser caracterizadas geotecnologias: [Sistemas de Informação Geográfica \(SIG\)](#), cartografia digital, Sensoriamento Remoto, [Sistema de Posicionamento Global \(GPS\)](#) e a topografia. Geoprocessamento, por sua vez, é um conceito mais abrangente e representa qualquer tipo de processamento de dados georreferenciados (ROSA, 2005).

Segundo (ROSA, 2005), um [SIG](#) se refere a um conjunto de ferramentas computacionais que integra dados, pessoas e instituições e que torna possível a coleta, armazenamento, processamento, análise e oferta de informação georreferenciada.

Os [SIGs](#) se desenvolveram consideravelmente durante as últimas décadas. Inicialmente, nos anos 80, cada sistema tinha um banco de dados próprio e o processamento era feito isoladamente, em softwares de código fechado. Nos anos 90, formatos de arquivos surgiram, o que facilitou a intercambialidade entre os programas disponíveis. A partir de 2000, surgiu a biblioteca [Geospatial Data Abstraction Layer \(GDAL\)](#), feita para ler e escrever dados geoespaciais. De 2010 até hoje em dia, os ambientes de computação em nuvem surgiram a fim de resolverem o problema de análise, mas novamente de maneira isolada, gerando problemas de reproduzibilidade. (PEBESMA, 2016)

#### 2.1.1 Dados geoespaciais

São considerados dados geoespaciais, dados georreferenciados, ou ainda dados espaciais, os dados que possuem uma localização definida. Computacionalmente, são representados por: pontos, que interligados podem formar linhas e polígonos que representam um objeto geográfico e são chamados de **dados vetoriais**; por matrizes de pontos, divididas em células de tamanhos iguais, que são os *pixels* de uma imagem, denominados como **dados raster**; ou ainda com metadados, que são textos, números e símbolos armazenados em tabelas e vinculados aos dados que possuem referência espacial. Todo dado espacial é composto por um [Sistema de Referência de Coordenadas \(SRC\)](#), que pode ser geográfico (esférico ou geodésico, ou seja, no formato da Terra) ou projetado (em duas dimensões).(LOVELACE; NOWOSAD; MUENCHOW, 2019; IBGE, 2019)

## 2.2 Sensoriamento Remoto

De maneira objetiva, SR pode ser definido como:

(...) uma ciência que visa o desenvolvimento da obtenção de imagens da superfície terrestre por meio da detecção e medição quantitativa das respostas das interações da radiação eletromagnética com os materiais terrestres (MENESES; ALMEIDA, 2012, p. 3)

Por essa definição, tem-se um sistema onde um **alvo** localizado na superfície da terra interage com a **energia** proveniente de uma **fonte** (como a luz solar) gerando uma resposta que é captada por um sensor (geralmente um satélite) e que por sua vez é processada e traduzida como uma **imagem**.

### 2.2.1 Energia

A **Radiação Eletromagnética (REM)** é caracterizada pela dualidade de comportamento na natureza: é ao mesmo tempo uma forma de onda e uma forma de energia que se propaga pelo espaço vazio. Segundo o modelo ondulatório, a radiação é definida como uma forma de onda propagada a partir da perturbação dos campos elétrico e magnético, gerados por uma partícula eletricamente carregada. As características das imagens de Sensoriamento Remoto são definidas pela intensidade com que um objeto reflete a **REM** em razão da textura de sua superfície e do comprimento de onda; essa interação é denominada **interação macroscópica**. Já o modelo corpuscular, define a **REM** como uma forma dinâmica de energia que se manifesta por suas interações com a matéria. As trocas de energia ocorrerão somente se a quantidade de energia da **REM** for igual à necessária para promover uma mudança nos níveis de energia dos átomos ou moléculas, caracterizando a **interação microscópica**. (MENESES; ALMEIDA, 2012)

A partir desses modelos, define-se a energia transportada  $E$ , o comprimento de onda  $\lambda$  relacionados pela equação:

$$E = \frac{hc}{\lambda} \quad (2.1)$$

Onde  $h$  é constante de Planck ( $6,624 \times 10^{-34}$  Joules.seg) e  $c$  a velocidade da luz de aproximadamente 300.000 km/s

#### 2.2.1.1 Interferências Atmosféricas

O nosso sistema solar tem como o próprio Sol a maior fonte de energia que chega até a Terra, que por sinal também emite **REM**, em menor quantidade mas que pode ser detectada por sensores. No caso do Sensoriamento Remoto orbital (ou seja, via satélite), a atmosfera é opaca à radiação para vários intervalos de comprimentos de onda. Isso ocorre devido aos efeitos de **absorção** e **espalhamento** causados pela interferência da interação entre a **REM** e as partículas e moléculas presentes na atmosfera terrestre. (MENESES; ALMEIDA, 2012)

### 2.2.1.2 Espectro Eletromagnético

O espectro eletromagnético representa a distribuição da REM por regiões espectrais conhecidas pelo homem. Da luz visível, por exemplo, cada cor tem seu comprimento de onda, portanto as imagens de SR são definidas em intervalos (ou **bandas**). Lembrando que a cor "real" dos alvos não é a mesma capturados pelos sensores, devido às interferências explicadas anteriormente. ([MENESES; ALMEIDA, 2012](#))

Os objetos (ou **alvos**) presentes na superfície terrestre refletem, absorvem e transmitem radiação de acordo com a característica de seu material de composição. As variações da energia refletida podem ser representadas através de curvas, que distinguem os alvos. A representação destes nas imagens vão variar, para cada banda, do branco (ou seja, refletem mais energia) ao preto (refletem pouca energia).

### 2.2.2 Sensores

O sensor é responsável por captar e converter para valores digitais a intensidade da radiância, ou seja, o fluxo radiante refletido pelo elemento da superfície. As imagens são capturadas e posteriormente pré processadas, e nesse processo geralmente são convertidos os valores da radiância para a reflectância, obtida pela divisão entre a radiância e a irradiância, que por sua vez representa a densidade do fluxo radiante solar incidente por área da superfície. O tipo mais comum de sensor, é o multiespectral. São sensores capazes de obter múltiplas imagens simultâneas da superfície em diversos comprimentos de ondas diferentes. ([MENESES; ALMEIDA, 2012](#))

#### 2.2.2.1 Resolução dos sensores

As características de uma imagem obtida por sensoriamento remoto podem ser resumidas em quatro resoluções: espacial (ou geométrica), espectral, radiométrica e temporal. A **resolução espacial**, dada em metros, é a área representada por um pixel na imagem final, ou seja, se é de 30 metros, significa que a largura de um pixel representa um espaço de 30 metros na superfície. A **resolução espectral** é definida por três características: o número de bandas, a largura do comprimento de onda de cada banda, e onde cada uma está posicionada no espectro. A largura da banda, vai definir por exemplo as feições de absorção de cada material, para aquela região do espectro. A intensidade da radiância da área de cada pixel é medida pela **resolução radiométrica**. O sinal que o sensor recebe é quantizado em valores digitais (*bits*), ou seja, essa resolução definirá quantos tons de cinza uma imagem consegue representar. Por último, a **resolução temporal** refere-se a frequência que um sensor revisita uma área, gerando imagens periódicas muito importantes para análises temporais. ([MENESES; ALMEIDA, 2012](#))

## 2.3 Aprendizado de Máquina

é uma área da inteligência artificial que se refere ao desenvolvimento de métodos que otimizam sua performance iterativamente aprendendo com dados. MITCHELL (1997, p.2) define como:

Um programa de computador é orientado a aprender da experiência  $E$ , com a uma tarefa  $T$  e uma medida de performance  $P$ , se sua performance em  $T$ , medida por  $P$ , melhora com a experiência  $E$ .

Os diversos métodos de AM podem ser categorizados em diversos critérios. Se no problema em questão é apresentado um conjunto de dados em que se sabe o resultado correto das predições, é chamado de **aprendizado supervisionado**. Caso não se tenha informações sobre os resultados, o problema é denominado como **aprendizado não supervisionado**.

Dado um conjunto de testes, o objetivo de um problema de aprendizado supervisionado é aprender uma função  $h(X|Y)$ , sendo  $h(x)$ , chamada de hipótese, um “bom” preditor do valor correspondente de  $y$ .

Dentro dos classificadores supervisionados, pode-se dividir em regressão e classificação. Em um **problema de regressão**, há a previsão de um resultado dentro de uma saída contínua, ou seja, é necessário mapear as variáveis de entrada em uma função contínua. No caso do **problema de classificação**, o objetivo é a previsão de um resultado em uma saída discreta, ou seja, mapeiam-se variáveis de entrada em categorias. (NG, 2017)

O foco deste trabalho foi o estudo de métodos classificadores supervisionados, a fim de predizer classes espectrais que representam diferentes alvos de uso e cobertura da terra. Portanto, segue uma explicação das principais técnicas utilizadas segundo a literatura, bem como as que foram implementadas.

### 2.3.1 Aprendizado supervisionado

No caso em que se assume que  $p(x|y)$  segue uma distribuição específica (gauss por exemplo), o método é chamado de **paramétrico**, pois é preciso estimar os parâmetros do modelo preditor. No caso dos métodos **não paramétricos**, não se utilizam parâmetros estatísticos para a modelagem da função. (WASKE, 2009)

#### 2.3.1.1 Alto e baixo viés

Quando há muitas variáveis na função hipótese  $h_\theta$ , corre-se o risco do modelo se ajustar bem aos exemplos de treinamento, porém, não é um bom preditor de novos exemplos, diz-se que é um problema de **alto viés**. O oposto também é problemático, se há poucas variáveis, corre-se o risco de acontecer **baixo-viés**. (NG, 2017)

### 2.3.1.2 Validação Cruzada

Validação cruzada é um método de reamostragem onde dividi-se o conjunto de dados repetidamente em conjuntos de treinamento, usados para ajustar o modelo, e conjuntos de teste, usados para verificar o desempenho das previsões. A validação é feita utilizando medidas de análise de precisão, e o resultado é um modelo preditor com viés reduzido, ou seja, tem maior capacidade de generalizar novos dados (LOVELACE; NOWOSAD; MUENCHOW, 2019; JAMES, 2013).

### 2.3.2 Aprendizado de Máquina e Sensoriamento remoto

Em sensoriamento remoto, algoritmos preditivos focam em classificações de cobertura da terra. Nesse contexto o algoritmo aprende a diferenciar tipos de padrões complexos, no caso, classes de cobertura da terra. (WASKE, 2009)

Classificadores não paramétricos aceitam diversos tipos de dados de treinamento de entrada, além de não fazerem suposições sobre a distribuição dos dados, que são características desejáveis para o problema. (MAXWELL; WARNER; FANG, 2018)

Quando se fala em classificação de imagens e reconhecimento de padrões, pode-se acrescentar mais um critério de divisão dos métodos. Se são utilizadas as informações espectrais de cada *pixel* de treinamento para encontrar regiões homogêneas, é um classificador classificador ***pixel a pixel***. Outro caso, é quando se realiza um agrupamento de *pixels* por métodos de segmentação de imagens em grupos que serão unidades a serem classificadas, então é um classificador **por região**, também conhecido como Object-based Image Analysis (OBIA), ou quando se fala em dados geográficos, *Geographic Object-Based Image Analysis* (GEOBIA). (MENESES; ALMEIDA, 2012; LU; WENG, 2007)

### 2.3.3 Algoritmos

Dois algoritmos de classificação foram utilizados neste trabalho: Máquina de Vetores de Suporte (MVS), ou *Support Vector Machine* e Floresta Aleatória (FA), ou *Random Forest*.

#### 2.3.3.1 Máquina de Vetores Suporte

Na classificação paramétrica, o objetivo é definir um espaço de características para cada classe. No caso da MVS (não paramétrica), o foco está apenas nos exemplos de treinamento que estão próximos do **limite de decisão** (*decision boundary*) ótimo que separa as classes. Estes exemplos definem os **vetores de suporte**. (MAXWELL; WARNER; FANG, 2018)

O objetivo é achar o limite de decisão ótimo entre duas classes, maximizando a margem entre os vetores de suporte. Originalmente, a MVS foi feita para identificar um limite de decisão linear (definindo um *hyperplano*), porém, essa limitação foi resolvida projetando o espaço de

características para uma dimensão maior. Essa projeção é feita com uma função denominada de **núcleo** (ou *kernel*). Em sensoriamento remoto, as funções de núcleo mais utilizadas são *Radial Basis Function* e também a polinomial. (MAXWELL; WARNER; FANG, 2018)

### 2.3.3.2 Florestas Aleatórias

O método de **Floresta Aleatória** (FA) é baseado em **Árvore de Decisão** (AD). Uma AD é definida como cortes recursivos nos dados de entrada. As divisões são feitas repetidamente, criando novas ramificações (como um tronco de uma árvore), sendo que ao chegar em uma “ponta” (ou folhas), é definida uma classe. Uma AD é um conceito simples de se entender e visualizar, e também podem ser boas preditoras, porém, correm o risco de se ajustarem bem demais para um conjunto de treinamento, caindo no problema do alto viés. (MAXWELL; WARNER; FANG, 2018)

As FA's utilizam em conjunto um número definido de AD's. Uma classe será definida a partir do “voto” da maioria das árvores presentes na floresta. Essa abordagem supera o problema de alto viés de uma única AD, chegando assim mais perto de uma solução global. (MAXWELL; WARNER; FANG, 2018)

Esse conceito é ainda ampliado: cada árvore é treinada com um único subconjunto de teste e variáveis, gerados aleatoriamente. Essa combinação significa que uma única árvore será menos precisa, porém, também estará menos correlacionada com todas as outras, tornando o conjunto mais confiável. (MAXWELL; WARNER; FANG, 2018)

### 2.3.4 Avaliação de Desempenho

Para verificar a acurácia após realizada a classificação, um método comumente empregado em sensoriamento remoto é o cálculo da matriz de erro (LU; WENG, 2007), utilizada para comparação entre os dados de referência e os dados classificados. Alguns fatores que podem influenciar na acurácia podem ser: a complexidade do terreno; o algoritmo utilizado; número de classes; conjunto de dados que representa a verdade (MENESES; ALMEIDA, 2012), que podem ser obtidos por exemplo através de outras classificações ou validação em campo.

A partir da matriz de erro são calculados índices de validação. Um deles é a avaliação geral, calculado a partir da divisão entre a soma dos elementos da diagonal principal (elementos classificados corretamente) e a soma do total de pontos. Outro índice muito utilizado é o *kappa*, proposto por LANDIS e KOCH (1977) que varia de 0 até 1, onde: 0 – 0,2 = ruim; 0,2 – 0,4 = razoável; 0,4 – 0,6 = boa; 0,6 – 0,8 = muito boa; e 0,8 – 1,0 = excelente. (MENESES; ALMEIDA, 2012)

# 3 Metodologia

## 3.1 Ferramentas e Dados

### 3.1.1 Linguagem R

R é uma linguagem e ambiente multiplataforma para análise estatística com ferramentas gráficas avançadas. A linguagem implementa o paradigma da programação funcional, bem como o da orientação a objetos. O projeto R é distribuído sob a [Licença Pública Geral \(GPL\)](#) do projeto GNU ([R Core Team, 2019](#)). O GNU foi criado como uma reação aos *softwares* proprietários e de código fonte fechado, marcando o início do movimento do *software* livre, ou seja, programas de computador com código fonte aberto e que está disponível para que qualquer um o estude, copie, modifique e o redistribua([TORRES, 2013](#)).

#### 3.1.1.1 Ambiente R

O desenvolvimento e crescimento da comunidade em torno do R se deu pela sua capacidade de integração com outros *softwares*, facilitando assim, por exemplo, a integração com diversas bibliotecas SIG. O uso da linguagem torna-se mais amigável com o Ambiente de Desenvolvimento Integrado (*IDE*) chamado *Rstudio*, que possui recursos como painéis de visualização interativos, acesso a documentação dos pacotes diretamente pela linha de comando ou painel, criação e manutenção de projetos, entre outros. ([LOVELACE; NOWOSAD; MUENCHOW, 2019](#))

Através da linha de comando, é muito simples instalar um pacote ou então acessar sua documentação, que é um dos pontos fortes do R. Todos os pacotes estão armazenados em uma rede de servidores chamada CRAN, ou *Comprehensive R Archive Network*, onde a comunidade desenvolvedora mantém um padrão de documentação muito bem organizado. Para que um pacote esteja disponível no CRAN, é necessário por exemplo um manual com referências teóricas para cada método implementado, e geralmente possuem exemplos simples e intuitivos, que também podem ser acessados com facilidade pela linha de comando.

#### 3.1.1.2 Rmarkdown

*Markdown* é uma linguagem de marcação simples cujo texto é convertido para o [Linguagem de Marcação de Hipertexto \(HTML\)](#). O *RMarkdown* é uma adaptação do *markdown* para o R que em conjunto com outros pacotes, pode ser facilmente convertida para formatos como [Portable Document Format \(PDF\)](#), arquivos de texto, [HTML](#), apresentação de slides, entre outros. Sua importância está na divulgação de resultados de análises que poderão ser facilmente reproduzidas por outras pessoas.

Com o *RMarkdown* é possível escrever texto no formato LaTeX, adicionar trechos de código (*chunks*) que por sua vez imprimirão seu resultado, seja ele um gráfico, uma tabela ou texto. De maneira integrada ao RStudio, os *chunks* também possuem suporte a outras linguagens como Python.

### 3.1.1.3 Bibliotecas

Para a realização deste trabalho, foram utilizadas, dentro do ambiente de desenvolvimento RStudio, bibliotecas que fazem a *interface* com a [GDAL](#), uma biblioteca tradutora para dados geoespaciais *raster* e vetoriais que é utilizada por muitos *software* de [SIG](#).

## 3.1.2 Google Earth Engine

O *Google Earth Engine* é uma plataforma de processamento geoespacial baseada em nuvem, feito principalmente para análises de dados ambientais em escala planetária. O acesso a plataforma foi utilizado para a obtenção e processamento de dados de treinamento. ([GORELICK, 2017](#))

## 3.1.3 Iniciativas

### 3.1.3.1 Mapbiomas

O Projeto de Mapeamento Anual da Cobertura e Uso do Solo do Brasil é uma iniciativa que envolve uma rede colaborativa com especialistas nos biomas, usos da terra, sensoriamento remoto, SIG e ciência da computação que utiliza processamento em nuvem e classificadores automatizados desenvolvidos e operados a partir da plataforma Google Earth Engine para gerar uma série histórica de mapas anuais de cobertura e uso da terra do Brasil. ([MAPBIOMAS, 2018](#))

O MapBiomas é uma plataforma aberta e colaborativa com uma metodologia de baixo custo que pode ser aplicada em diversos contextos. Grande parte deste trabalho teve essa iniciativa como referência, e também, parte dos dados que foram utilizados no modelo de classificação foram obtidos da Coleção 3.1 do projeto MAPBIOMAS (2018), através do próprio site e também pela plataforma do Google Earth Engine. Os detalhes estão explicados no próximo capítulo.

Além do mapeamento anual dos biomas brasileiros, também a outras iniciativas como o MapBiomas Alerta:

MapBiomas Alerta é um sistema de validação e refinamento de alertas de desmatamento, degradação e regeneração de vegetação nativa com imagens de alta resolução. ([MAPBIOMAS, 2018](#))

### 3.1.4 Satélites

Num contexto de Observação da Terra, a fim de monitorar os recursos terrestres, há uma grande quantidade de programas de satélites que orbitam o globo com a tarefa de fazer o imageamento da superfície terrestre.

O Instituto Nacional de Pesquisas Espaciais (INPE) possui um papel fundamental e histórico para o uso de Sensorimento Remoto em escala nacional. Foi pioneiro no desenvolvimento e formação nas áreas de interpretação de imagens e processamento digital ([MENESES; ALMEIDA, 2012](#)).

O instituto realiza a distribuição de imagens geradas por diversos desses programas através da sessão de Divisão de Geração de Imagens, como: o Landsat e TERRA, dos Estados Unidos; o RESOURCESAT, da Índia; o RapidEye da Alemanhã; bem como o AQUA, uma parceria entre Brasil e Japão; e ainda o CBERS, parceria entre Brasil e China que tem como objetivo o monitoramento de biomas, agricultura, crescimento urbano, gerenciamento hídrico e de desastres naturais. ([INPE](#), )

Vale o destaque para o Landsat, programa de origem norte americana gerenciado pela [Administração Nacional da Aeronáutica e Espaço \(NASA\)](#) e o [Serviço Geológico dos Estados Unidos \(USGS\)](#), que realizou uma série de lançamentos desde a década de 1970. Em um contexto, vindo da década anterior, da corrida espacial e as primeiras imagens da Terra capturadas por satélites, a história desse programa se confunde com o desenvolvimento das técnicas de SR e interpretação de imagens digitais. Também há o programa de Observação da Terra da União Europeia, o Copernicus, desenvolvido em parceria com a [Agência Espacial Europeia \(ESA\)](#), que possui a missão dos satélites Sentinel, com características semelhantes as do Landsat (ambos possuem resolução espacial média, por exemplo).

É importante destacar também o advento dos nano e microsatélites, que geralmente carregam sensores com uma resolução espacial maior e menor custo de lançamento, porém, com menor resolução radiométrica. Lembrando que cada característica dos sensores possuem diferentes fins de aplicações.

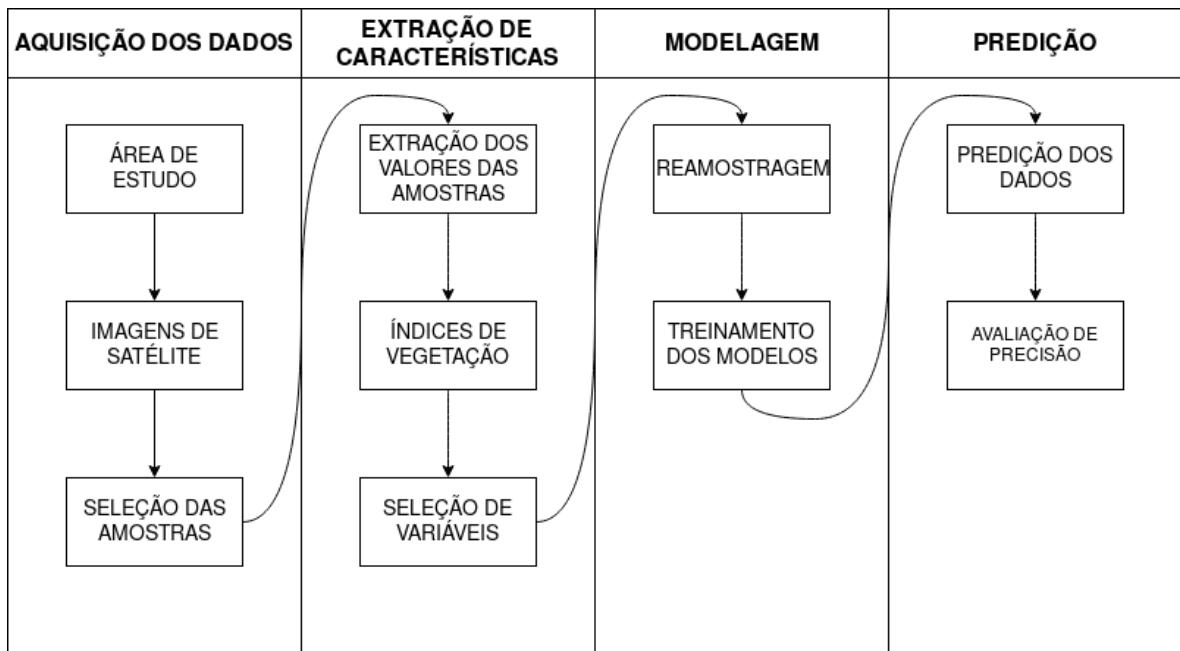
Os dados dos satélites do programa *Landsat* e *Sentinel* possuem acesso aberto desde 2008 e 2013, respectivamente. Estes são marcos importantes num contexto de Observação da Terra, gerando demanda para a computação em nuvem, que resolvem problema de processamento e armazenamento, enquanto o usuário pode focar no desenvolvimento do algoritmo. ([WULDER; COOPS, 2014](#))

## 3.2 Etapas

De acordo com LU e WENG (2007)([LU; WENG, 2007](#)), o sucesso da classificação de dados de SR em um mapa temático depende de fatores como: complexidade da área de

estudo, seleção dos dados, abordagens de processamento de imagens e seleção de sistema de classificação apropriado. Com base nos artigos de LU e WENG (2007) e MAXELL, WARNER E FANG (2018), foi elaborado o seguinte processo descrito no fluxograma da [Figura 6](#).

Figura 1 – Fluxograma do processo de desenvolvimento em etapas



Fonte: Elaborado pelo autor

# 4 Desenvolvimento

## 4.1 Etapas

### 4.1.1 Aquisição dos dados

#### 4.1.1.1 Área de Estudo

Primeiro foi feito o *download* de um arquivo no formato *shapefile*, da malha territorial do estado de São Paulo, com divisão por município, disponibilizado pelo [Instituto Brasileiro de Geografia e Estatística \(IBGE\)](#) (IBGE, 2018). A partir desse arquivo, foi selecionado apenas o polígono que representa o município de Bauru. A biblioteca no R utilizada para tal operação foi a *sf* (PEBESMA, 2018).

#### 4.1.1.2 Imagens de Satélite

Os dados do sensor Sentinel-2/MSI podem ser obtidos gratuitamente, mediante apenas a um cadastro realizado no *Copernicus Open Access Hub* ([ESA](#), ). No R, há diversas bibliotecas que fazem a interface com a API do *hub*, para este trabalho foi escolhida a *getSpatialData* ([SCHWALB-WILLMANN, 2018](#)).

O sensor multiespectral do Sentinel-2 possui as características descritas na [Tabela 1](#).

Tabela 1 – Características das bandas do sensor multiespectral MSI do Satélite Sentinel-2

banda	nome	lambda	resolucao
B1	Aerosol	0.44	60
B2	Azul	0.49	10
B3	Verde	0.56	10
B4	Vermelho	0.67	10
B8	NIR	0.84	10
B5	Red edge 1	0.70	20
B6	Red edge 2	0.74	20
B7	Red edge 3	0.78	20
B9	Vapor d'água	0.94	60
B10	Cirrus	1.38	60
B11	SWIR 1	1.61	20
B12	SWIR 2	2.19	20
B8A	Red edge 4	0.86	20

Fonte: Elaborado pelo autor

A partir dos filtros por data e cobertura de nuvem, foram selecionadas duas imagens e então realizada uma composição entre elas para que recobrissem a área de interesse completa-

mente. Posteriormente, foi feito um redimensionamento das bandas de 20m para 10m, para que fossem compiladas em uma única pilha de *raster*. Selecionando as bandas B04, B03 E B02, é possível visualizar a imagem nas cores vermelho, verde e azul, respectivamente, como visto na [Figura 2](#). A biblioteca utilizada que manipula arquivos *raster* no R é a *raster* ([HIJMANS, 2019](#)).

Figura 2 – Composição com as bandas vermelha, verde e azul do Sentinel-2/MSI para a região da região do município de Bauru



Fonte: Elaborado pelo autor

#### 4.1.1.3 Seleção de Amostras

Para a seleção das amostras de treinamento, era preciso que os dados representassem bem as classes de uso e cobertura do solo dentro do município. De acordo com o SEMMA (2015) e CAVASSAN (2013), as principais unidades fitogeográficas que ocorrem no município de bauru são as formações de Floresta Estacional Semidecidual e de Cerrado, apesar de a cobertura primitiva já tenha sido muito reduzida, assim como o Cerrado no estado de São Paulo, que já representou uma área maior.

Portanto, duas decisões foram tomadas: as amostras de treinamento seriam provindas do bioma Cerrado; como a extensão o Cerrado é grande, a fim de viabilizar o processamento, apenas a região de cerrado dentro do estado de São Paulo foi selecionada.

Para isso, foi utilizado um *raster* da coleção 3.1 do projeto MapBiomas ([MAPBIOMAS, 2018](#)) contendo os valores de cada classe para cada pixel, definidos pela metodologia aplicada

pelo projeto. As classes utilizadas estão referenciadas na Tabela 2.

Fonte: Elaborado pelo autor

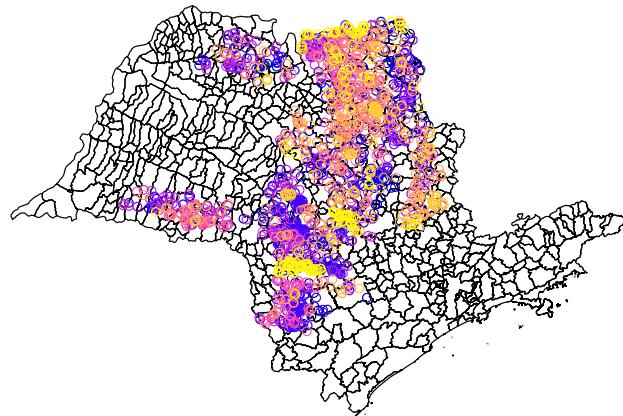
Tabela 2 – Classes de uso e cobertura da terra da coleção 3.1 do MapBiomas para o bioma Cerrado

value	name.full
3	1.1.1. Formação Florestal
4	1.1.2. Formação Savanica
9	1.2. Floresta Plantada
12	2.2. Formação Campestre
15	3.1. Pastagem
19	3.2.1. Cultura Anual e Perene
20	3.2.2. Cultura Semi-Perene
21	3.3. Mosaico de Agricultura e Pastagem
24	4.2. Infraestrutura Urbana
33	5.1 Rio, Lago e Oceano

Fonte: Elaborado pelo autor

Esse *raster* foi recortado para o tamanho do polígono do estado de São Paulo. Em seguida, foi feito a seleção de duzentas amostras aleatórias de cada classe, representadas na Figura 3, e em seguida foi exportado para um arquivo *shapefile*.

Figura 3 – Amostras aleatórias de pontos contidos na região de cerrado do Estado de São Paulo



Fonte: Elaborado pelo autor

#### 4.1.2 Extração de Características

A etapa da extração de características envolve a seleção das variáveis usadas no processo de classificação, que podem ser as assinaturas espectrais das bandas, índices de vegetação, informação de textura, entre outras ([LU; WENG, 2007](#))

##### 4.1.2.1 Extração dos Valores das Amostras

A próxima etapa, foi a de extração dos valores das amostras. Cada coordenada gerada na última etapa, contém a informação da classe que ela representa, então o próximo passo foi de extrair os valores de cada banda naquele ponto, e então compor a tabela com as variáveis preditoras para a modelagem de aprendizado de máquina.

Nessa etapa, foi utilizado o ambiente do *Google Earth Engine* ([GORELICK, 2017](#)) , onde o *shapefile* foi carregado, foi feita uma composição com imagens do satélite Sentinel-2, filtradas por baixa porcentagem de nuvens e para a região em que os pontos estão compreendidos. Em seguida, para cada ponto, foi extraído os valores das bandas, como mostra a Tabela 3, e exportado para um arquivo *geojson*, que foi importado de volta ao ambiente do RStudio.

Tabela 3 – Quantidade de exemplos de treinamento para cada classe

	Classes	Frequência
1	agricultura_pastagem	197
2	cultura_anual_perene	146
3	cultura_semi_perene	200
4	floresta_plantada	200
5	formacao_campestre	140
6	formacao_florestal	199
7	formacao_savanica	91
8	infra_urbana	105
9	pastagem	198
10	rio_lago_oceano	159

Fonte: Elaborado pelo autor

É importante notar que para algumas classes, na execução da amostragem, foram selecionados menos do que duzentas amostras, pois haviam poucos pixels representantes.

Em um sistema de classificação, antes da extração dos valores dos pontos das amostras, seria interessante que a imagem passasse por um processo de pré processamento. Essa etapa incluiria a aplicação de técnicas de processamento de imagens para correções atmosféricas e geométricas, eliminação de ruídos, entre outras (LU; WENG, 2007). Contudo, uma característica das imagens do Sentinel-2/MSI, é de que há uma série de etapas de pré processamento, como as mencionadas, que são realizadas antes das imagens serem disponibilizadas.

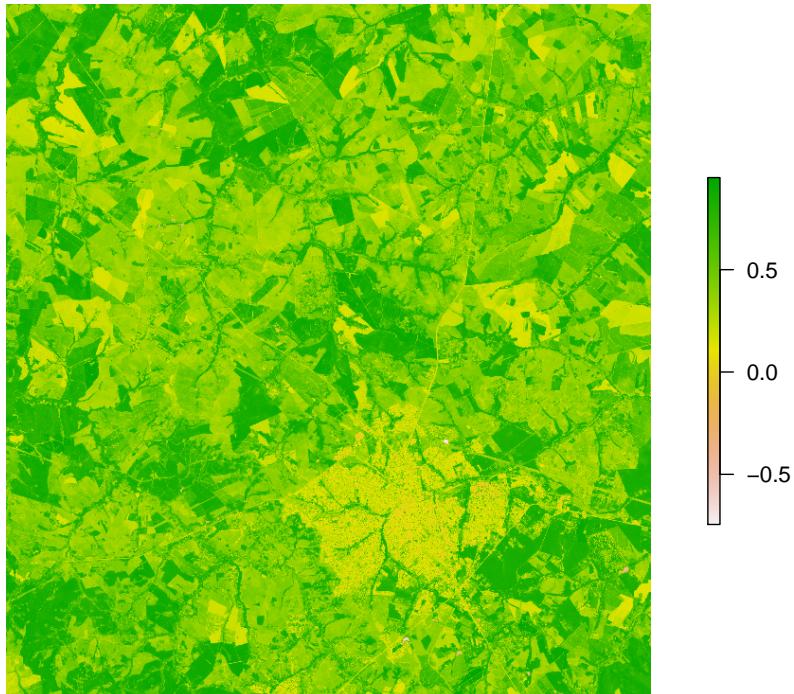
#### 4.1.2.2 Índices de Vegetação

Os índices de vegetação são obtidos através de operações aritméticas entre as bandas. Possuem a característica de realçar as variações de densidade da cobertura vegetal (MENESES; ALMEIDA, 2012). O NDVI é provavelmente o mais utilizado, ou pelo menos o mais conhecido. Esse índice possui a característica de evidenciar áreas da vegetação fotossinteticamente mais ativas. O cálculo do NDVI, que varia de 0 a 1, é dado por:

$$NDVI = \frac{NIR - RED}{NIR + RED} \quad (4.1)$$

Onde *NIR* é o valor para a banda na região do infravermelho próximo (a banda 8 do Sentinel-2/MSI) e *RED* é a banda na região do vermelho (banda 4). A Figura 4 mostra uma representação do cálculo do .

Figura 4 – Representação do cálculo do índice NDVI



Fonte: Elaborado pelo autor

#### 4.1.2.3 Seleção de Variáveis

Como variáveis preditoras, foram selecionadas as bandas “SWIR1”, “SWIR2”, “Azul”, “Verde”, “Vermelho”, “NIR”, “RE1”, “RE2”, “RE3”, “RE4” e uma nova coluna contendo o cálculo do NDVI para cada ponto da amostra.

#### 4.1.3 Modelagem

O pacote *caret* (*Classification And REgression Training*) (KUHN, 2019) R contém funções para otimizar o processo de treinamento de modelos para problemas de regressão e classificação. Ele funciona como um agregador de diversos pacotes que contém métodos de aprendizado estatístico, fazendo a interface entre as funções disponíveis no pacote para controle do processo de treinamento e avaliação de resultados.

#### 4.1.3.1 Reamostragem

A primeira etapa para a criação dos modelos, foi a divisão do conjunto de dados em dois subconjuntos: um de treinamento (75%) e outro de teste (25%). Utilizando a função de validação cruzada do *caret*, na criação do modelo, o conjunto de treinamento é reparticionado em 10, e então, a partir dessa repetição no ajuste, é escolhido o conjunto de parâmetros que teve melhor resultado para aquele método.

#### 4.1.3.2 Treinamento

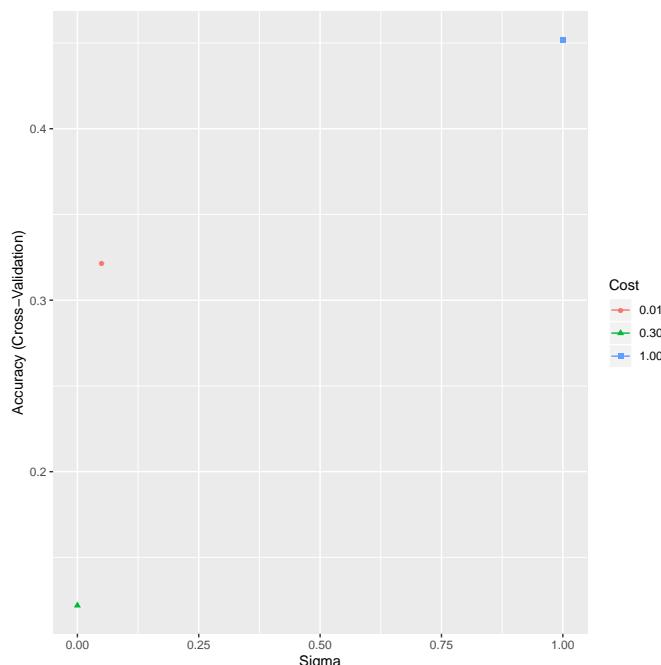
Na etapa do treinamento dos modelos, para a seleção dos parâmetros utilizados em cada algoritmo foi utilizada a técnica da validação cruzada *k-fold* (não é o mesmo que na última etapa), onde os dados de treinamento são subdivididos em  $k$  subconjuntos e então, o modelo é calculado  $k$  vezes, e então, o resultado é comparado para seleção do parâmetro (ou conjunto) que obteve melhor resultado, no caso, o método de comparação utilizado foi a acurácia geral a partir da matriz de confusão gerada para cada modelo, como mostra a Figura 5, Figura 6 e Figura 7.

O primeiro modelo foi treinado com o algoritmo MVS (*svmRadial* no *caret*), utilizado como núcleo a função de base radial dada por:

$$K(x, y) = \exp\left(-\frac{\|x - y^2\|}{\sigma^2}\right) \quad (4.2)$$

Após o treinamento do modelo, foram selecionados os parâmetros  $C = 1$  e  $Sigma = 1$ .

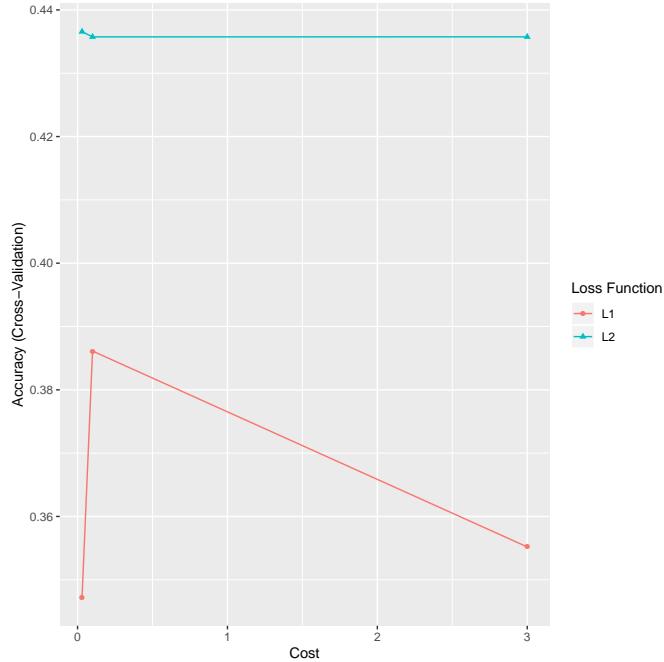
Figura 5 – Parâmetros selecionados após treinamento com o algoritmo maquina de vetores suporte com núcleo radial



Fonte: Elaborado pelo autor

O próximo modelo, foi MVS também, porém linear (*svmLinear3* no *caret*), ou seja, sem função de núcleo e com regularização, que significa aplicar um custo de penalização nos parâmetros  $\theta$ . Os parâmetros selecionados foram  $C = 0.03$  e Loss = L2.

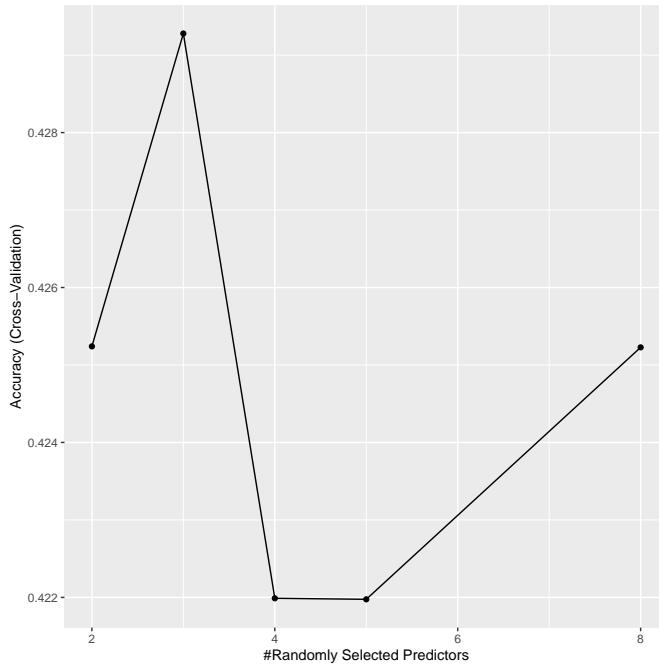
Figura 6 – Parâmetros selecionados após treinamento com o algoritmo maquina de vetores suporte linear com regularização



Fonte: Elaborado pelo autor

Por último, foi computado também um modelo utilizando Florestas Aleatórias (*rf* no *caret*), com o parâmetro selecionado  $mtry = 3$ .

Figura 7 – Parâmetros selecionados após treinamento com o algoritmo florestas aleatórias



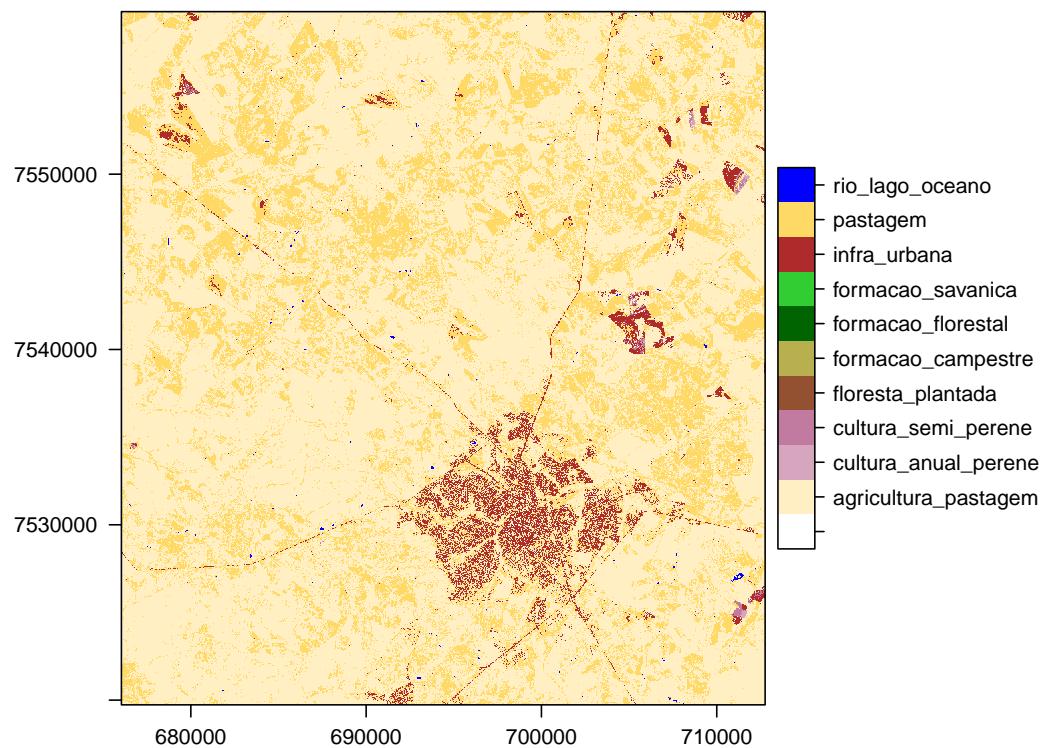
Fonte: Elaborado pelo autor

## 4.2 Resultados

### 4.2.1 Predição dos Dados

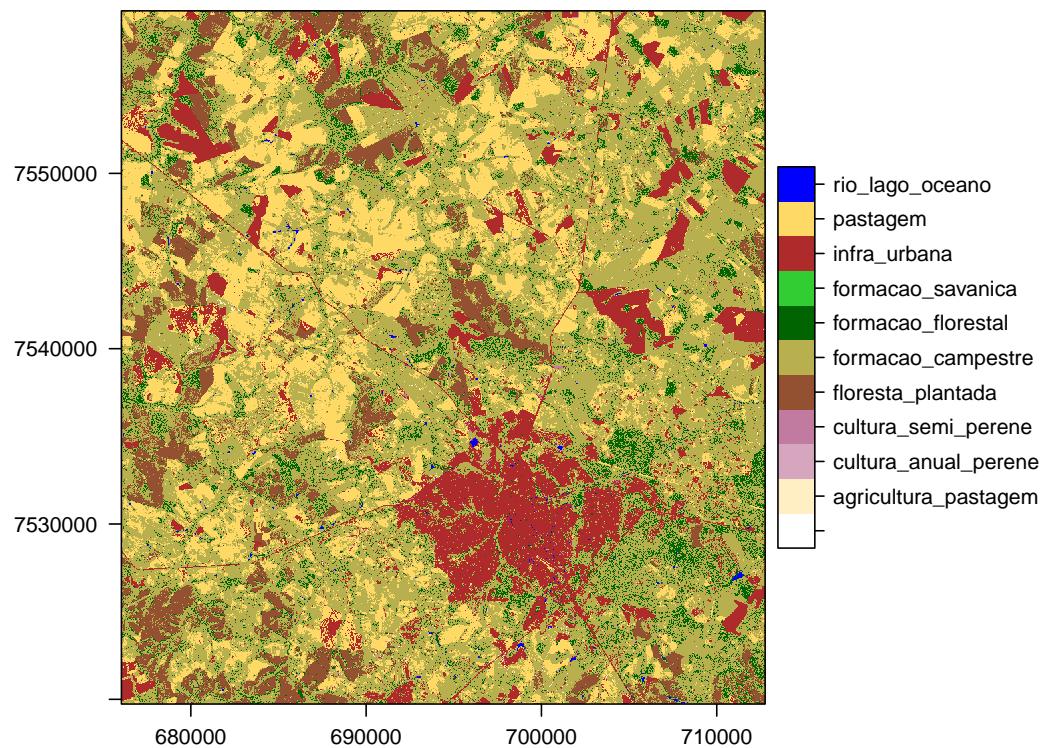
Para cada modelo treinado, foi realizada uma predição com a imagem selecionada anteriormente do sensor Sentinel-2/MSI para a região do município de Bauru, resultando em *rasters* com os valores para cada classes usadas para a criação dos modelos, demonstrados na Figura 8, Figura 9 e Figura 10.

Figura 8 – Mapa com classes de uso e cobertura da terra obtido a partir da predição do modelo utilizando MVS com núcleo radial



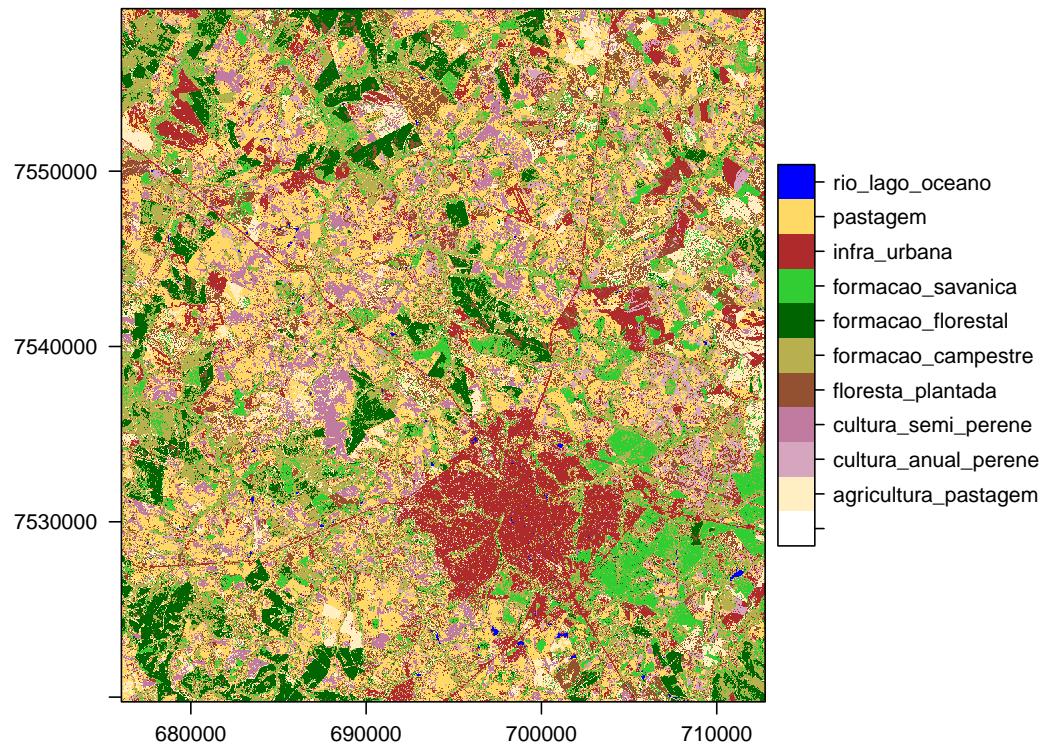
Fonte: Elaborado pelo autor

Figura 9 – Mapa com classes de uso e cobertura da terra obtido a partir da predição do modelo utilizando MVS linear com regularização



Fonte: Elaborado pelo autor

Figura 10 – Mapa com classes de uso e cobertura da terra obtido a partir da predição do modelo utilizando FA's

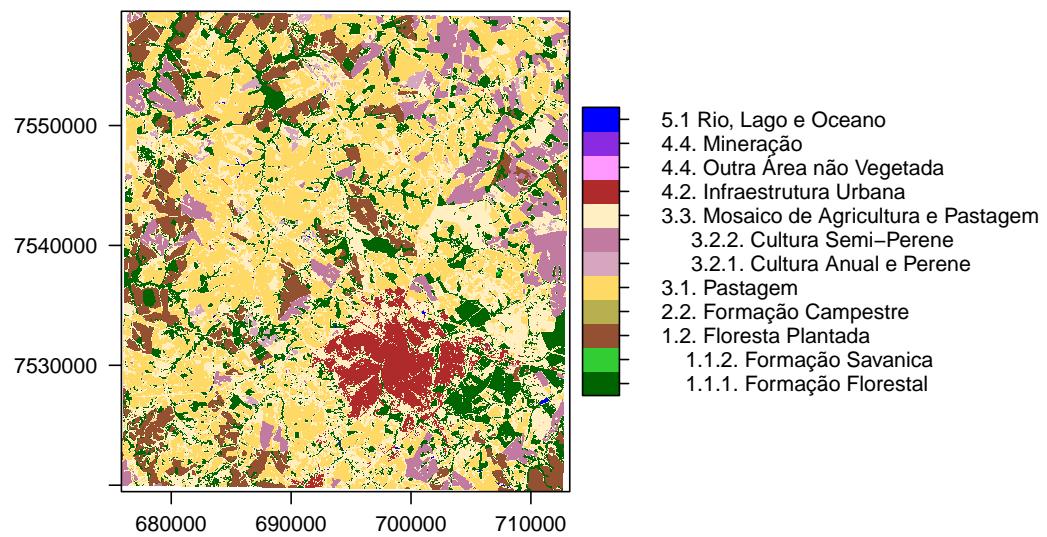


Fonte: Elaborado pelo autor

#### 4.2.1.1 Avaliação de Precisão

Para a avaliação da qualidade das predições realizadas, foi computado a matriz de confusão de cada uma, a partir dos índices calculados a partir da matriz de erro. A matriz de erro por sua vez foi montada a partir da seleção de 100 pontos de amostras para cada classe dos *rasters* obtidos na etapa da predição, e em seguida, os valores dos pontos foram extraídos de um conjunto verdade, que é um *raster* da mesma região com a classificação realizada pela iniciativa MapBiomas e disponibilizada para download. Para comparação visual com os mapas apresentados anteriormente, o *raster* do MapBiomas está representado na Figura 11.

Figura 11 – Mapa com classes de uso e cobertura da terra obtido a partir da classificação realizada pela iniciativa MapBiomas



Fonte: Elaborado pelo autor

Com destaque para os índices *kappa* e *accuracy* (acurácia geral), na Tabela 4 a seguir estão demonstrados os valores dos resultados:

Tabela 4 – Tabela de avaliação de precisão das previsões realizadas

	Modelo	Kappa	Acuracia
1	MVS linear	0.30	0.38
2	MVS radial	0.27	0.39
3	Florestas Aleatorias	0.13	0.23

#### 4.2.1.2 Considerações

Comparando visualmente os mapas obtidos com o mapa do MapBiomas além da matriz de erro gerada, pode-se notar algumas classes que sofreram certa confusão, como é

o caso da *formação\_florestal* na predição com FA, que na verdade representariam a classe *floresta\_plantada*. Levando em consideração que na metodologia do MapBiomas aplicada nos dados de treinamento são utilizadas outras variáveis de predição (como outros índices de vegetação), é provável que apenas as bandas selecionadas e o NDVI não seja suficiente para separar com precisão todas as classes utilizadas, como LU e WENG (2007) apontam no artigo o qual este trabalho se baseou. Ainda no caso dessa predição, a acurácia foi prejudicada pelo fato dos *pixels* apresentarem pouca homogeneidade nos espaços vizinhos. A predição realizada com o algoritmo MVS com núcleo não linear, apesar de apresentar uma acurácia mediana, resultou em poucas classes que predominaram. A classificação que obteve melhor resultado foi a que utilizou MVS linear para criação do modelo.

# 5 Conclusão

O estudo de metodologias de análise e inferência de dados gerados a partir da Observação da Terra permite que seja realizado o monitoramento dos recursos terrestres. A característica dos sistemas de Sensoriamento Remoto permite o imageamento da superfície terrestre em escala global e de maneira frequente. Nesse contexto, métodos de classificação de imagens digitais tornam-se essenciais para extração de dados que poderão ser usados em outros estudos de diversas áreas.

Além disso, atualmente, plataformas baseadas em nuvem - que computam os dados em servidores remotos e com alta capacidade de processamento - viabilizam a coleta, distribuição e processamento de dados coletados por Sensoriamento Remoto. Com a utilização das imagens de satélite distribuídas gratuitamente bem como as ferramentas e softwares livres disponíveis, juntamente (porém de maneira opcional) com as plataformas de nuvem, as aplicações tornam-se acessíveis.

Métodos de classificação a partir de aprendizado de máquina vêm sendo amplamente utilizados em diversas áreas do conhecimento. Em Sensoriamento Remoto, esses métodos são interessantes por fazerem pouca ou nenhuma suposição dos dados e dão bons resultados. Neste trabalho, foi realizado um estudo de um processo de classificação, passando pela aquisição dos dados, extração de características, modelagem e por fim, avaliação de precisão.

Pôde-se perceber como cada uma das etapas exige uma série de considerações precisam ser avaliadas para se obterem bons resultados, e muitas delas exigem conhecimento especializado. Portanto, como resultado final, este trabalho não alcançou todos os objetivos propostos de maneira satisfatória.

## 5.1 Trabalhos Futuros

A partir de um modelo preditivo bem ajustado, uma aplicação interessante é a análise temporal de uma região de interesse, a partir da comparação das alterações nas classes ao longo de um determinado período. Outras aplicações podem ser por exemplo: planejamento urbano, monitoramento ambiental, cruzamento com outros dados, entre outras. No escopo deste trabalho, alguns ajustes e aprofundamento nas etapas também podem ser realizados como por exemplo: utilização de outros dados que não os do MapBiomass, a fim de comparação; seleção de amostras de outras regiões; etapas adicionais de pré processamento para extração de valores das amostras de treinamento; extração de outras características, como outros índices; utilização de outros algoritmos como Redes Neurais para comparação.

# Referências

- ESA. *Sentinel-2 MSI Technical Guide*. Disponível em: <<https://web.archive.org/web/20190517092002/https://sentinels.copernicus.eu/web/sentinel/technical-guides/sentinel-2-msi>>. Acesso em: outubro de 2019.
- FLORENZANO, T. G. Imagens de satélite para estudos ambientais. In: *Imagens de satélite para estudos ambientais*. [S.I.]: Oficina de Textos, 2002.
- GORELICK, N. Google earth engine: Planetary-scale geospatial analysis for everyone. *Remote Sensing of Environment*, Elsevier, 2017.
- GREGORIO, A. D. Land cover classification system: Classification concepts. software version 3. FAO, 2016.
- HIJMANS, R. J. *raster: Geographic Data Analysis and Modeling*. [S.I.], 2019. R package version 3.0-7.
- IBGE. *Malha municipal digital do Brasil: situação em 2018*. 2018. Disponível em: <[ftp://geoftp.ibge.gov.br/organizacao\\_do\\_territorio/malhas territoriais/malhas\\_municipais/municipio\\_2018/UFs/SP/sp\\_municipios.zip](ftp://geoftp.ibge.gov.br/organizacao_do_territorio/malhas territoriais/malhas_municipais/municipio_2018/UFs/SP/sp_municipios.zip)>. Acesso em: agosto de 2019.
- IBGE. *Acesso e uso de dados geoespaciais*. [S.I.]: IBGE, Coordenação de Cartografia, 2019.
- INPE. *Divisão de Geração de Imagens*. Disponível em: <<http://www.dgi.inpe.br/>>. Acesso em: maio de 2019.
- JAMES, G. *An introduction to statistical learning*. [S.I.]: Springer, 2013. v. 112.
- KUHN, M. *caret: Classification and Regression Training*. [S.I.], 2019. R package version 6.0-84.
- LOVELACE, R.; NOWOSAD, J.; MUENCHOW, J. *Geocomputation with R*. [S.I.: s.n.], 2019.
- LU, D.; WENG, Q. A survey of image classification methods and techniques for improving classification performance. *International Journal of Remote Sensing*, Taylor Francis, v. 28, n. 5, p. 823–870, 2007.
- MAPBIOMAS, P. *Projeto MapBiomass – Coleção 3.1 da Série Anual de Mapas de Cobertura e Uso de Solo do Brasil*. 2018. Disponível em: <<https://web.archive.org/web/20190828220533/http://mapbiomas.org/>>. Acesso em: agosto de 2019.
- MATHIEU, P. P.; AUBRECHT, C. *Earth Observation Open Science and Innovation*. [S.I.]: Springer, 2018. v. 15.
- MAXWELL, A. E.; WARNER, T. A.; FANG, F. Implementation of machine-learning classification in remote sensing: An applied review. *International Journal of Remote Sensing*, Taylor & Francis, v. 39, n. 9, p. 2784–2817, 2018.
- MENESES, P. R.; ALMEIDA, T. d. Introdução ao processamento de imagens de sensoriamento remoto. *Universidade de Brasília, Brasília*, 2012.

- NG, A. Y. *Machine Learning*. 2017. MOOC offered by the Stanford University. Disponível em: <<https://pt.coursera.org/learn/machine-learning>>. Acesso em: maio de 2019.
- PEBESMA, E. *OpenEO: a GDAL for Earth Observation Analytics*. 2016. Disponível em: <<https://web.archive.org/web/20190427030002/https://www.r-spatial.org/2016/11/29/openeo.html>>. Acesso em: maio de 2019.
- PEBESMA, E. Simple Features for R: Standardized Support for Spatial Vector Data. *The R Journal*, v. 10, n. 1, p. 439–446, 2018.
- R Core Team. *R: A Language and Environment for Statistical Computing*. Vienna, Austria, 2019. Disponível em: <<https://www.R-project.org/>>.
- ROSA, R. Geotecnologias na geografia aplicada. *Revista do Departamento de Geografia*, v. 16, p. 81–90, 2005.
- SCHWALB-WILLMANN, J. *getSpatialData: Get different kinds of freely available spatial datasets*. [S.I.], 2018. R package version 0.0.4. Disponível em: <<http://www.github.com/16eagle/getSpatialData/>>.
- TORRES, A. L. *A tecnoutopia do software livre: uma história do projeto técnico e político do GNU*. Tese (Doutorado) — Universidade de São Paulo, 2013.
- WASKE, B. Machine learning techniques in remote sensing data analysis. *Kernel methods for remote sensing data analysis*, Wiley, p. 3–24, 2009.
- WULDER, M. A.; COOPS, N. C. Satellites: Make earth observations open access. *Nature News*, v. 513, n. 7516, p. 30, 2014.