MINISTÉRIO DA CIÊNCIA, TECNOLOGIA E INOVAÇÃO
INSTITUTO NACIONAL DE PESQUISAS ESPACIAIS

sid.inpe.br/mtc-m21d/2023/12.19.20.17-TDI

INTEGRAÇÃO DE MÉTODOS DE ANÁLISE DURANTE A COLETA DE AMOSTRAS DE USO E COBERTURA DA TERRA

Abner Ernâni dos Anjos

Dissertação de Mestrado do Curso de Pós-Graduação em Computação Aplicada, orientada pelos Drs. Karine Reis Ferreira Gomes, e Gilberto Ribeiro de Queiroz, aprovada em 30 de novembro de 2023.

URL do documento original:
<http://urlib.net/8JMKD3MGP3W34T/4ADHDKE>

INPE
São José dos Campos
2023

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

MINISTÉRIO DA CIÊNCIA, TECNOLOGIA E INOVAÇÃO
INSTITUTO NACIONAL DE PESQUISAS ESPACIAIS

sid.inpe.br/mtc-m21d/2023/12.19.20.17-TDI

INTEGRAÇÃO DE MÉTODOS DE ANÁLISE DURANTE A COLETA DE AMOSTRAS DE USO E COBERTURA DA TERRA

Abner Ernâni dos Anjos

Dissertação de Mestrado do Curso de Pós-Graduação em Computação Aplicada, orientada pelos Drs. Karine Reis Ferreira Gomes, e Gilberto Ribeiro de Queiroz, aprovada em 30 de novembro de 2023.

URL do documento original:
<http://urlib.net/8JMKD3MGP3W34T/4ADHDKE>

INPE
São José dos Campos
2023

Dados Internacionais de Catalogação na Publicação (CIP)

Anjos, Abner Ernâni dos.

An59i Integração de métodos de análise durante a coleta de amostras de uso e cobertura da terra / Abner Ernâni dos Anjos. – São José dos Campos : INPE, 2023.

xxii + 117 p. ; (sid.inpe.br/mtc-m21d/2023/12.19.20.17-TDI)

Dissertação (Mestrado em Computação Aplicada) – Instituto Nacional de Pesquisas Espaciais, São José dos Campos, 2023.

Orientadores : Drs. Karine Reis Ferreira Gomes, e Gilberto Ribeiro de Queiroz.

1. Uso e cobertura da terra. 2. Aprendizado de máquina. 3. Análise de dados de observação da terra. 4. Séries temporais de imagens de satélite. I.Título.

CDU 004.942

MINISTÉRIO DA CIÊNCIA, TECNOLOGIA E INOVAÇÃO

INSTITUTO NACIONAL DE PESQUISAS ESPACIAIS

DEFESA FINAL DE DISSERTAÇÃO ABNER ERNÂNI DOS ANJOS
BANCA Nº 269/2023, REG. 865826/2021

No dia 30 de novembro de 2023, as 09h, por teleconferência, 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, com a exigência de que o trabalho final a ser publicado deverá incorporar as correções sugeridas pela Banca Examinadora, com revisão pelo(s) orientador(es).

Título: "Integração de Métodos de Análise Durante a Coleta de Amostras de Uso e Cobertura da Terra."

Membros da Banca:

Dr. Rafael Duarte Coelho dos Santos – Presidente - INPE
Dra. Karine Reis Ferreira Gomes – Orientadora - INPE
Dr. Gilberto Ribeiro de Queiroz – Orientador – INPE
Dr. Júlio César Dalla Mora Esquerdo – Membro Externo - Embrapa

Documento assinado eletronicamente por **Rafael Duarte Coelho dos Santos**, **Pesquisadora**, em 04/12/2023, às 08:40 (horário oficial de Brasília), com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Julio cesar dalla mora esquerdo (E)**, **Usuário Externo**, em 04/12/2023, às 08:55 (horário oficial de Brasília), com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Gilberto Ribeiro de Queiroz**, **Tecnologista**, em 04/12/2023, às 09:12 (horário oficial de Brasília), com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Karine Reis Ferreira Gomes**, **Pesquisadora**, em 04/12/2023, às 09:55 (horário oficial de Brasília), com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020.

A autenticidade deste documento pode ser conferida no site https://sei.mcti.gov.br/verifica.html, informando o código verificador **11522329** e o código CRC **C8BF8E06**.

Referência: Processo nº 01340.009582/2023-04 SEI nº 11522329

"A ciência é uma vela no escuro [...] Neste mundo possuído por demônios que habitamos em virtude de seres humanos, possivelmente seja isso o único que nos isola da escuridão que nos rodeia.".

Carl Sagan em
"O Mundo Assombrado pelos Demônios", 1997

AGRADECIMENTOS

Agradeço a minha mãe Esmeralda Alves Pereira dos Anjos por ser minha protetora, cuidadora, professora, conselheira, psicóloga, ouvinte, amiga e outras centenas de funções que só as mães exercem quando precisamos.

Agradeço a minha avó materna Maria José Alves que me contava diversas histórias, e ao meu avô materno Geraldo Alves Pereira que me ensinou o valor da paciência e que todo trabalho tem resultado.

Agradeço ao meu irmão mais velho Vinícius Elivelton dos Anjos por ser meu melhor amigo e ter me proporcionado uma infância repleta de imaginação e fantasia. Imaginação essa que nunca foi embora e que hoje me inspira a produzir novos contos e histórias.

Agradeço a todas as minhas madrinhas e tias de consideração, mulheres fortes que me ofereceram amparo e o aconchego de um lar com uma boa refeição, além de uma boa conversa com excelentes conselhos que apenas a experiência traz.

Agradeço a todos meus colegas de mestrado que de alguma forma me ajudaram a chegar nesta linha de chegada. Também aos que me proporcionaram momentos de descontração durante esta jornada prejudicada por episódios de tensão devido à pandemia do Covid-19.

Agradeço aos pesquisadores e membros do projeto *Brazil Data Cube* por todo o conhecimento adquirido e compartilhado. Em especial aos nomes da equipe de desenvolvedores: Fabiana Zioti que me aconselhou durante esta jornada como co-orientadora; Felipe Carvalho, Claudinei de Camargo, Gabriel Sansigolo e Raphael Costa que me proporcionaram um excelente ambiente de trabalho em equipe.

Agradeço aos meus orientadores Dra. Karine Reis Ferreira e Dr. Gilberto Ribeiro de Queiroz pelos valiosos ensinamentos, conselhos, muita paciência e complacência durante a minha experiência no mestrado.

Um agradecimento a todos que contribuíram de alguma forma para desenvolvimento deste trabalho.

RESUMO

Os métodos de aprendizado de máquina tornaram-se ferramentas amplamente usadas para o mapeamento do uso e cobertura da terra. Para obter resultados de maior acurácia, estas técnicas exigem um conjunto de amostras de treinamento rotuladas *a priori* com boa qualidade e que demonstram uma variabilidade significativa. A coleta de amostras para treinamento de modelos preditivos a partir de imagens de média e alta resolução é uma etapa custosa e pode demandar tempo, além de ser uma atividade sujeita a erros na rotulação. Além de parte do esforço ser desperdiçado com a coleta não informativa ou pela aquisição de recursos cuja contribuição é marginal. Por conta disso, há disponível uma gama de abordagens para explorar, analisar e avaliar os dados a serem usados como amostras de treinamento em modelos de aprendizado de máquina que, posteriormente, serão usados na classificação de imagens e geração de mapas de uso e cobertura da terra. Diversos métodos têm sido desenvolvidos e empregados com sucesso para analisar, avaliar e questionar a qualidade das amostras antes da classificação. Estes métodos possuem abordagens envolvendo a análise exploratória e testes com modelos preditivos para calcular valores que quantificam a relevância de certas amostras. Contudo, estes métodos exigem conhecimentos técnicos específicos de linguagem de programação e algoritmos. Entre eles, destacam-se os baseados na análise exploratória de dados, *Transferring Learning*, *Semi-Supervised Learning*, *Active Learning* e redução de ruído com *Self-organizing maps*. Este conhecimento técnico se deve a necessidade de especialização destes métodos que devem se adequar aos dados que foram coletados, todavia é discutível que não há uma forma otimizada de escolher o melhor método, e até mesmo realizar esta análise devido à quantidade de opções, abordagens e possibilidade na variação de parâmetros usados na análise. Além do mais, as ferramentas existentes para coleta não os integram para automatizar esta etapa, fazendo com que a maioria dos pesquisadores migrem dados de uma plataforma à outra. O objetivo principal deste trabalho é propor uma arquitetura de *software* para a integração de métodos de análise de amostras de uso e cobertura da terra na plataforma de coleta *TerraCollect*, buscando auxiliar na produção de amostras com boa qualidade. A arquitetura de integração foi implementada como uma extensão das funcionalidades desta plataforma. O *TerraCollect* está em desenvolvimento na infraestrutura digital do projeto *Brazil Data Cube*, que por sua vez fornece ferramentas de acesso e análise de dados de observação da Terra.

Palavras-chave: Uso e Cobertura da Terra. Aprendizado de Máquina. Análise de Dados de Observação da Terra. Séries Temporais de Images de Satélite.

INTEGRATION OF ANALYSIS METHODS DURING LAND USE AND LAND COVER SAMPLING

ABSTRACT

Machine learning methods have become widely adopted tools for land use and land cover mapping. To achieve higher accuracy, these techniques necessitate a set of *a priori* labeled training samples of high quality and significant variability. Collecting samples for predictive model training from medium to high-resolution images is a resource-intensive and time-consuming process, susceptible to labeling errors. Moreover, a portion of the effort is often squandered on uninformative collection or the acquisition of resources with marginal contributions. Consequently, a range of approaches is available for exploring, analyzing, and evaluating data to be utilized as training samples in machine learning models subsequently employed in image classification and land use/land cover map generation. Various methods have been successfully developed and employed to analyze, evaluate, and scrutinize sample quality before classification. These methods involve exploratory data analysis and tests with predictive models to calculate values quantifying the relevance of certain samples. However, these methods require specific technical knowledge of programming languages and algorithms. Notable among them are those based on exploratory data analysis, Transferring Learning, Semi-Supervised Learning, Active Learning, and noise reduction using Self-organizing Maps. This technical knowledge is due to the need for specialization of these methods, as some methods must be adapted to the data that was collected. Nevertheless, it is debatable that there is no optimized way to choose the best method, or even to conduct this analysis, given the plethora of options, approaches, and parameter variations used in the analysis. Furthermore, existing tools for collection do not integrate these methods to automate this step, leading most researchers to migrate data from one platform to another. The primary objective of this work is to propose a software architecture for integrating land use and land cover sample analysis methods into the TerraCollect collection platform, aiming to assist in producing high-quality samples. The architecture was implemented as an extension of the functionalities of this platform. TerraCollect is under development in the digital infrastructure of the Brazil Data Cube project, which, in turn, provides tools for accessing and analyzing Earth observation data.

Keywords: Land Use and Land Cover. Machine Learning Classification. Earth Observation Data Analysis. Time Series Data.

LISTA DE FIGURAS

Pág.

LISTA DE TABELAS

Pág.

LISTA DE ABREVIATURAS E SIGLAS

AED	–	Análise Exploratória de Dados
AL	–	*Active Learning*
API	–	*Application Programming Interface*
ARD	–	*Analysis-Ready Data*
BDC	–	*Brazil Data Cube*
BBOX	–	*Bounding Box*
CBERS-4	–	*China–Brazil Earth Resources Satellite 4*
CRUD	–	*Create, Read, Update and Delete*
DETER	–	Sistema de Detecção de Desmatamento em Tempo Real
EO	–	*Earth Observation*
FAO	–	*Food and Agriculture Organization of the United Nations*
GEE	–	*Google Earth Engine*
LULC	–	*Land Use and Land Cover*
IDW	–	*Inverse Distance Weighting*
INPE	–	Instituto Nacional de Pesquisas Espaciais
MODIS	–	Moderate Resolution Imaging Spectrodiameter
NASA	–	National Aeronautics and Space Administration
PRODES	–	Projeto de Monitoramento do Desmatamento na Amazônia Legal por Satélite
SOM	–	*Self-organizing maps*
SIG	–	Sistema de Informação Geográfica
sits	–	*R Package for Satellite Image Time Series Analysis*
STAC	–	*Spatial Temporal Asset Catalog*
USAID	–	*States Agency for International Development*
WFS	–	*Web Feature Service*
WLTS	–	*Web Land Trajectory Service*
WMS	–	*Web Map Service*
WTSS	–	*Web Time Series Service*

SUMÁRIO

Pág.

1 INTRODUÇÃO

Mapas de uso e cobertura da terra, em inglês *Land Use and Land Cover* (LULC), são produzidos, principalmente, a partir de imagens de sensoriamento remoto coletadas por satélites de observação da Terra (ALMEIDA et al., 2016; SIMOES et al., 2021a). Mapas LULC são instrumentos essenciais para a compreensão das ações humanas na superfície terrestre. Esses mapas são fundamentais para que órgãos governamentais desenvolvam políticas públicas relacionadas, por exemplo, à gestão de recursos naturais, à agricultura e ao planejamento urbano, e com isso prevenir impactos negativos no meio ambiente (HANSEN; LOVELAND, 2012).

As atividades humanas transformam o meio ambiente em escala regional a global, gerando mudanças que estão cada vez mais dinâmicas e imprevisíveis ao longo do tempo. Mudanças que se devem ao rápido crescimento de áreas urbanas, industriais e áreas destinadas à atividades agrícolas. O crescimento sem precedentes destas áreas é ocasionado pela alta demanda por recursos como água, alimento, minérios e energia. Estas atividades classificam-se em ações irreversíveis que resultam em alterações no clima, nos sistemas hidrológicos e na biogeoquímica (GRIMM et al., 2008).

Os padrões das mudanças no solo possuem características dinâmicas, no caso de áreas agrícolas, há a influência de diversas variáveis como o clima e métodos usados no cultivo. A supervisão não eficaz da expansão de atividades irreversíveis como a agricultura pode resultar no desmatamento de florestas. Dado que a retirada de cobertura vegetal como florestas provoca a redução da biodiversidade e extinção de espécies animais (RWANGA; M., 2017; ROSSONI; MORAES, 2020).

Segundo o relatório publicado pela FOOD AND AGRICULTURE ORGANIZATION (FAO) (2021), áreas destinadas à práticas agrícolas ocupam uma grande parcela da superfície terrestre do planeta. A agricultura usa cerca de 4,7 milhões de hectares de terra para o cultivo de culturas e criação de animais contrastando com áreas florestais. As principais variáveis socioeconômicas que impulsionam esta expansão são crescimento populacional e econômico.

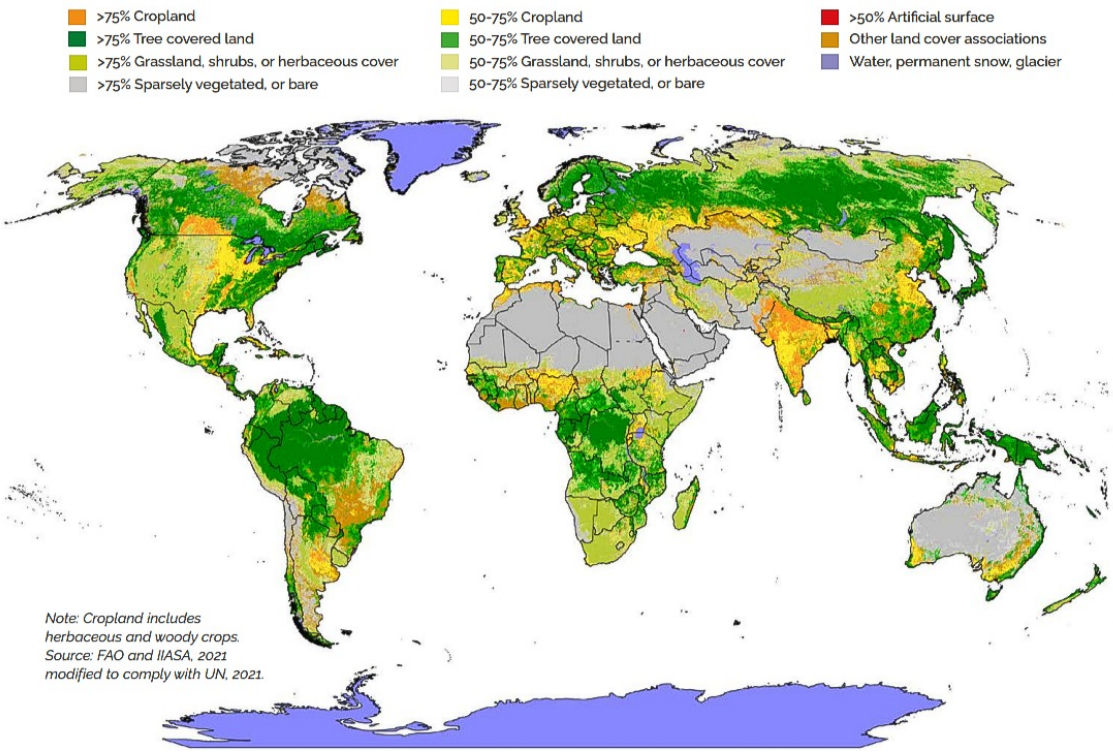
Os resultados desse relatório (Tabela 1.1 e Figura 1.1) demonstram que o rápido crescimento de áreas urbanas deslocou diversos tipos de usuários agrícolas. Logo, causando mudanças nos padrões de chuva e períodos de seca que exigem ajustes na produção. Por isso, o estudo das dinâmicas e padrões do uso e cobertura da terra são importantes, principalmente em áreas agrícolas, pois as condições agroclimáticas modificam-se rapidamente.

Tabela 1.1 - Mudanças no uso e cobertura da terra por classe nos anos 2000 e 2019 em milhões de hectares.

Classe	2000	2019	Mudança
Outras terras	3,968	4,188	220
Terrenos sob prados e pastagens permanentes	3,387	3,196	-191
Terras agrícolas (total de terras agrícolas e prados e pastagens permanentes)	4,880	4,752	-128
Área florestal	4,158	4,064	-94
Terras agrícolas (aráveis e culturas permanentes)	1,493	1,556	63
Área de terreno equipada para irrigação	289	342	53
Terra sob culturas permanentes	134	170	36
Terra arável (terra com culturas temporárias)	1,359	1,383	24

Fonte: Traduzido de FOOD AND AGRICULTURE ORGANIZATION (FAO) (2021).

Figura 1.1 - Mapa global de uso e cobertura da terra com as classes dominantes.



Fonte: FOOD AND AGRICULTURE ORGANIZATION (FAO) (2021).

Os Dados de Observação da Terra, ou *Earth Observation Data* (*EO Data*) são extremamente usados no monitoramento de mudanças na superfície terrestre. Estes dados possibilitam o estudo da dinâmica do solo através do tempo, permitindo mapear, estimar e prever o comportamento de culturas em áreas agrícolas e das ações humanas com alta precisão (MACIEL et al., 2020). Em 2008, com a nova política de dados abertos da missão *Landsat*, as agências espaciais disponibilizaram o acesso livre às imagens provenientes do sensoriamento remoto resultando em um aumento de *EO Data* (WULDER et al., 2012; GIULIANI et al., 2017).

Atualmente, diferentes missões estão lançando satélites para obter imagens sobre uma mesma área com uma taxa de revisita frequente, ao ponto que novos instrumentos estão sendo desenvolvidos com diferentes bandas para observar alvos específicos de forma mais eficaz. Os avanços e inovações referentes a produção de *EO Data* resultam em grandes volumes de dados abertos de imagens consistentes ao longo do tempo e do espaço (FERREIRA et al., 2020). Alguns países organizam projetos para lidar e dar suporte ao *EO Data*, como: a Austrália com o *Digital Earth in Australia* (GAVIN et al., 2018), a Suíça com o *Swiss Data Cube* (GIULIANI et al., 2017) e o Brasil com o *Brazil Data Cube* (FERREIRA et al., 2020).

O *Brazil Data Cube* (BDC)[1] é um projeto que está em desenvolvimento no Instituto Nacional de Pesquisas Espaciais (INPE). O objetivo do projeto é criar cubos de dados multidimensionais no formato *Analysis-Ready Data* (ARD) a partir de imagens de satélite para o território brasileiro. O projeto BDC desenvolve ferramentas computacionais para o processamento, armazenamento, visualização e análise destas imagens para demais projetos do INPE. Além disso, busca gerar informações sobre uso e cobertura da terra aplicando aprendizado de máquina e análise de séries temporais (FERREIRA et al., 2020). ARD são dados pré-processados que possibilitam a sua análise imediata com mínimo esforço adicional (SIQUEIRA et al., 2019).

Cubo de dados ou *EO Data Cubes* possibilitam a análise de imagens consistentes no espaço e no tempo, esta consistência ao longo dos anos permite a geração de mapas LULC com excelentes resultados. A metodologia usada na geração desses mapas é baseada na análise e classificação de séries temporais extraídas de imagens de satélite. O pacote em linguagem de programação R chamado *sits* foi desenvolvido pela equipe do projeto BDC para auxiliar na produção de mapas LULC oferecendo métodos para a extração, análise e classificação usando aprendizado de máquina de séries temporais provenientes de *EO Data Cubes* (SIMOES et al., 2021a).

[1] <http://brazildatacube.org/>

Diversos estudos relacionados desenvolveram metodologias para a classificação de imagens de satélite e geração de mapas LULC em escala local, regional e continental (MACIEL; VINHAS, 2017; LU et al., 2016; PELLETIER et al., 2019). A maioria dos métodos usados por estes estudos são supervisionados, ou seja, necessitam de um conjunto representativo de amostras rotuladas a ser subdividido em treinamento, validação e teste. Para obter o conjunto de treinamento são necessárias campanhas de coleta de amostras em campo ou com o auxílio de *softwares* para a visualização de imagens de alta resolução espacial (HUANG et al., 2020).

O uso de Sistemas de Informação Geográfica (SIG's), e demais *softwares* relacionados à visualização de *EO Data*, tem sido uma estratégia muito usada para as campanhas de coleta (BEY et al., 2015). Para atender a estas campanhas, exite uma demanda por SIG's voltados especialmente à coleta de amostras com o objetivo de otimizar o processo de interpretação de *EO Data*. O *TerraCollect* possui como um dos objetivos atender a esta demanda, sendo uma plataforma em desenvolvimento no projeto BDC para a coleta e análise de amostras LULC com base na interpretação de imagens de satélite e séries temporais de imagens (ANJOS et al., 2022). Além desta demanda, há uma carência por metodologias e ferramentas de análise eficientes para avaliar a qualidade das amostras de treinamento (MILLARD; RICHARDSON, 2015).

Alguns métodos para melhorar a qualidade do conjunto de amostras têm sido desenvolvidos e empregados com sucesso, uma vez que a qualidade dos dados é crucial na classificação, pois leva a mapas com melhor acurácia (TUIA et al., 2009; PELLETIER et al., 2017; PENGRA et al., 2020; SANTOS et al., 2021a). Porém tais métodos exigem conhecimentos específicos de linguagem de programação, algoritmos e processos de análise. Este conhecimento técnico se deve a necessidade de especialização das funcionalidades que devem se adequar aos dados que foram coletados. Não há uma forma otimizada de realizar esta análise ou mesmo avaliar quais métodos se adéquam ao interesse da pesquisa em questão de forma menos custosa. Este problema pode resultar em descarte de amostras que poderiam gerar resultados importantes para certas áreas de estudo.

1.1 Motivação

Coletar amostras de treinamento é um processo custoso e pode demandar uma boa parcela do tempo do grupo de especialistas envolvidos. Em grandes áreas, a variabilidade das classes LULC é alta e intrínseca em diferentes regiões e períodos devido à heterogeneidade da biodiversidade, bem como condições climáticas e práticas de manejo distintas. As amostras, sejam coletadas em campo ou por interpretação vi-

sual a partir de *softwares* com imagens de alta resolução, podem conter ruídos e estão sujeitas a erros na rotulação (HOSTERT et al., 2015; MERONI et al., 2021). Geralmente parte do esforço de amostragem é desperdiçado pela coleta de amostras não informativas e pouco representativas ou por recursos de amostragem cuja contribuição é marginal (TUIA et al., 2011).

O uso de SIG's para a coleta com a interpretação de *EO Data* facilita a avaliação na distribuição da cobertura da terra pois disponibilizam métodos de processamento de imagens e geometrias. Sendo que cada SIG possui características específicas com vantagens e desvantagens (RWANGA; M., 2017). Definir qual método de análise é o mais adequado é uma tarefa complexa devido à quantidade de opções, abordagens e possibilidades na variação de parâmetros. SIG's de propósito geral, como o QGIS, possuem um conjunto completo de ferramentas analíticas, porém exigem experiência e conhecimento específico para a configuração e utilização. Os *softwares* desenvolvidos especialmente para o processo de coleta, como o *CollectEarth* (BEY et al., 2016), possuem funcionalidades mais restritas a coleta de amostras e não integram muitas ferramentas analíticas para auxiliar no processo.

1.2 Objetivos

O objetivo principal deste trabalho é propor uma arquitetura de *software* para a integração de métodos de análise de amostras de uso e cobertura da terra (amostras LULC) na plataforma de coleta *TerraCollect*. Estes métodos possuem uma abordagem baseada no processamento e análise de séries temporais de imagens de satélite. Os métodos usados são disponibilizados pelo pacote *sits* e foram adaptados e separados em: análise exploratória de amostras LULC (WICKHAM; GROLEMUND, 2017); análise de métricas com aprendizado de máquina e técnicas de *Active Learning* (TUIA et al., 2009); e por fim o controle de qualidade e redução do ruído de classe (SANTOS et al., 2021a). A integração busca facilitar a identificação de amostras catalogadas erroneamente, exploração e refinamento do conjunto de dados durante e após o processo de coleta.

As abordagens foram integradas em uma ferramenta computacional desenvolvida no ambiente da plataforma *TerraCollect* como uma extensão de suas funcionalidades. A plataforma *TerraCollect* está usando a infraestrutura digital para testes e desenvolvimento no projeto BDC. A arquitetura para a integração das abordagens nesta plataforma usa o modelo cliente-servidor com dois componentes principais: um serviço *web* e um componente de visualização na interface de coleta do *TerraCollect*.

A extensão de análise para o *TerraCollect* fornece uma interface gráfica para a execução das abordagens em um painel de controle na *web* para subsidiar os especialistas durante a coleta de amostras LULC, buscando automatizar as etapas para a extração e exploração de dados relacionados às amostras. Desta forma, os usuários que não possuem conhecimentos específicos como a construção de algoritmos e técnicas com linguagens de programação poderão visualizar e compartilhar resultados da avaliação de forma interativa com uma interface simples, tornando o processo mais eficiente e otimizado unindo a coleta com a análise.

1.3 Contribuições

O presente trabalho contribui para a evolução das tecnologias e serviços disponíveis no projeto BDC, ao avançar na pesquisa e desenvolvimento da plataforma *TerraCollect* de coleta e análise de amostras LULC com base em séries temporais de imagens. Baseado-se na arquitetura proposta nesta dissertação, foi feito um protótipo da ferramenta de análise como uma extensão desta plataforma de coleta. Além disso, um artigo foi publicado no Simpósio Brasileiro de Geoinformática (GEOINFO)[2] edição 2022 (SANTOS; PEREIRA, 2022). GEOINFO é uma conferência anual para explorar a pesquisa e o desenvolvimento de aplicações inovadoras em ciência da informação geográfica e áreas afins.

O artigo "Integrando métodos de análise durante a coleta de amostras de uso e cobertura da terra" (ANJOS et al., 2022) descreve a arquitetura, a metodologia usada e detalhes do desenvolvimento do protótipo de extensão no *TerraCollect*. A extensão para análise de amostras foi testada em um ambiente real de campanha de coleta com possíveis usuários para estudar a viabilidade, elicitar requisitos funcionais e ideias para a interface.

Como o projeto BDC está alinhado com as demandas do INPE, espera-se que os resultados do presente trabalho tragam inovações para demais projetos do instituto. Esta arquitetura de integração de métodos de análise durante a coleta de dados LULC também contribuirá na evolução das tecnologias e serviços para auxiliar programas de monitoramento nos biomas brasileiros como *DETER*, TerraClass e *PRODES*.

[2] <http://www.geoinfo.info/>

2 REVISÃO DA LITERATURA

2.1 Mapas de uso e cobertura da terra

Mapas de uso e cobertura da terra, ou mapas LULC, são representações das ações humanas, recursos terrestres e suas interações na superfície do planeta Terra para a compreensão da interação humana com o ambiente (GREGORIO; JANSEN, 2000; HANSEN; LOVELAND, 2012; NEDD et al., 2021). A cobertura da terra é definida como a cobertura biofísica observada em uma determinada área (GREGORIO; JANSEN, 2000). O uso da terra são as atividades humanas que se desenvolvem na superfície terrestre (MATTILA et al., 2011). Os recursos terrestres abrangem componentes ecológicos, como clima, água, solo, relevo, flora, fauna e sistemas socioeconômicos interagindo com a agricultura, silvicultura e outros usos da terra (ESTES et al., 2017; ANANDHI et al., 2020).

Em suma, mapas LULC resumem-se a interpretações feitas por especialistas com base em *EO Data*, usando atributos para definir e identificar elementos e recursos terrestres. Esses mapas possuem legendas com classes e cores para representar a natureza observada e, por se tratar de valores atribuídos categoricamente, necessita-se de padronização com sistemas de classificação. O *Land Cover Classification System* (LCCS) é um sistema de classificação *a priori* abrangente e padronizado em exercícios de mapeamento, independentemente da escala ou dos meios usados para mapear. Alguns exemplos de classes LULC são: florestas, pastagens, corpos d'água e área urbana (GREGORIO; JANSEN, 2000).

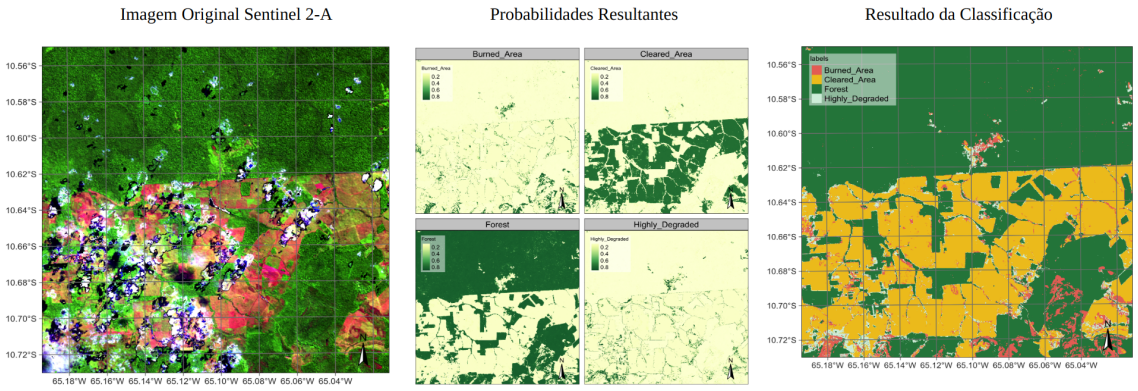
Existem diversas metodologias para a geração de mapas LULC, porém a comumente usada tanto em pequenas ou grandes áreas envolve a interpretação de imagens de sensoriamento remoto coletadas por satélites. Nesta metodologia há várias etapas desde o estudo da área alvo, aquisição de dados e decisão do uso do sistema de legendas, etc. (SIMOES et al., 2021a). As etapas para a podem variar de acordo com o objeto de estudo e dos resultados esperados (OLOFSSON et al., 2014).

Para a classificação de áreas nas imagens são usados critérios que permitem a correlação das observações e sistemas de legendas existentes. Geralmente as interpretações são influenciadas por contextos socioeconômicos e políticos relacionados à área de estudo (GREGORIO; JANSEN, 2000). Algumas classes, por exemplo, podem possuir uma mesma nomenclatura e apresentar comportamentos distintos em diversas áreas como no caso de produtos agrícolas: o desenvolvimento das culturas, o máximo vigor vegetativo e a época da colheita (MERONI et al., 2021).

A geração de mapas LULC trata da análise de múltiplas variáveis, por isso, atualmente, esta atividade está sendo automatizada com a união da interpretação de imagens de sensoriamento remoto por especialistas humanos com métodos de aprendizado de máquina. *EO Data* são gerados e armazenados diariamente em bancos de dados históricos, o que facilita a análise temporal das dinâmicas do uso do solo. Por essa razão o uso deste recurso se tornou indispensável devido à demanda por alta precisão nos mapas resultantes (FERREIRA et al., 2020).

A Figura 2.1 apresenta um exemplo de classificação de imagem de sensoriamento remoto usando técnicas de aprendizado de máquina. Neste caso a variável de resposta é qualitativa pois com treinamento de modelos busca-se predizer qual a classe em uma dada área na imagem, dado um valor probabilístico. Assim, as sessões de treinamento do modelo são realizadas de maneira supervisionada, onde se tem acesso a um conjunto de dados rotulados por especialistas com base nas imagens, os chamados dados de entrada ou amostras, e o resultado obtido a partir do treinamento e classificação, chamado dado de saída (GOODFELLOW et al., 2016).

Figura 2.1 - Exemplo de classificação de imagens usando aprendizado de máquina.



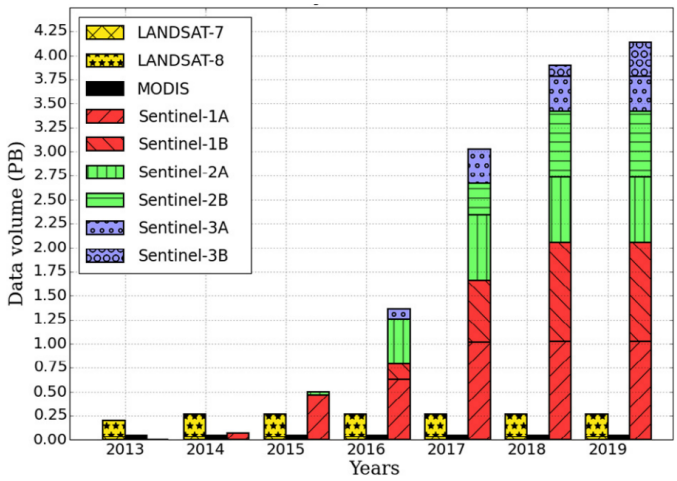
Fonte: Simoes et al. (2021b).

2.2 Dados de observação da terra

Fatores como a implementação de políticas relacionadas à disponibilização de dados abertos pelas agências espaciais e aos avanços tecnológicos em dispositivos para satélites de observação têm culminado no substancial aumento da disponibilidade de dados espaço-temporais de imagens da superfície terrestre, conhecidos como *EO Data*. Como resultado, há uma densa massa de dados para auxiliar o monitoramento e geração de mapas LULC (GIULIANI et al., 2017; SOILLE et al., 2018; FERREIRA et al., 2020; GIULIANI et al., 2020).

A Figura 2.2 ilustra uma estimativa do volume de dados gerados pelas três primeiras missões Sentinel, bem como pelas missões Landsat 7 e 8, além do MODIS (*Moderate Resolution Imaging Spectroradiometer*), presente nos satélites TERRA e AQUA da NASA (Administração Nacional de Aeronáutica e Espaço) / USGS (Serviço Geológico dos Estados Unidos).

Figura 2.2 - Estimativa para o volume de dados abertos por sensor em 2018.



Fonte: Soille et al. (2018).

Anualmente, o programa *Landsat* é responsável pela geração de cerca de 0.25 *petabytes* de dados. Em consonância com o acréscimo dos volumes de dados oriundos do programa *Copernicus* (Programa Europeu de Observação da Terra), idealizado pela ESA (Agência Espacial Europeia), Soille et al. (2018) procederam a uma projeção estimativa até o ano de 2019 referente à geração de dados inerente a todas as missões *Sentinel*. Conforme resultado desta análise, os volumes de dados gerados superaram a marca dos *petabytes*. A nova geração de satélites produz grandes volumes de *EO Data*, trazendo o conceito de *Big Data* com o acrônimo *Big Earth Observation Data*.

O novo conceito de *Big EO Data* traz consigo novos desafios para a administração dos dados envolvendo o armazenamento e processamento. As abordagens tradicionais não são adequadas para o contexto de *Big Data*, desafios que geram uma demanda por novas ideias e métodos que facilitem a distribuição dos dados para aproveitá-los ao máximo (FERREIRA et al., 2020). Essa grande massa de dados possibilita o mapeamento da dinâmica do uso e cobertura da terra com alta precisão (MACIEL et al., 2020). Contudo, as aplicações de *Big EO Data* no contexto de mapeamento e monitoramento da terra, os pesquisadores precisam considerar como facilitar a interoperabilidade e análise desses dados (GOMES et al., 2020).

Para extrair informações na classificação e mapeamento LULC do *Big EO Data* é necessária a análise de *terabytes* de dados. Cada dado com diferentes resoluções espaciais, temporais e espectrais obtidos por diversos sensores. Processar esses *terabytes* de forma eficiente requer uma estrutura computacional robusta que a maioria dos pesquisadores não têm acesso. Por essa razão, o suporte ao *Big EO Data* tornou-se o objetivo de muitas iniciativas e projetos como o BDC, o *Swiss Data Cube* (GIULIANI et al., 2017) e o *Digital Earth Australia* (GAVIN et al., 2018).

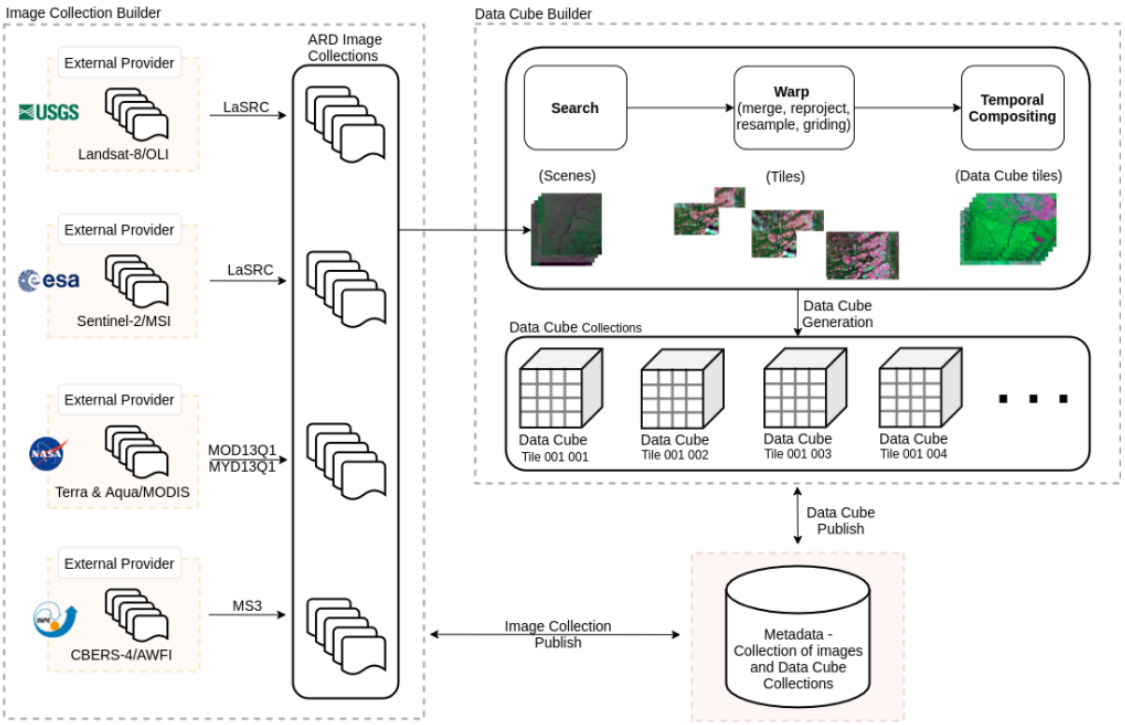
2.2.1 Cubos de dados

O termo cubo de dados ou *EO Data Cubes* refere-se a um conjunto de imagens de sensoriamento remoto alinhadas espacialmente no tempo. A produção e o fornecimento de *EO Data Cubes* facilita a disseminação e a análise dos *EO Data* (GIULIANI et al., 2020). O projeto BDC tem como objetivo distribuir dados prontos para a análise (ARD) e desenvolver ferramentas computacionais para o armazenamento e processamento de um grande volume de imagens em *EO Data Cubes*, facilitando o acesso ao *Big EO Data*. Para isso é necessário uma sequência de etapas como a aquisição das imagens dos fornecedores, pré-tratamento, recorte, composição, regularização dessas imagens, além da validação dos dados e metadados (FERREIRA et al., 2020).

Conforme a Figura 2.3, as imagens coletadas com os fornecedores originais como os satélites *Landsat* (USGS), *Sentinel* (ESA) e *CBERS* (INPE) são tratadas e pré-processadas em produtos de refletância de superfície. Após a aquisição são usados os algoritmos *LaSRC* e *MUX* para a correção atmosférica dependendo das características de cada sensor. As nuvens são tratadas usando o algoritmo *Fmask*, para cada imagem, onde máscaras de nuvens e sombra de nuvens são calculadas. Após este pré-processamento, o conjunto de dados resultante é catalogado e são usados para gerar os cubos de dados (FERREIRA et al., 2020).

Após o ajuste das imagens, um método de composição de tempo é usado, considerando apenas observações válidas, onde não são considerados valores como nuvens ou sombras de nuvens. Este método define a abordagem para selecionar ou gerar um valor que representa a série de observações disponíveis durante o intervalo de tempo. Por fim, o mosaico é recortado na extensão do bloco de grade. Cubos de dados regulares requerem a definição de uma etapa de tempo regular (por exemplo, um mês ou 16 dias) que orientará a composição temporal de múltiplas imagens disponíveis em cada etapa de tempo.

Figura 2.3 - Aquisição de dados de observação da Terra, pré-tratamento e geração de cubos de dados pelo projeto *Brazil Data Cube*.



Fonte: Ferreira et al. (2020).

As principais aplicações dos *EO Data Cubes* são a análise temporal e a extração de séries temporais de imagens (SIMOES et al., 2021a). Em paralelo, com o objetivo de otimizar estas aplicações, o projeto BDC explora os benefícios proporcionados pelo desenvolvimento de serviços *web* e ferramentas computacionais. Como por exemplo: o serviço *Spatial Temporal Asset Catalog* (STAC) para catalogação, busca, visualização e análise temporal das imagens; o serviço *web Web Time Series Service* (WTSS) para extração de séries temporais; e a plataforma *Data Cube Explorer*[1] que oferece uma interface gráfica para a descoberta e visualização dos produtos de cubos e coleções de imagens produzidos pelo projeto.

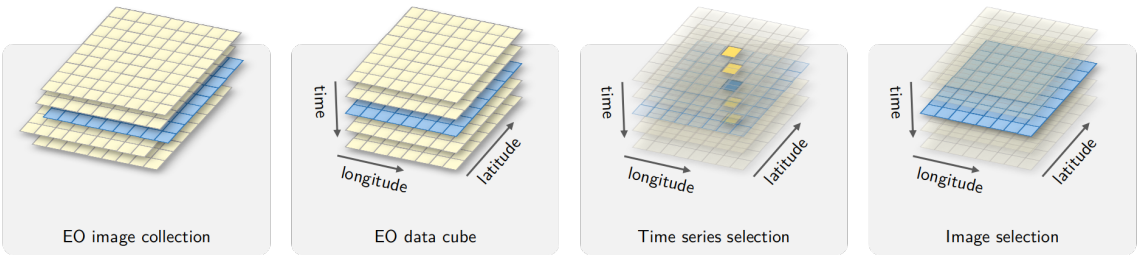
[1] <https://brazildatacube.dpi.inpe.br/portal>

2.2.2 Séries temporais de imagens de satélite

Séries temporais provenientes de cubo de dados são sequências de valores observados ao longo de um período dada uma coordenada geográfica, no contexto de imagens digitais, são os valores de um *pixel* xy através de um período de tempo t (VINHAS et al., 2017). Após o tratamento das imagens em *EO Data Cubes*, os dados podem ser compreendidos usando o conceito de matriz de duas dimensões. Cada matriz possui coordenadas xy em colunas e linhas, além de uma ou mais camadas que identificam as bandas espectrais. Em uma imagem digital formada por *pixels*, cada *pixel* contém a coordenada xy e um valor numérico da observação, traduzido como a energia refletida e captada pelo sensor do satélite. *EO Data Cubes* são padronizados no tempo $T(t_1, t_2, t_3, ..., t_n)$ e cada tempo t_i tem uma imagem digital associada.

Como as imagens de sensoriamento remoto estão organizadas em matrizes, para extrair a série temporal de uma dada localização é necessário converter as coordenadas longitude e latitude em xy da imagem para encontrar o *pixel* correspondente, conforme a representação na Figura 2.4. A coordenada da localização alvo é convertida em coluna e linha usando os metadados salvos da imagem. Logo que o *pixel* alvo é encontrado é possível extrair uma série temporal de uma ou mais bandas em um intervalo de tempo (SIMOES et al., 2021b). Este processo pode ser otimizado usando o WTSS, onde pode-se recuperar as séries temporais das imagens de um sensor dada uma coordenada espacial e um período de tempo selecionado.

Figura 2.4 - Visão conceitual de cubos de dados e séries temporais de imagens.



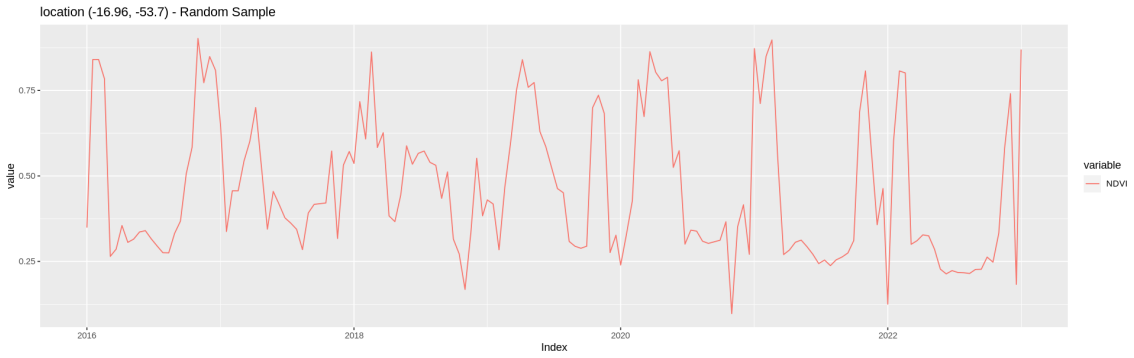
Fonte: Simoes et al. (2021b).

No WTSS, a conversão de coordenadas geográficas em linhas e colunas para a localização de *pixels* é feita com base no centro dos *pixels* das extremidades da imagem. A imagem possui um atributo chamado resolução espacial referente à área coberta pelo *pixel*, ou seja, a capacidade da imagem em representar características do espaço

real. O centro do *pixel* é o centro desta área possuindo uma coordenada geográfica. As coordenadas usadas pelas ferramentas do projeto BDC seguem o sistema de coordenadas '*EPSG*: 4326" que utiliza o *datum WGS* 84 (*World Geodetic System* 1984). Esse sistema de coordenadas é amplamente adotado em sistemas de informação geográfica. A extensão temporal encontra-se no formato *ISO* 8601 para representação da data e hora das observações (VINHAS et al., 2017).

A Figura 2.5 demonstra um exemplo de série temporal para o cubo de dados do satélite *CBERS-4* com imagens do sensor WFI. O gráfico apresenta uma sequência de valores para o índice *Normalized Difference Vegetation Index* (NDVI) em uma coordenada exemplo no estado do Mato Grosso (longitude -53.7 e latitude -16.96 em *EPSG*: 4326 – *WGS84*) no período de 2016 à 2023. Este cubo tem 64 metros de resolução espacial e usa a composição temporal de 16 dias usando a melhor abordagem de pixel.

Figura 2.5 - Exemplo de série temporal do cubo de dados do satélite *CBERS-4* para o índice espectral $NDVI$.



Fonte: Adaptado de BRAZIL DATA CUBE (BDC) (2022).

O WTSS preenche a lacuna entre as ferramentas para lidar com grandes volumes de dados e a aplicação de séries temporais com uma *Application Programming Interface* (API). Além do serviço *web*, também há pacotes clientes em linguagem *Python* e R que consomem a API. Com os pacotes é possível extrair séries temporais de forma eficiente com poucas linhas de código (QUEIROZ et al., 2015). Entretanto as funcionalidades do WTSS estão limitadas somente à extração de séries, a manipulação e aplicação de filtros fica a cargo dos conhecimentos do usuário na linguagem de programação escolhida. O desenvolvimento de ferramentas que lidam, não apenas com a extração, mas também com a análise de séries temporais está se tornando cada vez mais necessário (VINHAS et al., 2017).

2.3 Amostras de uso e cobertura da terra

A geração de mapas LULC com base no aprendizado de máquina aplicado à imagens de sensoriamento remoto é um método que requer a classificação de imagens por modelos previamente treinados. Este tipo de classificação, usualmente baseia-se no método supervisionado, onde há o treinamento, validação e teste de um modelo classificador a partir de um conjunto de dados rotulados. Para esta etapa, são necessárias amostras de uso e cobertura da terra (amostras LULC) que são, neste contexto, coordenadas geográficas associadas a um período de ocorrência e uma determinada classe (GOODFELLOW et al., 2016; FERREIRA et al., 2020; SANTOS et al., 2021a; SIMOES et al., 2021a).

O *design* amostral define como as amostras serão coletadas, existem diversos métodos e protocolos para esta definição como a teoria probabilística e sistemática. Há diversos meios para a coleta de amostras LULC, por exemplo, a coleta em campo (SANTOS et al., 2021b), a análise de dados públicos (BELGIU et al., 2020) e interpretação visual de imagens de sensoriamento remoto (SIMOES et al., 2021a). Um estudo, apresentado por Olofsson et al. (2014), traz um guia para obter melhores resultados em mapas de classificação de mudanças ao estudar as características essenciais nos protocolos de amostragem. Uma recomendação exposta como extremamente importante no estudo é se basear na coleta de forma aleatória.

As boas práticas apresentadas por Olofsson et al. (2014) não procuram definir regras ou a melhor prática e sim discutir os processos que envolvem a coleta, sendo as boas práticas, um conjunto de estudos de vários outros autores que obtiveram melhores resultados ao utilizá-las. Ou seja, a definição do *design* amostral deve se adequar ao tipo de estudo e ao resultado esperado. Por exemplo, nem sempre é possível utilizar a teoria probabilística como *design* amostral devido a problemas físicos ou técnicos. Como por exemplo o custo ou perigo na coleta, quando o *design* amostral escolhido necessita de amostras de campo. Outro exemplo, quando a imagem selecionada aleatoriamente possui alta cobertura de nuvem e se torna dispensável para a análise.

No estudo apresentado por Belgiu et al. (2020) ocorre a união de dois métodos, a coleta por dados históricos públicos e a interpretação de imagens de forma automática. A coleta em campo é mais usada quando a área de estudo é menor e possui mais subclasses da mesma cultura, por exemplo a análise do ciclo fenológico da soja que possui diferentes estágios como florescimento e maturação. Na coleta em campo, o pesquisador deve conduzir até a área de interesse e anotar as classes encontradas, porém requer o transporte à área de estudo.

A coleta por análise de dados públicos é baseada em dados históricos armazenados em órgãos governamentais, programas ambientais ou projetos e sistemas de monitoramento, como por exemplo no Brasil, o PRODES, DETER e TerraClass. Esta coleta é usada para uma área de estudo maior, mesmo sendo um método menos custoso e mais rápido em relação aos outros, este método pode acarretar em amostras ruidosas devido aos diferentes sistemas de classificação usados pelos programas, necessitando de ajustes e de padronização (BELGIU et al., 2020; ZIOTI et al., 2021).

Atualmente, os satélites de observação da Terra fazem a cobertura de uma mesma área com uma taxa de revisita frequente, proporcionando imagens consistentes ao longo do tempo. Estas imagens possibilitam o estudo das características do uso e cobertura da terra através do tempo de forma mais eficiente (FERREIRA et al., 2020). A crescente quantidade de *EO Data* disponível aliada com SIG's permite a coleta e análise de amostras de forma eficaz. Neste tipo de coleta, com interpretação visual de imagens, o pesquisador faz a análise dos *pixels* combinada com a extração de séries temporais (SIMOES et al., 2020; SIMOES et al., 2021a).

Um exemplo de amostra LULC é descrito na Tabela 2.1. Esta amostra é proveniente dos testes realizados com o Projeto TerraClass para a ferramenta *TerraCollect*, em desenvolvimento no projeto BDC. Os atributos de localização definem a posição no espaço geográfico, usualmente são coordenadas como longitude e latitude baseadas em uma projeção cartográfica da superfície esférica da Terra. Os atributos de período descrevem o espaço de tempo a qual classe permaneceu, a data de início e fim, (definidas em inglês como *start_date* e *end_date*). A classe LULC descreve a característica observada, sendo baseada em um sistema de classificação LCCS.

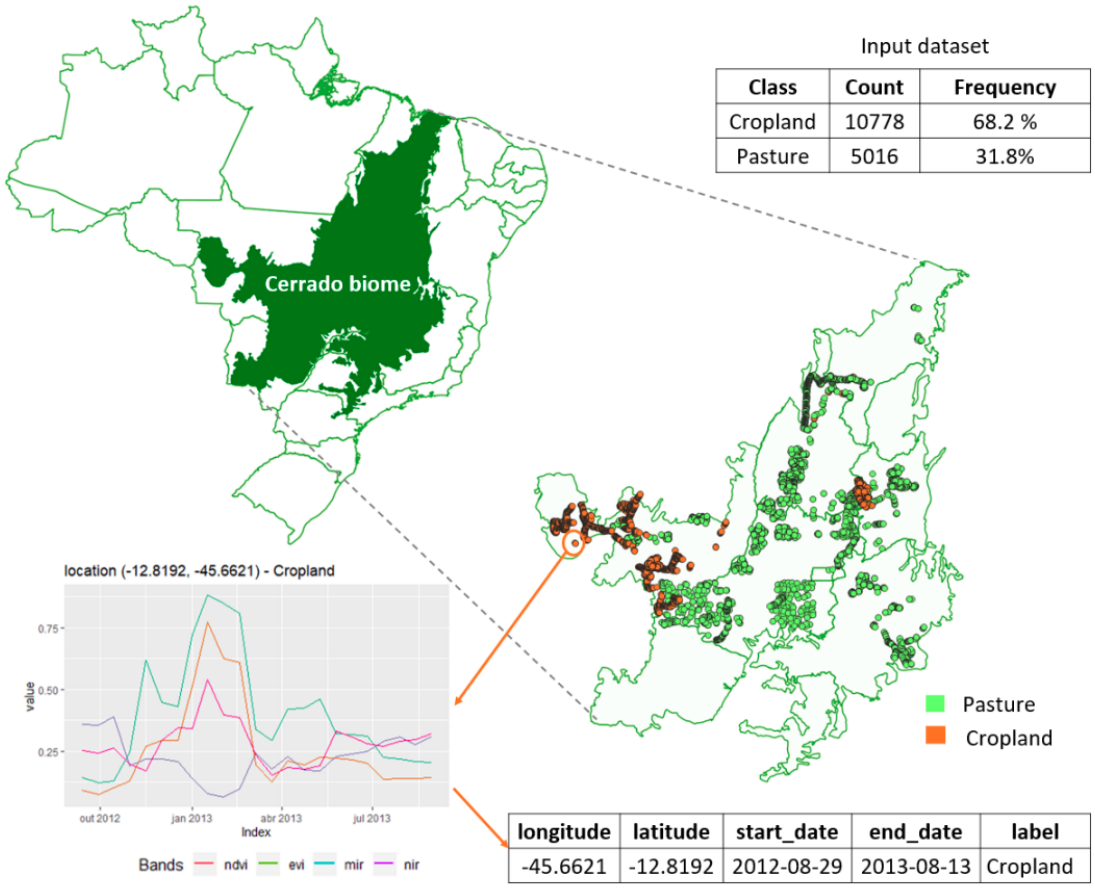
Tabela 2.1 - Exemplo de amostra de uso e cobertura da terra coletada com o *TerraCollect*.

ID	Longitude	Latitude	Data Inicial	Data Final	Classe	Cor
1548	-51.18920	-17.12320	2019-08-13	2020-08-28	Silvicultura	#A8A800

Fonte: Produção do autor.

Outro exemplo é conjunto de dados apresentado pela Figura 2.6, um produto do estudo descrito por Santos et al. (2021b) para identificação de padrões espaço temporais nas amostras no Cerrado brasileiro. Os dados usados no estudo são conjuntos de amostras coletadas em campo e a partir de interpretação de imagens por especialistas em sensoriamento remoto. Para o estudo dos padrões foram usadas séries temporais do produto *MOD13Q1* do sensor *MODIS* associadas a cada amostra.

Figura 2.6 - Exemplo de conjunto de amostras de uso e cobertura da terra.



Fonte: Santos et al. (2021b).

2.4 Análise de séries temporais de imagens de satélite

A abordagem comumente usada para a detecção de mudanças em sensoriamento remoto é comparar duas imagens classificadas do mesmo local em datas consecutivas e derivar uma matriz de transição. Segundo Camara et al. (2016) esta abordagem tradicional é caracterizada como "*space-first*, *time-later*", onde a prioridade é a avaliação espacial e depois a análise temporal, sendo a alternativa "*time-first*, *space-later*", onde a prioridade é a análise temporal. Alguns autores concluem que os métodos "*time-first*, *space later*" são adequados para mapear mudanças mais precisamente do que a abordagem tradicional (CAMARA et al., 2016; PELLETIER et al., 2019; SIMOES et al., 2021a).

Nesta abordagem alternativa todos os valores espectrais observados são entradas no treinamento de métodos de aprendizado de máquina e posteriormente para a classificação, ou seja, toda a sequência de valores da série temporal são considerados. Séries temporais de imagens permitem uma visão mais ampla das mudanças da superfície terrestre, capturando mudanças graduais à abruptas. Assim, os modelos de aprendizado de máquina classificam primeiro cada série individualmente e depois aplicam pós-processamento espacial para capturar informações da vizinhança. A autocorrelação temporal dos dados é mais forte do que a espacial, em outras palavras, dados com repetibilidade adequada indicam que um *pixel* está mais relacionado aos seus vizinhos temporais do que aos seus vizinhos espaciais (CAMARA et al., 2016).

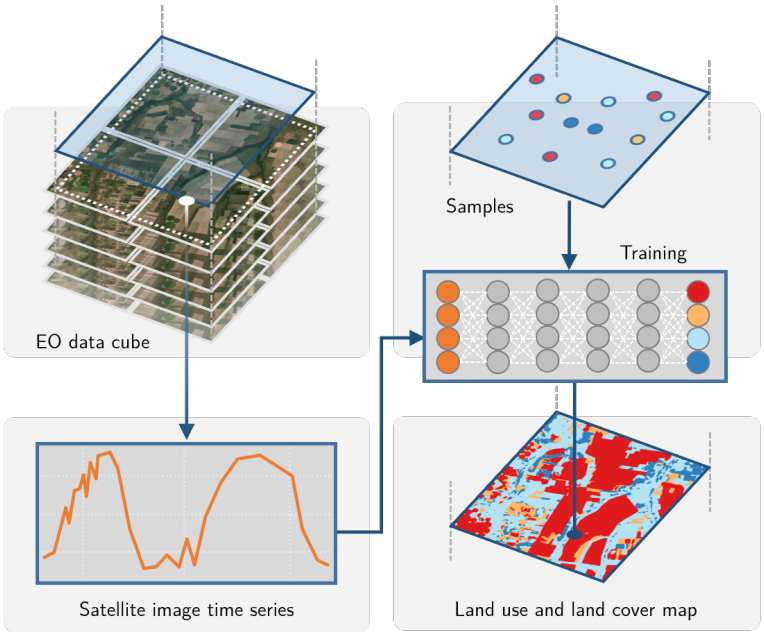
Como benefícios, a nova abordagem permite o acompanhamento de tendências sazonais e de longo prazo, bem como a identificação de eventos, padrões e anomalias a partir dos dados como: incêndios florestais, inundações ou secas. Para conjunto de dados com séries temporais de imagens, uma das principais características é a resolução temporal dos dados, com uma alta quantidade de observações pode-se capturar mudanças importantes e cruciais de uma dada classe LULC (SIMOES, 2021).

A análise e classificação de séries temporais é uma metodologia comumente usada para a geração de mapas LULC. Desta forma, modelos de aprendizado de máquina são treinados usando amostras em conjunto com estas séries para classificar os *pixels* das imagens de satélite (SIMOES et al., 2020). Este método requer um conjunto denso de amostras para o treinamento, surgindo a necessidade de ferramentas para auxiliar a tarefa de extração de dados das imagens de forma eficiente (VINHAS et al., 2017). O pacote denominado *sits*, do acrônimo em inglês "*Satellite Image Time Series*" (SITS), fornece um conjunto extenso de ferramentas para lidar com a extração e análise de séries temporais de imagens (SIMOES et al., 2021b).

O pacote em linguagem de programação R *sits*, descrito por (SIMOES et al., 2021a), fornece um conjunto de funcionalidades para manipular e extrair informações de *EO Data Cubes* com comandos simples de alto nível. O pacote oferece o suporte para extração de dados, análise e classificação de imagens de satélite usando aprendizado de máquina. Por ser um pacote robusto e com suporte à programação paralela, é recomendado quando se lida com um número alto de pontos para a extração de séries temporais. Dentre as ferramentas oferecidas estão: a criação de um cubo de dados local; cálculo de índices espectrais; aplicação de regularização, normalização e suavização nas séries; métodos para análise de amostras; treinamento de modelos com séries temporais para a classificação de imagens.

Conforme demonstrado pela Figura 2.7, o *sits* funciona em quatro etapas principais. A primeira etapa é a preparação dos dados, onde o usuário seleciona a área de interesse para a busca no provedor de imagens. O pacote possui suporte a vários provedores de cubos de dados como o BDC, AWS e customizados criados pelo usuário. Assim, o usuário fornece as amostras LULC para realizar a extração e análise das séries temporais e treinar os modelos de aprendizado de máquina fornecidos pelo *sits*. Por fim, ocorre a classificação das imagens de sensoriamento remoto, nesta fase, é aplicado um algoritmo de suavização *bayesiana* que considera a vizinhança espacial de cada *pixel* e melhora os resultados da classificação, e como produto final obtêm-se o mapa LULC (SIMOES, 2021).

Figura 2.7 - Visão geral do pacote *sits* em linguagem de programação R.



Fonte: Simoes et al. (2021b).

A metodologia usada pelo pacote *sits* é separa em algoritmos de aprendizado de máquina (*Machine Learning*) e aprendizado profundo (*Deep Learning*). Todos os algoritmos foram adaptados para trabalhar com o formato de dados apresentados como amostras LULC combinadas com séries temporais de imagens de forma supervisionada. Para isso, é necessário que as séries do conjunto de treinamento estejam alinhadas no tempo com a mesma quantidade de observações no mesmo período. Assim, o pacote também possui ferramentas de identificação de nuvens e interpolação de valores nas séries para resolver os problemas como dados nulos.

Os algoritmos suportados pelo *sits* são: *Random Forest* (RF), *Support Vector Machines* (SVM), *Convolutional Neural Networks* (CNN's), *Extreme Gradient Boosting*, *Multi-Layer Perceptrons* (MLP's), *Residual Neural Networks* (*ResNET*) e *Temporal Convolutional Neural Networks* (*TempCNN*). Os algoritmos de *Machine Learning* como RF, SVM's e MLP's, na adaptação *sits*, não levam em consideração a componente temporal das séries, em outras palavras, a ordem na sequência de valores não influencia nos resultados. Os algoritmos de *Deep Learning* suportam o processamento das características temporais da sequência de valores, como os algoritmos CNN's, *ResNET* e *TempCNN* (SIMOES et al., 2021b).

O pacote suporta o ciclo completo de análise de dados abstraindo as etapas para classificação LULC em diferentes ambientes de produção do usuário, com possibilidade do uso em ambientes para computação em nuvem sem custos de *hardware* de alto desempenho. O *sits* resume um pacote em linguagem R com comandos simples, mas eficiente em análises de *Big EO Data*. O pacote possui código aberto e uma documentação[2] simples e compreensiva do básico ao intermediário.

2.5 Métodos para análise da qualidade de amostras

O *design* amostral, mesmo usando as boas práticas e abordagem aleatória, não oferece soluções completas em relação à qualidade das amostras (OLOFSSON et al., 2014). As técnicas de coleta, em campo ou por interpretação de imagens, podem gerar amostras com ruído amostral, pois alguns especialistas tendem a agrupar amostras de caraterísticas diferentes com a mesma classe LULC (SANTOS et al., 2021a). Na coleta, muitas vezes perde-se a perspectiva do algoritmo subjacente à classificação, usualmente, o estudo de confusão entre os dados não ocorre, sendo os ruídos mais comuns problemas semânticos, erros nos rótulos, dados duplicados e incoerências com a distribuição espacial (TUIA et al., 2011).

Durante a coleta, principalmente no contexto da análise de séries temporais, um dos maiores desafios é a aquisição de amostras que produzam uma variabilidade significativa ao longo do tempo, (SANTOS et al., 2021b). Pesquisas anteriores propuseram técnicas para a geração de amostras representativas com o uso de métricas estatísticas e estudos com relação ao treinamento e predição de modelos. Alguns destes métodos são apresentados na Tabela 2.2.

[2] <https://e-sensing.github.io/sitsbook>

Tabela 2.2 - Lista de métodos para a geração semi-automática de amostras de treinamento.

Autoria	**Título do Estudo**	**Método Usado**
Fritz et al. (2009)	Geo-Wiki.Org: The Use of Crowdsourcing to Improve Global Land Cover	*Crowdsourcing*
Hu et al. (2010)	The migration of training samples towards dynamic global land cover mapping	Uso de inventários existentes para orientar a rotulagem das novas amostras de treinamento
Tuia et al. (2011)	A Survey of Active Learning Algorithms for Supervised Remote Sensing Image Classification	*Transferring Learning*
Malambo e Heatwole (2020)	Automated training sample definition for seasonal burned area mapping	Análise das assinaturas espectrais e temporais das classes alvo para uma área de estudo usando agrupamento *c-means*
Belgiu et al. (2020)	Phenology-based sample generation for supervised crop type classification	Análise das assinaturas espectrais de dados históricos com Time-Weighted Dynamic Time Warping (TWDTW) e refinamento com *Random Forest*

Fonte: Produção do autor.

Existem estudos, como Brian et al. (2011) que aplicam a análise de atributos das amostras como fase anterior ao treinamento por algoritmos com aprendizado de máquina. Neste estudo, são usadas variáveis geograficamente ponderadas nas amostras para reduzir o conjunto de dados sem diminuir a qualidade das amostras. A hipótese é que amostras de mesma classe geograficamente próximas possuem maior peso para a classificação. Outros estudos como Tuia et al. (2011) utilizam análise de métricas de probabilidades diretamente nas sessões de treinamento para o modelo definir e selecionar amostras pela representatividade.

No uso de modelos de aprendizado de máquina supervisionados, fatores como a representatividade dos dados de entrada influenciam diretamente na acurácia do modelo resultante (GOODFELLOW et al., 2016). Considera-se extremamente necessário o uso de métodos analíticos para verificar a qualidade das amostras como etapa anterior ao treinamento. A compreensão das características dos dados e dos resultados esperados, pode indicar qual o método de análise é o mais adequado ao problema (WICKHAM; GROLEMUND, 2017).

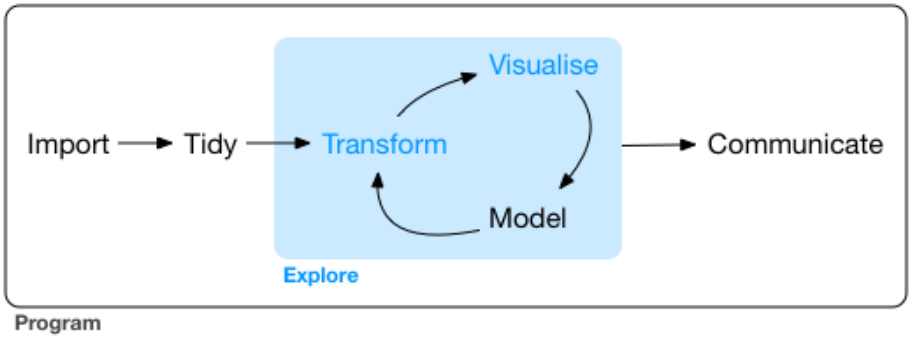
No contexto da coleta e análise de amostras na geração de mapas LULC, deve-se pensar nas possíveis classes a serem estudadas, quais bandas ou índices espectrais que melhor descrevem a natureza destas classes, selecionar a coleção de imagens que se ajustem ao problema e posteriormente o método a ser usado para a classificação (RWANGA; M., 2017). Tais métodos requerem dados precisos e rotulados para o treinamento como um fator chave na precisão dos mapas resultantes, contudo os estudos relacionados à análise dos dados de entrada são diversos e apresentam benefícios e danos. Neste capítulo serão explorados os atuais métodos de análise para mensurar, avaliar e buscar a qualidade dos dados.

2.5.1 Análise exploratória de dados

Conforme Wickham e Grolemund (2017), a Análise Exploratória de Dados (AED) é um método usado para resumir e sumarizar as principais características do conjunto de dados, geralmente usando técnicas de visualização com gráficos ou tabelas. Desta forma é possível identificar padrões, detectar anomalias e verificar se as informações atendem aos objetivos esperados ou se necessitam de ajustes. Nesta etapa não há exigências ou critérios a serem seguidos, o que ocorre é baseado na experiência e curiosidade do analista. A principal ideia é transformar, criar visualizações usando os dados e comunicar informações relevantes para demonstrações, conforme a Figura 2.8 ilustrando as etapas da AED.

O método foi primeiramente explorado por Tukey (1977) com a finalidade de examinar o problema e as informações disponíveis previamente à qualquer aplicação de técnica estatística. AED resume-se a gerar hipóteses e formular perguntas buscando a exploração e visualização dos principais atributos dos dados, a fim de ter uma melhor perspectiva desde o início ou identificar áreas ou padrões onde pode-se aprofundar. Desta forma o analista consegue um entendimento das relações entre as variáveis existentes nos dados, identificando as falhas que, por exemplo, podem poluir e diminuir a qualidade de um conjunto de amostras de treinamento.

Figura 2.8 - Etapas da análise exploratória de dados.



Fonte: Wickham e Grolemund (2017).

As técnicas para a AED diferem do que é feito na Estatística Clássica e Estatística *Bayesiana*, pois não há a imposição de um modelo para descrição, mas sim um trabalho de mineração que pode eventualmente indicar qual o melhor modelo estatístico. O objetivo de um modelo é fornecer um resumo simples de um conjunto de dados, o foco da modelagem está na inferência, ou na confirmação de que uma hipótese é verdadeira. A AED vai além do uso descritivo da estatística, procura olhar de forma mais profunda, no entanto sem resumir muito a quantidade de informações sobre os dados, o que pode resultar em possíveis problemas para a definição das técnicas de visualização e afins.

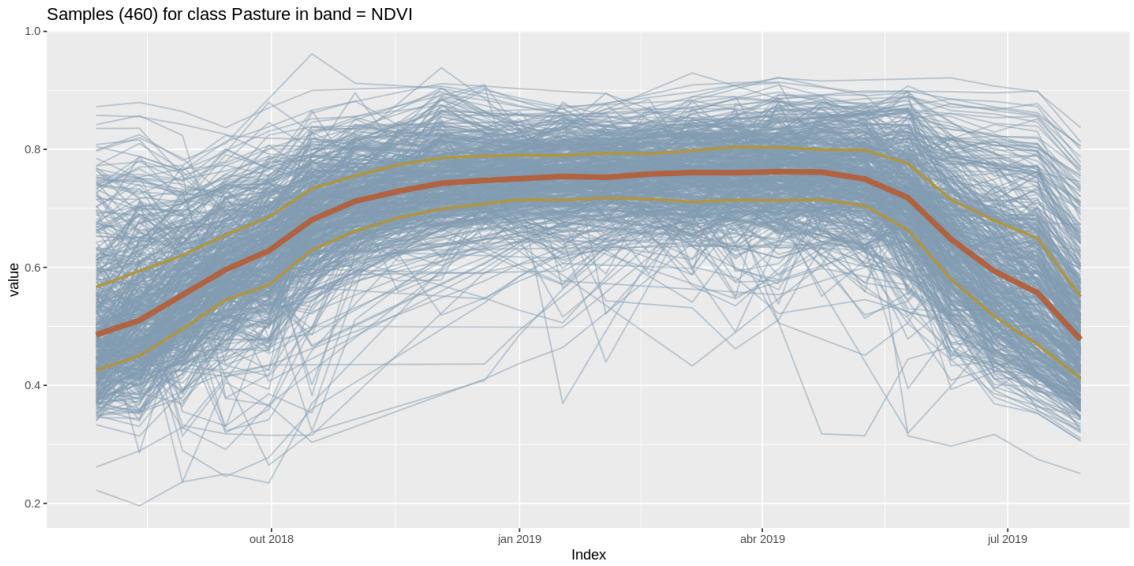
As técnicas de visualização de dados desempenham um papel fundamental na AED, enquanto as tabelas fornecem uma ideia mais precisa e possibilitam um tratamento mais rigoroso aos dados, os gráficos são mais indicados em situações cujo objetivo é representar as variáveis com ilustrações. Os gráficos constituem uma das formas mais eficientes de apresentação sendo composto, essencialmente, por uma figura constituída a partir de uma tabela. Portanto, a qualidade na representação gráfica deve ser pautada na clareza, simplicidade e auto explicação para que a informação seja comunicada ao analista (WICKHAM; GROLEMUND, 2017).

Indiferente do campo de estudo, a AED é fundamentalmente um processo criativo essencial para ciência de dados como uma uma área interdisciplinar que usa cálculos estatísticos para extrair conhecimento e informação. A chave é gerar uma grande quantidade de perguntas, contudo é difícil ser extremamente objetivas no início de uma análise, pois essas visualizações são generalizadas (WICKHAM; GROLEMUND, 2017). Em contrapartida, as perguntas expõem cada vez mais o real estado dos dados, e mesmo que sejam simples de serem respondidas, sempre é recomendado investigar e questionar a veracidade dos dados (TUKEY, 1977).

No contexto das amostras LULC, algumas questões essenciais podem ser levantadas como: quantas amostras cada classe possui? Qual o padrão da série temporal de cada classe? Quais são as amostras que se distanciam deste padrão? O pacote *sits* fornece opções para diversas visualizações que sumarizam um conjunto de amostras LULC e satisfazem estas questões. Dentre elas, o gráfico da distribuição de frequência de amostras por classe, a visualização de todas as séries temporais de uma classe e o cálculo do padrão para as bandas e índices espectrais (SIMOES et al., 2021b).

Como por exemplo, na Figura 2.9 gerada usando pacote *sits* que apresenta 460 séries temporais de NDVI coletadas do satélite *Sentinel-2* para as amostras de pasto no estado de Rondônia. Estas amostras têm o mesmo período de 12 de julho de 2018 a 28 de julho de 2019. Esta visualização representa os dados brutos, a real situação da disposição dos valores nas séries. Os traços que definem a sequência de valores são empilhados, assim é possível perceber se esta classe tem amostras que distanciam da média (em vermelho), do segundo e terceiro quartil (em laranja) e ruídos causados pelo sensor ou presença de nuvens.

Figura 2.9 - Visualização no *sits* para todas as séries temporais de um conjunto de dados de amostras de "Pastagem".



Fonte: Produção do autor.

2.5.2 Análise de métricas e estimativas com aprendizado de máquina

As técnicas de aprendizado de máquina usualmente possuem as seguintes fases principais: coleta de dados, treinamento, avaliação e predição. O treinamento torna o modelo capaz de predizer novos dados, nesta etapa a máquina aprende com erros e acertos. A fase da predição ocorre quando o modelo classificador pode efetivamente ser usado para responder à perguntas para as quais foi treinado, baseando-se na teoria probabilística, onde são calculadas as probabilidades do novo dado pertencer a uma dado grupo ou classe (GOODFELLOW et al., 2016).

A análise de métricas e estimativas com aprendizado de máquina possui uma abordagem não convencional que consiste em usar os resultados da predição de um modelo pré-treinado para calcular a certeza e a confusão dos dados e quantificar a sua qualidade e representatividade (TUIA et al., 2011). Nesta seção serão apresentados alguns métodos que usam este tipo de abordagem na busca de boa qualidade em amostras e mais precisão nos resultados.

2.5.2.1 Aprendizado por transferência

Aprendizado por transferência, em inglês *Transferring Learning*, é o processo de utilizar conhecimentos adquiridos de uma tarefa ou domínio para aprimorar o desempenho de uma tarefa ou domínio relacionado. Nesse processo, um modelo é inicialmente treinado em uma tarefa de origem que possui uma grande quantidade de dados rotulados e, em seguida, o conhecimento adquirido é transferido para uma tarefa de destino que possui dados rotulados limitados. Em resumo, na aprendizagem por transferência como também é conhecida, primeiro uma rede base é treinada, em seguida, transfere-se os recursos aprendidos para uma segunda rede de destino para ser treinada (TUIA et al., 2011).

A aplicação do *Transferring Learning* é especialmente valiosa quando existem semelhanças ou padrões subjacentes compartilhados entre as tarefas ou domínios de origem e destino. Esse mecanismo auxilia o modelo a generalizar de maneira mais eficaz e alcançar um desempenho superior na tarefa de destino, mesmo quando há uma escassez de dados rotulados. Este processo tenderá a funcionar se os recursos forem gerais, ou seja, adequados para as tarefas base e alvo, em vez de específicos para a tarefa base (TUIA et al., 2011).

Belgiu et al. (2020) propõe soluções alternativas com métodos de *Transferring Learning*, com a geração de dados por meio da análise de inventários existentes sobre fenologia de séries temporais para orientar a rotulagem de novas amostras de treinamento. Cada classe possui uma fenologia, um comportamento distinto que descreve eventos sazonais como ciclos de desenvolvimento em resposta às variações ambientais. O método envolve amostragem estratificada em uma área de origem com dados históricos e rotulados, e posteriormente a classificação de amostras não rotuladas de uma área alvo usando como base esses dados de origem com o algoritmo *Time-Weighted Dynamic Time Warping* (TWDTW) e, por fim, ocorre o refinamento usando medidas de proximidade com *Random Forest* (RF).

O *Transferring Learning* é paradigma do aprendizado de máquina que faz um estudo da predição de valores com a reutilização de modelos pré-treinados. Métodos de análise que usam as estimativas calculadas com o aprendizado de máquina necessitam de conjunto dados históricos base para pré-treinar os modelos. No contexto de amostras LULC, a geração de *EO Data Cubes* fornece anos de séries temporais de imagens de sensoriamento remoto. Estes dados históricos podem ser usados como dados base para estabelecer características fundamentais na fenologia de classes LULC como eventos sazonais e variações ambientais, assim pode-se prever o comportamento de uma dada classe em diversas áreas (BELGIU et al., 2020).

2.5.2.2 Aprendizado semi-supervisionado

Aprendizado semi-supervisionado, em inglês *Semi-supervised Learning*, é uma abordagem do aprendizado de máquina que combina tanto dados rotulados quanto não rotulados para treinar um modelo. No aprendizado supervisionado tradicional, apenas dados rotulados são usados para treinar o modelo nas sessões de treinamento, enquanto no aprendizado semi-supervisionado ocorrem diversas sessões com a união de dados rotulados e não rotulados. O *Semi-supervised Learning* aproveita as informações adicionais fornecidas pelos dados não rotulados para aprimorar o desempenho do modelo durante as sessões (LI et al., 2010).

O aprendizado semi-supervisionado também é um paradigma que se relaciona com a estimativa de probabilidade e predição de valores, porém difere-se das etapas comuns usadas para o aprendizado supervisionado. Neste paradigma ocorre primeiro o treinamento de um modelo classificador com um conjunto de dados rotulados de referência, geralmente com um número mínimo de amostras disponíveis e, posteriormente, a predição de novos dados não-rotulados para entrar no conjunto total e treinar o modelo de referência (CHAPELLE et al., 2006).

Normalmente, os algoritmos baseados em aprendizado semi-supervisionado tentam melhorar o desempenho utilizando informações geralmente associadas à outra. Por exemplo, ao lidar com um problema de classificação, os dados adicionais para os quais o rótulo é desconhecido podem ser usados para auxiliar na nova sessão de treinamento e atualizar o modelo. Portanto, o procedimento de aprendizagem pode se beneficiar do conhecimento de novas feições a cada atualização do modelo, resultando em uma classificação mais precisa (GOODFELLOW et al., 2016).

Essa abordagem é útil em cenários nos quais obter dados rotulados é custoso em questão de tempo e recursos necessários (LI et al., 2010). No entanto, métodos que baseiam-se na certeza da predição de dados carecem de mais estudos e adaptações, pois podem gerar problemas de *overfitting* ou especificidade do modelo. Este problema ocorre principalmente pela falta de representatividade no conjunto de dados, uma amostra se diz representativa, quando a mesma possui uma variação na mesma classe. O problema de *overfitting* é identificado quando um modelo obtêm bons resultados no treinamento, porém na avaliação obtêm-se maior confusão na predição devido ao super ajuste do modelo a base de dados de treinamento (MACKAY, 2003).

O uso de métricas e estimativas de probabilidades para a coleta de amostras pode resultar neste problema. Pois quando se usa a certeza da predição na hora de rotular novos dados, o resultado é a especificação do modelo para um determinado conjunto de dados, portanto estudos usando a incerteza da predição com as técnicas de *Active Learning* como apresentado por Tuia et al. (2009) são necessários. A fenologia das classes é influenciada pela região, assim, uma mesma classe pode apresentar diferentes padrões em cada região ou bioma (MERONI et al., 2021). Por causa disso, o uso de modelos pré-treinados de um bioma para outro, no contexto da classificação de imagens no território brasileiro, nem sempre é possível de ser aplicado a todas as classes, pois as práticas agrícolas diferenciam-se pela condição climática.

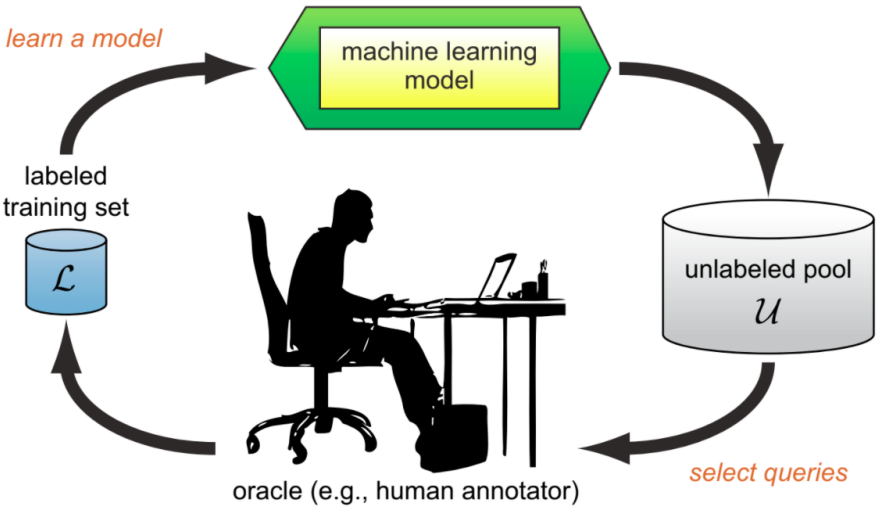
2.5.2.3 Aprendizado ativo

Segundo Tuia et al. (2009), o aprendizado ativo, em inglês *Active Learning*, é uma abordagem em que o modelo classificador consulta ativamente os pontos de dados mais informativos de um grande conjunto de dados não rotulados usando cálculos de incerteza. Sendo um processo iterativo em que o modelo interage com um anotador humano ou especialista para rotular os pontos de dados selecionados. Os dados rotulados são então usados para atualizar e melhorar o desempenho do modelo. O objetivo do *Active Learning* é reduzir a quantidade de dados necessários para o treinamento, ao mesmo tempo em que mantém ou melhora a precisão do modelo.

A metodologia usada por Tuia et al. (2009), baseia-se na definição automática de amostras durante o treinamento usando dois algoritmos que visam melhorar a eficiência e adaptabilidade do modelo classificador: *Margin Sampling by Closest Support Vector* (MS-CSV) e *Entropy Query by Bagging* (EQB). O estudo de caso apresentado aplica esses algoritmos a amostras de imagens ópticas de alta resolução, começando com um conjunto de treinamento pequeno e não ideal e selecionando iterativamente os *pixels* mais valiosos para rotulagem com base em heurísticas predefinidas, logo o processo é repetido até que um critério ideal seja alcançado. Os resultados experimentais mostram que os algoritmos propostos podem reduzir significativamente o número de amostras de treinamento necessárias, mantendo o mesmo nível de precisão de conjuntos de dados mais densos.

Conforme é ilustrado pela Figura 2.10, o *Active Learning* usa um modelo de aprendizado de máquina pré-treinado com um conjunto base de treinamento para analisar as divergências nas previsões do modelo ao predizer novas amostras candidatas, ou seja a confusão é usada para definir dados mais informativos para o modelo, sendo etapa denominada de *query*. Este método é derivado do mesmo princípio da estimativa de probabilidades, porém é realizado um cálculo com base na predição que mede a incerteza para avaliar e ranquear a representatividade das amostras. Usando estes resultados, o oráculo define se as amostras devem entrar no conjunto total ou não, esta etapa pode ser automática ou manual.

Figura 2.10 - O ciclo de *Active Learning* baseado em *pool*.



Fonte: Settles (2010).

A etapa de *query* para selecionar as amostras mais confusas é o que diferencia o *Active Learning* dos métodos semi-supervisionados. Na literatura, os cenários requerem algum tipo de medida de informatividade das instâncias não rotuladas. Existem quatro principais métricas a serem usadas na operação de *query* como apresentado na Tabela 2.3. A amostragem com base na entropia é a mais conhecida, pois são usadas todas as probabilidades para as classes possíveis. A fórmula de entropia (Tabela 2.3) é aplicada a cada instância e o maior valor é consultado. Esta métrica mede a incerteza em uma classificação, logo as amostras com maior entropia são as melhores candidatas por possuírem maior representatividade.

Tabela 2.3 - Lista de abordagens para as métricas de *Active Learning*.

Abordagem	**Descrição**	**Fórmula**
Ratio of Confidence	A taxa ou razão de confiança usa a proporção entre as duas probabilidades mais altas.	$\mathbf{RC} = \frac{p(X_1)}{p(X_2)}$
Margin of Confidence	A análise de confiança usa as classes mais confiáveis e desconsidera as probabilidades de outras classes.	$\mathbf{MC} = 1 - (p(X_1) + p(X_2))$
Least Confidence	A confiança mínima é usada para listar as amostras com a menor confiança para o rótulo previsto. a	$\mathbf{LC} = (1 - p_{max}(X_1))\frac{n}{n-1}$
Information Entropy	A entropia mede a incerteza na classificação, quanto maior a entropia, maior a incerteza da classe.	$\mathbf{H} = -\sum p(X) \log p(X)$

Fonte: Tuia et al. (2011), Walpole et al. (2012), Goodfellow et al. (2016).

Existem outras métricas menos conhecidas, como por exemplo, a explorada por Hu et al. (2010) em *Exploration Guided Active Learning* (EGAL). O EGAL classifica as amostras com base em sua densidade e diversidade, logo as amostras com maior EGAL devem ser selecionadas para entrar no conjunto total.

A ideia principal do *Active Learning* resume-se a encontrar as instâncias em que o modelo de aprendizado de máquina tem maior probabilidade de aprender e tornar-se mais robusto a erros. O conceito carrega um aprimoramento da aprendizagem semi-supervisionada baseado no estudo da confusão dos resultados. Este método pode ser usado para auxílio na coleta de novas amostras ou para refinamento na busca por dados representativos.

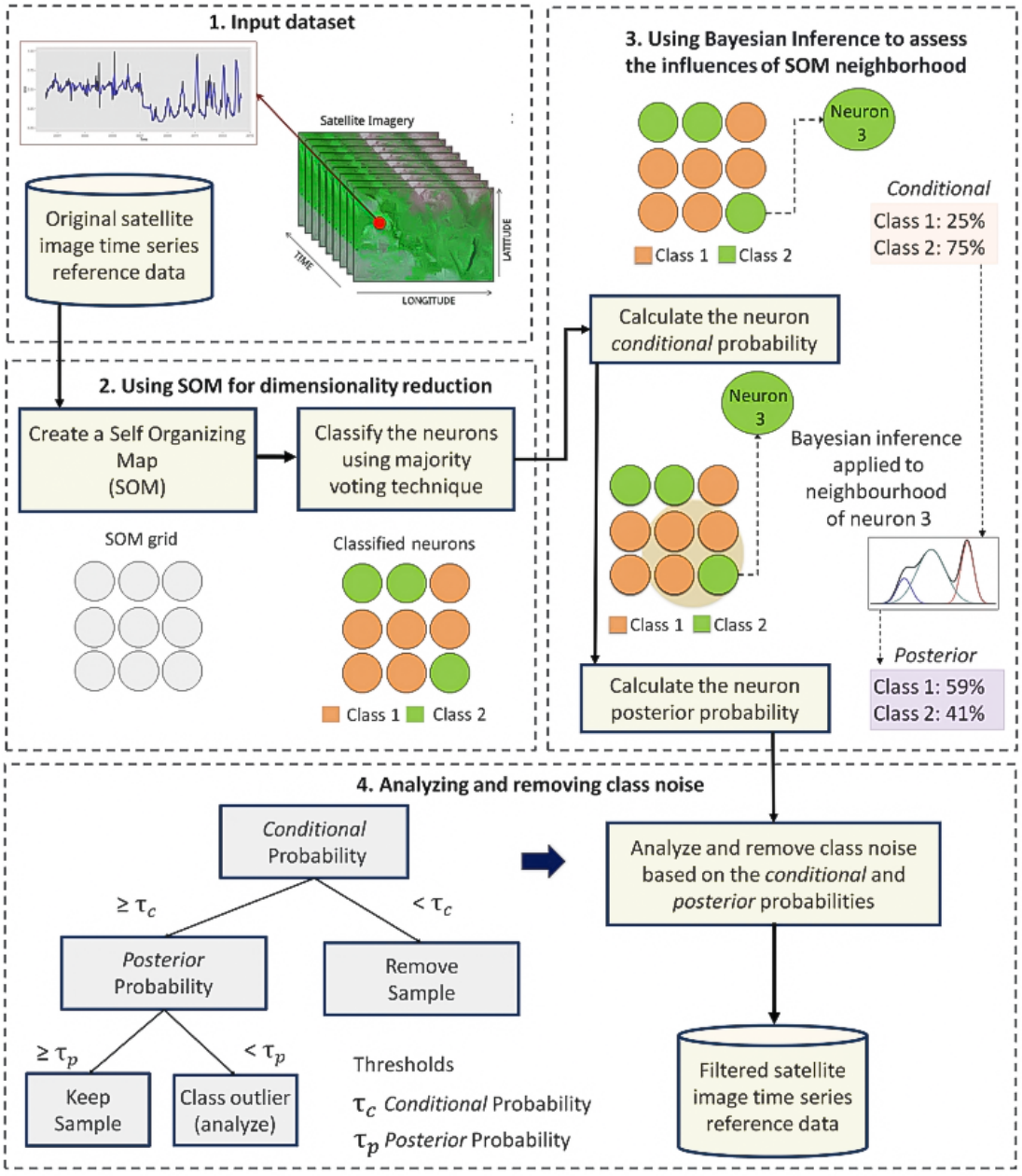
2.5.3 Controle de qualidade e redução de ruído de classe

Santos et al. (2021a) propôs um método que avalia a qualidade das amostras LULC com base na inferência *bayesiana* aplicada no agrupamento com *Self-organizing maps* (SOM). Este artigo aborda o problema de ruído de classe em séries temporais e propõe um método para reduzir esse ruído e melhorar a qualidade dos dados de treinamento. O método usa o SOM para redução de dimensionalidade e agrupamento. A inferência *bayesiana* é usada para avaliar a similaridade e consistência de amostras nos grupos criados e mensurar a probabilidade do rótulo ter sido atribuído corretamente. A Figura 2.11 ilustra de forma esquemática o método para controle de qualidade e redução de ruído.

O SOM (ou *Kohonen Map*) é um método de aprendizado não supervisionado que permite a projeção de conjuntos de dados de alta dimensão em uma representação de menor dimensão. Por meio desse processo de agrupamento, os dados são mapeados em uma grade bidimensional, na qual cada unidade é denominada neurônio (KOHONEN, 1990). No caso de amostras LULC, como o SOM possui a propriedade de preservar a topologia, de modo que padrões similares são agrupados em proximidade na grade de neurônios, onde cada neurônio será associado a várias séries temporais, como resultado obtêm-se o mapa 2D.

Neste contexto, dado que o SOM preserva a estrutura topológica de vizinhanças em múltiplas dimensões, os neurônios que contêm amostras de uma determinada classe geralmente serão vizinhos no espaço 2D. Por causa da estrutura em grade no SOM, os vizinhos de cada neurônio fornecem informações sobre a variabilidade intra-classe e inter-classe. Nesse contexto, a inferência *bayesiana* é aplicada considerando a vizinhança de cada neurônio. Ao analisar as probabilidades resultantes, é possível identificar amostras com rótulos incorretos e neurônios atípicos, indicando ruído de classe. Neurônios atípicos são aqueles cuja classe majoritária difere da de seus vizinhos, sugerindo variabilidade ou rotulagem incorreta.

Figura 2.11 - Metodologia para o controle de qualidade e redução de ruído de classe.



Fonte: Santos et al. (2021a).

A avaliação de qualidade é feita com base nos filtros de *threshold* aplicado nas probabilidade *a priori* e *posteriori* resultantes do processo de agrupamento com a inferência. Ao aplicar esses filtros de limite, o método sinaliza amostras que devem ser mantidas ou removidas ou podem exigir uma análise mais aprofundada. Esta sinalização faz uso de três rótulos para identificar o *status*: "*clean*" (permanece no conjunto), "*analyze*" (análise necessária) e "*remove*" (remoção aconselhável).

O estudo de caso apresentado por Santos et al. (2021a) se concentra em um conjunto de amostras de treinamento do bioma Cerrado no Brasil. Os dados apresentados no artigo são um conjunto de amostras coletadas por especialistas através de interpretação visual, além da coleta em campo com a observação e entrevistas com agricultores locais fornecidas pelo INPE. Foram utilizadas imagens de sensoriamento remoto de um período de sete anos (2010-2017) do sensor *MODIS* associadas com as classes de cultivo (*Cropland*) e pasto (*Pasture*). O bioma Cerrado é o segundo maior bioma da América do Sul, cobrindo uma área de mais de 2 milhões de quilômetros quadrados, por conta disso há grande variabilidade nos padrões das classes LULC neste bioma (BRASIL, MINISTRO DO MEIO AMBIENTE, 2022).

Os resultados mostraram um impacto positivo na precisão geral da classificação usando o conjunto de dados filtrado que passou pelo processo de redução de ruído, em comparação com o conjunto de dados original. Amostras consideradas ruins causam ruído e confusão no treinamento, por isso devem ser identificadas e removidas, o que revelou-se uma etapa essencial, apesar de adicionar um custo extra na análise.

Contudo, nem sempre as amostras identificadas com erro precisam ser removidas, é necessário um estudo profundo sobre os padrões da classe para identificar se a mesma representa a classe ou uma subclasse no conjunto. Uma amostra representativa é aquela que possui diversidade ou pequenas variações na sequência de valores da série temporal. Como apresentado por Santos et al. (2021b), as técnicas de redução de dimensionalidade do SOM podem ser usadas para agrupar amostras similares e, posteriormente, o agrupamento hierárquico para identificar subclasses.

2.6 Ferramentas para a coleta e análise de amostras

Os Sistemas de Informação Geográfica ou pela sigla SIG's, em inglês *Geographic Information Systems* (GIS), são amplamente utilizados na pesquisa científica devido à sua capacidade de mapear interações humanas com o meio ambiente. Um SIG consiste em um conjunto de ferramentas de hardware e software que processam, armazenam e analisam dados geoespaciais para a representação digital da Terra (WIECZOREK; DELMERICO, 2009). Para descentralizar o processamento, SIG's de propósito geral oferecem conexões com serviços *web*, fortalecendo o desenvolvimento de ferramentas para atender demandas cada vez mais específicas entre os usuários (GOMES et al., 2020).

A *Open Geospatial Consortium* (OGC) desempenha um papel essencial no estabelecimento de padrões nos serviços *web*, dentre eles o Web Map Service (WMS) e o Web Coverage Service (WCS) (OPEN GEOSPATIAL CONSORTIUM (OGC), 2022). Serviços *web* são soluções eficazes para o compartilhamento e processamento de informações, abstraindo métodos de busca e acesso a dados geoespaciais e diminuindo o uso de infraestrutura do cliente. Além de que, estes serviço facilitam a interoperabilidade, para que os usuários lidem apenas com o necessário ao realizar as análises em questão de armazenamento (VINHAS et al., 2017).

A coleta de amostras por interpretação visual de imagens de satélite através de SIG's de propósito geral, como o QGIS, ou desenvolvidos especificamente para a coleta, como o *CollectEarth* está se tornando uma estratégia muito usada (BEY et al., 2016). Estes sistemas oferecem ferramentas para a visualização e análise de mudanças usando conjuntos de imagens, permitindo uma visão mais ampla das características de uma determinada área. A aquisição de amostras é uma etapa complexa e necessita de cuidados, como por exemplo, o estudo da área-alvo deve ser explorada como um conjunto de variáveis e suas características próprias que podem influenciar diretamente nas classes alvo (MERONI et al., 2021).

Existem diversos SIG's que permitem a coleta de amostras seguindo os princípios para a interpretação de imagens como, por exemplo, o *QGIS*, o *Google Earth Engine* (GEE), o *CollectEarth* e o *TerraCollect*. *Softwares* como o QGIS e GEE são exemplos de SIG's que contemplam ferramentas analíticas completas, porém requerem algumas habilidades que muito dos especialistas não possuem, como construção de algoritmos ou técnicas de visualização usando linguagens de programação. O *CollectEarth*, por exemplo, mesmo oferecendo mais abstração com uma interface gráfica, não integra métodos analíticos para o processamento de séries temporais e auxiliar na produção de amostras representativas (BEY et al., 2015; FLENNINKEN et al., 2020; GOOGLE, 2023).

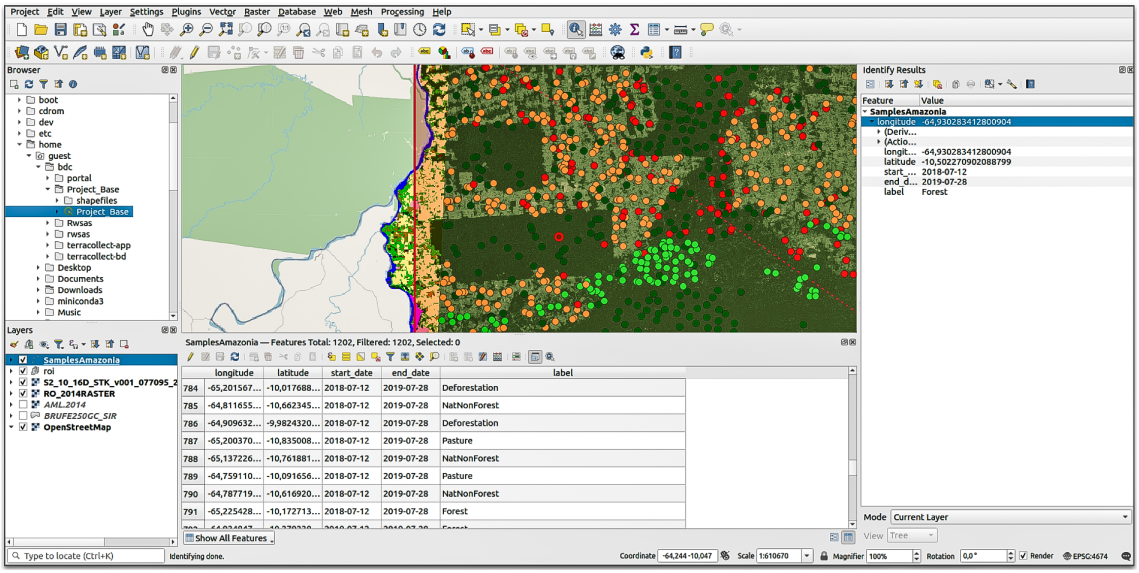
A falta do suporte à análise de dados e de ferramentas completas em SIG's, comumente, força os especialistas a migrarem o conjunto de dados de uma aplicação para outra para realizar algum tipo de análise mais profunda. O suporte ao *Big EO Data* ainda configura um desafio para a infraestrutura tradicional na maioria dos SIG's, por isso sistemas mais robustos capazes de armazenar, processar e analisar densas bases de dados estão sendo desenvolvidos para atender a esta demanda. Nesta seção serão analisadas plataformas e serviços *web* relacionados direta ou indiretamente à coleta de amostras por interpretação visual de imagens.

2.6.1 QGIS

O QGIS é um *software* livre criado em 2002 produto de um projeto dirigido por voluntários que aceita contribuições de usuários na forma de código, desenvolvimento de ferramentas, extensões, relatórios, correções de problemas, documentação e suporte, tendo como vantagem o compartilhamento de conhecimento. Diferente de sistemas como o GEE e *Sentinel Hub* com versões gratuitas e funções limitadas, o QGIS é um *software* gratuito com funções completas. Usualmente, custos de software podem impedir o desenvolvimento e a expansão de novas ideias dentro do ecossistema de SIG's, a solução pode ser o uso e o desenvolvimento de *softwares* gratuitos e de código aberto (FLENNINKEN et al., 2020).

A Figura 2.12 apresenta a interface gráfica da versão mais recente do *software* QGIS. Este exemplo apresenta a visualização de 1.202 amostras LULC no estado de Rondônia, no período de 12 de julho de 2018 a 28 de julho de 2019. As amostras estão sobrepostas a um recorte de 28 de Julho de 2018 do satélite *Sentinel-2* e a um mapa LULC de Rondônia em 2014 proveniente do TerraClass. A análise espacial é de extrema importância para a classificação de amostras e geração de mapas, por isso o QGIS dispõe de ferramentas de edição para as imagens *raster* e mais o recorte e desenho de geometrias (QGIS DEVELOPEMENT TEAM, 2023).

Figura 2.12 - Visualização de amostras de uso e cobertura da terra sobrepostas a uma imagem de alta resolução Sentinel 2 no QGIS.



Fonte: Produção do autor.

No exemplo (Figura 2.12), o QGIS foi usado para o recorte da área e definir um conjunto de dados para a análise. O projeto TerraClass no período selecionado de 2019 a 2020 possuía mais de 4.000 amostras, das quais 500 amostras foram selecionadas baseada no método retângulo envolvente de uma dada localização. Há na parte inferior da tela a visualização da tabela de atributos destas amostras, onde pode-se analisar os dados de forma tabular e realizar pesquisas. Uma função muito usada no QGIS é a conexão com o banco de dados que torna o armazenamento mais padronizado e eficiente, outras conexões também são oferecidas como a conexão FTP e local usando sistemas de arquivos.

É notório que o QGIS é popular com algumas desvantagens em relação aos outros SIG's, porém, em contrapartida por ser gratuito fomenta o desenvolvimento de extensões de *software* (conhecidas pelo acrônimo *Plug-in*). O desenvolvimento de *Plug-ins* permite que pesquisadores construam programas baseados em necessidades específicas para a solução de um problema, utilizando as funções pré-programadas do QGIS (QGIS ASSOCIATION, 2020).

Os *plug-ins* no QGIS são possíveis devido ao suporte à linguagem de programação *Python*. Por ter uma pequena curva de aprendizagem e ser multi-paradigma, a linguagem *Python* é muito usada para a ciência de dados. Como por exemplo, o *Plug-in QCircularStats* que auxilia de forma abrangente, compreensível e visual realizar análises estatísticas circulares sobre os dados extraídos de imagens de sensoriamento remoto (CUARTERO et al., 2023).

O *plug-in Q-LIP* também é um bom exemplo de uma extensão para auxiliar análises no QGIS. Este *plug-in* foi criado com o objetivo de auxiliar os usuários em tarefas que normalmente exigiriam algum conhecimento em códigos e linguagem de programação. O *Q-LIP* foi desenvolvido por Sebbah et al. (2021) para o cálculo automático de índices espectrais específico para o satélite *Landsat-8*. Entre suas funcionalidades, estão o download, o cálculo e a visualização dos índices com base nas imagens. Os métodos são abstraídos pela interface gráfica do QGIS.

O QGIS em conjunto com suas extensões oferecem ferramentas completas para a análise e coleta de dados com várias funcionalidades, desenvolvidos unindo diferentes tecnologias. Outro tópico em questão é a complexidade dos SIGs, que envolve a integração de ferramentas em diversas linguagens de programação, conhecidos como *Stacks* ou conjuntos de tecnologias (FOWLER; LEWIS, 2014). Frequentemente tornando-os complexos demais para o usuário, portanto uma interface limpa e objetiva é um requisito indispensável.

A documentação do QGIS é considerada limitada e precária em relação às suas diversas funcionalidades e aplicações. Outra limitação comum de softwares de código aberto em geral é a falta de suporte ao usuário na forma de manuais e tutoriais dependendo da comunidade de desenvolvimento. Além disso, algumas funcionalidades mais atuais em SIG's como a visualização em três dimensões e a criação de animações podem levar tempo até serem integradas, em comparação a demais *softwares* pagos (FLENNINKEN et al., 2020).

2.6.2 *Google Earth Engine*

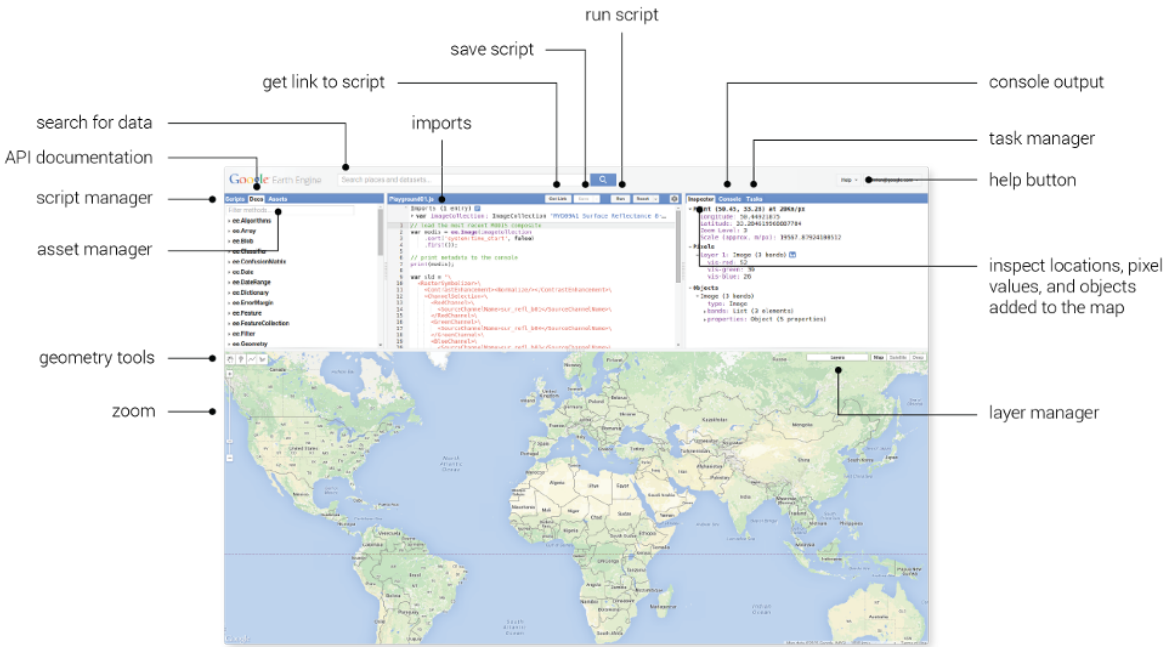
O *Google Earth Engine*, ou GEE, é uma plataforma *web* que permite aos usuários realizar a análise e o processamento de dados geoespaciais na infraestrutura computacional do *Google*. O GEE oferece uma interface simples baseada em um editor de código com uma linguagem de programação adaptada especialmente para esta infraestrutura. A plataforma dispõe de dois componentes principais, o editor de código e o catálogo de dados. O editor de código é um ambiente de desenvolvimento integrado baseado na *web* para escrever e executar *scripts*. Já o catálogo possui um aplicativo *web* leve para explorar os dados e executar análises simples. O GEE também oferece bibliotecas clientes em *Python* e *JavaScript* para os serviços *web* disponibilizados pelo projeto (GOOGLE, 2023).

Segundo Gorelick et al. (2017) a computação em nuvem em grande escala e o acesso ao *Big Data* está universalmente disponível como uma mercadoria. Contudo, o proveito desse recurso por meio de SIG's baseados em *softwares* instalados de forma local e centralizada como o *QGIS*, ainda requer um poder de gerenciamento em computação. Se torna indispensável o conhecimento técnico no gerenciamento básico de tecnologia da informação, tais como a aquisição de computadores com alta capacidade de CPU, o armazenamento em banco de dados, alocações de máquinas e organização de redes.

Até então, na discussão sobre SIG's como ferramentas de coleta e análise de dados amostrais foram citados *softwares* centralizados e instalados na própria infraestrutura do usuário. Esta infraestrutura inclui banco de dados e recursos de *hardware*, porém o GEE usa a computação em nuvem descentralizada que permite que usuários acessem tais funcionalidades apenas com o acesso à internet. Essa característica da arquitetura do GEE é uma vantagem com relação aos outros SIG's de propósito geral (GORELICK et al., 2017).

A Figura 2.13 lista e indica as principais funcionalidades na plataforma *web* do GEE. A interface principal é dividida em três componentes: O primeiro corresponde ao gerenciamento de *scripts* e base de dados; o segundo é o próprio editor de *scripts*; O terceiro dispõe a visualização dos resultados da execução dos comandos. O GEE inclui ferramentas e abstrações para o armazenamento e processamento na nuvem, ou seja, o sistema oferece uma capacidade de processamento escalável e rápida.

Figura 2.13 - Funcionalidades do *Google Earth Engine*.



Fonte: Google (2023).

Explorando o GEE como um sistema para a coleta e análise de amostras LULC, há uma interface interativa que permite desenhar geometrias diretamente no mapa, como pontos, linhas e polígonos, sendo o usuário capaz de criar, editar e excluir essas geometrias. Há também a importação e exportação de informações em formatos como *GeoJSON* e *Shapefile*. Isso permite trabalhar o *design* amostral em outra plataforma, por exemplo, e migrar os pontos para o GEE com áreas de interesse personalizadas. As funcionalidades no GEE são listadas em sequência:

- **Acesso a dados geoespaciais:** O GEE oferece vários conjunto de dados geoespaciais, incluindo dados de sensoriamento remoto como imagens de satélite e climatologia;

- **Ferramentas de visualização:** ferramentas de visualização para explorar e exibir dados geoespaciais em mapas interativos, incluindo a capacidade de criar animações, gerar imagens compostas e visualização de resultados por meio de links ou incorporação em páginas *web*;

- **Linguagem de programação:** A plataforma possui semelhanças com ambientes de desenvolvimento integrado, o editor de código usa a linguagem *JavaScript* para escrever os comandos para a análise e permite o compartilhamento de *scripts*;

- **Análise geoespacial:** é possível realizar análises como cálculos de índices, classificação de imagens, detecção de mudanças e modelagem de fenômenos geográficos. É possível usar geometrias como filtros para consultar dados geoespaciais ao filtrar imagens de satélite e demais dados;

- **Operações com geometrias:** A plataforma permite o desenho livre ou importação de geometrias, onde pode-se realizar cálculos espaciais, como interseção, união, diferença e mapas de distância (*buffer*).

Algumas funcionalidades como a extração e o processamento de séries temporais, por exemplo, ficam à cargo do usuário, este tipo de dado é um requisito para métodos de análise de amostras LULC. Esta operação requer conhecimentos em linguagem de programação e análise de dados geoespaciais. Mesmo dispondo de ferramentas para a aplicação de algoritmos estatísticos, criar *scripts* no GEE exige uma curva muito extensa de aprendizagem. O GEE ganha em processamento por causa da computação em nuvem, mas em contrapartida possui a desvantagem de necessitar deste tipo de conhecimento. Este fardo pode colocar esta ferramenta fora do alcance de muitos pesquisadores que desejam aproveitar o máximo destes recursos.

2.6.3 *Collect Earth*

Segundo Bey et al. (2015), o *Collect Earth* é sobretudo um sistema de informação para a coleta de dados LULC com base na visualização e interpretação de imagens de satélite de código aberto desenvolvido pela SERVIR em parceria com organizações técnicas regionais como a *Food and Agriculture Organization of the United Nations* (FAO). O código-fonte é compartilhado por meio da *Open Foris Initiative* da FAO. O SERVIR é uma iniciativa de desenvolvimento conjunto da Administração Nacional de Aeronáutica e Espaço dos EUA (NASA) e da Agência dos EUA para o Desenvolvimento Internacional (USAID).

O *Collect Earth* possui plataformas em duas modalidades: uma versão *online* e *offline*. A versão *online* chama-se *Collect Earth Online* (CEO), um ambiente *web* onde os processos são realizados na infraestrutura do servidor e a *offline*, um *software* instalado que utiliza a infraestrutura do usuário. Diferente de SIG's como o QGIS e GEE que oferecem apenas uma modalidade. Uma das vantagens do *Collect Earth* é a integração com os bancos de dados SQLite e PostgreSQL, permitindo que a plataforma se adapte ao ambiente do usuário (BEY et al., 2016).

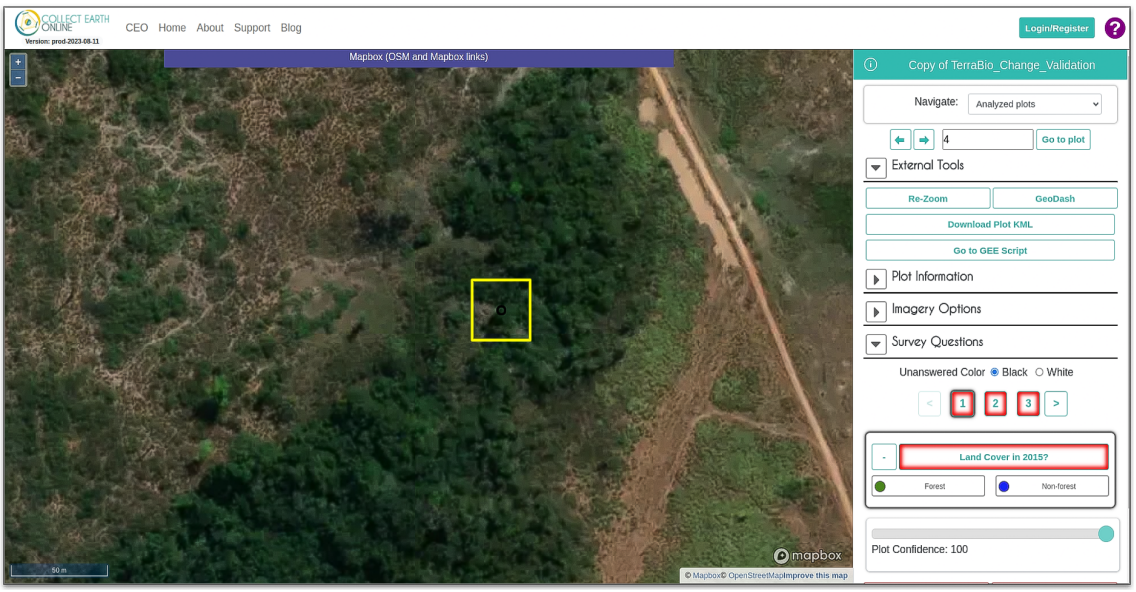
O objetivo do *Collect Earth* é fornecer ferramentas de *software* gratuitas para o monitoramento de mudanças na superfície terrestre com base na interpretação visual de imagens de satélite. O *Collect Earth* oferece funcionalidades para coletar dados LULC com a entrada simultânea de dados por vários usuários. Ele foi projetado para promover a consistência na localização, interpretação e rotulagem de parcelas de dados de referência para uso na classificação e monitoramento de mudanças. A versão *online* CEO também é baseada na computação em nuvem, portanto, o *software* não requer instalação na infraestrutura do usuário e exige apenas uma conexão com a internet para ser utilizada. Esta versão inclui a conexão de bancos de dados e coleta de pontos de referência (SAAH et al., 2019).

O CEO permite a coleta de dados de forma sistemática usando como base o GEE e produtos derivados. As imagens de alta resolução multi-temporais são provenientes de projetos como o *Sentinel Hub*, *Bing Maps* e coleções de imagens fornecidas pelo GEE. As ferramentas disponíveis permitem a realização de inventários florestais, monitoramento de uso da terra, análise de cobertura do solo, entre outras aplicações. O CEO possui componentes projetados para aprimorar o processo de interpretação, fornecendo dados adicionais e visualizações para auxiliar na classificação das categorias de cobertura de terra e uso da terra.

O CEO oferece uma interface somente para a visualização e análise de dados, o *Geo-Dash*, onde os usuários têm acesso a um conjunto de ferramentas interativas para auxiliar na interpretação e classificação LULC. Essas ferramentas fornecem acesso a vários tipos de informações, como coleções de imagens, gráficos de séries temporais, estatísticas, coleções de imagens duplas ou ativos de imagens pré-processadas, além de que, o usuário pode personalizar a interface para mostrar informações específicas relevantes para seu projeto. Porém, o *Geo-Dash* é voltado apenas na visualização, não permitindo o processamento, aplicação de filtros ou métodos estatísticos nos dados (SAAH et al., 2019).

A Figura 2.14 apresenta a interface para a coleta de amostras usando a versão *online*, o CEO. Esta tela apresenta dados públicos do projeto *TerraBio* da *Catalyzing and Learning through Private Sector Engagement for Biodiversity Conservation* (CAL–PSE), em parceria com a USAID/Brasil, Aliança para Bioversidade e *Spatial Informatics Group*. *TerraBio* é uma abordagem para monitoramento, avaliação e geração de relatórios para fornecer avaliação ambiental orientando para a conservação e responsabilização de empresas do setor privado. O projeto tem foco em empresas que comercializam produtos agrícolas e sustentáveis ou investem em modelos de negócios sustentáveis como desenvolvimento lucrativo e orientado para a conservação (CAL-PSE, 2023).

Figura 2.14 - Tela de exemplo para coleta de amostras no *Collect Earth Online*.



Fonte: CAL-PSE (2023).

Ao contrário de SIG's como o GEE e o QGIS de propósito geral, o *Collect Earth* é uma especificação para suprir a necessidade por uma plataforma de coleta de dados LULC, tendo as funcionalidades limitadas somente à coleta e interpretação de imagens. Isso o torna mais acessível aos usuários por ser menos complexo em relação aos demais. Os usuários podem traçar e delinear áreas de interesse nas imagens de satélite, associando atributos a essas áreas por meio de formulários personalizados e armazenar estas informações, e posteriormente realizar o *download* dos dados coletados em arquivos *shapefile*.

A interface intuitiva e recursos flexíveis tornam o *Collect Earth* uma ferramenta valiosa para estudos relacionados à conservação ambiental, manejo de recursos naturais e monitoramento de áreas protegidas. Uma vez que o uso da plataforma envolve etapas simples como a instalação se necessário, definição do projeto, escolha das imagens a serem usados para o objeto de estudo. Após esta etapa é possível coletar os dados e exportá-los posteriormente. Em sequência há uma lista das principais operações oferecidas pelo *Collect Earth*:

- **Coleta de dados LULC:** Os usuários podem inserir informações por meio de formulários personalizáveis sobre o uso da terra, cobertura do solo, características do terreno, entre outros parâmetros relevantes para o estudo em questão;
- **Desenho de geometrias:** A ferramenta oferece recursos para desenhar geometrias diretamente nas imagens de satélite, permitindo a delimitação precisa das áreas de interesse para a coleta de dados;
- **Acesso à imagens de satélite:** A ferramenta integra imagens de satélite de alta resolução proveniente do *Sentinel Hub*, *Bing Maps* e *Google Earth Engine*, isto permite a análise detalhada das áreas de interesse;
- **Análise e monitoramento de mudanças:** Os usuários podem comparar imagens de satélite de diferentes períodos e registrar as mudanças observadas nas áreas estudadas;
- **Colaboração e compartilhamento:** Os dados podem ser exportados em vários formatos, como CSV, KML ou *Shapefile*. O *Collect Earth* facilita a colaboração entre pesquisadores, permitindo o compartilhamento de projetos e dados entre membros de uma equipe ou instituição.

Como uma ferramenta de código aberto, o *Collect Earth* é constantemente atualizado e aprimorado por uma comunidade de desenvolvedores e usuários. Ele é usado tanto por organizações governamentais quanto por instituições acadêmicas. A ferramenta é especialmente útil para estudos de monitoramento de florestas tropicais, onde a combinação de dados de satélite com informações coletadas em campo permitem uma avaliação mais precisa das mudanças na cobertura florestal, identificação de áreas desmatadas e monitoramento de processos ambientais.

No entanto, no *Collect Earth* a coleta é sistemática e baseada em grades e áreas de estudo, não permitindo desenhos livres ao usuário. Além disso, o sistema em si não possui recursos nativos para extrair séries temporais de dados com um grande conjunto de amostras, o que torna a análise temporal de *pixels* uma tarefa custosa com relação às demais ferramentas. Para aplicar filtros e processar séries temporais, o usuário deve considerar o uso de plataformas como o GEE, neste contexto pode ser uma opção mais adequada do que o *Collect Earth*. No entanto, é importante observar que o GEE ainda sofre com problemas de complexidade com relação às suas funcionalidades.

2.6.4 *TerraCollect*

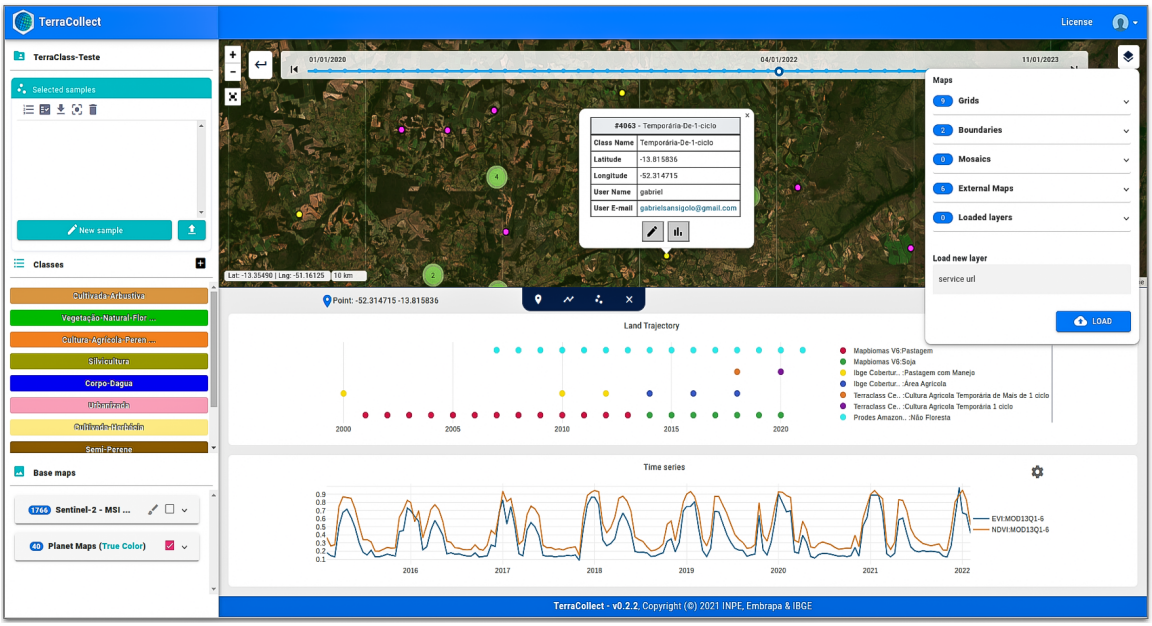
O *TerraCollect* é uma plataforma *web*, atualmente em fase de desenvolvimento no projeto BDC, voltada para a coleta e análise de amostras LULC com base em séries temporais de imagens. Uma ferramenta essencial para o Programa de Monitoramento Ambiental do INPE, bem como projetos nacionais, como o TerraClass. A plataforma integra as tecnologias fornecidas pelo BDC para visualização e análise temporal de *EO Data*, disponibilizando o acesso aos produtos dos satélites *CBERS-4*, *CBERS-4A*, *Landsat-8*, *Sentinel-2* e *MODIS* catalogados com o serviço STAC. O sistema para o *TerraCollect* integra componentes como a interface gráfica na aplicação *web*, os serviços de gerenciamento, controle de usuários e banco de dados para armazenar os dados coletados na infraestrutura do servidor (ANJOS et al., 2022).

O *TerraCollect* fornece insumos para o apoio na tomada de decisão durante a coleta como acesso à mapas de referências disponibilizados via padrão OGC *Web Map Service* (WMS) provenientes de projetos como TerraClass e PRODES. Há também o acesso aos gráficos de séries temporais e trajetórias LULC, respectivamente pelos serviços WTSS e o *Web Land Trajectory Service* (WLTS) (ANJOS et al., 2022). O WLTS fornece o acesso à trajetórias LULC através de mapas fornecidos pelo PRODES, TerraClass e *MapBiomas*, desta forma é possível visualizar como um dado ponto foi classificado ao longo do tempo (ZIOTI et al., 2021).

Esta plataforma *web* permite criar projetos de coleta de dados com configurações definidas pelo usuário. Configurações como o sistema de legendas de classes com a simbologia para representar as amostras, a lista para o acesso aos cubos e coleções de imagens alinhadas no tempo, e mais a lista dos gráficos para a análise temporal. Com o projeto criado é possível realizar a coleta usando desenho livre ou fazer o *upload* de pontos a serem rotulados na plataforma. Sendo um ambiente multiusuário, cada projeto é formado por um conjunto de especialistas.

Durante a coleta, as amostras são salvas em um banco de dados utilizando o modelo do *Sample Database Model* (*Sample-DB*) através do serviço *Sample Web Service* (*Sample-WS*). Com essas ferramentas, o *TerraCollect* consegue armazenar, consultar e acessar as amostras coletadas. O modelo Sample-DB permite descrever os dados e seus metadados de forma organizada facilitando a reprodutibilidade de experimentos e compartilhamento de informações. A Figura 2.15 apresenta a interface para a coleta de amostras no *TerraCollect*, onde é possível criar novos pontos e visualizar a série temporal, a trajetória LULC e ao mesmo tempo visualizar as imagens de satélite como mapas base.

Figura 2.15 - Interface para a coleta de amostras no *TerraCollect*.



Fonte: Produção do autor.

Em sequência há a lista completa das principais funcionalidades do *TerraCollect*:

- **Acesso a mapas base:** a interface gráfica fornece a visualização de mapas base como plano de fundo do mapa interativo como a *Planet*, *Google Maps* e demais produtos do BDC como mosaicos;

- **Acesso a mapas auxiliares:** a aplicação permite a visualização de limites municipais, biomas e estados brasileiros, além de ser possível acessar mapas classificados de outros projetos como o PRODES, DETER e TerraClass;

- **Visualização de imagens de sensoriamento remoto:** a aplicação permite que os usuários realizem a busca por imagens de satélite selecionando uma área de estudo e visualizem os resultados incluindo os satélites *CBERS-4*, *CBERS-4A*, *Landsat-8*, *Sentinel-2* e *MODIS*;

- **Análise e monitoramento de mudanças:** no *TerraCollect*, a visualização dos resultados das imagens e dos mapas base é feita por meio de uma linha temporal, onde pode-se selecionar a data da imagem a ser visualizada;

- **Acesso à séries temporais e trajetórias LULC:** selecionando as amostras ou qualquer ponto no mapa, é possível retornar os dados de séries temporais e trajetórias em gráficos para uma visualização interativa;

- **Exportação de séries temporais:** a aplicação possui suporte para a extração, análise, processamento e exportação das séries temporais associadas às amostras em diferentes formatos de arquivos como *CSV* e *Rdata*;

- **Gerenciamento de amostras:** O *TerraCollect* possui o suporte para o desenho de pontos e associação destes pontos à classes, além de um modelo de bando de dados desenvolvido especialmente para amostras LULC, o que facilita o gerenciamento dos dados.

Uma característica importante do *TerraCollect* é a execução na infraestrutura dos servidores do projeto BDC. Desta forma o usuário só precisa ter acesso à internet e acessar via *web* sem instalar dependências. Além disso, a plataforma oferece uma interface gráfica para todas as ferramentas citadas abstraindo as etapas para o acesso aos dados, ou seja, não há a necessidade de conhecimentos prévios em algoritmos e linguagem de programação como o GEE ou o QGIS. A visualização das séries temporais, por exemplo, requer apenas uma seleção de um ponto no mapa interativo conforme demonstrado na Figura 2.15.

Tanto a plataforma *TerraCollect* quanto o *Collect Earth* possuem funcionalidades especializadas para a coleta e análise de amostras, permitindo o usuário realizar estudos profundos nos dados. A principal diferença entre estas plataformas é que o *TerraCollect* é focado na análise de séries temporais, onde durante a coleta cada amostra possui uma série temporal associada de acordo com o cubo de dados selecionado pelo usuário. Neste contexto, as aplicações são extensas como o monitoramento e controle de safras, a análise do ciclo fenológico na agricultura e o monitoramento de mudanças ao longo do tempo com a análise temporal das imagens.

O *TerraCollect* pode ser uma opção mais adequada quando o objetivo final é usar as amostras associadas com a série temporal no treinamento de modelos de aprendizado de máquina. Os modelos treinados são usados na classificação de imagens para a geração de mapas LULC. O diferencial da plataforma, em relação às demais, é o suporte a extração, processamento e exportação de séries temporais para um grande conjunto de dados. Contudo, o *TerraCollect* ainda não está aberto ao público geral, sendo testado apenas com projetos privados, além de ainda estar na fase de desenvolvimento e elicitação de requisitos pela equipe do projeto BDC.

2.7 Aplicações *web* para ciência de dados com R

Ciência de dados é uma área abrangente e multidisciplinar, tornando possível a transformação de dados brutos em conhecimento e informação. A comunidade R é colaborativa com o compartilhamento de conhecimento. A linguagem R é uma ferramenta estratégica e eficiente para análise estatística de dados. Uma ferramenta de código aberto com pacotes e funcionalidades para manipulação, visualização e modelagem de dados (R CORE TEAM, 2013).

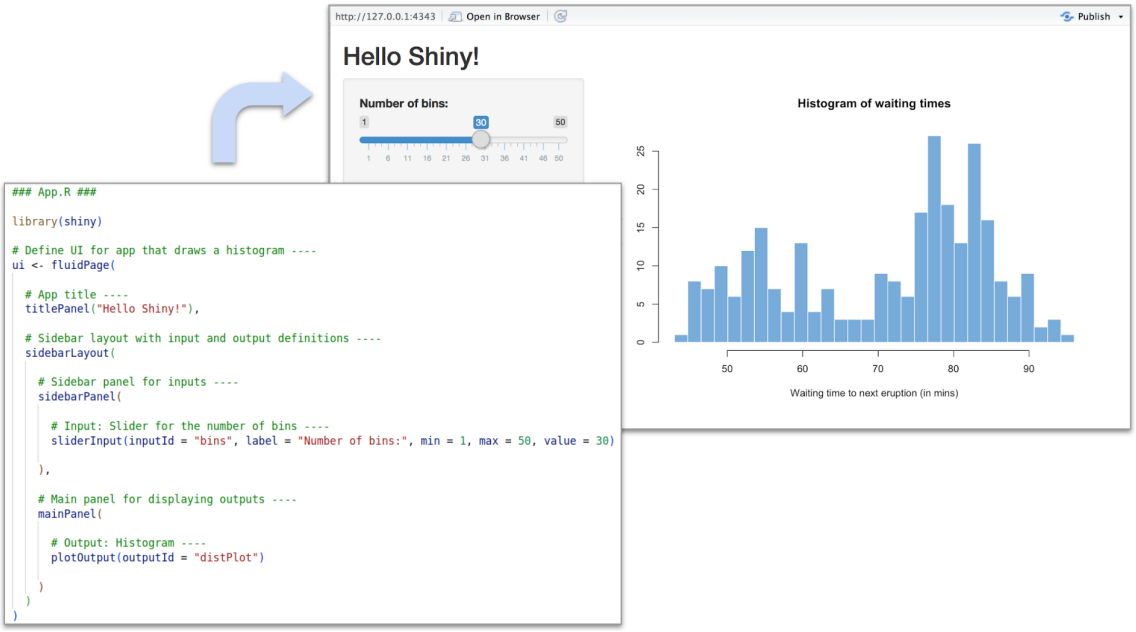
Esta linguagem leva o pesquisador a ter habilidades em análises complexas e tomada de decisões embasadas em evidências. Porém, quando se aborda questões como aplicações e ferramentas gráficas para análise na *web*, o R é deixado de lado e pesquisadores optam por linguagens melhor fundamentadas nesta área de desenvolvimento como *Python* ou *Java*.

2.7.1 *Shiny Rstudio*

Shiny é um pacote para desenvolvimento de aplicações *web* interativas disponível na linguagem de programação R e *Python*. A versão para o R chama-se *Shiny Rstudio* e permite a criação de interfaces gráficas complexas para a manipulação, busca e visualização de dados com poucas linhas de código e usando apenas *scripts* em R (CHANG et al., 2021).

A Figura 2.16 apresenta um exemplo de estrutura de *script* e o exemplo de uma interface gerada com o *Shiny*. O exemplo "*Hello Shiny*"gera um histograma do conjunto de dados "*faithful*"do R com um número configurável de intervalos. O pacote compila o código R em $HTML$, CSS e *JavaScript*, tornando-o acessível até para não programadores.

Figura 2.16 - Exemplo de aplicação com *Shiny RStudio*.



Fonte: Chang et al. (2021).

Neste contexto, o estudo Jia et al. (2021) faz uma revisão do desenvolvimento de aplicações com *Shiny R* na área de ciência de dados na biologia. Abordando etapas básicas de como construí-las, pacotes R comuns para a interface e servidor, além da implantação em nuvem e recursos online. Mesmo com a proposta de facilitar o desenvolvimento, o *Shiny* se torna limitado em relação a integração com demais servidores de dados e ferramentas de visualização para mapas, gráficos e imagens.

2.7.2 *Plumber R*

Plumber R é um pacote na linguagem R focado no desenvolvimento de serviços e aplicações na *web*. O pacote converte as funções implementadas em *scripts* R pré-existente em um serviço usando uma coleção de comentários especiais para encapsula-los nos métodos *GET*, *POST* e *DELETE*. Ampliando significativamente as possibilidades de aplicação do R em diferentes cenários, tornando-o uma escolha versátil para o desenvolvimento de aplicações interativas e serviços de análise de dados (SCHLOERKE; ALLEN, 2022). A Figura 2.17 apresenta um exemplo de *script* para exibir uma tela simples com *HTML*.

Figura 2.17 - Exemplo de aplicação com *Plumber R*.

Fonte: Schloerke e Allen (2022).

Um exemplo breve da aplicação desta ferramenta é o *Turing Geovisualisation Engine* (TGVE) apresentado por Hama et al. (2023). O TGVE é uma aplicação para visualização e análise interativa de dados geoespaciais como geometrias e imagens. O TGVE possui dois componentes principais o *React JS* para a interface gráfica e a API em *Plumber R* como gerenciador e servidor de dados. Os autores fornecem uma visão geral das capacidades do pacote para a área geoespacial.

Na ciência de dados, o *Plumber R* permite a criação de visualizações complexas para gráficos, tabelas e mapas usando o R. A aplicação resultante possui uma estrutura flexível e pode ser distribuída em diferentes formatos, tanto para somente a distribuição de dados quanto para a visualização. Esta flexibilidade torna possível a exposição e compartilhamento de funcionalidades com diversos sistemas e aplicações. Porém esta flexibilidade aumenta a complexidade de aprendizado do *Plumber R* tornando-o inacessível a alguns usuários.

3 ARQUITETURA DE INTEGRAÇÃO DOS MÉTODOS DE ANÁLISE DE AMOSTRAS DE USO E COBERTURA DA TERRA

Este capítulo descreve a arquitetura de *software* proposta nesta dissertação e sua implementação para a integração de métodos de análise durante a coleta de amostras LULC. Esta arquitetura foi implementada como uma extensão das funcionalidades da plataforma *web* de coleta *TerraCollect* (Seção 2.6.4). Um dos objetivos é contribuir na classificação LULC disponibilizando a abstração de ferramentas para análise em uma interface gráfica com base nos métodos apresentados por Wickham e Grolemund (2017), Santos et al. (2021a) e Tuia et al. (2009).

A arquitetura usa o modelo cliente-servidor com dois componentes principais: um serviço *web* e um componente de visualização na interface gráfica da plataforma de coleta. Como a arquitetura foi implementada como uma extensão do *TerraCollect*, em desenvolvimento no projeto BDC, foram usados os recursos da infraestrutura computacional do projeto. Recursos como: servidores de alto desempenho; sistemas gerenciadores de banco de dados; serviços e ferramentas de busca, extração e visualização de informações.

As abordagens para os métodos de análise foram implementadas usando as funcionalidades do pacote *sits* e estruturadas em um serviço *web* a ser consumido pela extensão no ambiente *TerraCollect*. O protótipo da extensão desenvolvido para o teste da arquitetura fornece um painel de controle na *web* para subsidiar os especialistas durante e após o processo de coleta de amostras LULC. Desta forma, etapas como a identificação de amostras catalogadas erroneamente, exploração e refinamento do conjunto de dados podem ser automatizadas.

A extensão possui um conjunto de funções para abstrair a busca e visualização de dados, assim o usuário não precisa de poder computacional ou conhecimento específico de algoritmos complexos para análise. Os dados e métricas resultantes ficam disponíveis de forma colaborativa em um ambiente baseado em projetos, permitindo o compartilhamento dos resultados obtidos. Esta arquitetura permitirá ao usuário a produção de dados de boa qualidade unindo a coleta com a análise de dados.

Esta arquitetura requer a comunicação entre diversos componentes com instâncias e ambientes em diferentes tecnologias e linguagens de programação para otimizar o acesso e armazenamento dos dados. Neste capítulo serão explorados os requisitos, a visão geral, a metodologia, a arquitetura para a conexão com os recursos do projeto BDC e por fim a abordagem para cada um dos métodos de análise.

3.1 Requisitos para a arquitetura

O *TerraCollect* possui uma arquitetura composta de componentes como o gerenciamento de usuários, projetos e amostras respectivamente usando os serviços e ferramentas *OAuth*, *Sample-WS* e *Sample-DB*. Para auxiliar na etapa desenvolvimento e implementação da arquitetura para a integração dos métodos de análise na extensão *TerraCollect*, diversos testes foram feitos e alguns requisitos foram levantados:

- **Conexão com o pacote *sits* R:** Os métodos de análise foram implementados usando o pacote *sits*, ou seja, torna-se necessário a comunicação entre *TerraCollect* e um ambiente em linguagem de programação R para a execução dos métodos;
- **Arquitetura cliente-servidor:** O *TerraCollect* foi desenvolvido com o objetivo de integrar os serviços do BDC no contexto da coleta de amostras, o que torna a implementação de extensões e ferramentas neste ambiente menos custosa usando o modelo cliente-servidor;
- **Separação de seções na interface:** O componente da arquitetura para a interface gráfica deve ser baseado em um *dashboard* separado em duas seções principais como: entrada (formulários) e saída (visualização em gráficos e mapas). Também deve haver formulários para a extração de séries temporais, seleção de modelos de aprendizado de máquina e mapas SOM;
- **Visualização interativa de resultados:** A interação deve permitir a seleção de amostras para visualizar séries temporais e métricas de probabilidades no mapa de referência. O mapa SOM, por exemplo, deve permitir selecionar neurônios e amostras agrupadas no conjunto de peso;
- **Compartilhamento de resultados:** A extensão para a análise deve ser um ambiente compartilhado entre os usuários de um mesmo projeto, e com isso permitir a exportação de resultados como imagens, gráficos e arquivos de diversos formatos com os dados das amostras;
- **Processamento paralelo:** A extração de séries temporais com base nas amostras armazenadas no banco de dados do *TerraCollect* demanda tempo e não pode ser executada a cada requisição a um método de análise. Por essa razão se torna necessário uma estratégia para o processamento paralelo em segundo plano. Permitindo que o usuário requisite a extração e seja notificado quando o processamento estiver concluído;

- **Armazenamento de séries temporais:** O conjunto de séries temporais é um requisito para a aplicação dos métodos de análise, e devido a quantidade de amostras necessárias e do tempo de processamento ao extrair cada série temporal se torna necessário um sistema de cache para armazenar temporariamente as séries diminuindo o tempo de espera entre as requisições. Além de estratégias para a verificação de novas amostras e atualização do banco de séries;

- **Gerenciamento de modelos de aprendizado de máquina:** para a estimativa de probabilidades é necessário que o modelo treinado seja salvo e atualizado conforme novas amostras são adicionadas no banco de dados do *TerraCollect*. Desta forma o tempo de processamento para cada requisição da estimativa é reduzido;

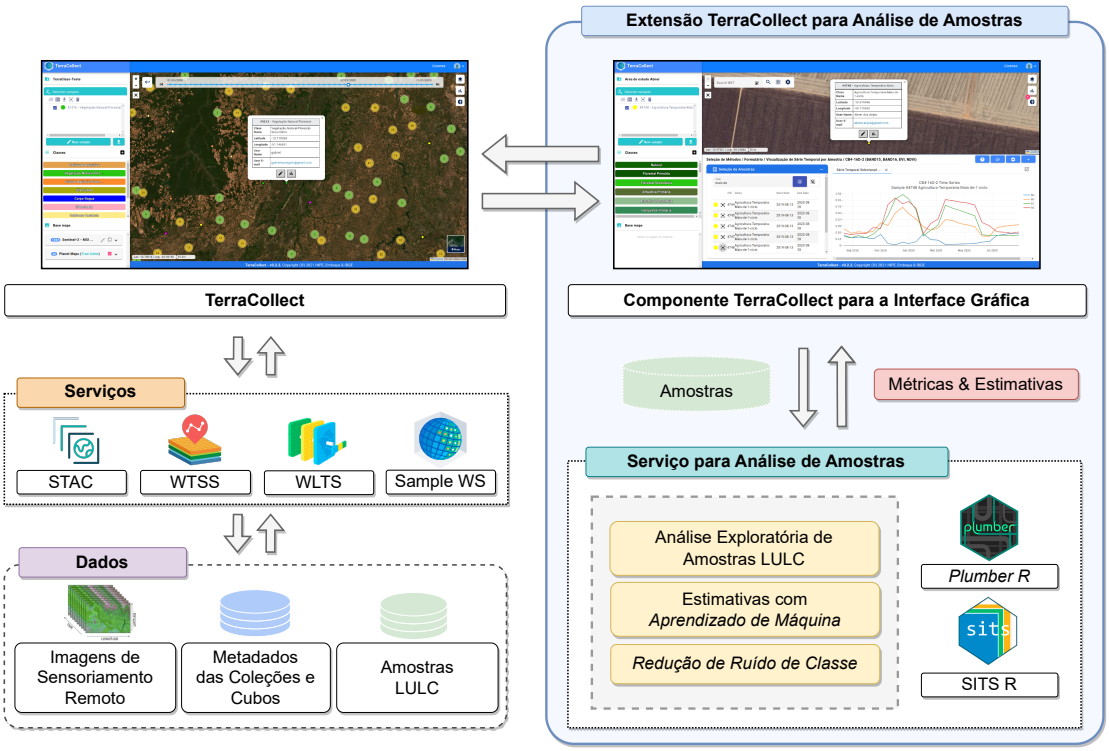
- **Armazenamento de agrupamentos SOM:** Para implementar o método apresentado por Santos et al. (2021a) na plataforma é necessário que os dados e metadados do mapa SOM sejam armazenados. Assim é possível a análise de qualidade com o filtro para a probabilidade *priori* e *posteriori*. Pois este tipo de análise demanda tempo e requer testes exaustivos para validação ao aplicar diferentes combinações de parâmetros.

A descrição fundamental da arquitetura para a integração dos métodos de análise pode ser resumida a uma sequência de etapas. Cada etapa possui um conjunto definido de entradas e saídas e pode ser traduzida em um fluxograma como a visão geral da arquitetura e das abordagens para cada um dos métodos. As abordagens de análise precisam ser estruturadas na extensão de forma a não prejudicar a experiência de coleta na plataforma *TerraCollect* e sim apenas adicionar novas funções.

3.2 Visão geral da arquitetura

A Figura 3.1 apresenta a visão geral da arquitetura de *software* para a integração dos métodos. Em resumo, a aplicação *web* principal do *TerraCollect* consome o serviço *web* com a API para a análise de amostras e gera as visualizações interativas dos resultados na interface. A API de análise faz a extração e o armazenamento de séries temporais com cada amostra associada. A extração de dados é feita com os serviços do projeto BDC que distribuem os dados. Estes dados constituem imagens de sensoriamento remoto, metadados das coleções e cubos e amostras que estão armazenados na infraestrutura do projeto. Com cada amostra associada a uma série temporal, a API usa o *sits* R para executar os métodos e dispor os resultados usando requisições $HTTP$ e respostas em $JSON$.

Figura 3.1 - Arquitetura geral para a integração dos métodos de análise de amostras.



Fonte: Produção do autor.

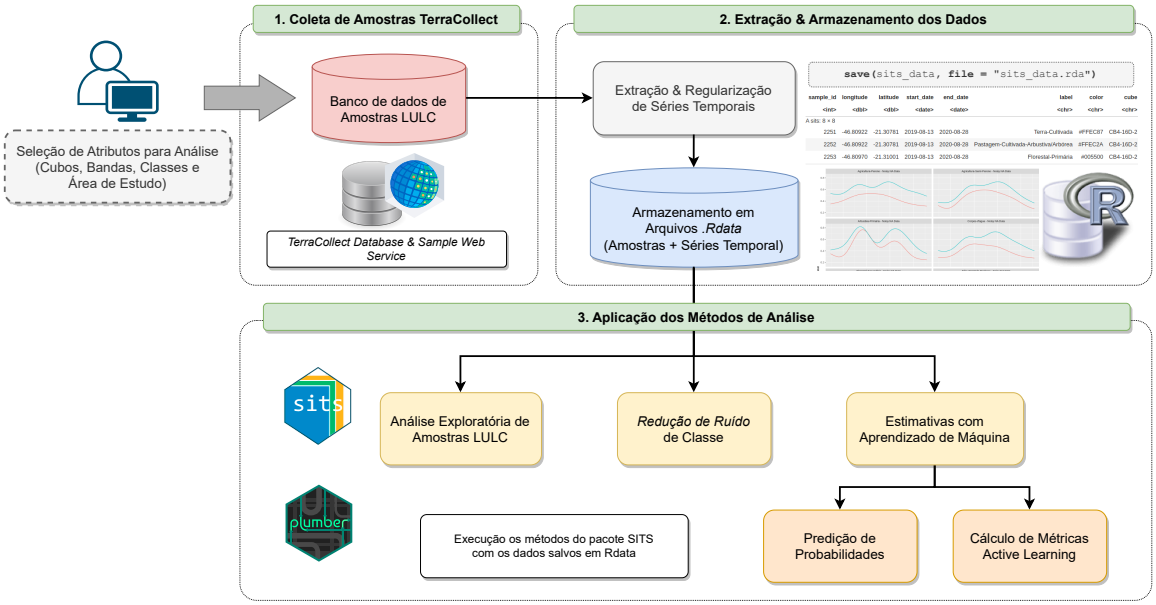
O **componente para a interface gráfica** como uma extensão do *TerraCollect* aproveita as conexões com os serviços do projeto BDC e demais *frameworks* da plataforma. Com isso, há disponível componentes para facilitar a geração de formulários e visualização de gráficos que abstraem funções como acesso, visualização e processamento de dados. Nesta extensão o usuário seleciona as visualizações e gráficos para os métodos de análise, e o que é enviado na requisição na API de análise são metadados sobre as amostras e o método selecionado. Estes dados são passados ao *sits* que faz a leitura de amostras e executa os métodos, assim como resposta têm-se as métricas e estimativas.

A arquitetura usa o pacote *Plumber R* para abstrair as funções do pacote *sits* R e implementar a API em um **serviço de análise de amostras**. O *sits* é fundamental na execução dos métodos, é necessário que o acesso às séries temporais seja otimizado para cada amostra. Este pacote oferece funções prontas para a aplicação de filtros, busca e seleção de dados. Ao todo, as abordagens para a análise estão separadas em três tópicos principais: Análise exploratória de amostras; Análise de métricas com aprendizado de máquina e técnicas de *Active Learning*; e por fim o controle de qualidade e redução de ruído de classe.

3.3 Metodologia para a integração dos métodos de análise de amostras

A Figura 3.2 apresenta a metodologia geral para lidar com a integração dos métodos de análise. Para executar as operações no *sits*, é requisitado que as amostras alvo tenham uma série temporal associada. A primeira fase envolve a seleção de atributos pelo usuário para a extração das séries temporais usando o banco de dados de amostras e as configurações de projeto do *TerraCollect*.

Figura 3.2 - Metodologia para a integração dos métodos de análise de amostras.



Fonte: Produção do autor.

O *TerraCollect* faz o gerenciamento dos dados com base em um projeto, onde cada projeto tem um grupo de usuários associados. Este projeto contém as configurações como o sistema de classificação, mapas base, cubos e coleções para visualização. A arquitetura geral para a análise faz a adição de novas configurações relacionadas à extração de séries temporais no *sits*, esta adição resulta no chamado "projeto de análise". Para iniciar este tipo de projeto é necessário que o usuário forneça atributos como o cubo, as bandas, as classes e a área de estudo. Desta forma o usuário pode escolher quais amostras serão usadas para a extração de séries temporais e posteriormente para a execução dos métodos.

Com os atributos do projeto de análise, a API faz a seleção de amostras no banco de dados do *TerraCollect* e o processo de extração das séries é iniciado usando o *sits*. Esta operação ocorre com o auxílio do serviço de gerenciamento de amostras, o *Sample WS*. Os parâmetros para a seleção são enviados no formulário de requisição no *Sample WS* e a API formata os dados para a compatibilidade no *sits* (Seção 2.3).

Todo o tratamento dos dados e processos descritos na Figura 3.2 é feito usando a linguagem de programação R, inclusive as requisições para os demais serviços auxiliares, a formatação das amostras, a extração dos dados e a execução dos métodos de análise. O pacote *Plumber R* é usado como última camada para abstrair este conjunto de operações como API em um serviço *web*. Este serviço é responsável por fazer a integração dos métodos no ambiente R *sits* e a extensão no *TerraCollect*.

Após a seleção e formatação das amostras, o *sits* extrai os dados das imagens de satélite usando as configurações salvas. Em seguida as séries são regularizadas e armazenadas em *Rdata* no sistema de arquivos do ambiente R. Assim, as amostras com as séries respectivamente associadas estão disponíveis para serem passadas à análise exploratória, a redução de ruído e para o treinamento de modelos de aprendizado de máquina para estimar probabilidades. Com as séries temporais salvas, o pacote *sits* é usado para aplicar estes métodos conforme as requisições do usuário. Cada método utiliza uma abordagem para adaptar sua funcionalidade em uma API na *web* com a entrada e saída de informações, cada abordagem com etapas específicas que serão descritas nas próximas seções.

3.4 Extração e regularização de séries temporais

A fase de extração e regularização de séries temporais para a análise (Figura 3.2) é performada em segundo plano usando estratégias da linguagem R para gerenciar subsessões e sub-processos com processamento paralelo. Enquanto isso, o usuário pode visualizar as mensagens e saídas geradas pela execução dos processos e será notificado quando for concluído. A extração é feita em conjunto com a regularização para otimizar o tempo de processamento, já o armazenamento, por ser menos complexo, é feito como última etapa. A regularização é um processo para formatar a estrutura de dados para as séries, verificar e preencher possíveis lacunas na linha temporal.

O pacote *sits* foi feito para auxiliar a análise pós coleta de amostras, todavia, este trabalho busca integrar os métodos durante e pós coleta. É possível que o conjunto de dados seja incrementado com mais amostras no decorrer da análise usando esta abordagem. Por esta razão, a metodologia para a extração usada para este trabalho

faz a união das duas ferramentas *sits* e WTSS. Foi necessário implementar soluções para regularizar os dados provenientes destas duas ferramentas, pois cada ferramenta possui uma estratégia diferente para a extração. A regularização é uma etapa necessária para tratar as séries temporais padronizando-as para que sejam suportadas pelos métodos no pacote *sits*. Há requisitos a serem seguidos para que as séries possam ser utilizadas na análise que são aplicados no processo regularização, como:

- O conjunto de amostras com séries temporais deve ser separado e organizado pelo identificador do cubo, para não haver divergências com a resolução espacial e temporal durante a leitura;
- Cada conjunto de amostras com séries temporais, baseando-se no cubo, precisa possuir o mesmo número de observações e a mesma linha temporal, ou sequência de datas, em cada série;
- Caso haja valores nulos, ao alinhar as séries temporais de cada amostra, o mesmo valor deve ser interpolado ou removido para não haver perda de dados ou confusão na análise;
- As bandas em cada série temporal associada a um registro de amostra devem possuir as nomenclaturas, o nome da banda segundo os metadados do cubo, padronizadas, ordenadas e em caixa alta.

No *sits*, a extração das séries é feita a partir da criação de um cubo de dados local, e todas as operações consequentes orientam-se com base nos metadados do cubo como as bandas e a linha temporal que define a sequência das imagens. Usando este objeto "cubo de dados" para a extração pode exigir um alto poder de processamento e armazenamento. As etapas para a regularização como quantidade de observações, ordenação das bandas e aplicação do filtro de nuvem são omitidas durante o processo. No WTSS não há estratégias para regularização de dados, este serviço apenas retorna os dados brutos sem tratamento, visto que não é um de seus objetivos oferecer métodos para trabalhar com séries temporais como no caso do *sits*. Foram estudadas algumas abordagens para a regularização com base no conjunto de amostras ao invés da criação de um cubo de dados local.

Existem duas abordagens para a regularização das amostras discutidas durante o desenvolvimento do presente trabalho. As abordagens devem priorizar o registro das amostras armazenadas no banco de dados do *TerraCollect*, já que somente o usuário pode excluir tais dados. A primeira é baseada na interpolação de dados e a segunda na "não interpolação". Usando como exemplo a seguinte situação: "um

conjunto de 500 amostras onde a maioria (496 amostras) possui uma série temporal com 25 observações e algumas (o restante 4 amostras) com 23 observações". A primeira abordagem irá interpolar as duas observações faltando para as 4 amostras, preservando os dados das demais 496 para regularizar o número de observações. Neste caso todas as amostras serão regularizadas com 25 observações cada.

A segunda abordagem baseada na "não interpolação", onde para o mesmo caso, ao invés de interpolar (estimar dados), esta abordagem irá deletar as datas das demais 496 amostras para regularizar com base nas 4 amostras com datas faltando. O conjunto terá agora 23 observações para cada amostra, contudo perde-se os dados das demais 496, preservando apenas os dados reais observados. Serão perdidas 992 observações, por esta razão a segunda abordagem não é muito recomendada.

As abordagens para a regularização, durante o uso da plataforma ficam a critério do usuário caso deseje a interpolação ou não. Existe mais um tratamento para as séries temporais que é a aplicação do filtro de nuvens, baseando-se na banda de qualidade dos metadados do cubo. Esta etapa para lidar com a cobertura de nuvens é abstraída pelo *sits*, onde é identificado valores com nuvens e sombras de nuvens e os mesmos são interpolados.

Após a extração é feito o armazenamento das séries temporais já formatadas para o pacote *sits* com arquivos *Rdata*, assim a cada requisição pode-se realizar a leitura direta de dados e aplicar os métodos de análise. Os arquivos *Rdata* criam uma cópia da tabela de amostras em conjunto com a série temporal. Há um custo para sincronizar os dados salvos em *Rdata* e o banco de dados do sistema *TerraCollect*, já que a API de análise não possui conexão direta com este componente. Quando uma nova amostra é adicionada no *TerraCollect*, a API deve verificar esta nova amostra, para posteriormente extrair e regularizar a série para a linha temporal base e armazenar no arquivo *Rdata* correspondente.

3.5 Armazenamento e gerenciamento de arquivos *Rdata*

O acesso aos dados das séries temporais e demais resultados, como os produtos dos métodos, deve ser feito de forma eficiente, um dos objetivos do presente trabalho é facilitar e otimizar o procedimento da análise. Os bancos de dados relacionais como o *PostgreSQL*, comumente usado em SIG's, não satisfazem com eficiência o requisito essencial para o *sits* que é a associação de uma amostra a uma série. No *sits*, as séries temporais são armazenadas diretamente no registro da amostra em uma coluna, pois a estrutura de dados em R possibilita esta abordagem.

O armazenamento de séries temporais se tornou um requisito essencial para análise durante as etapas de desenvolvimento da plataforma. A extração de dados das imagens é uma etapa custosa dependente da quantidade de amostras em termos de tempo e poder de processamento. É inviável a extração de séries a cada requisição de um método ou a cada inicialização de um projeto. O formato de dados necessário para armazenar estes dados não é compatível com bancos de dados relacionais. Portanto, há uma demanda por técnicas de armazenamento não usuais como sistemas de arquivos *Rdata*.

O uso de bancos de dados é recomendado por oferecer técnicas de gerenciamento padronizadas e ferramentas prontas para lidar com atualizações e demais eventos. Usar o sistema de arquivos com extensão *Rdata* oferece a liberdade para criar estruturas próprias, todavia aumenta a complexidade para gerenciar os dados. Essa complexidade é devido a necessidade de *scripts* auxiliares para lidar com adição, exclusão e atualização de arquivos.

Em um banco de dados relacional, a associação de uma amostra a uma série temporal pode ser traduzida pelo relacionamento de duas tabelas usando as respectivas chave-primária e chave-estrangeira. Neste sentido, há a ligação de um registro, ou amostra, a outra tabela com o conjunto de séries temporais. Foram feitos diversos testes com o armazenamento em bancos relacionais usando formatos JSON e até mesmo serviços auxiliares com *Python*. Entretanto, o tempo de leitura e formatação dos dados para a aplicação dos métodos no *sits* fica custoso e pode ser até mais custoso que a própria extração das séries a cada requisição. Além do aumento da complexidade da arquitetura ao adicionar mais componentes. Por isso os arquivos *Rdata* são a melhor solução para armazenar, não havendo necessidade de formatação de dados para o *sits* a cada requisição.

Com o formato *Rdata* é possível armazenar a *R tibble* gerada para o *sits* diretamente no sistema de arquivos. *Rdata*, arquivos com extensão *.RDA*, suportam objetos da linguagem R como *Kohonen Maps*, modelos de aprendizado de máquina e as séries temporais no formato *sits*, todos objetos acoplados ao seus metadados como projeto e usuários com acesso. Então, a arquitetura para a análise de amostras usa o sistema de arquivos *Rdata* para salvar temporariamente os dados da análise e principalmente as séries temporais. Isto permite não só o compartilhamento dos resultados, mas também a análise de forma mais eficiente. Por exemplo, uma vez salvo um agrupamento SOM, todos os usuários de um projeto poderão visualizar e fazer testes com os filtros para a análise de qualidade.

A complexidade no gerenciamento para os arquivos *Rdata* foi resolvida com a definição de um modelo lógico de dados para cada arquivo. A Figura 3.3 apresenta os modelos lógicos para a estrutura de arquivos *Rdata* usado no serviço de análise de amostras. Cada tabela descreve um *DataFrame* contendo os atributos e metadados de um objeto salvo. O serviço de análise conecta seus objetos salvos com versionamento de arquivos e com as chaves primárias identificadoras de projetos e amostras no banco de dados do *TerraCollect*. Essa dinâmica foi escolhida por causa do esquema baseado em chaves primárias e estrangeiras, que conectam outros componentes de sua arquitetura como o LCCS e *Sample-DB* que usam o *PostgreSQL*.

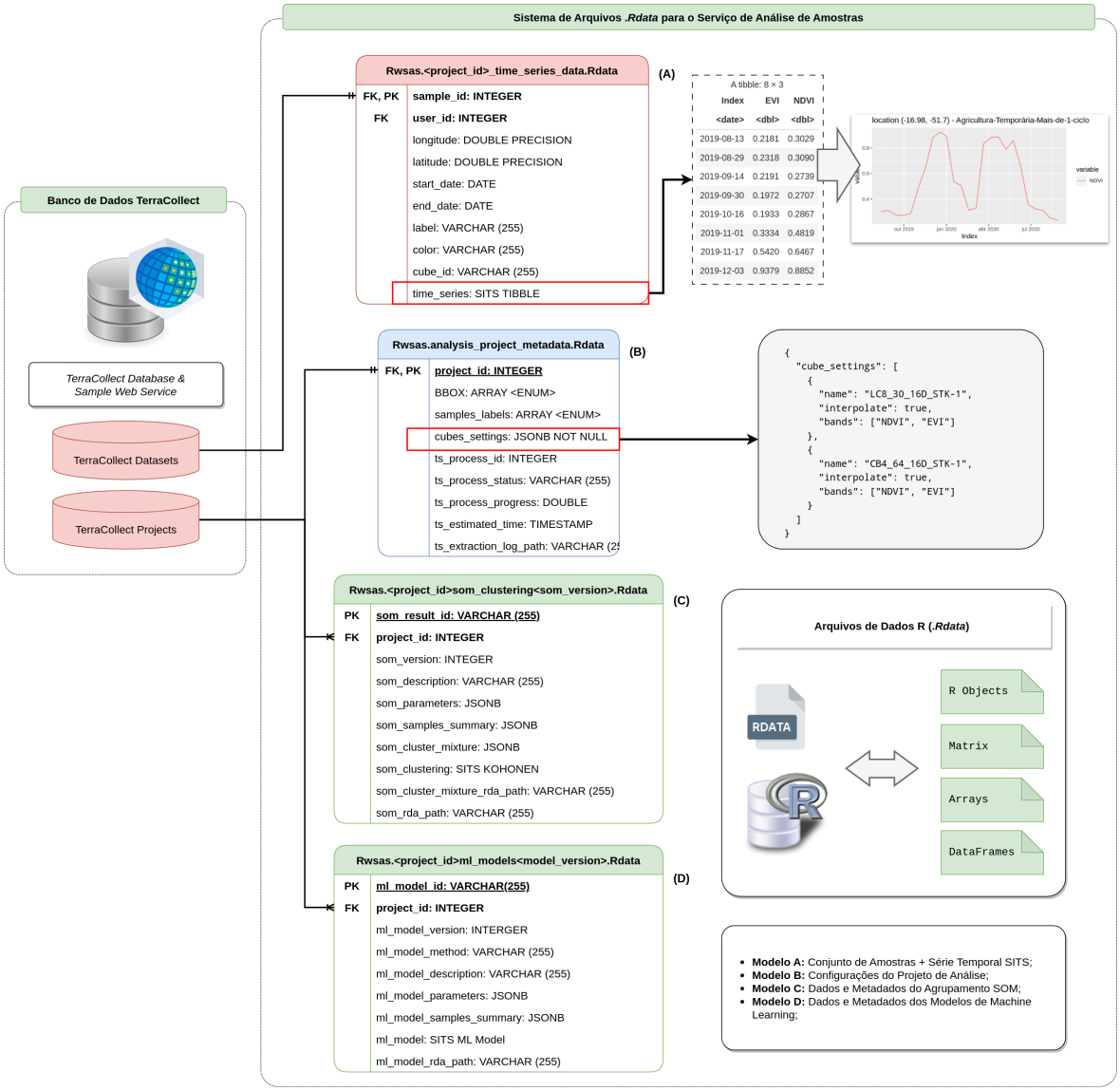
As técnicas para lidar com eventos de atualização que necessitam da formatação de dados, foram centralizadas e organizadas separadamente dos métodos de análise com *scripts* auxiliares. Os arquivos são salvos em um volume criado para o *container Docker* da aplicação. A dinâmica para o acesso a este armazenamento baseia-se no acesso ao projeto, todos os usuários presentes no projeto têm acesso às séries temporais e aos resultados da análise gerados. Estas informações são provenientes do serviço de autenticação do BDC *OAuth* implementado no *TerraCollect* para gerenciar usuário e projetos.

3.5.1 Amostras & séries temporais de imagens

O arquivo "*analysis_project_metadata.Rdata*" (Figura 3.3.B) armazena as configurações dos cubos selecionados para a extração de cada projeto de análise cadastrado no serviço de análise. Esta informação é necessária para manter o padrão e controle dos dados armazenados no "$< project\_id > \_time\_series\_data.Rdata$" (Figura 3.3.A) que contém todas as séries temporais de um projeto. Este arquivo é criado após a extração de dados de um cubo e é identificado pelo ID do projeto.

O arquivo "$< project\_id > \_time\_series\_data.Rdata$" é uma cópia da tabela com os dados das amostras do *TerraCollect*, porém no formato *sits* com a associação da série temporal de cada cubo. Este objeto possui dois atributos para descrever os dados das séries associadas às amostras, o atributo identificador do cubo de imagens e a coluna com a tabela de dados brutos das séries, onde cada índice e banda representa uma coluna com as observações. O objetivo de criar esta cópia é facilitar a entrada de dados no pacote *sits* com as amostras prontas para análise, assim quando algum método for selecionado é necessário apenas a leitura das amostras. As operações para armazenar, atualizar e excluir são organizadas separadamente pois precisam das funções para conversão de dados JSON em *sits*.

Figura 3.3 - Modelo lógico para a estrutura de arquivos em *Rdata* do serviço de análise de amostras.



Fonte: Produção do autor.

A atualização deste "banco de séries temporais" não é feita de forma automática, o usuário, caso deseje, deve selecionar esta opção. A atualização faz uma verificação no banco do *TerraCollect* relacionando o ID da amostra na tabela no *Sample-DB* com o ID da amostra no objeto *Rdata sits*. Assim, se necessário, a API exclui e adiciona novas séries usando a configuração do projeto armazenada no "*analysis_-project_metadata.Rdata*". Todas as operações que lidam com a edição e atualização do banco de séries temporais usam a configuração do projeto de análise como base para a verificação de amostras novas ou excluídas.

3.5.2 Resultados da análise

Após a fase de preparação dos dados com a extração, regularização e armazenamento das amostras acopladas às séries temporais, a API de análise passa os dados para o *sits* que gera os resultados. São apenas dois tipos de resultados, o agrupamento SOM e o objeto modelo de aprendizado de máquina. Tanto para o resultado do SOM quanto para o modelo, foram usados *DataFrames* com atributos descrevendo os metadados e mais uma coluna contendo o objeto R gerado com a identificação usando o versionamento e o ID do projeto. Cada arquivo contém um único objeto para facilitar a leitura quando o mesmo é selecionado na requisição. Os arquivos contém os dados necessários para recriar a visualização no lado da interface no *TerraCollect*, assim os dados são lidos e os gráficos são gerados usando o *Plotly*.

O arquivo "$< project\_id > \_som\_clustering < som\_version >$" (Figura 3.3.C) armazena o objeto gerado pelo agrupamento SOM no *sits*. Cada arquivo possui um único SOM identificado pelo conjunto de atributos ID do projeto e versão do objeto. Este arquivo também armazena os metadados como configurações usadas para o agrupamento como o tamanho da grade e o número de interações. O armazenamento deste objeto facilita a avaliação de qualidade, pois permite que vários testes com o *threshold* sejam aplicados em mesmo agrupamento SOM. Desta forma não há a necessidade de geração de diversos mapas SOM, o que pode alterar resultados.

O arquivo "$< project\_id > \_ml\_model\_ < model\_version >$" (Figura 3.3.D) armazena o objeto gerado pelo treinamento de modelos no R. Cada arquivo armazena um objeto com o modelo de aprendizado de máquina e seus metadados, onde cada modelo é identificado com o ID do projeto e sua versão. O armazenamento dos modelos é essencial, pois o treinamento é algo aleatório que depende das condições atuais do processamento, ou seja, mesmo que dois modelos sejam treinados com os mesmos atributos obterão resultados diferentes. Isto é o mesmo caso da predição onde os resultados podem apresentar divergências dependentes do momento da execução. Por isso, armazenar o modelo pode trazer mais consistência na predição de várias amostras ou em testes com o *Active Learning*.

Tanto para o SOM quanto para os modelos, a implementação com cada objeto salvo em seus respectivos arquivos otimiza as requisições no serviço. Uma vez que para selecioná-los é necessário apenas fornecer o ID do projeto e a versão do objeto, o serviço monta o nome do arquivo, realiza a leitura e retorna seus metadados. O versionamento também é ideal para atualizações e para testar a execução desta análise diversificando atributos e hiper-parâmetros.

3.6 Análise exploratória de amostras de uso e cobertura da terra

A análise exploratória de dados (AED) é um processo criativo sem pré-requisitos para responder hipóteses com gráficos e visualizações, ocupando uma parcela de tempo na etapa de análise. No contexto de amostras LULC, as principais indagações podem ser resolvidas com gráficos pré-configurados. Na extensão para análise no *TerraCollect*, o componente para o método AED fornece visualizações para resumir o conjunto de dados baseando-se em suas características essenciais. Conforme ilustrado pela Figura 3.4, são três opções disponíveis: A distribuição das amostras por classe, os padrões de série temporal e a seleção de série temporal por amostra.

Há dois tipos de visualização para a distribuição de frequência de amostras por classe. O primeiro é baseado no gráfico de barras e o segundo no gráfico de pizza. Estes dois gráficos possuem características distintas, como por exemplo, em certos problemas com poucas classes LULC o gráfico de pizza pode ser usado para avaliar a proporção e o balanço das classes. Entretanto, para analisar um conjunto de dados com muitas classes, este tipo já não é recomendado. Este gráfico pode dificultar na avaliação do problema, pois o cálculo da porcentagem omite classes de menor proporção em comparação às demais, o que não ocorre na distribuição no gráfico de barras, onde é demonstrada a quantidade real das amostras por classe.

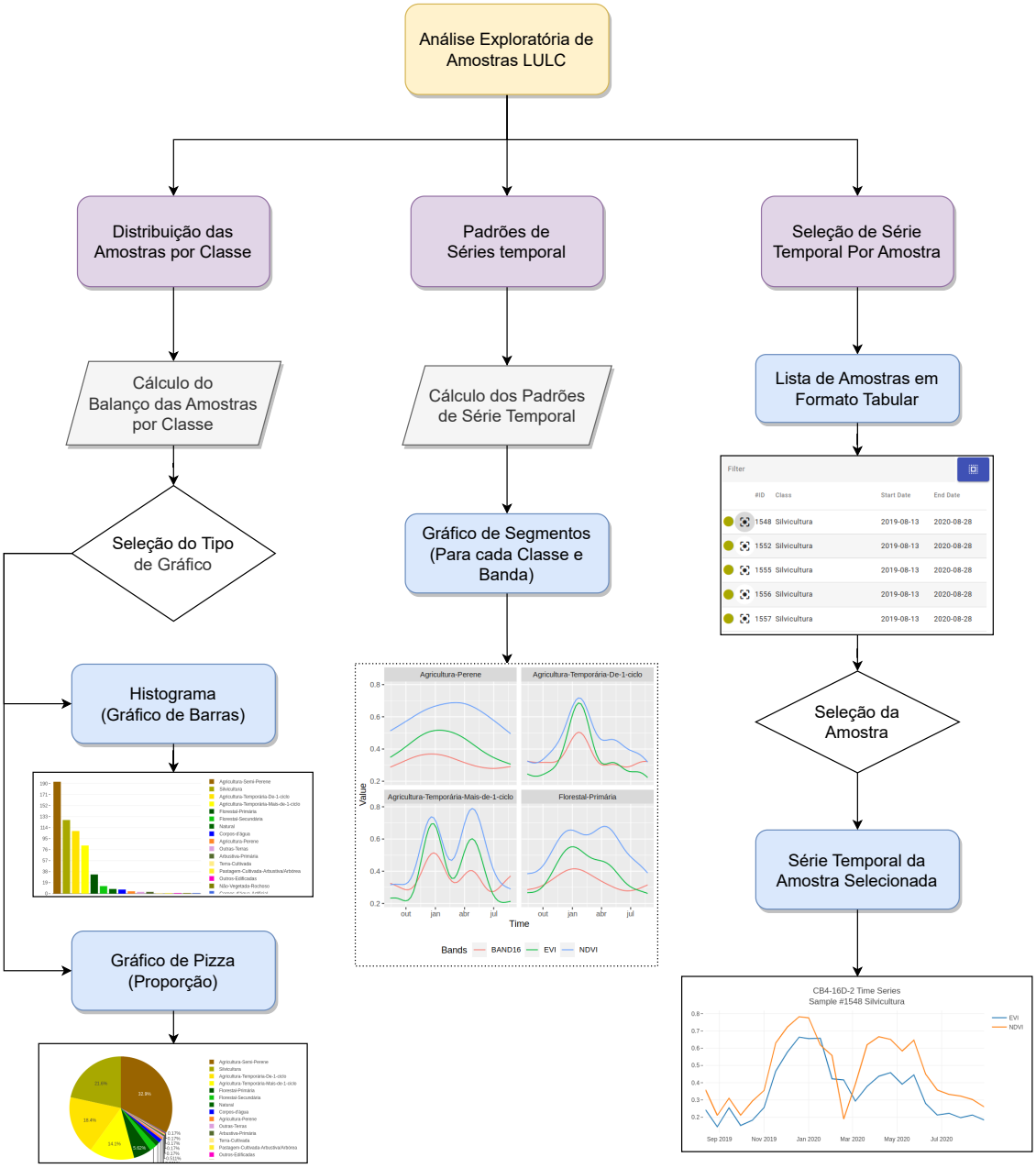
Na visualização para os padrões de série temporal, o usuário pode definir filtros e atributos para o cálculo do padrão. Por exemplo, pode-se escolher quais classes LULC, quais bandas e qual área será a base (usando uma caixa delimitadora ou "*Bounding Box*"). Com estes filtros, o *sits* calcula o padrão das bandas e índices espectrais escolhidos baseando-se no conjunto de amostras de cada classe. O *sits* usa *Generalized Additive Models* (GAM) para suavizar a série temporal, e assim, estimar um padrão de comportamento no período para uma determinada classe.

A visualização do padrão das séries temporais possui gráficos de segmentos exibindo as bandas e índices para resumir o comportamento do conjunto de amostras disponíveis por classe (Figura 3.4). Assim é possível, por exemplo, calcular o padrão de índices de vegetação com base em uma área específica. Com estas visualizações o analista consegue distinguir as características de cada classe, pois cada uma apresenta uma curva com mínimos e máximos distintos.

A seleção de série temporal por amostra oferece a visualização em formato tabular, onde pode-se buscar e selecionar séries temporais para visualização no gráfico de segmentos. Nesta opção, não é necessário a aplicação de filtros e tratamentos nas

séries, o resultado é o dado real salvo no arquivo *Rdata*. Esta opção tem o objetivo de fornecer ao usuário uma interface para explorar os dados das séries que serão passados para os demais métodos, como o treinamento de modelos de aprendizado de máquina e o agrupamento com SOM.

Figura 3.4 - Análise exploratória de amostras de uso e cobertura da terra.



Fonte: Produção do autor.

A metodologia usada no componente para a AED, não há o uso de técnicas avançadas com modelos de aprendizado de máquina, apenas a exploração das características essenciais dos dados como parte de um conjunto total. Os gráficos da AED para amostras LULC fornecem respostas que podem auxiliar os usuários na tomada de decisão durante a coleta, como: sobre quais atributos dentre bandas, índices e filtros de suavização das séries são mais interessantes para a análise ou quais atributos representam com mais precisão uma dada classe LULC. A principal ideia da interface é fornecer ao usuário um ambiente onde seja possível responder estas questões, disponibilizando as ferramentas e os filtros necessários.

3.7 Estimativas com aprendizado de máquina

A metodologia usada para as estimativas com aprendizado de máquina explorado no presente trabalho baseia-se nas técnicas de aprendizado semi-supervisionado e nas técnicas de *Active Learning* (Seção 2.5.2). Onde para iniciar a predição de probabilidades é necessário que um modelo seja treinado com um conjunto mínimo de amostras com as séries temporais associadas. Estas amostras de treinamento serão um conjunto base para o modelo e as probabilidades serão baseadas nas classes de referência que foram usadas neste conjunto base. Assim, quando uma nova amostra é cogitada para entrar no conjunto base, ocorre a classificação e a predição de probabilidades usando o modelo treinado e salvo na API de análise.

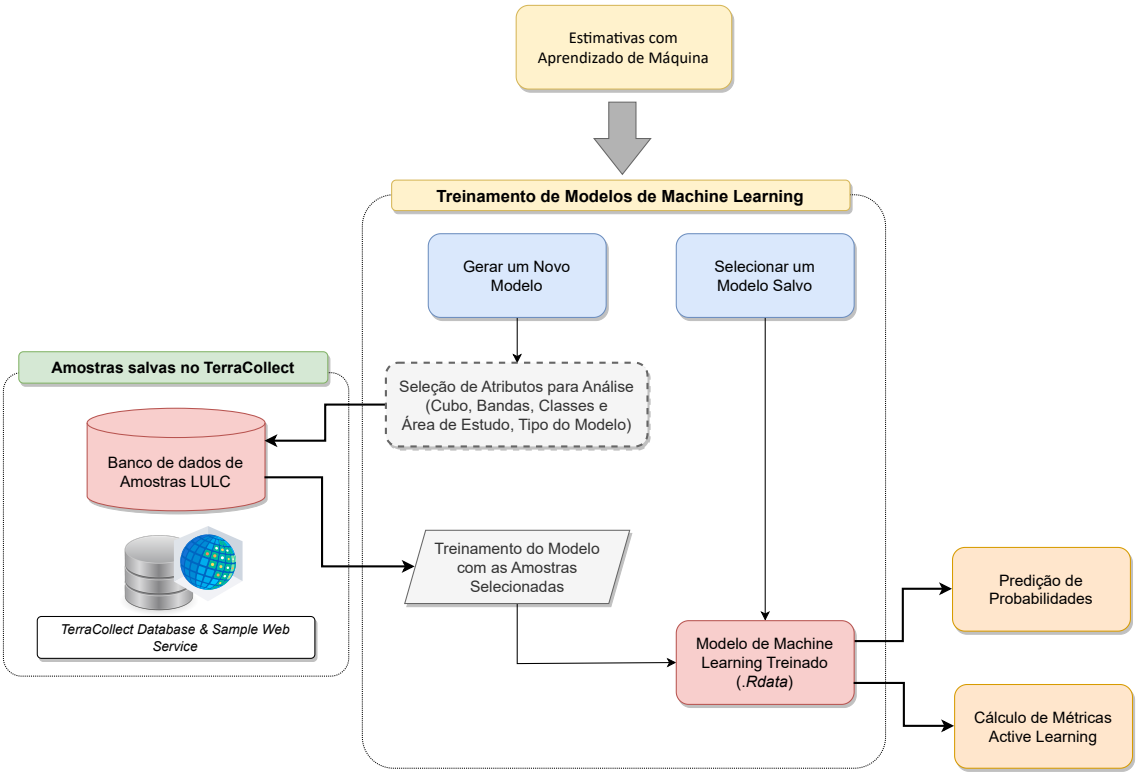
A classificação irá predizer qual a classe esta nova amostra pertence com um conjunto de probabilidades para cada classe, mas a decisão final a respeito do rótulo, a nomenclatura da classe LULC, fica a critério do usuário, o oráculo (TUIA et al., 2009). Há duas abordagens principais para rotular: usar a certeza do modelo usando a classe de maior probabilidade ou a confusão na predição medida pelas métricas do *Active Learning*. Após atribuir um rótulo à amostra, a mesma é salva no conjunto base e uma nova sessão de treinamento deve ser feita para atualizar o modelo com os dados da nova amostra.

3.7.1 Treinamento de modelos de aprendizado de máquina

Como demonstrado pela Figura 3.5, ao escolher a opção de estimativas com aprendizado de máquina, o usuário pode treinar um novo modelo ou escolher um dentre os modelos salvos em arquivos *Rdata*. Para o treinamento de um novo modelo deve haver um conjunto mínimo de amostras salvas no *TerraCollect*, ao menos uma amostra para cada classe alvo. Assim, o usuário deve fornecer à API, por meio do formulário de requisição, os atributos para seleção de amostras definindo o conjunto base,

como as classes LULC e mais os cubo e bandas para selecionar as séries temporais. Também é possível a aplicação de filtros de suavização nas séries temporais como *Savitzky–Golay* (CHEN et al., 2004) e *Whittaker* (ATZBERGER; EILERS, 2011).

Figura 3.5 - Treinamento de modelos de aprendizado de máquina.



Fonte: Produção do autor.

Em seguida, deve-se escolher o tipo do modelo de aprendizado de máquina e seus hiper-parâmetros, estes são baseados nos principais métodos implementados no *sits* para a classificação LULC em imagens, como: *Support Vector Machines* (SVM's), *Random Forests*, *Extreme Gradient Boosting*, *Multi-Layer Perceptrons* (MLP's), *Residual Neural Networks* (*ResNET*) e *Temporal Convolutional Neural Networks* (*TempCNN*). Todos estes métodos foram adaptados para lidar com amostras LULC associadas com séries temporais a partir do pacote *sits* R.

Com os atributos selecionados é feita a leitura das amostras para a definição do conjunto base de treinamento no banco de dados do projeto alvo no *TerraCollect* usando o serviço *Sample WS*. Estas amostras serão usadas no treinamento do modelo escolhido de forma supervisionada usando os hiper-parâmetros fornecidos pelo usuário. No caso do *Random Forest*, por exemplo, o número de árvores, ou no caso de redes convolucionais, o número de camadas. Em seguida, o modelo e seus meta-

dados, como data de criação, data de atualização, configurações, hiper-parâmetros e a descrição do conjunto base são salvos usando o sistema de arquivos *Rdata*. Após a primeira sessão de treinamento do modelo, o mesmo é salvo e pode-se selecioná-lo para a predizer probabilidades e calcular métricas para a tomada de decisão na atribuição dos rótulos em novas amostras.

A hipótese é que as métricas de probabilidades provenientes do treinamento de modelos de aprendizado de máquina podem auxiliar na aquisição de amostras representativas e com boa qualidade. Amostras que posteriormente serão usadas na classificação de imagens e geração de mapas LULC. Desta forma, durante o processo de coleta, as estimativas dos modelos treinados podem ser utilizadas para testar e levantar questões sobre o *status* das amostras a respeito da representatividade para a classificação. A metodologia é iterativa, onde o modelo salvo é enriquecido com novas informações a cada atualização na adição de novas amostras ao conjunto base.

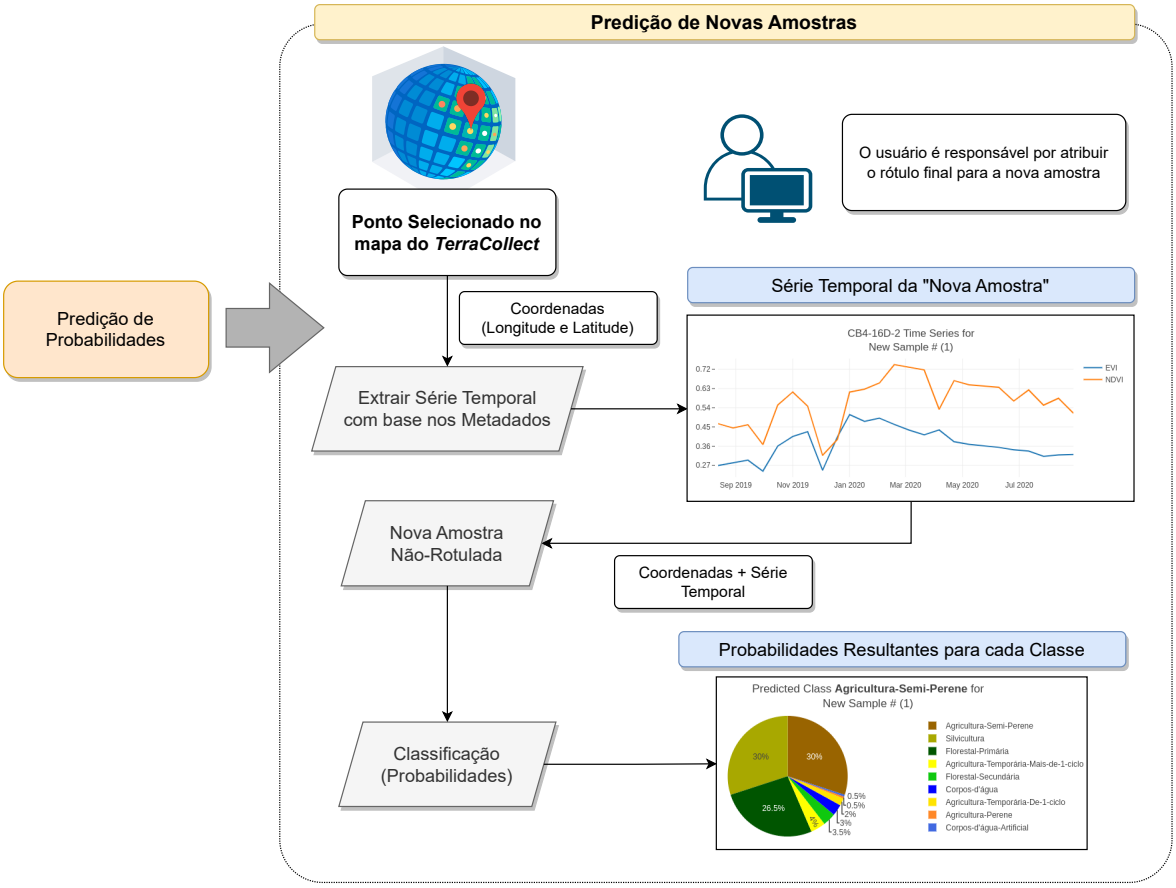
Com a adição ou mesmo remoção de uma ou mais amostras é recomendável ao pesquisador atualizar o modelo com uma nova sessão de treinamento. A atualização pode ser feita de forma manual conforme desejo do usuário ou automática usando a própria adição como evento para atualizar o modelo. Neste contexto, atualizar é passar o conjunto de dados base enriquecido com as novas amostras a uma nova sessão de treinamento usando os mesmos hiper-parâmetros salvos nos metadados do modelo. Isto irá trazer novas informações de novas amostras ao modelo salvo, aumentando sua precisão e capacidade de identificar características nos dados.

3.7.2 Predição de probabilidades

A metodologia usada para a predição de probabilidades com base nos modelos de aprendizado de máquina salvos usa a seleção de uma nova amostra para iniciar o processo. Conforme a Figura 3.6, a seleção de um novo ponto no mapa de referência é descrito na metodologia como uma nova amostra não-rotulada. Esta amostra não-rotulada ainda não está salva no banco de dados do *TerraCollect* e, até que seja atribuída uma classe, não faz parte do conjunto de dados base de treinamento.

Com as coordenadas do novo ponto (longitude e latitude) e as configurações do cubo e bandas do conjunto base, a série temporal do novo ponto é extraída. Estas configurações do cubo ficam salvas nos metadados do arquivo *Rdata* do modelo de aprendizado de máquina. A série temporal associada à localização da amostra não-rotulada é passada ao modelo treinado que irá calcular a probabilidade da mesma pertencer a uma certa classe LULC, usando o método de classificação no *sits*.

Figura 3.6 - Predição de probabilidades para as amostras de uso e cobertura da terra.



Fonte: Produção do autor.

Esta metodologia permite somente a demonstração da estimativa de probabilidades, fica a critério do pesquisador usar os valores resultantes para rotular e salvar a amostra não-rotulada no conjunto de dados base do modelo. O pesquisador como usuário pode tomar a decisão do rótulo usando o conjunto de visualizações como a série temporal e as probabilidades no gráfico. Após a atribuição de um rótulo, o conjunto de dados base é atualizado com a adição desta nova amostra já rotulada.

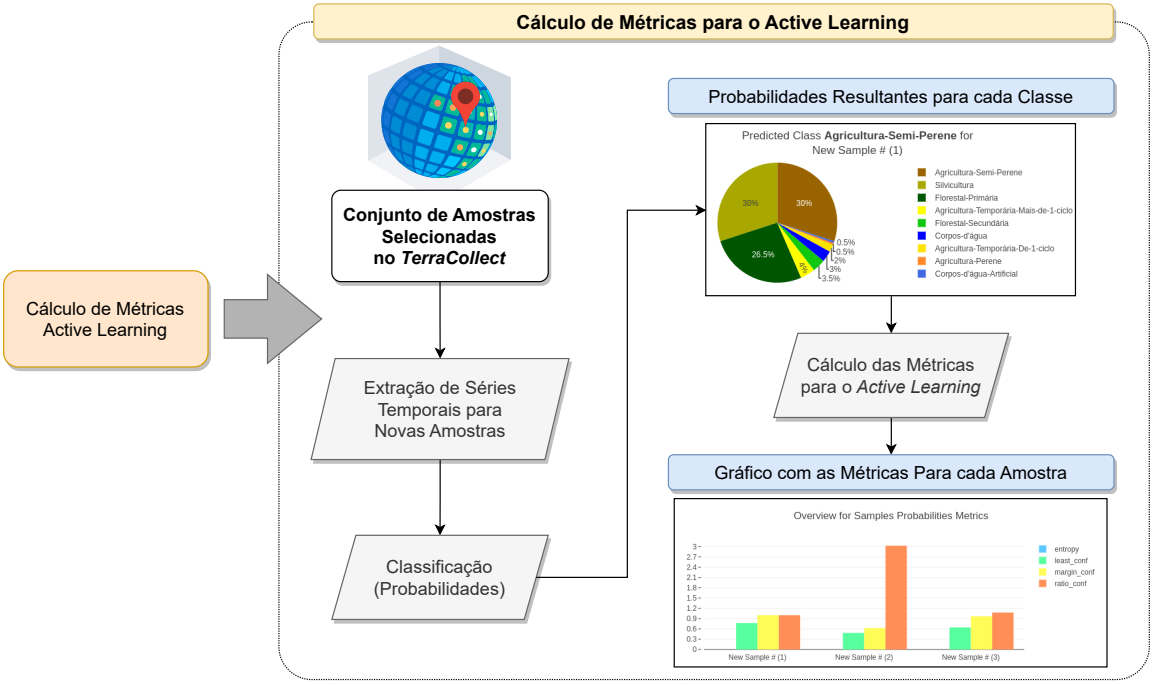
3.7.3 Cálculo de métricas com *Active Learning*

A um certo nível, a predição de probabilidades pode ajudar na aquisição de amostras mais precisas com boa qualidade. Porém utilizar a certeza de um modelo para rotular novos pontos cotados para o conjunto base não é sempre recomendado, pois acarreta em problemas de *overfitting*. O *Overfitting*, ou super ajuste, do modelo em relação à um conjunto de amostras pode gerar maior confusão na classificação de novos dados e posteriormente prejudicar a classificação das imagens e geração de mapas

LULC, por essa razão a metodologia usada para as estimativas com aprendizado de máquina também oferece métodos para o cálculo e visualização de métricas com as técnicas do *Active Learning*.

Conforme a Figura 3.7, o princípio do *Active Learning* é baseado nas mesmas etapas da predição de probabilidades. Contudo, o evento para iniciar o cálculo das métricas de *Active Learning* é a seleção de três ou mais pontos no mapa como um conjunto de amostras não-rotuladas. Cada nova amostra não-rotulada terá uma série temporal associada e será formatada para o *sits*, a probabilidade de cada uma será estimada, seguindo os mesmos passos descritos na seção anterior (Seção 3.7.2).

Figura 3.7 - Cálculo das métricas de probabilidades com *Active Learning*.



Fonte: Produção do autor.

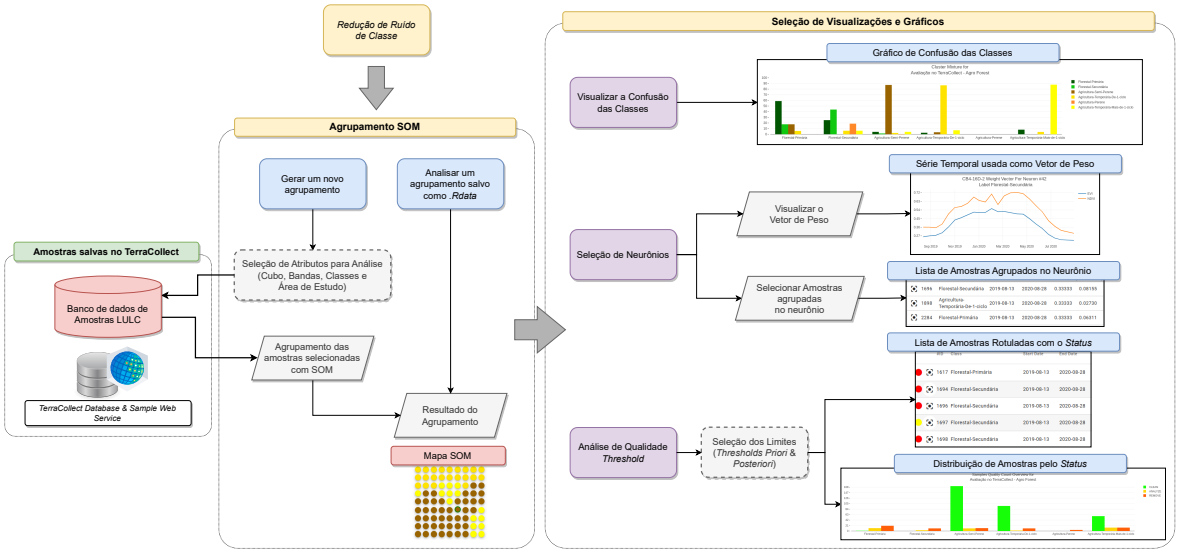
Com os valores das probabilidades, é feito o cálculo das métricas como entropia, razão e margem de confiança que quantificam a confusão do modelo em relação a classificação das novas amostras não-rotuladas. As métricas buscam mensurar a representatividade de uma amostra com base na confusão na predição. Observa-se que esta metodologia foca na operação de *query*, onde o "oráculo" seleciona as amostras com maior valor para o modelo aprender certas variâncias e ramificações nos dados. Como resultado, é obtido um gráfico de barras com as métricas para cada amostra indicando quais são as mais valiosas para o aprendizado do modelo.

A metodologia usada no *Active Learning* busca ranquear as amostras mais representativas usando as métricas calculadas com as probabilidades dentre um conjunto de novas amostras não-rotuladas. Fica a critério do usuário rotular as novas amostras ou descartá-las de acordo com a análise do gráfico contendo os valores das medidas de probabilidade e métricas do *Active Learning*, e mais a análise da série temporal. Com novas amostras salvas, na sessão de treinamento para a atualização do modelo de aprendizado de máquina, espera-se beneficiá-lo enriquecendo-o com novas informações, assim o modelo torna-se mais robusto a futuros erros.

3.8 Redução de ruído de classe

O método para a redução do ruído amostral proposto por Santos et al. (2021a) usa o SOM para realizar a redução da dimensionalidade preservando a topologia dos conjuntos de dados originais e medir a confusão das classes. Este tipo de avaliação nas amostras é baseado em etapas sequenciais como: geração do agrupamento SOM, análise dos neurônios no mapa SOM, análise das probabilidades *priori* e *posteriori* e por fim a detecção de ruído com a avaliação da qualidade. Conforme a Figura 3.8, para iniciar esta avaliação da qualidade nas amostras com a metodologia implementada na API de análise, o usuário deve selecionar um agrupamento salvo nos arquivos *Rdata* ou gerar um novo.

Figura 3.8 - Detecção de ruído amostral pós-coleta com *Class Noise Reduction*.



Fonte: Produção do autor.

Para gerar um novo agrupamento, o usuário deve fornecer à API os atributos para a definição das amostras a serem agrupadas e os hiper-parâmetros desejados para o SOM. A definição dos atributos para as amostras seguem o mesmo princípio do treinamento de modelos, onde ocorre a seleção de classes LULC para o conjunto base e mais o cubo e as bandas para selecionar as séries temporais. As amostras selecionadas para o agrupamento são classificadas e ajustadas no tamanho da grade selecionada. Este processo usa a classe majoritária e o cálculo das probabilidades condicionais conforme descrito na Seção 2.5.3.

Na geração de um novo agrupamento, ou na seleção de um previamente salvo, o resultado é o mapa SOM 2D e cada neurônio contendo as amostras com as probabilidades *priori* e *posteriori*, além das opções para a visualização do vetor de peso e de cada amostra e série temporal. Desta forma, é possível a análise dos resultados obtidos como a confusão das classes LULC que demonstra a proporção de confusão e mistura para cada classe dentre as usadas no agrupamento. Pode-se analisar também um neurônio manualmente, selecionando-o para visualizar as amostras agrupadas e o vetor de peso usado para classificá-lo, e posteriormente, pode-se realizar a detecção de ruído com a avaliação da qualidade usando os filtros *priori* e *posteriori*.

3.8.1 Interação com o objeto *self-organizing maps*

Ao selecionar o método para a redução de ruído amostral na API, a interface para a extensão de análise no *TerraCollect* fornece uma visualização interativa para o mapa SOM, onde o analista consegue explorar os neurônios e recuperar as amostras de cada grupo. Com a amostra selecionada pode-se analisar a sua série temporal e compará-la com o padrão da classe e o vetor de peso. Levando em consideração que a interface geral para a extensão de análise possui uma área dedicada somente a visualização de resultados, o usuário pode selecionar mais de um método dentre os disponíveis no menu de opções.

Este tipo de análise interativa pode ajudar o pesquisador a entender características e diferenças entre as classes LULC e semelhanças entre as amostras de uma mesma classe. Cada amostra é associada ao seu neurônio, e cada neurônio é rotulado com a classe majoritária, por isso dados que pertencem à mesma classe serão normalmente vizinhas no espaço 2D. Esta abordagem fornece informações adicionais sobre a variabilidade intraclasse e interclasse nas amostras. Devido à variabilidade entre séries temporais da mesma classe, os dados podem ser mapeados em neurônios diferentes, e a visualização interativa facilita a identificação destas amostras mais confusas.

3.8.2 Detecção de ruído de classe pós-coleta

Conforme a Figura 3.8, a metodologia para a detecção de ruído amostral pós-coleta envolve a aplicação de filtros *threshold*, fornecidos pelo usuário, nas probabilidades *priori* e *posteriori* associadas a cada amostra após o agrupamento SOM com a inferência *bayesiana*. A aplicação deste *threshold* irá classificar as amostras com *status* "*clean*", "*analyze*" ou "*remove*" para a avaliação da qualidade. Esta avaliação possui duas visualizações, a lista de amostras coloridas de acordo com o *status* e o gráfico de distribuição do *status* para cada classe LULC.

A visualização da lista de amostras com *status*, permite a busca e a seleção com base no *status*. Assim, o usuário pode selecionar todas as amostras classificadas com *remove* para removê-las do conjunto de dados base ou selecionar um conjunto de amostras com *analyze* para verificar a série temporal e reajustar os rótulos. Com o gráfico de distribuição do *status*, o usuário pode analisar a quantidade de amostras classificadas como "*clean*", "*analyze*" e também verificar classes com maior número de amostras ruidosas como "*remove*". Também é possível visualizar as amostras coloridas com *status* no mapa interativo no *TerraCollect*, onde pode-se, por exemplo, analisar qual área possui dados ruidosos e corrigi-los.

3.9 Serviço para análise de amostras

O serviço para análise de amostras possui a estrutura baseada em uma *Application Programming Interface* (API) desenvolvida com o pacote *Plumber R* (SCHLOERKE; ALLEN, 2022). Uma API permite a comunicação entre diferentes aplicações e serviços *web*, neste trabalho um dos requisitos essenciais é a comunicação entre a aplicação *TerraCollect* e o *sits*. O *sits* fornece um conjunto de ferramentas para a extração, análise e classificação de séries temporais com aprendizado de máquina, este pacote é a base da metodologia de análise fornecendo as implementações das abordagens descritas na Seção 3.5 (SIMOES et al., 2021a).

O pacote *Plumber R* converte as funções implementadas em *scripts* R pré-existentes em uma API na *web* usando uma coleção de comentários especiais para encapsula-los nos métodos *GET*, *POST* e *DELETE* (SCHLOERKE; ALLEN, 2022). Desta forma, qualquer função em R pode ser um método desta API *Plumber R* contendo entradas e saídas controlados por requisições. Contudo os métodos *sits* ainda precisam de adaptação para serem disponibilizados em um serviço *web*, principalmente na questão do formato de dados a serem passados como entradas e saídas. No caso,

os *scripts* são adaptados no serviço com funções complementares para converter dados em formato *JSON* para os formato *DataFrame* do *sits*, conforme demonstrado pela Figura 3.9. Esta conversão é necessária para facilitar a leitura e listagem de amostras e permitir a aplicação dos métodos de análise de forma eficiente.

O formato *JSON* se tornou popular para a comunicação de aplicações *web*, pois é mais leve e menos complexo que o *XML*. O pacote *Plumber R* possui algumas abstrações para lidar com a conversão de dados, contudo o *sits* possui estruturas de dados *DataFrame* com atributos específicos, um dos mais essenciais é a coluna contendo as as séries temporais em um registro. Por isso a necessidade de adaptação com *scripts* completares, onde se faz a conversão de *JSON* para *DataFrame sits* de forma automática a cada requisição da interface para o serviço. Também há soluções para converter as amostras provenientes do serviço *Sample WS* do *TerraCollect* na leitura para a extração das séries temporais, neste caso os dados do *Sample WS* possuem os mesmos atributos com nomenclaturas diferentes no *sits*.

Figura 3.9 - Conversão de amostras em formato JSON para o objeto *DataFrame* do pacote *sits* R.

Fonte: Produção do autor.

As abordagens para análise de amostras estão organizadas no serviço seguindo uma especificação[1]. O resumo desta especificação é apresentado na Figura 3.10 com a descrição de cada método de requisição HTTP e a abordagem correspondente.

Figura 3.10 - Definição de métodos de requisição HTTP para a API de análise de amostras.

Abordagem	Método de Requisição HTTP	Descrição

Extração, Regularização, Armazenamento & Gerenciamento de Séries Temporais	`POST samples/extract`
	`GET samples/extract/status`
	`GET samples/extract/stop`
	`GET samples/extract/rebuild`
		Iniciar o processo de extração, regularização e armazenamento. Permite verificar o progresso e *status* da extração. Além de ser possível parar o processo e reiniciar.

	`GET samples/sync/check` `GET samples/sync`	Este grupo de métodos é usado para verificar atualizações no banco de dados e a necessidade de extração em novas amostras.
	`POST samples/extract/add` `DELETE samples/extract/remove`	O método “add” adiciona novos conjuntos de dados baseado no cubo e nas bandas. Já o método “remove”, remove conjunto de dados.

Análise Exploratória de Amostras LULC	`GET samples/`
	`GET samples/summary`
	`GET samples/patterns`
	`GET samples/download`
		Grupo de métodos para a análise exploratória, incluindo a busca e download de amostras mais séries temporais.

Estimativas com Machine Learning	`POST model/train`
	`GET model/update`
	`DELETE model/remove`
	`GET model/estimation`
	`POST model/active_learning`
		Treinamento de novos modelos, descrição de metadados, atualização com novos dados, remoção e predição de probabilidades. Permite também o cálculo das métricas de *active learning*.

SOM Clustering
 	
Class Noise Reduction
	`POST clustering/`
	`DELETE clustering/remove`
	`GET clustering/description`
	`POST samples/quality`
		Gerenciamento de agrupamentos SOM. Permitem a criação de novos agrupamentos, descrição, remoção e análise de qualidade.

Fonte: Produção do autor.

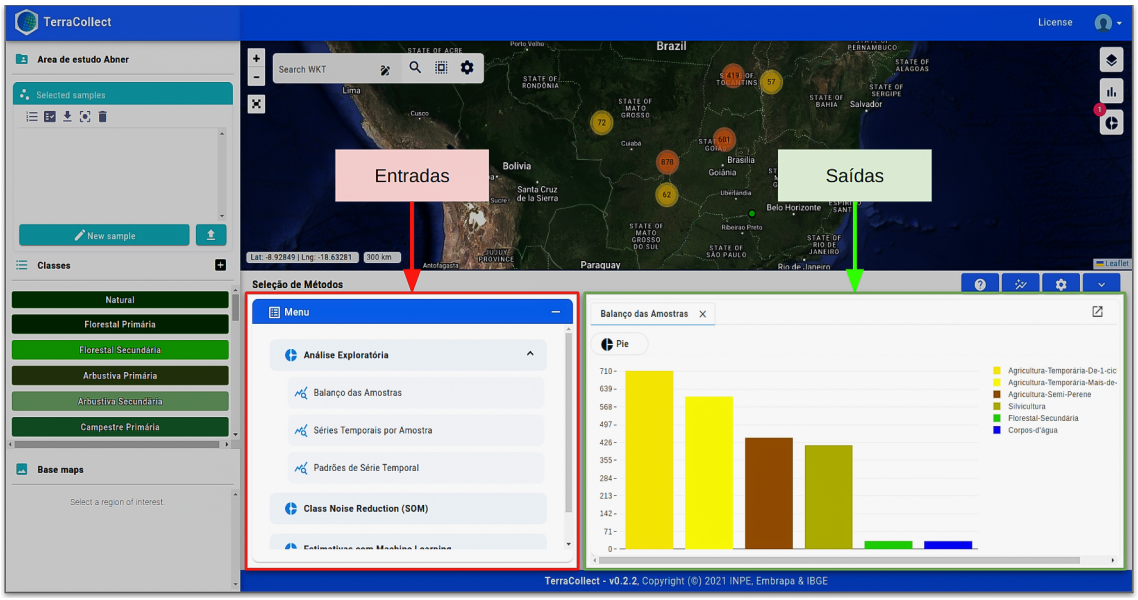
[1] <https://brazildatacube.dpi.inpe.br/dev/terracollect/wsas/___docs___>

3.10 Interface para análise de amostras

A aplicação *web* da plataforma *TerraCollect* foi desenvolvida com o *framework Angular* (JAIN et al., 2014) para criar uma interface gráfica com visualizações interativas de amostras, produtos de imagens BDC, gráficos de trajetórias LULC e séries temporais. O *Angular* possui soluções e componentes prontos para gráficos, mapas, criação de formulários e construção de *dashboards*. Como por exemplo, os gráficos são gerados usando a biblioteca *Plotly* na qual os estilos e dados são configurados usando apenas *JSON*. A extensão, ou *plug-in*, para análise constitui um *dashboard* acoplado a interface no *TerraCollect* e oferece os controles para o envio de requisições no serviço de análise para executar os métodos.

Conforme Figura 3.11, a interface para análise de amostras foi desenhada baseando-se nas funcionalidades do pacote *sits*, com áreas separadas para entrada de dados e saída de resultados. A área de entrada contém os formulários e a área de saída possui a estrutura em abas, onde conforme as opções são selecionadas, gráficos, tabelas e demais visualizações são acumulados nesta área. Esta interface traz algumas semelhanças em comparação ao GEE, contudo o intuito é oferecer automação, então não há necessidade de inserir ou editar *scripts* ou até mesmo gerenciar arquivos.

Figura 3.11 - Exemplo para demonstração da interface para a plataforma de análise.



Fonte: Produção do autor.

A interface para a extensão de análise separa as abordagens em diferentes componentes disponibilizando-os em um menu opções para a execução. Cada opção possui um formulário para a seleção de atributos e filtragem dos dados. A metodologia para a extensão permite que todos os filtros estejam disponíveis a serem aplicados nas amostras para o critério do usuário, como: filtragem por classes LULC, cubos, bandas, aplicação de suavização de séries e até a seleção de uma área de estudo. Por exemplo, para todos os métodos de análise é possível a aplicação de filtros para a definição de um conjunto base que irá gerar um resultado, seja para executar a análise exploratória ou mesmo treinar um modelo de aprendizado de máquina.

A extensão para análise comunica-se diretamente com o serviço *web Plumber R* para a integração dos métodos usando apenas *JSON*, com isso as técnicas de visualização sofrem algumas alterações em comparação da interface para os *scripts* R. Não há como levar os pacotes gráficos do R para a aplicação *web* do *TerraCollect*, por isso o pacote *Plotly* foi usado em algumas visualizações e em outras foram feitas adaptações. Uma dessas adaptações foi feita para exibir o mapa SOM interativo na extensão, onde foi desenvolvido um *script* em *JavaScript* exclusivo para implementar esta funcionalidade.

As etapas para análise na interface gráfica não diferenciam muito das apresentadas no manual oficial do pacote *sits*, caso o usuário deseje migrar da interface para *scripts* R. A extensão de análise possui opções para a exportação de amostras em formatos como *Rdata* e *.RDS* para a leitura das séries temporais e execução dos métodos usando o pacote *sits* em algum editor de código da escolha do usuário. Também há a exportação de arquivos *Rdata* contendo o objeto agrupamento SOM ou modelos treinados para realizar testes, gerar resultados e visualizações mais específicos fora do escopo da extensão.

4 RESULTADOS E EXPERIMENTOS

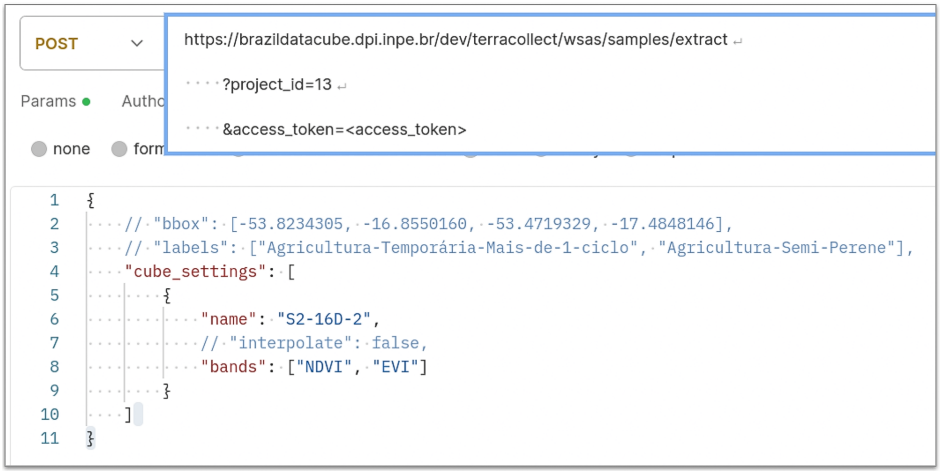
Como prova de conceito, a arquitetura de integração dos métodos de análise de amostras LULC foi implementada como uma extensão das funcionalidades da plataforma *TerraCollect*. Foram feitos diversos testes e implementações para estudar as diferentes abordagens para a análise de amostras. O objetivo principal dos testes foi verificar a viabilidade da integração em uma aplicação *web*. Em sequência, serão apresentados os resultados obtidos com a implementação de um protótipo da extensão de análise no *TerraCollect*.

Primeiramente será apresentado a versão atual e o funcionamento do protótipo da extensão no ambiente de desenvolvimento do *TerraCollect*. Prosseguindo com exemplos de aplicação como extração e busca de séries temporais usando o serviço de análise e a interface gráfica. Após estas seções será apresentado um estudo de caso demonstrando a aplicação de todos os métodos de análise e, por fim, serão apresentados alguns pontos para a discussão.

4.1 Aplicações do serviço *web* para análise de amostras

A base para a extensão no *TerraCollect* é o serviço *Web* para análise de amostras. Esta API *Plumber R* segue as mesmas etapas descritas no manual *sits* para iniciar o processo de análise. A primeira etapa é a extração de séries temporais, e com isso a criação da extensão do projeto *TerraCollect* com as configurações do cubo armazenadas em arquivos *Rdata*. A Figura 4.1 apresenta um exemplo para realizar a requisição de extração diretamente no serviço.

Figura 4.1 - Exemplo de requisição para extração de séries temporais de imagens no serviço *web* de análise de amostras.



Fonte: Produção do autor.

Para a extração de séries temporais, é necessário que o projeto tenha um número mínimo de amostras salvas, ao menos uma para cada classe. Com isso, deve-se fornecer um projeto *TerraCollect* usando o ID e, opcionalmente, conforme as escolhas do usuário, selecionar classes LULC e uma área de estudo para as amostras alvo. Estes atributos são usados para selecionar e recuperar amostras no banco de dados do *TerraCollect* e definir a base de dados para os métodos de análise. Para a configuração da série temporal, é necessário fornecer o ID do cubo no BDC e as bandas e índices espectrais conforme as configurações no modelo lógico descrito na Seção 3.5.

Pode-se notar que a Figura 4.1, no corpo da requisição em *JSON*, apresenta a seleção de área (atributo "*BBOX*" ou "*Bounding Box*") e classes LULC (atributo *labels*) definidos como comentários apenas para exemplificar. Uma vez que esses atributos são opcionais, em caso do usuário desejar delimitar uma área e/ou usar classes específicas. A interpolação é selecionada por padrão, para não interpolar as séries, o atributo "*interpolate*" deve ser configurado como "*False*".

Quando uma requisição à extração é feita, o resultado é um *JSON* com a descrição do projeto mais os metadados do sub-processo criado, conforme a Figura 4.2. A API *Plumber R* recebe as configurações e inicia um sub-processo em uma sub-sessão no ambiente R onde serão executadas as etapas descritas na Seção 3.4 e 3.5 para extração, regularização e armazenamento de séries temporais. Este projeto é apenas um exemplo, contendo amostras fornecidas pelo TerraClass em conjunto com amostras coletadas pelos próprios desenvolvedores do *TerraCollect*.

Nesta resposta demonstrada na Figura 4.2, o usuário pode visualizar os metadados do projeto como configurações do cubo e data de criação e atualização, o *status* do processo e demais notificações sobre a extração. No exemplo nota-se que o processo está extraindo dados do satélite *Sentinel-2* para as bandas EVI e NDVI, seguindo as configurações do cubo. As notificações e saídas de texto do processo são adicionadas a chave *output*, como o processo acabou de ser iniciado, neste exemplo o atributo está como "*No output!*" indicando que não há notificações.

Neste exemplo pode-se observar os principais atributos para verificar o estado atual do processo de extração. A chave "*process_status*" apresenta apenas duas mensagens "*Extracting...*", para representar o processo ativo, ou "*Not working!*", representando que não está em execução. A chave "*process*" contém os dados do sub-processo no ambiente R e a chave "*process_id*" contém o ID que o identifica. A chave "*process_progress*" demonstra o progresso da extração com a porcentagem para a conclusão variando de 0 (não iniciado), entre 0 e 1 (em andamento) e 1 (concluído).

Figura 4.2 - Exemplo de resposta para a requisição de extração de séries temporais no serviço *web* de análise de amostras.

response.json

```
/*
    https://brazildatacube.dpi.inpe.br/dev/terracollect/wsas/
        samples/extract
            ?project_id=13
            &access_token=<access_token>
*/
{
    "project_status": {
        "process": {
            "pid": 21265,
            "ppid": 20417,
            "name": "R",
            "status": "running",
            "system": 0.62,
            "rss": 75456512,
            "vms": 1159966720,
            "created": "2023-08-30 17:51:05"
        },
        "message": "The time series extraction process has started!",
        "output": [
            "No output!"
        ],
        "project": {
            "project_id": 13,
            "title": "Area de estudo Abner",
            "name": "Area-de-estudo-Abner",
            "description": "avaliaçao de amostras 09-2021.",
            "start_date": "2019-08-13",
            "end_date": "2020-08-28",
            "bbox": "None",
            "labels": [ ... ],
            "cube_settings": [
                {
                    "name": "S2-16D-2",
                    "interpolate": true,
                    "bands": [
                        "EVI",
                        "NDVI"
                    ]
                }
            ],
            "process_id": 21265,
            "process_status": "Extracting ... ",
            "process_progress": 0,
            "estimated_time": "0h 0m 0s",
            "created_at": "2023-08-30 14:51:05",
            "updated_at": "2023-08-30 14:51:06"
        }
    }
}
```

Fonte: Produção do autor.

A chave "*estimated_time*" exibe o tempo estimado para a conclusão, o cálculo não é exato pois é feito usando testes de conexão com o serviço WTSS e testes com a execução das etapas e sub-etapas de tratamento nos dados usando uma amostra aleatória. Enquanto a extração é executada em segundo plano, o armazenamento é feito por cubo, quando a extração de dados de um cubo é concluída o mesmo é tratado e armazenado e o processo inicia o próximo cubo selecionado pelo usuário. Caso o processo esteja em andamento, até que algum conjunto de dados seja salvo no arquivo *Rdata* correspondente, não é possível visualizar amostras com a séries temporal ou executar os métodos.

Para o protótipo, ainda não foi incluído o suporte para eventuais erros de rede, processamento ou limitação da memória que podem prejudicar a execução, caso algum erro seja encontrado o processo é interrompido. Os dados já processados são mantidos e o que estavam em processamento são excluídos para não haver arquivos corrompidos. O usuário pode manualmente verificar o *status* do processo se algum dado foi salvo, e reiniciá-lo. A Figura 4.3 apresenta a resposta para a requisição do *status* do processo de extração. A visualização do *status* do processo de extração fica sempre disponível em "*samples/extraction/status*" usando o serviço de análise.

Na Figura 4.3, observa-se que o atributo *output* contém algumas notificações e mensagens do processo em execução. O exemplo está indicando que ainda está extraindo as séries temporais do *Sentinel-2*, o processo já fez o pré-tratamento nos dados como a leitura das amostras do *TerraCollect* e já salvou as configurações do projeto de análise em *Rdata*. O tempo para a execução da extração e tratamento dos dados, no exemplo, está estimado para ser concluído em 24 minutos e 9 segundos.

Figura 4.3 - Exemplo de resposta para a requisição do *status* da extração de séries temporais no serviço *web* de análise de amostras.

response.json

```
/*
    https://brazildatacube.dpi.inpe.br/dev/terracollect/wsas/
        samples/extract/status
            ?project_id=13
            &access_token=<access_token>
*/
{
    "project_status": {
        "process": { ... },
        "message": "The process did not save any time series yet!",
        "output": [
            "Color configurations found in /home/guest/miniconda3/envs/rwsas-dev/lib/R/library/sits/extdata/config_colors.yml",
            "To provide additional configurations, create an YAML file and inform its path to environment variable 'SITS_CONFIG_USER_FILE'.",
            "Using raster package: terra",
            "Loading project cube settings ... ",
            "Initializing extraction ... ",
            "Getting samples from TerraCollect ... ",
            "Filtering samples ... ",
            "Getting saved time series settings ... ",
            "Preparing to save Rdata file ... ",
            "Calculating estimated time to conclusion ... ",
            "Extracting time series from S2-16D-2 ... "
        ],
        "cube": [ ... ],
        "project": {
            "project_id": 13,
            "title": "Area de estudo Abner",
            "name": "Area-de-estudo-Abner",
            "description": "avaliaçao de amostras 09-2021.",
            "start_date": "2019-08-13",
            "end_date": "2020-08-28",
            "bbox": "None",
            "labels": [ ... ],
            "cube_settings": [ ... ],
            "process_id": 21265,
            "process_status": "Extracting ... ",
            "process_progress": 0.35,
            "estimated_time": "24min 9s",
            "created_at": "2023-08-30 14:51:05",
            "updated_at": "2023-08-30 14:51:25"
        }
    }
}
```

Fonte: Produção do autor.

O *status* do processo de extração é uma resposta padronizada e contém as notificações para todos os métodos da API com ligação ao armazenamento de séries temporais separados em extração, adição, atualização e remoção. A "extração" é o início de um projeto de análise, logo depois quando novas amostras são salvas ou removidas no *TerraCollect*, a base do projeto de análise requer "atualização". O processo de "adição" é quando o usuário deseja adicionar outros cubos à base do projeto. Também é possível atualizar os dados de um cubo já salvo, isso ocorre quando as imagens de satélite sofrem algum tipo de atualização no servidor BDC. Ao requisitar qualquer um destes métodos, o usuário pode então verificar o andamento de cada um dos processos com a requisição ao *status*.

Esta implementação não permite a execução simultânea de métodos com ligação ao armazenamento para manter a integridade dos dados, uma vez que estes métodos lidam com o sistemas de arquivos *]Rdata*. Caso o usuário requisite a atualização e logo depois a adição, por exemplo, com a estratégia usada de verificação do *status*, haverá uma notificação indicando que a atualização está em andamento. Não é possível, neste exemplo, adicionar novos cubos à base até que a atualização seja concluída. Com o processo de armazenamento concluído tanto para atualização quanto para a adição, a base de dados para a análise fica disponível.

Para facilitar a criação de formulários na interface gráfica, as informações fundamentais de cada projeto de análise são organizadas em um único método na API. A Figura 4.4 apresenta a resposta para a requisição deste método. Para obter esta resposta é necessário a seleção de um projeto usando o ID, o mesmo usado pelo *TerraCollect*. Nesta resposta *JSON* é possível ter a lista e os dados de cada modelo e agrupamento SOM salvo pela API. Assim, pode-se verificar quantos modelos de aprendizado de máquina estão salvos e quantos agrupamentos foram feitos, os cubos que foram salvos, a descrição das bandas espectrais, etc.

Cada item da lista de cubos salvos em arquivos *Rdata* contém a descrição, título, ID, a linha temporal para as séries, os metadados do cubo e mais uma lista com a descrição de cada banda ou índice espectral. A cada requisição o arquivo *Rdata* pronto para a leitura é recuperado, o *sits* é aplicado e a API monta a resposta. Alguns dados como configurações do projeto são armazenadas, porém dados como o período e metadados do cubo são buscados a cada requisição usando o STAC, WTSS e *sits*. Dados como o balanço das amostras, retângulo envolvente para o *BBOX* e linha temporal das séries temporais são calculados usando os métodos do *sits*.

Figura 4.4 - Exemplo de resposta para a requisição da descrição e resumo do projeto de análise salvo pelo serviço *web* de análise.

response.json

```
/*
    https://brazildatacube.dpi.inpe.br/dev/terracollect/wsas/summary
        ?project_id=13
        &access_token=<access_token>
*/
{
    "summary": {
        "project": {
            "project_id": 13,
            "title": "Area de estudo Abner",
            "name": "Area-de-estudo-Abner",
            "description": "avaliaçao de amostras 09-2021.",
            "created": "2022-02-15",
            "start_date": "2019-08-13",
            "end_date": "2020-08-28",
            "available_mlmodels": [
                "RandomForest"
            ],
            "available_distances": [
                "euclidean"
            ],
            "saved_ml_models": [ ... ],
            "saved_som_results": [ ... ]
        },
        "samples": {
            "labels_summary": [
                {
                    "label": "Agricultura-Temporária-De-1-ciclo",
                    "color": "#FFE300",
                    "count": 108,
                    "prop": 0.1855670103093
                },
                ...
            ],
            "bbox": {
                "xmin": -55.112993,
                "ymin": -21.483741,
                "xmax": -46.809225,
                "ymax": -9.622414
            },
            "time_interval": {
                "start_date": "2019-08-13",
                "end_date": "2020-08-28"
            },
            "size_grid": {
                "grid_xdim": 9,
                "grid_ydim": 9
            }
        },
        "cube": [
            {
                "id": "S2-16D-2",
                "title": "Sentinel-2 - 10m - 16 days - v2",
                "description": "This datacube was generated with all available surface
                    reflectance images processed using Sen2cor. The data is provided with 10
                    meters of spatial resolution, reprojected and cropped to BDC_SM grid Version 2 (BDC_SM V2),
                    considering a temporal compositing function of 16 days using the
                    Least Cloud Cover First (LCF) best pixel approach.",
                "interpolate": true,
                "timeline": [ ... ],
                "bands": [
                    ... ,
                    {
                        "name": "NDVI",
                        "common_name": "ndvi",
                        "description": "",
                        "min": -10000,
                        "max": 10000,
                        "nodata": -9999,
                        "scale": 0.0001,
                        "data_type": "int16"
                    }
                ]
            }
        ],
        "users": [ ... ]
    }
}
```

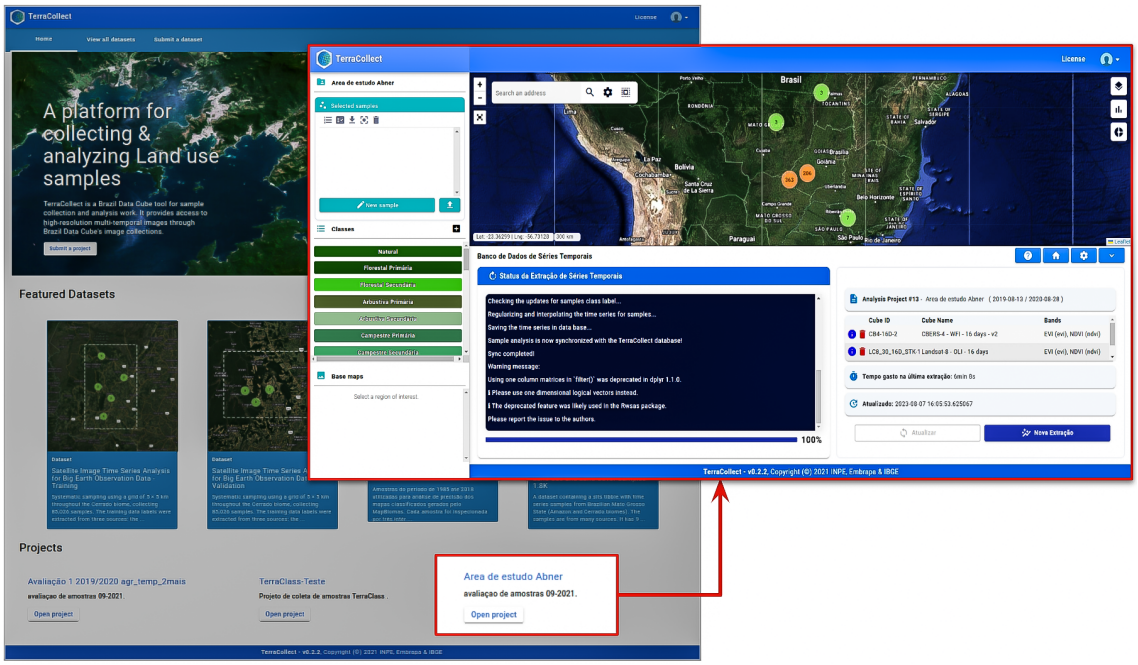
Fonte: Produção do autor.

4.2 Extensão *TerraCollect* para análise de amostras

O *TerraCollect* é uma plataforma *web* separada em serviços, componentes e suas conexões, onde cada componente possui o controle de uma funcionalidade como coleta, busca de imagens, visualização de dados em gráficos, etc. A extensão para análise de amostras foi implementada como um componente do *TerraCollect* contendo a interface gráfica mais a conexão com o serviço *web* de análise conforme a arquitetura descrita na Seção 3.2. Em sequência serão apresentadas as principais características deste componente e da interface para análise.

Para começar a analisar amostras no *TerraCollect*, o usuário, no primeiro acesso, deve criar um projeto seguindo o formulário de submissão na página inicial. Após a criação, o usuário pode selecioná-lo dentre a lista de projetos disponíveis, conforme a Figura 4.5. Com o projeto aberto, é um requisito para iniciar a análise que o mesmo tenha um conjunto mínimo de amostras salvas para extrair a série temporal e salvar as configurações do cubo associadas ao projeto. Para isso o usuário pode coletar amostras usando o desenho livre em conjunto com a visualização de imagens ou fazer o *upload* de amostras já coletadas em outra plataforma.

Figura 4.5 - Demonstração da seleção de um projeto na plataforma *web* *TerraCollect* para iniciar a análise de amostras.



Fonte: Produção do autor.

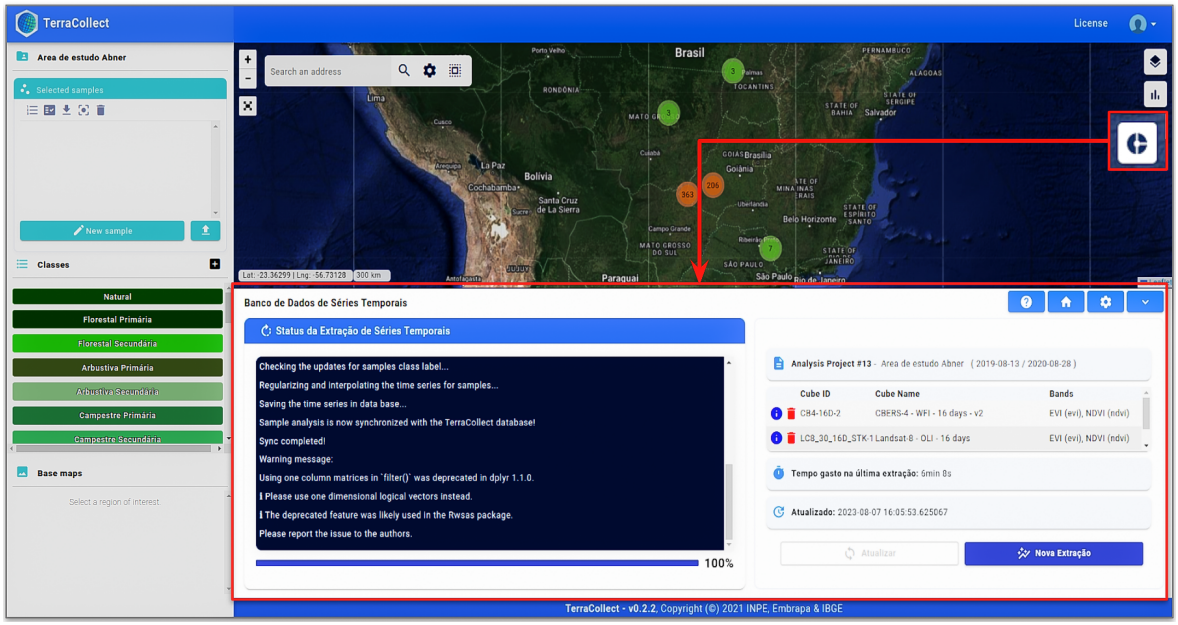
Com um conjunto de amostras salvo no banco de dados do *TerraCollect*, pode-se abrir o componente para análise com o botão no canto superior direito da plataforma, como ilustrado pela Figura 4.6. A primeira tela exibe os controles da primeira etapa da análise, a extração de séries temporais, sendo um requisito essencial para a aplicação dos métodos. Caso seja o primeiro acesso a plataforma, esta tela exibirá somente um botão para abrir o formulário de seleção de cubos e bandas do WTSS. Como neste exemplo as séries temporais já foram configuradas, há a descrição do último processo de extração e armazenamento das séries.

A Figura 4.7 apresenta a descrição da primeira interface no caso do processo de extração já ter sido configurado. Esta tela mostra as informações separadas em seções sendo o *status* do processo de extração e mais os metadados e configurações salvas para o projeto de análise. O *status* possui uma área para a visualização das saídas como notificações do processamento em segundo plano sendo executado no ambiente R no lado do servidor. A seção com as configurações salvas contém a descrição do projeto de análise, conjunto de dados mais opções para atualizar a base de dados e adicionar novos conjuntos.

Na seção de visualização do *status* é exibido uma caixa de texto com as notificações e mensagens do processo de extração, o *output*, e mais uma barra de progresso com a porcentagem de etapas concluídas. Durante o processo de extração esta seção exibe o nome do cubo em processamento e mais a opção de interromper o processamento. A verificação do *output* é essencial para identificar possíveis erros durante a execução, dentre os mais graves como erros de conexão de rede no servidor e mais comuns como falhas na extração de uma série com uma coordenada inválida ou na interpolação de dados. Caso houver problemas dos mais graves o processamento é interrompido, erros mais leves, por exemplo em uma amostra em específico, surgirá apenas uma mensagem indicando o ID da amostra e mais o cubo com erro no *output*, o processo continuará em execução.

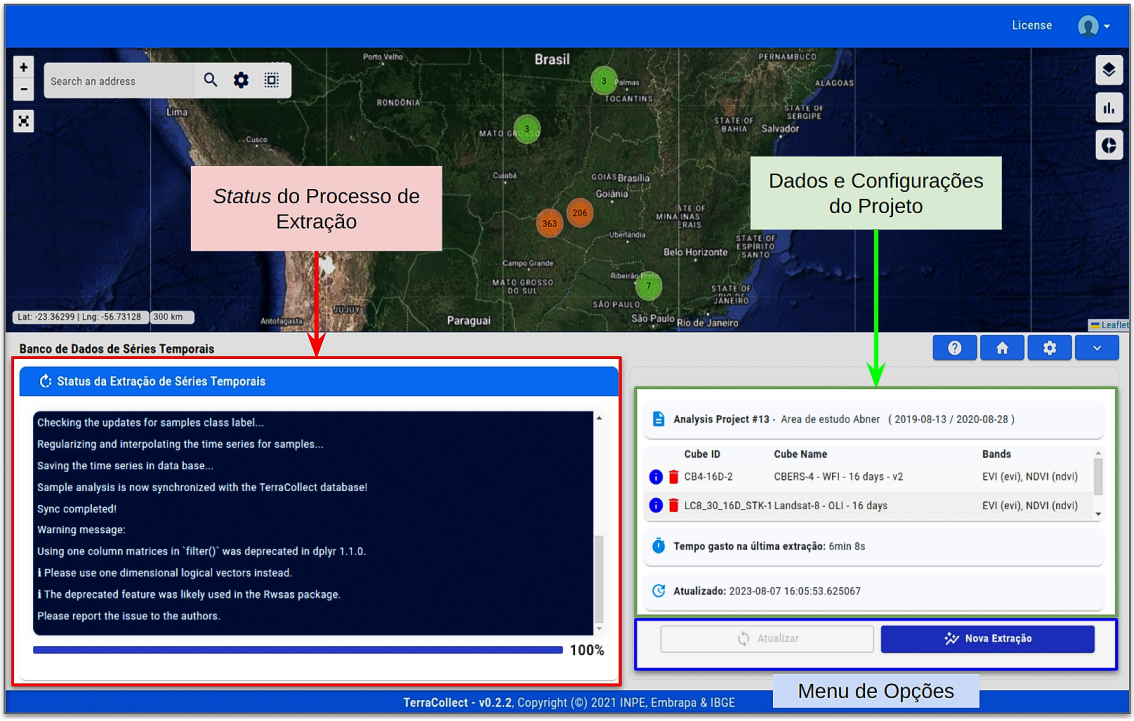
A seção com a descrição das configurações salvas apresenta metadados como ID, nome, descrição e período definindo a linha temporal do projeto de análise. Nesta seção há a listagem dos conjuntos de dados salvos organizados pelo cubo, cada item contém opções onde pode-se remover um conjunto ou exibir informações sobre, em sequência o ID, o nome do cubo no BDC e a lista de bandas e índices espectrais. Esta listagem tem o objetivo de oferecer ao usuário um resumo dos conjuntos de dados de séries temporais disponíveis para análise. Abaixo desta lista, segue-se a data da última atualização e mais o tempo gasto no último processamento de extração.

Figura 4.6 - Tela inicial para extração de séries temporais na extensão de análise no *TerraCollect*.



Fonte: Produção do autor.

Figura 4.7 - Descrição da interface gráfica para a extração de séries temporais na extensão de análise no *TerraCollect*.

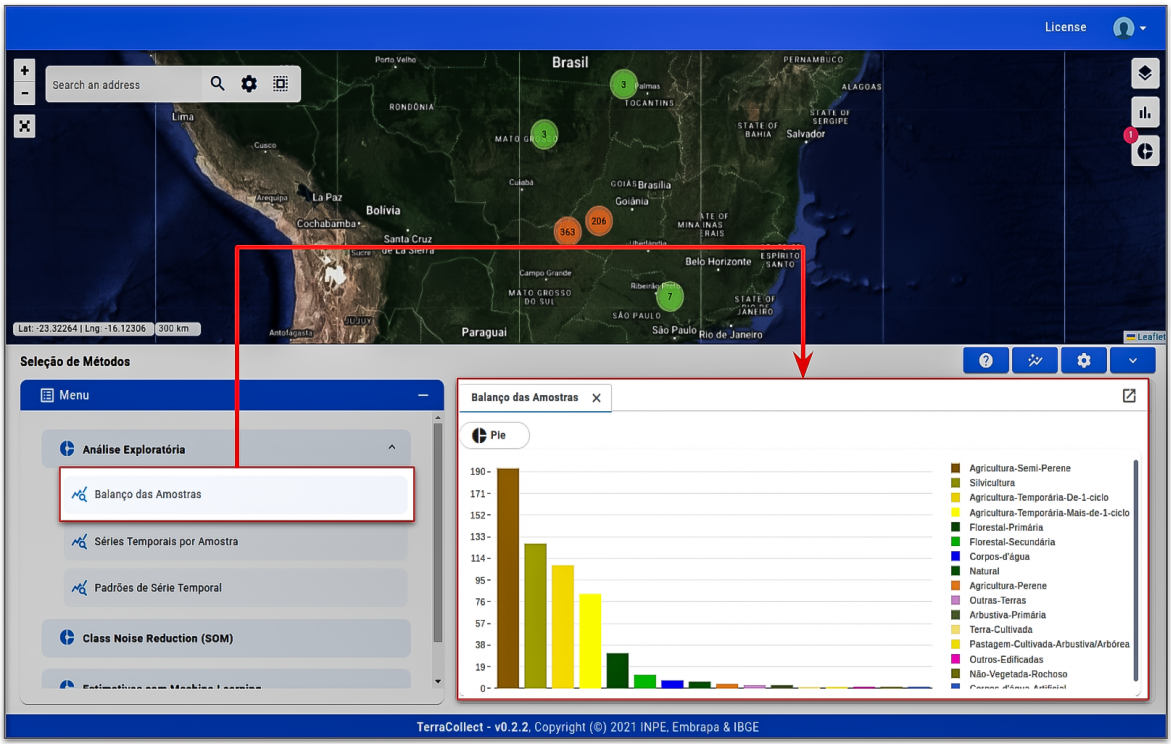


Fonte: Produção do autor.

Ainda nesta seção, há o menu com duas opções “atualizar” e “nova extração”. A opção “atualizar” irá verificar se novas amostras foram adicionadas ou removidas do banco de dados do *TerraCollect* para atualizar o banco de séries temporais orientando-se pelas configurações salvas. Por exemplo, caso uma nova amostra seja salva, esta opção fará a extração da série temporal de forma automática apenas para esta amostra seguindo a configuração dos cubos listados. A opção “nova extração” é utilizada para adicionar novos cubos ao conjunto, abrindo novamente o formulário para a seleção de cubos e bandas para a extração das séries do cubo selecionado.

Após esta primeira etapa de extração de dados, a tela para análise fica disponível no menu de opções no canto superior direito do componente, conforme a Figura 4.8. Neste menu é possível configurar opções globais consideradas avançadas como a seleção de área, seleção de amostras coletadas por usuário e o tipo de visualização de amostras no mapa. Nesta interface, ao lado esquerdo é exibido a lista de métodos disponíveis e ao lado direito as saídas em geral como visualizações, tabelas e mapas. No caso da Figura 4.8, é ilustrado a seleção do balanço das amostras que não necessita de formulário, o resultado é o gráfico de barras à direita.

Figura 4.8 - Descrição da interface gráfica para execução dos métodos de análise na extensão no *TerraCollect*.

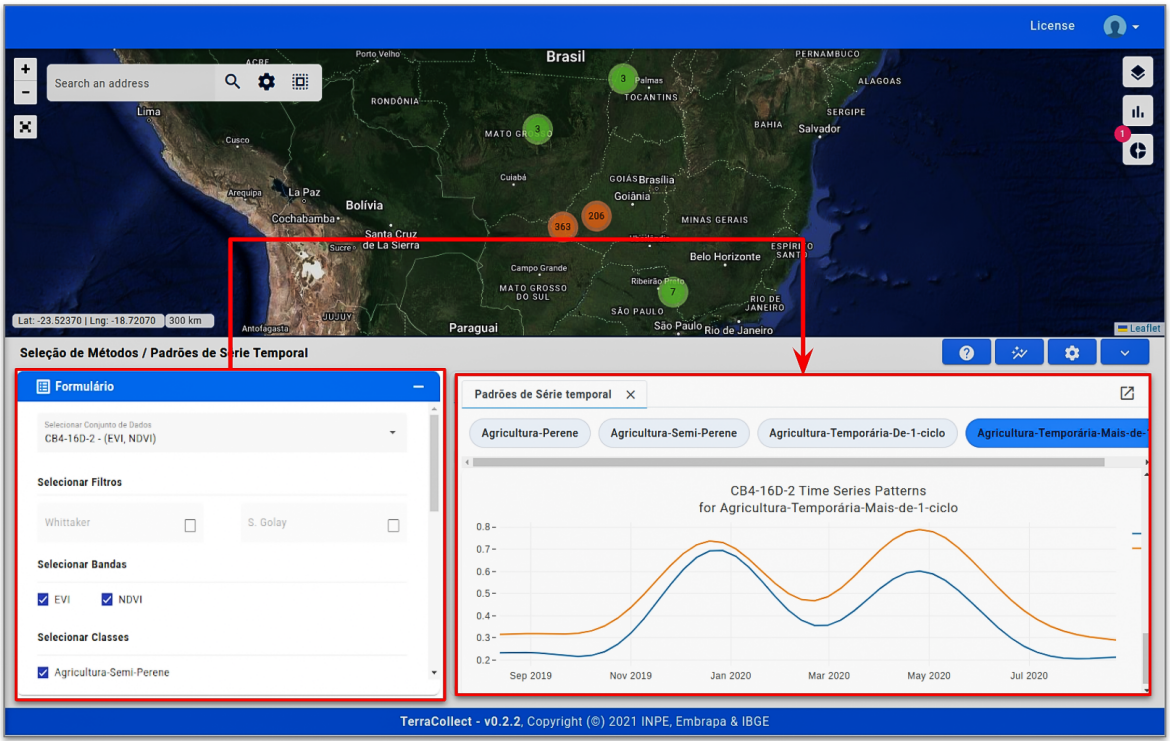


Fonte: Produção do autor.

Nesta tela (Figura 4.8), a lista de métodos a esquerda é organizada conforme a metodologia descrita na Seção 3.3. As visualizações para a análise exploratória estão listadas como subtópicos, esta estruturação deve-se ao fato destas opções serem independentes e requisitarem apenas a seleção do cubo. Outra observação é que há uma breve descrição na parte superior do componente, indicando qual foi o método selecionado e qual a etapa que está em aberto. Esta descrição serve para auxiliar o usuário, sendo essencial para organizar métodos com muitas etapas, tais como a avaliação de qualidade com SOM.

A Figura 4.9 apresenta a seleção de um método com formulário, o cálculo dos padrões de série temporal. Para este método da análise exploratória é necessário selecionar o cubo, bandas e classes para o cálculo, os resultados são exibidos ao lado direito com a organização em abas. Algumas visualizações, tais como esta, apresentam a seleção de classes LULC no resultado para organizar a exibição dos gráficos de segmentos. Neste exemplo, foram selecionadas apenas classes de agricultura, e o padrão selecionado na tela foi para a classe de "Agricultura mais de 1 ciclo".

Figura 4.9 - Demonstração da interface gráfica com formulário para execução dos métodos de análise na extensão no *TerraCollect*.



Fonte: Produção do autor.

4.3 Estudo de caso TerraClass Cerrado

Este estudo de caso tem como região de interesse a extensão do bioma Cerrado, o segundo maior bioma da América do Sul com uma área de mais de 2 milhões de km^2 aproximadamente 22% do Brasil [1]. Esta região foi escolhida pela dinâmica no uso e cobertura da terra e por causa da intensa atividade agropecuária. Existem vários estudos sobre a análise e classificação dessas atividades e suas mudanças nesta região como apresentado por Santos et al. (2021a) e Simoes et al. (2021a).

A Tabela 4.1 contém a distribuição de amostras por classe para a criação do conjunto de dados neste estudo de caso. Estas amostras foram obtidas com o projeto teste no *TerraCollect* para o TerraClass. O projeto teve como objetivo testar a interface e as funcionalidades da plataforma com dados e usuários reais, envolvendo a coleta com base na interpretação visual de imagens de satélite. Estas amostras possuem o mesmo período de Agosto de 2019 à Agosto de 2020 e foram selecionadas por região de interesse usando os limites do bioma Cerrado (Figura 4.10). Em sequência, uma breve descrição das classes LULC alvo segundo o *Geoportal TerraClass Cerrado*[2]:

- **Agricultura Temporária de 1 Ciclo:** culturas agrícolas temporárias, apresentando apenas um ciclo de produção no ano de referência, sobretudo de grãos e fibras;
- **Agricultura Temporária de Mais de 1 Ciclo:** culturas agrícolas temporárias que apresentam mais de um ciclo de produção no ano de referência;
- **Agricultura Semiperene:** culturas agrícolas Semi perenes que podem apresentar o ciclo de produção superior ao período de referência, sobretudo cana-de-açúcar;
- **Silvicultura:** culturas de espécies florestais naturais ou exóticas, de interesse comercial, geralmente representadas por formações arbóreas;
- **Corpos D'água:** corpos d´água naturais ou artificiais, decorrentes de represamentos de cursos d´água, tais como represas, açudes, etc;
- **Formação Florestal:** Formação vegetal natural com predominância de floresta savânica; onde foi feito corte raso, vegetação caracterizada por árvores e arbustos com troncos tortuosos.

[1] <https://www.gov.br/mma/pt-br/assuntos/ecossistemas-1/biomas/cerrado>
[2] <https://www.terraclass.gov.br/geoportal-cerrado/>

Tabela 4.1 - Distribuição de amostras por classe estudo de caso TerraClass Cerrado.

Classe LULC	**Número de Amostras**
Agricultura Temporária de 1 Ciclo	712
Agricultura Temporária de Mais de 1 Ciclo	609
Agricultura Semiperene	445
Silvicultura	415
Formação Florestal	33
Corpos D'água	32
Total	**2.246**

Fonte: Produção do autor.

4.3.1 Extração de séries temporais de imagens de satélite

Conforme a Figura 4.11, as séries temporais para as amostras foram extraídas do cubo de dados com imagens de refletância de superfície do sensor WFI a bordo do satélite CBERS-4. Estas imagens possuem uma resolução espacial de 64 metros e temporal de 16 dias, usando a abordagem de melhor *pixel*. Foram selecionadas as bandas *RED* e *Near Infra-red* (NIR), mais os índices espectrais como *Enhanced Vegetation Index* (EVI) e *Normalized Difference Vegetation Index* (NDVI). As séries temporais deste estudo de caso possuem uma linha temporal de 13 de Agosto de 2019 à 28 de Agosto de 2020, seguindo as datas das imagens do CBERS-4.

Foi selecionada a interpolação linear de dados para o processo de regularização das séries, localizando observações tratadas como "*No Data*" e estimando valores para cobrir as datas na linha temporal. Diversas amostras tiveram a série temporal interpolada para a criação do conjunto de dados, a descrição da interpolação pode ser observada no *output* do processo em andamento na Figura 4.11.

A extração das séries seguiu as mesmas etapas descritas na Seção 3.4 e Seção 3.5 para a extração, regularização e armazenamento em *Rdata*. Todo este processo obteve uma previsão de conclusão de 2 horas e 17 minutos, conforme descrito no canto inferior direito na Figura 4.11. No entanto, o processo para 2.246 amostras, para duas bandas (RED e NIR), dois índices espectrais (EVI e NDVI) e 24 observações cada, levou uma hora, isto é, foram processadas 37 amostras por minuto. Logo após, o conjunto de dados ficou disponível para a aplicação dos métodos de análise.

4.3.2 Análise exploratória

A Figura 4.12 apresenta a interface inicial para a execução dos métodos de análise de amostras com foco na análise exploratória. Esta figura busca demonstrar a seleção para visualizar a distribuição da quantidade de amostras por classe, onde é possível escolher dentre dois tipos de gráficos, o gráfico de barras e o gráfico de pizza. Como explorado na Seção 3.6, estas duas formas de visualizar são importantes, pois pode-se analisar diferentes pontos de vista, o qual respectivamente representam a proporção e o valor numérico da quantidade. Neste estudo de caso pode-se notar que as classes que representam agricultura possuem maior quantidade de amostras mais de 75% das amostras. As classes de "Agricultura Temporária de 1 Ciclo" e "Agricultura Temporária Mais de 1 Ciclo" tem maior cobertura que as demais classes, respectivamente com 34.4% e 21.9%, ou 712 e 609 amostras.

A Figura 4.13 apresenta a seleção de série temporal por amostra dentre as opções para a análise exploratória. Nesta opção é possível escolher qual o conjunto de dados para a busca e seleção de amostras, neste caso, foi armazenado apenas um conjunto com os dados do *CBERS-4*. Nesta tela a lista de amostras é apresentada de forma tabular, onde pode-se visualizar a série temporal com a seleção de um item da lista. No exemplo foi selecionado a série temporal de uma amostra de "Agricultura Temporária Mais de 1 Ciclo" em destaque no mapa interativo. É possível também buscar pela nomenclatura da classe e pelo ID da amostra, além de selecionar todas as amostras resultantes desta busca, nesta opção é possível editar a classe de um conjunto de amostras.

A Figura 4.14 apresenta a visualização do padrão de série temporal para a classe de "Agricultura Temporária Mais de 1 Ciclo", nesta tela pode-se observar os padrões dos índices espectrais NDVI e EVI que representam as principais características e feições do conjunto de séries temporais paras esta classe, a partir do cálculo do pacote *sits*. Esta opção possui um formulário, o qual pode-se selecionar as bandas e as classes alvo, conforme o canto inferior esquerdo. A Figura 4.15 demonstra a capacidade da interface para visualizar dois resultados com a comparação dos gráficos, neste caso há a visualização da série temporal de uma amostra de "Agricultura Temporária Mais de 1 Ciclo" com o padrão da classe relacionada. Neste contexto, esta opção pode ser usada para explorar amostras que não possuem semelhanças com o padrão, podendo ser removidas ou reclassificadas.

4.3.3 Predição de probabilidades & cálculo de métricas

Foi treinado um modelo de aprendizado de máquina baseado no método *Random Forest* com 200 árvores usando o conjunto de dados do *CBERS-4* neste estudo de caso. A Figura 4.16 apresenta a predição de probabilidades usando este modelo pré-treinado para uma nova amostra não rotulada selecionada no mapa interativo do *TerraCollect*. Esta interface apresenta à esquerda, a descrição do modelo e suas opções e à direita o resultado da predição, onde pode-se visualizar as probabilidades em forma tabular, gráfico de barras, de pizza, e mais a série temporal desta nova amostra. No mapa também é possível analisar este ponto, neste caso, a coordenada (Longitude: -48.71723, Latitude: -14.30697), baseado nas séries temporais do *CBERS-4* apresentou maior probabilidade, 55.5%, de pertencer à classe de "Agricultura Semiperene". A principal ideia desta visualização é demonstrar ao usuário como um modelo está interpretando as suas amostras coletadas, é recomendado sempre questionar a resposta e não tratar esta probabilidade como verdade absoluta.

Por um lado, a predição pode ser útil na coleta por usar com base dados de amostras já coletadas. Contudo, a coleta de muitas amostras que apresentam uma alta probabilidade de serem da classe correta pode acarretar no *overfitting* do modelo. Usar o cálculo das métricas com *Active Learning* pode reduzir esse problema, pois se baseia na representatividade do conjunto. A Figura 4.17 apresenta a predição de probabilidades e o cálculo de métricas para um conjunto de amostras já rotuladas, mas ainda não foram salvas no conjunto de análise. Esta interface possui opções parecidas com a tela de predição, apresentando a lista de amostras selecionadas com as métricas em formato tabular. Seguindo os conceitos do *Active Learning* e a regra de cada métrica para ranquear, pode-se notar que neste exemplo as amostras Sample # (6992) e Sample # (6998) apresentam maior entropia, ou maior confusão na predição, sendo as melhores candidatas a entrar no conjunto de dados.

4.3.4 Detecção de ruído amostral

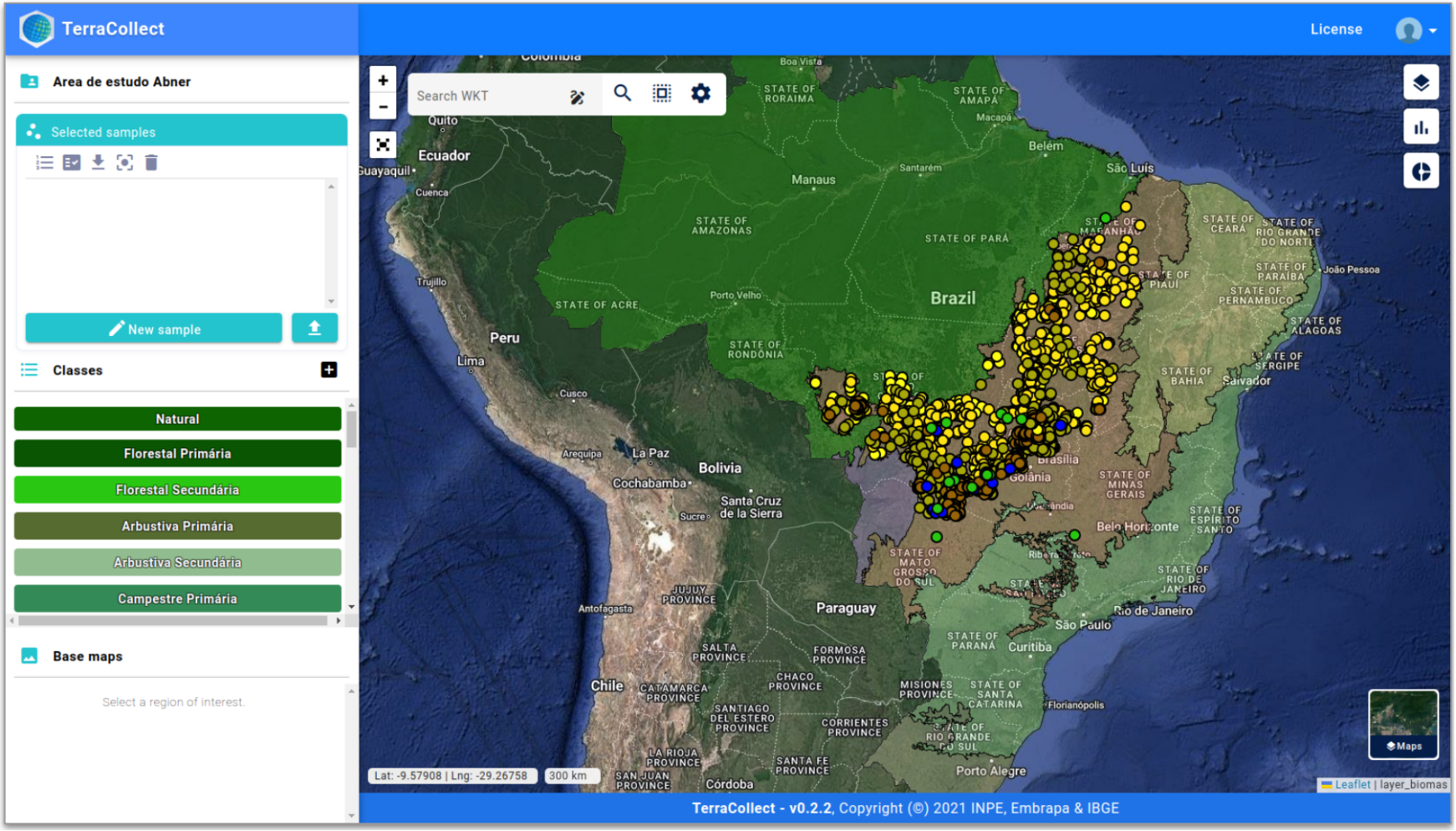
Para testar a implementação do método apresentado por Santos et al. (2021a), foi gerado um agrupamento SOM com grade 10x10 usando a distância euclidiana, selecionando todas as amostras do conjunto do estudo de caso. A Figura 4.18 apresenta o resultado deste agrupamento, nesta interface a descrição dos parâmetros do SOM e do conjunto base usado está à esquerda e o mapa SOM interativo à direita. Neste resultado, há a seleção de um neurônio considerado confuso ou atípico pois sua classe majoritária, "Agricultura Semiperene", difere da de seus vizinhos, sugerindo uma certa variabilidade ou amostras com rótulos incorretos.

Para investigar esta variabilidade e amostras consideradas ruidosas, há três opções principais para o agrupamento SOM que foi armazenado, conforme ilustrado pela Figura 4.19. Há a visualização interativa do SOM com seleção de neurônios, amostras agrupadas e o vetor de peso, a visualização da confusão no agrupamento e a distribuição da quantidade de amostras por *status* para cada classe. Nesta figura há a ilustração destas visualizações para o estudo de caso, onde pode-se observar no gráfico de confusão que a classe mais confusa foi a de "Corpos D'água" e "Formação Florestal". Estas duas classes apresentam menor número de amostras, um número insuficiente para capturar suas feições principais, o que não ocorre para as classes de Agricultura e Silvicultura.

Na avaliação de qualidade, o gráfico de distribuição da quantidade de amostras por *status* para cada classe é gerado após a aplicação dos filtros de *threshold* nas probabilidade *a priori* e *posteriori*. No formulário para a aplicação destes filtros foi selecionado o valor 0.6 tanto para a probabilidade *a priori* quanto para a *posteriori*. Após isso, as amostras agrupadas com o SOM foram sinalizadas com "*clean*", "*analyze*" e "*remove*", conforme Seção 2.5.3. A Figura 4.20 apresenta o resultado desta avaliação, com a lista de amostras contendo a sinalização. Neste resultado, observa-se que a classe de "Agricultura Temporária Mais de 1 Ciclo" apresentou mais amostras ruidosas, sinalizadas com "*remove*" em comparação à quantidade de amostras com "*clean*". As demais classes "Corpos D'água" e "Formação Florestal" podem ser desconsideradas nesta avaliação, pois possuem baixo número de amostras e possuirão maior número de amostras ruidosas.

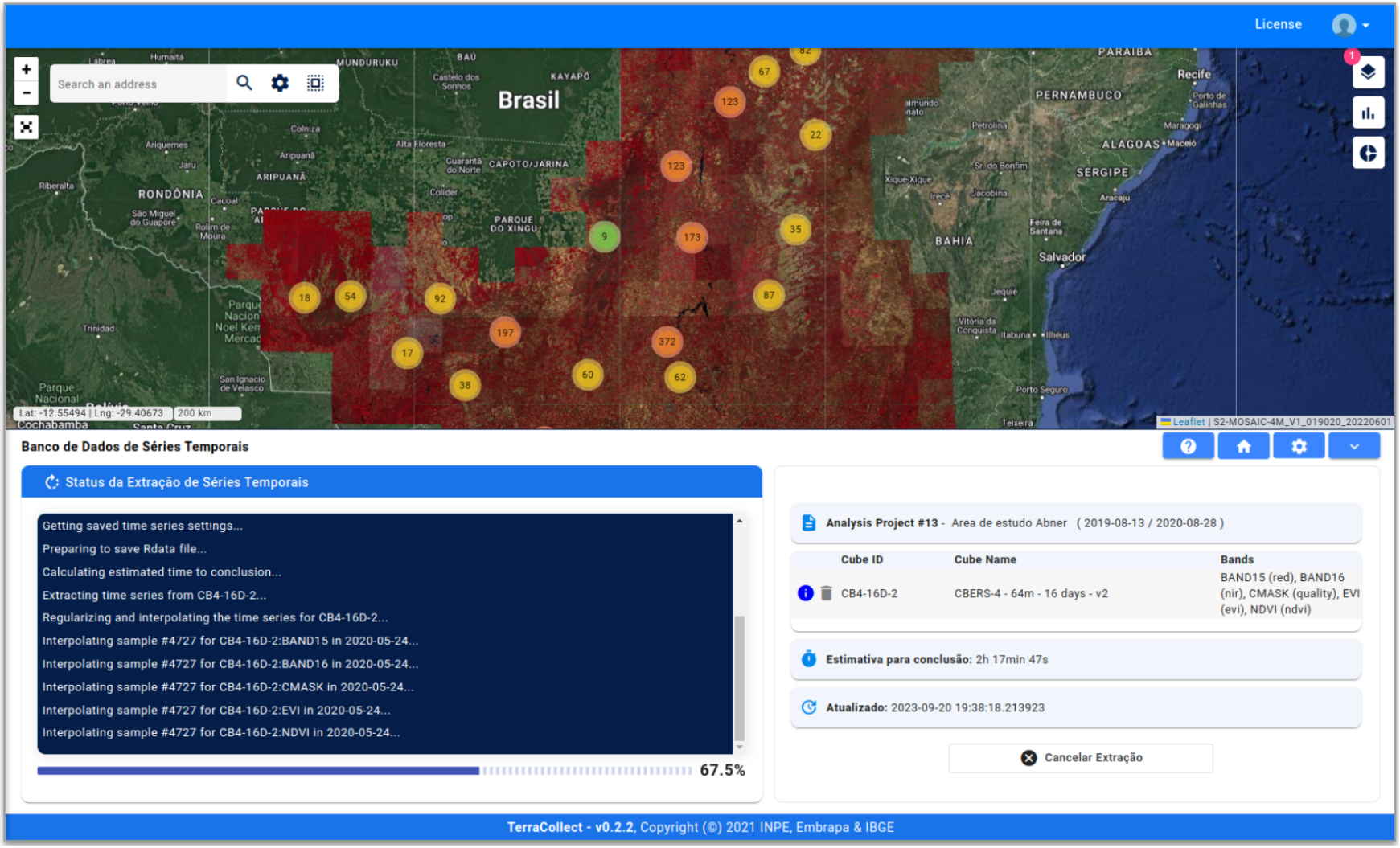
No exemplo, as duas opções para o resultado da avaliação de qualidade foram selecionadas, a visualização do gráfico de distribuição do *status* e a visualização dos *status* com as probabilidades no mapa. As opções para esta avaliação seguem o mesmo princípio da análise exploratória, onde há uma lista de amostras em formato tabular para a seleção da série temporal por item, mais uma barra de busca por *status*, ID e classe. Com a visualização no mapa, observa-se a distribuição geográfica das amostras sinalizadas com *clean*, *analyse* e *remove*. Na lista em formato tabular, ao selecionar uma amostra, a mesma recebe um foco no mapa interativo e sua série temporal pode ser analisada, conforme Figura 4.21. Nesta interface é possível selecionar todas as amostras com *remove* de uma classe, no exemplo foram selecionadas todas as amostras sinalizadas como *remove*.

Figura 4.10 - Área de interesse do estudo de caso no projeto criado *TerraCollect*.



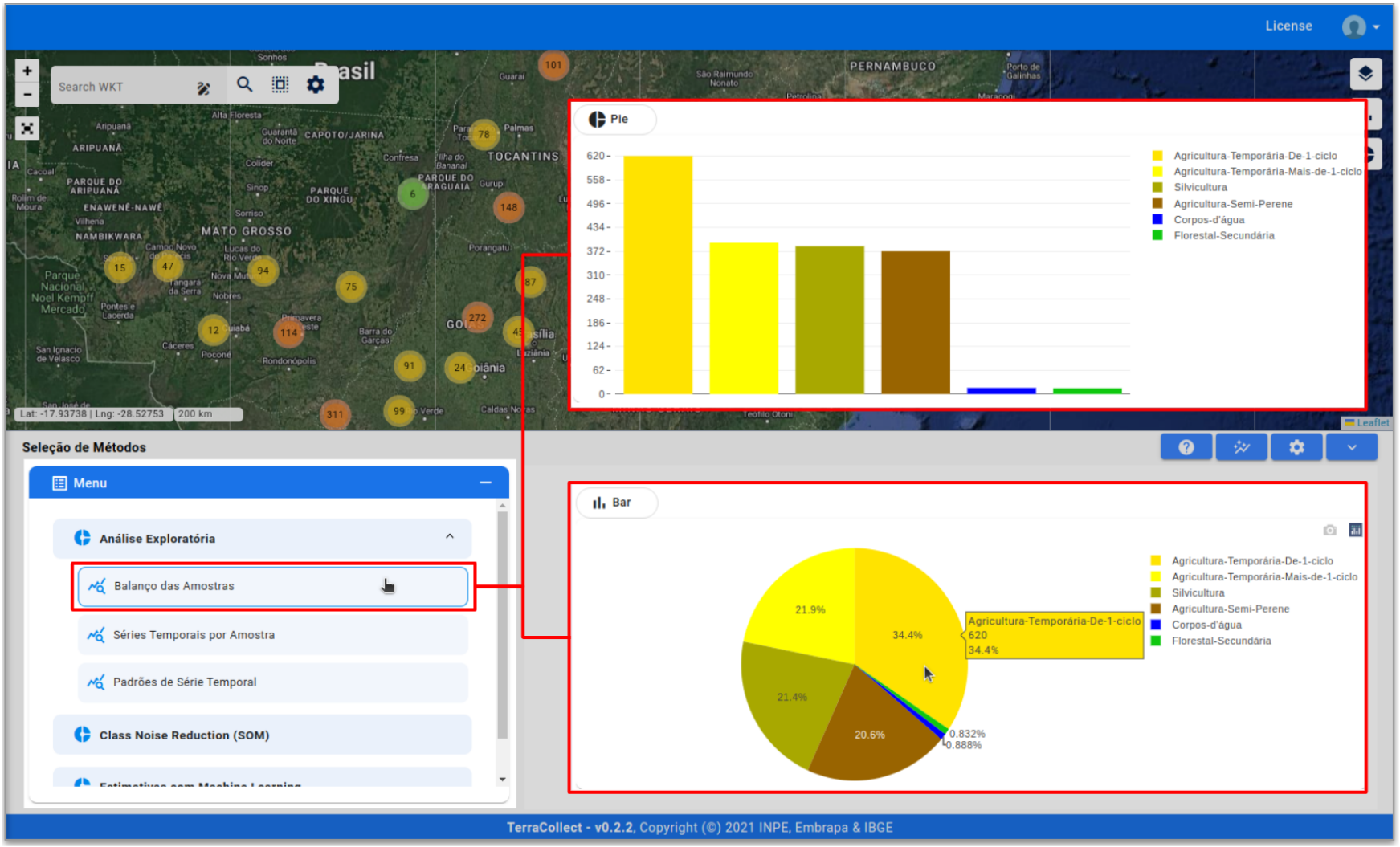
Fonte: Produção do autor.

Figura 4.11 - Extração de séries temporais de imagens para as amostras do estudo de caso no projeto *TerraCollect*.



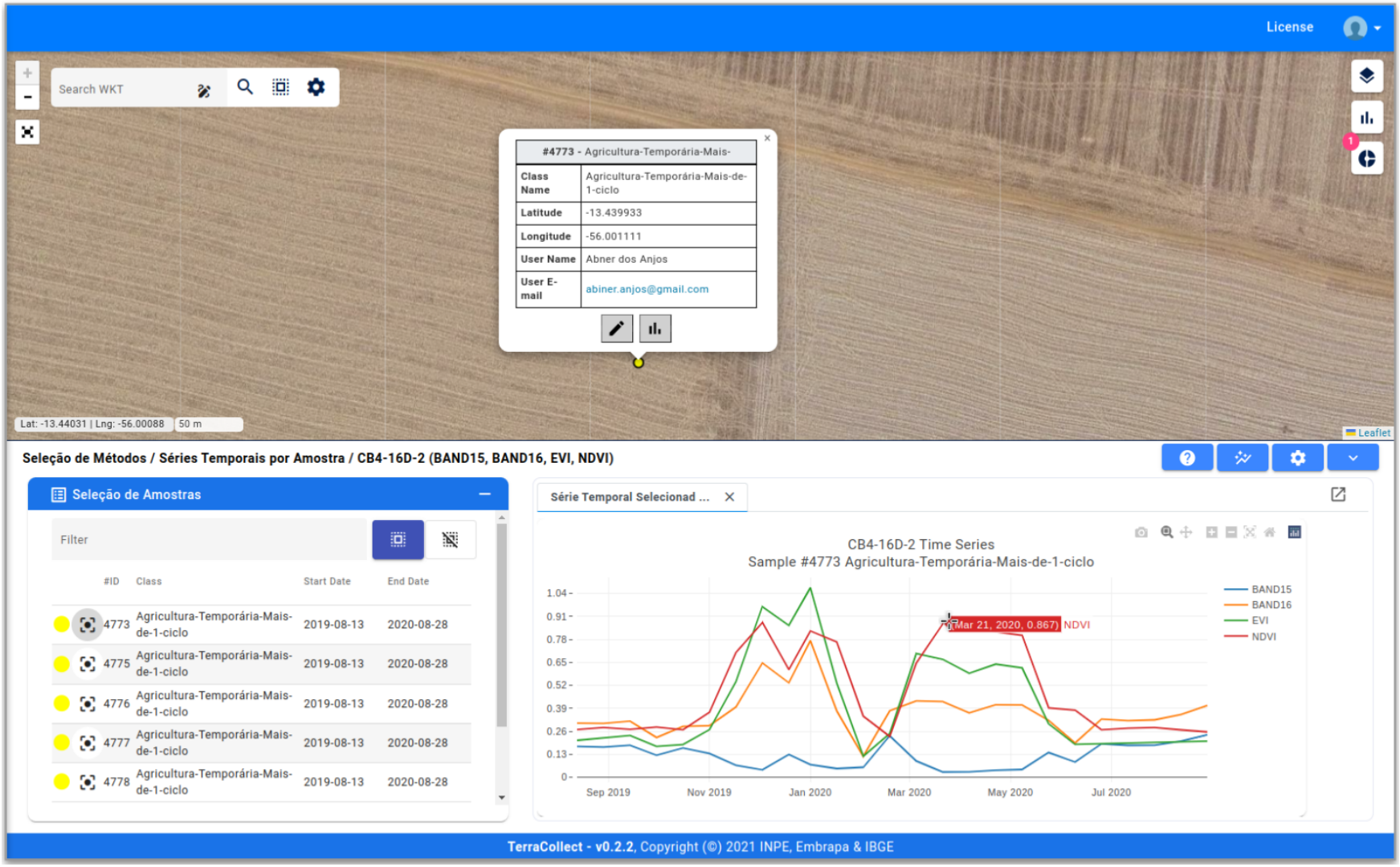
Fonte: Produção do autor.

Figura 4.12 - Seleção de gráficos para o balanço das amostras do estudo de caso no projeto *TerraCollect*.



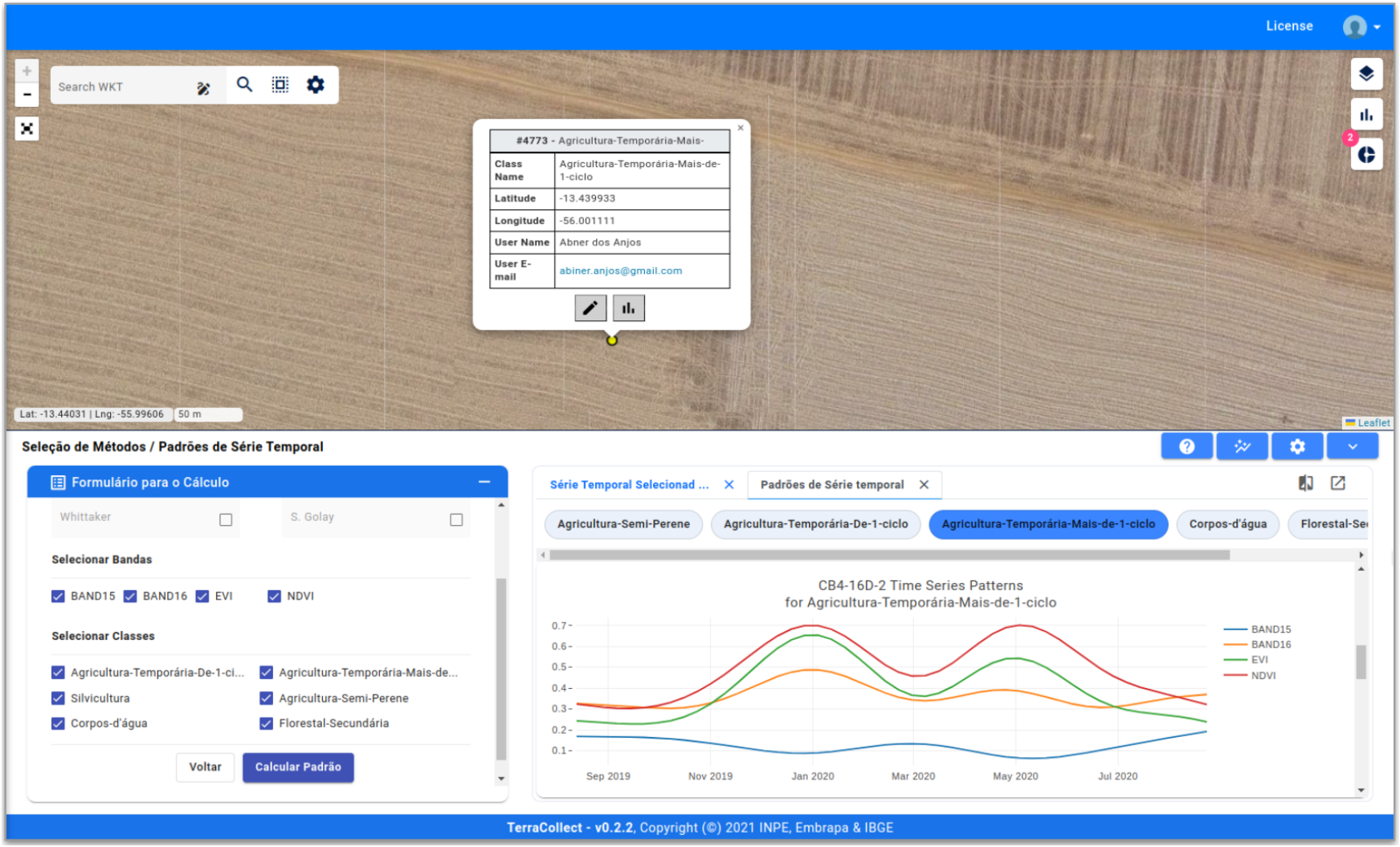
Fonte: Produção do autor.

Figura 4.13 - Seleção de série temporal para uma amostra do conjunto de dados do estudo de caso.



Fonte: Produção do autor.

Figura 4.14 - Visualização dos padrões de série temporal para as classes de uso e cobertura da terra do estudo de caso.

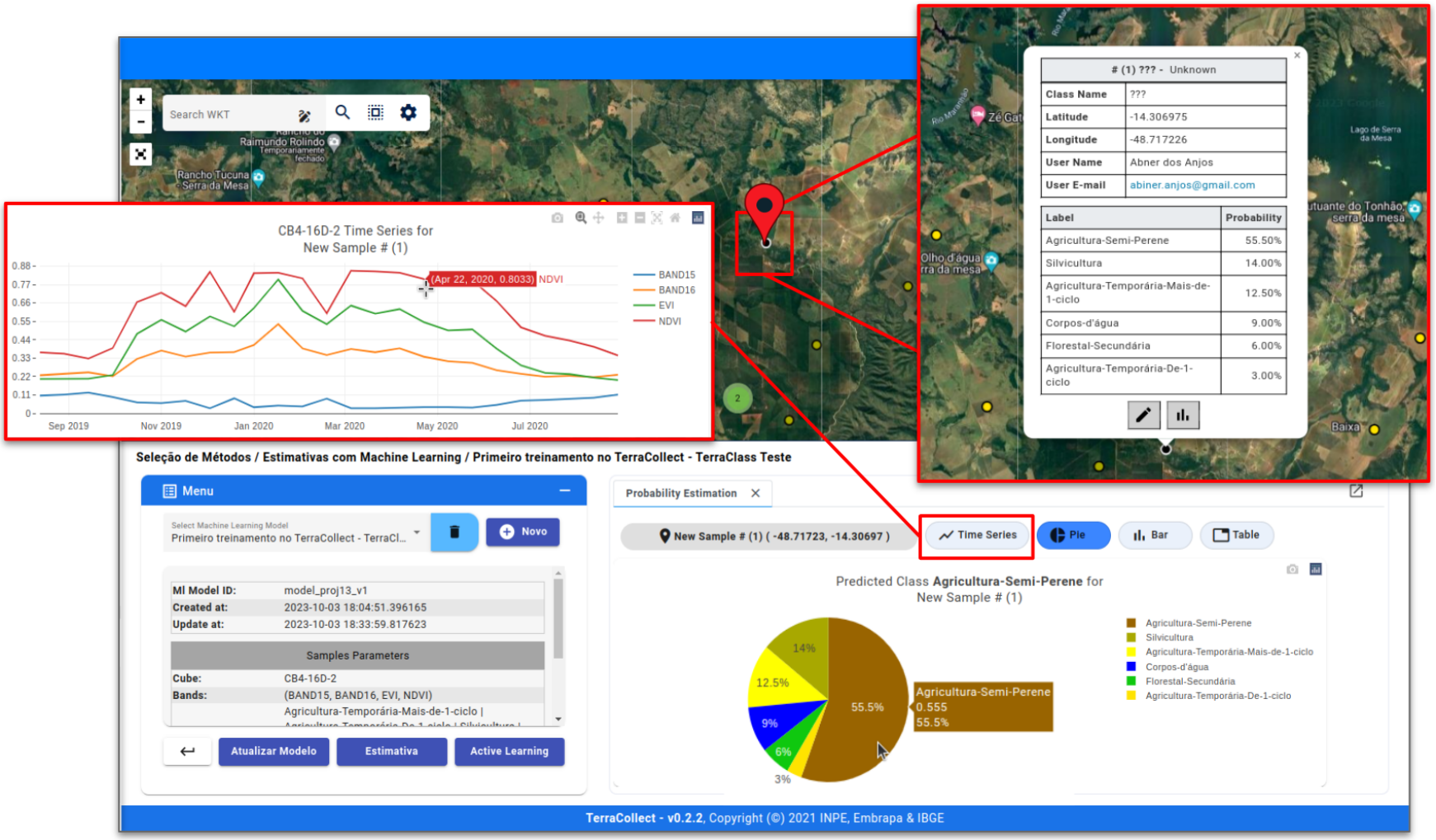


Fonte: Produção do autor.

Figura 4.15 - Comparação de uma série temporal com o Padrão da Classe com base no conjunto de dados do estudo de caso.

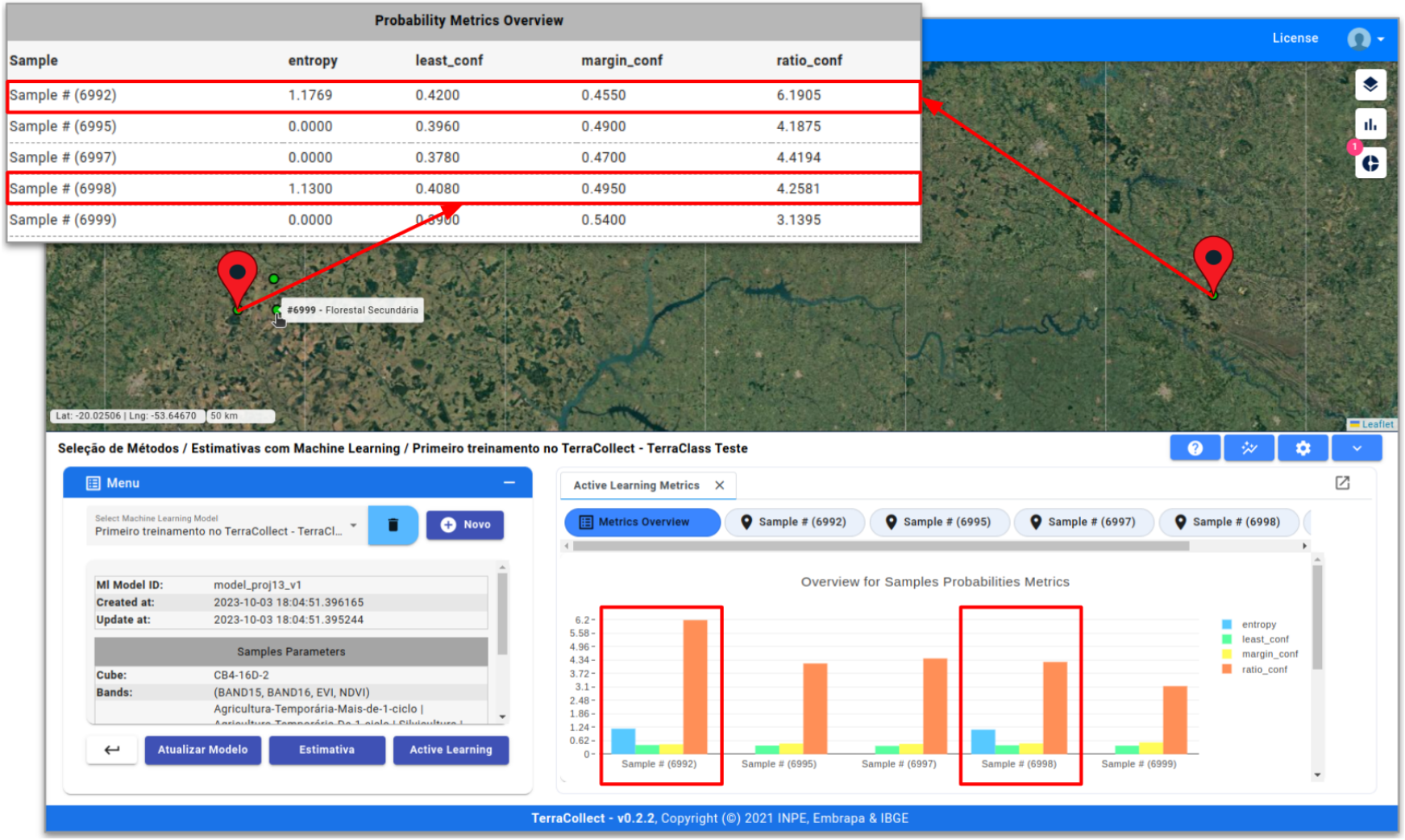
Fonte: Produção do autor.

Figura 4.16 - Predição de probabilidades para uma nova amostra não-rotulada coletada para o estudo de caso.



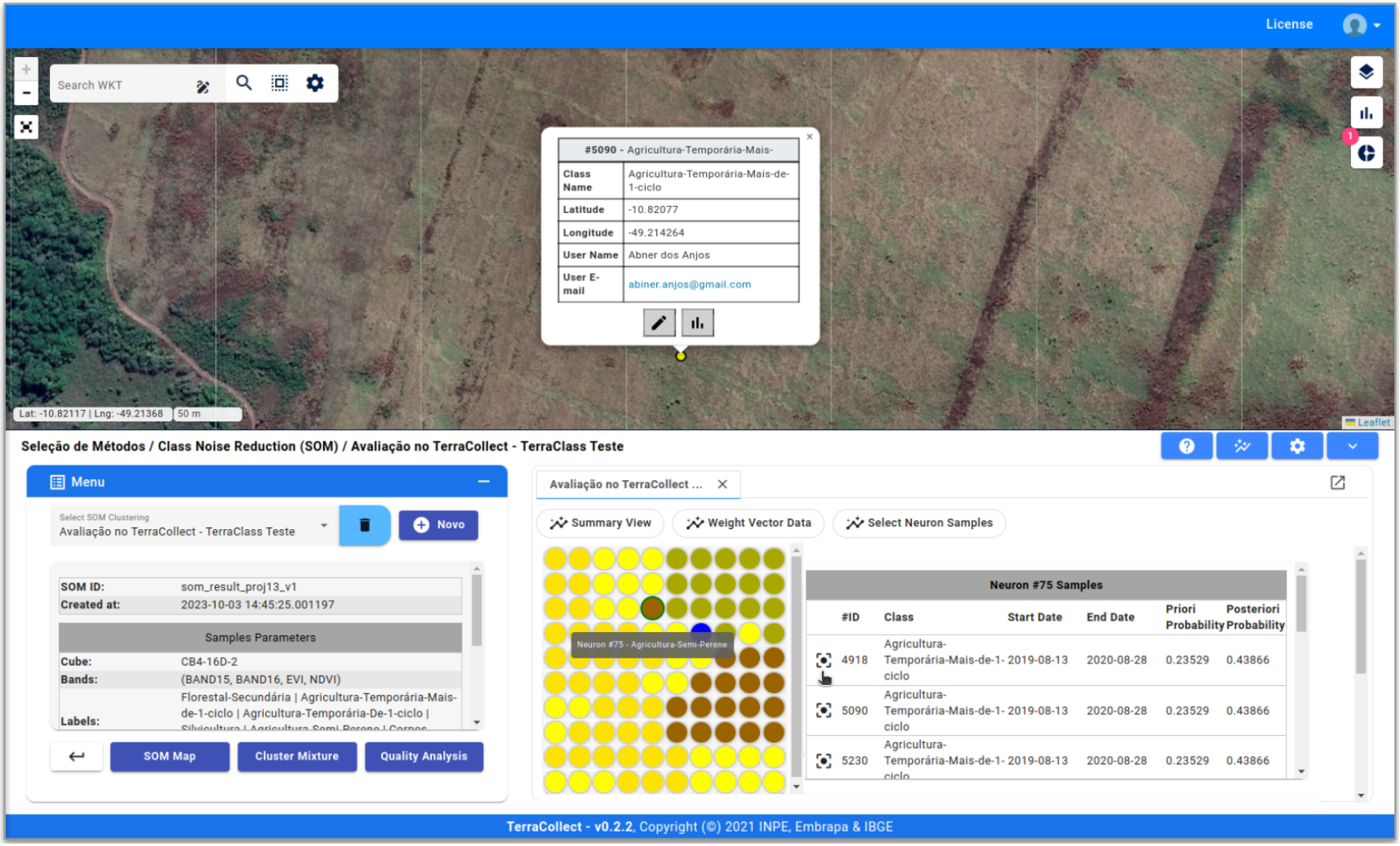
Fonte: Produção do autor.

Figura 4.17 - Resultado do cálculo de métricas *Active Learning* para as amostras de Floresta do estudo de caso.



Fonte: Produção do autor.

Figura 4.18 - Visualização do Mapa SOM com seleção de amostras agrupadas em um neurônio com os metadados do agrupamento.

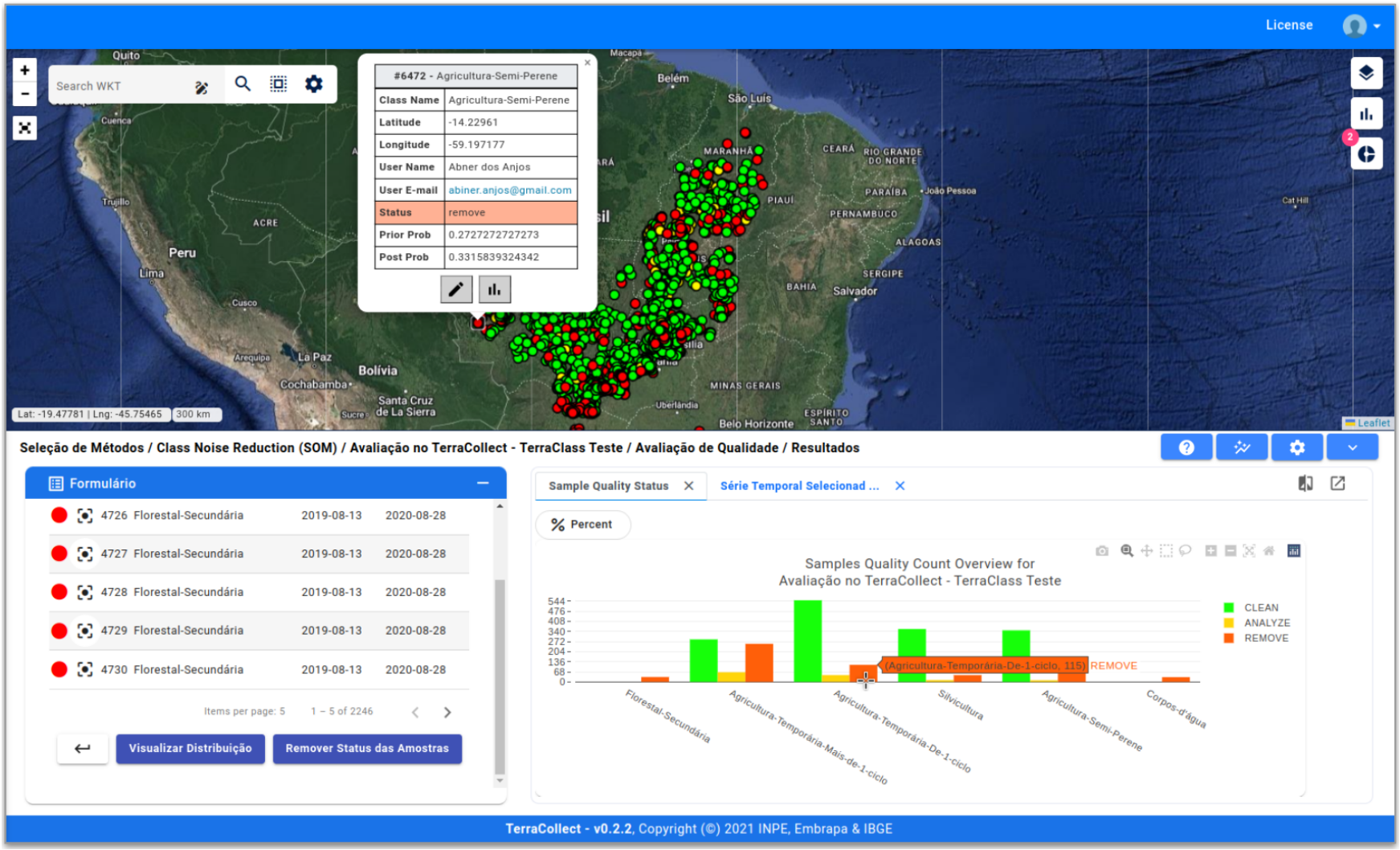


Fonte: Produção do autor.

Figura 4.19 - Opções disponíveis para a visualização no menu de seleção dos Mapas SOM armazenados no projeto *TerraCollect*.

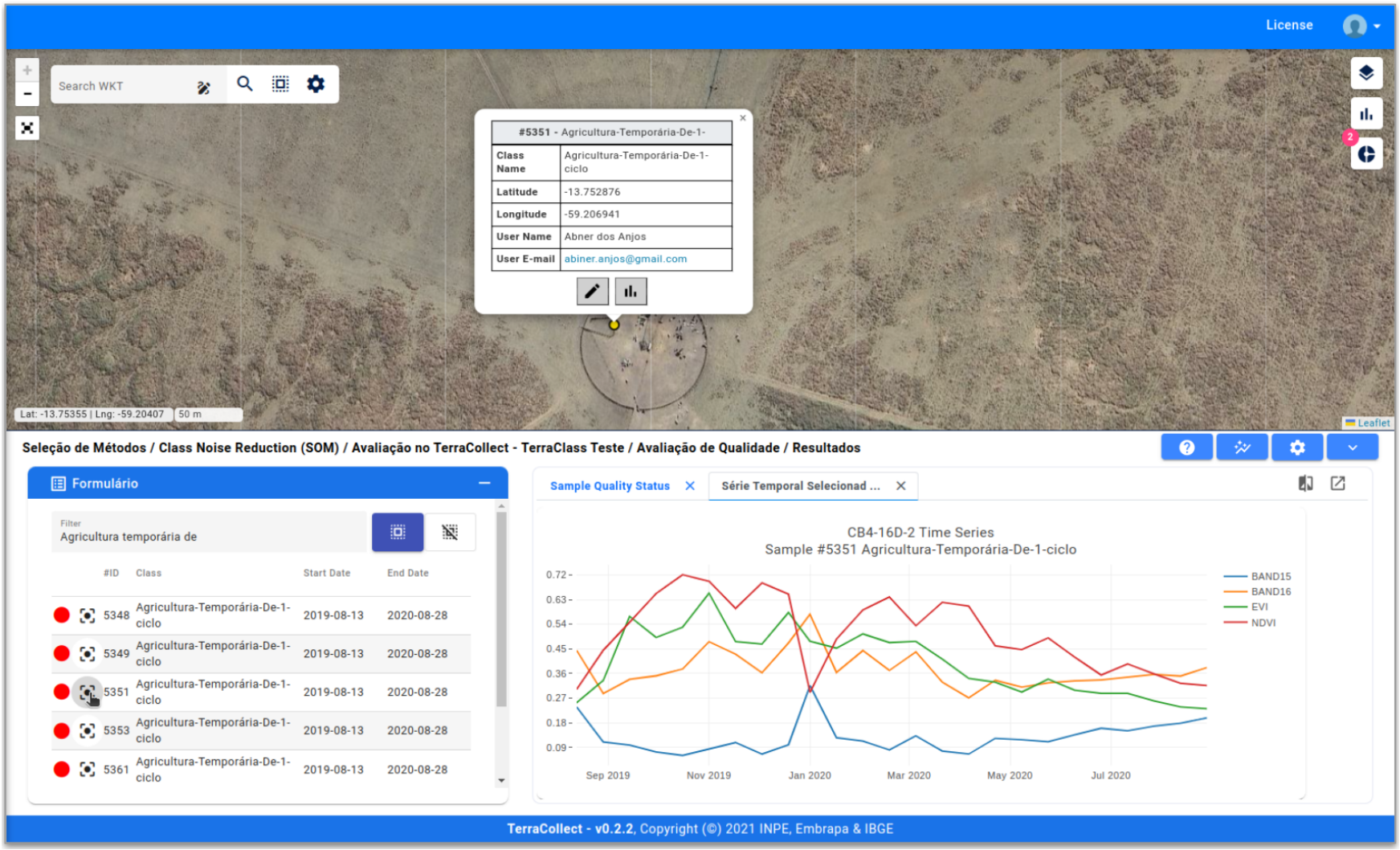
Fonte: Produção do autor.

Figura 4.20 - Visualização da distribuição de amostras para o *status* da qualidade com base das probabilidades resultantes do Mapa SOM.



Fonte: Produção do autor.

Figura 4.21 - Seleção de amostra com a série temporal com base no resultado da avaliação de qualidade para o estudo de caso.



Fonte: Produção do autor.

4.4 Discussão dos resultados

As seções anteriores demonstraram de forma breve o processo de análise e avaliação da qualidade de amostras para o estudo de caso TerraClass com foco no bioma Cerrado. Para a comparação e discussão dos resultados, foi feito um teste com uma classificação e validação simples usando *K-fold* com o conjunto de amostras sem análise (S) e com a análise (C). Foi feito o *download* das amostras logo após a extração de séries temporais do *CBERS-4* para o conjunto sem análise. Depois da análise, amostras foram removidas de acordo com os resultados da análise exploratória, comparação com os padrões de classe, predição de probabilidades, cálculo de métricas e avaliação de qualidade com o SOM, e foi feito o *download* das amostras restantes, representando o conjunto com análise.

A Tabela 4.2 apresenta a comparação da distribuição de amostras por classe antes e depois da análise. A coluna "S" representa a distribuição no conjunto de dados sem a análise (S), a coluna "C" com análise (C) e a coluna "D" o número de amostras removidas de cada classe (D). O conjunto S possui 2.246 amostras e apresenta os padrões de série temporal apresentados na Figura 4.22. O conjunto C possui 1.801 amostras e apresenta os padrões de série temporal apresentados na Figura 4.23.

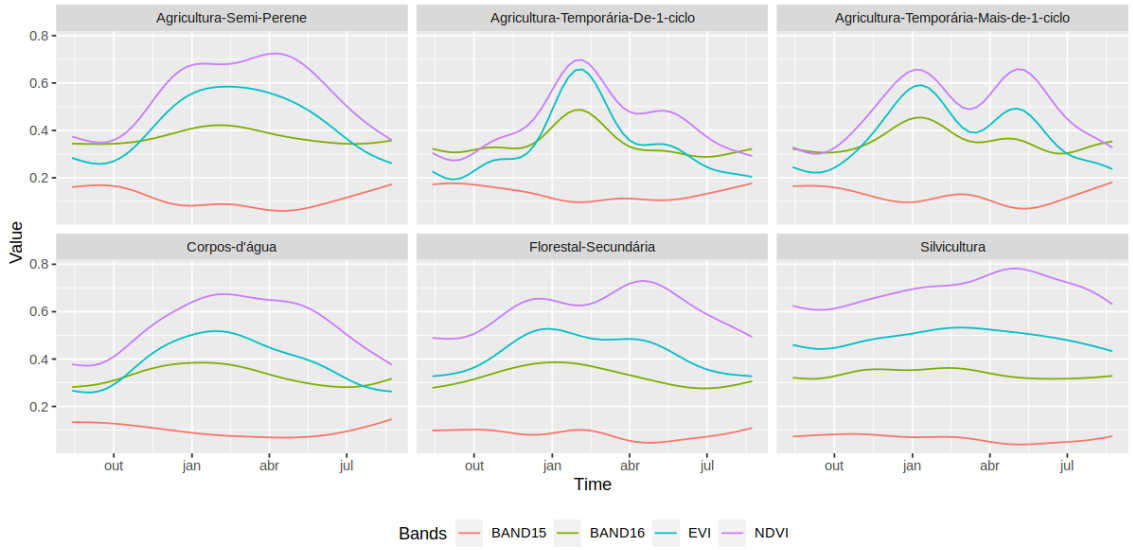
Tabela 4.2 - Distribuição das amostras por classe sem análise e com análise para o estudo de caso *TerraCollect*.

Classe LULC	**S**	**C**	**D**
Agricultura Temporária de 1 Ciclo	712	620	-92
Agricultura Temporária de Mais de 1 Ciclo	609	394	-215
Agricultura Semiperene	445	371	-74
Silvicultura	415	385	-30
Formação Florestal	33	15	-18
Corpos D'água	32	16	-16
Total	**2.246**	**1.801**	**-445**

(S) Conjunto de amostras sem análise; (C) Conjunto de amostras com análise; (D) Número de amostras removidas de cada classe depois da análise.

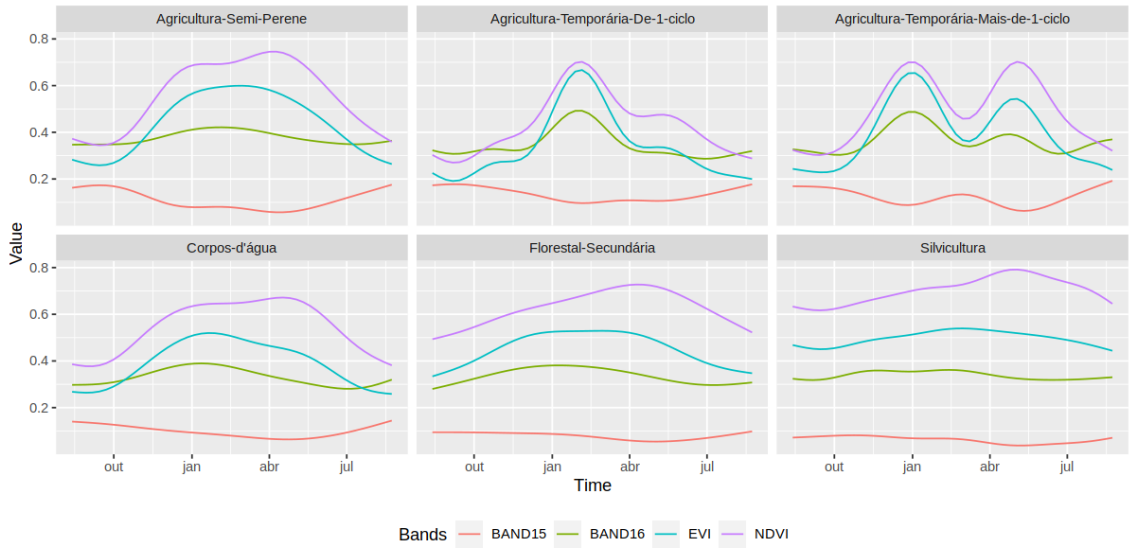
Fonte: Produção do autor.

Figura 4.22 - Padrões de série temporal para o conjunto de dados sem a etapa da análise no estudo de caso.



Fonte: Produção do autor.

Figura 4.23 - Padrões de série temporal para o conjunto de dados com a análise no estudo de caso.



Fonte: Produção do autor.

Ao todo foram descartadas 445 amostras depois da análise, a classe “Agricultura Temporária Mais de 1 Ciclo” teve mais amostras ruidosas que foram removidas. Usando como base a análise do controle de qualidade e redução de ruído, por se tratar de uma classe que possui um padrão bem definido, as demais variações na série temporal desta mesma classe podem ter sido consideradas ruidosas durando a aplicação do método. Pode-se notar que esta classe no conjunto de dados com análise, possui o padrão de série temporal com curvas mais acentuadas como picos e vales (Figura 4.23), no conjunto sem análise as linhas são mais suaves (Figura 4.22). Nas demais classes LULC não há considerável mudança nos padrões do conjunto S para o conjunto C, apenas algumas linhas mais acentuadas e uma leve suavização. No caso desta classe de agricultura, pode ter sido uma questão de limitação do dado, do conjunto de observações, da resolução espacial e temporal, não sendo suficientes para identificar variações na mesma classe.

Para os dois casos sem análise e com análise dos conjuntos S e C, foi definido um modelo de aprendizado de máquina *Random Forest* com 100 árvores, apenas para fins de teste. Nos dois casos 70% das amostras foram definidas para treinamento e o restante 30% para validação e teste. A Tabela 4.3 apresenta a comparação dos resultados para a validação dos dois modelos gerados com o treinamento em cada caso. O modelo com o conjunto S obteve uma acurácia de 0.88, já a mesma definição do modelo para o conjunto C obteve a acurácia de 0.94.

A classificação com o conjunto C, que passou por etapas como a análise exploratória e a remoção de ruído na avaliação de qualidade com o SOM, obteve uma acurácia maior e precisa segundo o intervalo de confiança. Foram removidas 445 amostras consideradas ruidosas do conjunto original para gerar um conjunto de qualidade. A proposta da análise de amostras é a identificação e remoção de ruído, obtendo um conjunto limpo com representação e qualidade, sem prejudicar a acurácia no resultado final ao descartar certas amostras consideradas imprecisas.

Tabela 4.3 - Resultados da validação dos modelos de aprendizado de máquina treinados usando cada um dos dois conjuntos com e sem análise.

Conjunto de Amostras	**Acurácia**	***Kappa***	**Intervalo de Confiança**
(S) sem análise	0.8735	0.8299	(0.8462, 0.8976)
(C) com análise	0.9415	0.9215	(0.9184, 0.9596)

Fonte: Produção do autor.

5 CONCLUSÕES E TRABALHOS FUTUROS

O presente trabalho discursou sobre o desenvolvimento de uma arquitetura para a integração de métodos de análise de amostras baseados em séries temporais durante a coleta de amostras LULC. Esta arquitetura foi implementada com uma extensão das funcionalidades da plataforma de coleta *TerraCollect*. A implementação da arquitetura compreende uma ferramenta para a produção de amostras de boa qualidade unindo a coleta com a análise de dados do pacote *sits* dentro da plataforma *TerraCollect*. A ferramenta busca contribuir na geração de amostras e produção de mapas LULC além de auxiliar os pesquisadores que não possuem conhecimentos de algoritmos em linguagem de programação para aplicar os métodos em um interface gráfica, abstraindo desde a extração à análise de séries temporais.

A dinâmica da implementação da arquitetura foi baseada na estrutura cliente-servidor, onde há uma interface gráfica na aplicação *Angular* do *TerraCollect* enviando e consumindo dados de um serviço *web* com uma API em *Plumber R*. Esta estrutura possibilita o uso otimizado dos serviços auxiliares como Sample-WS, STAC, WTSS e WLTS e o pacote *sits*. Estes serviços fornecidos pelo projeto BDC auxiliam ao gerenciar e manipular *EO Data*, permitindo o gerenciamento de usuários e projetos com serviços de autenticação, busca e leitura no banco de dados de amostras.

Os métodos para análise de amostras com séries temporais foram implementados seguindo a especificação do *sits* em linguagem de programação R. A API com o pacote *Plumber R* encapsula as funções em R do *sits* nos métodos *GET*, *POST* e *DELETE* em um serviço na *web*. O *Plumber R* possibilita o acesso ao ambiente R por meio de solicitações HTTP. Desta forma o *TerraColect* envia as requisições e recebe os resultados da aplicação dos métodos. Neste componente foram implementados as abordagens para: análise exploratória de amostras LULC (WICKHAM; GROLEMUND, 2017); análise de métricas com aprendizado de máquina e técnicas de *Active Learning* (TUIA et al., 2009); e por fim o controle de qualidade e redução do ruído de classe (SANTOS et al., 2021a).

Para a integração dos métodos de análise foi necessário implementar e adaptar as funções e algoritmos do pacote *sits* para extrair, regularizar, armazenar e gerenciar dados de séries temporais usando arquivos *Rdata* com processamento paralelo em segundo plano. O formato de arquivo *Rdata* foi usado para armazenar objetos específicos do R como o formato *sits* para amostras em conjunto com séries temporais, modelos de aprendizado de máquina e agrupamentos SOM.

As séries temporais são armazenadas pois devem estar disponíveis a cada requisição e há um alto tempo de espera para o processo de extração e tratamento dos dados antes da aplicação nos métodos. A estratégia com processamento paralelo, mesmo otimizando o pré-tratamento das séries, ainda requer atenção e revisão. Não foram realizados testes exaustivos para viabilizá-la, como por exemplo, a avaliação de pior caso. É preciso revisar os modelos lógicos usados nos arquivos, avaliar o poder de processamento necessário para estas operações.

Após o armazenamento das séries temporais as opções de aplicação dos métodos ficam disponíveis na interface gráfica. A interface disponibiliza os meios de visualização dos dados e objetos salvos em *Rdata* de forma interativa e colaborativa permitindo o compartilhamento dos resultados obtidos. Resume-se uma ferramenta na *web* capaz de gerar gráficos como padrões de séries temporais, distribuição de amostras por classe e gerar resultados com base em treinamento de modelos e agrupamentos SOM. A visualização é um aspecto importante para a análise exploratória, neste contexto, permite que o usuário questione a veracidade de amostras LULC de forma mais eficiente. É possível identificar padrões, detectar anomalias, verificar se tais dados atendem aos objetivos esperados e se necessitam de ajustes.

Ademais, o presente trabalho discutiu a viabilidade das técnicas de aprendizado de máquina aplicado à geração e análise de amostras LULC. Esta abordagem baseia-se nos métodos semi-supervisionados e *Active Learning* explorados por Tuia et al. (2011) que realizou uma pesquisa abordando seus principais aspectos como benefícios e desvantagens. Os experimentos e testes para a implementação destes métodos na arquitetura, no contexto de uma abordagem de análise de amostras, ainda não permitiram um estudo mais elaborado. Além do mais, deve-se avaliar o suporte da API para o treinamento de modelos mais robustos baseados em *SVM* e *TempCNN*.

Ainda não foi comprovada a eficácia das técnicas de *Active Learning* em um ambiente de coleta com pesquisadores e amostras reais. É necessário realizar mais experimentos especializados no cálculo e análise das métricas de probabilidades para demonstrar a sua viabilidade no refinamento em um conjunto de amostras ruidosas. Os testes apresentados neste trabalho exploraram isoladamente a operação de *query*, o que é apenas uma parcela do contexto geral do *Active Learning*.

Futuramente, espera-se que as operações com aprendizado de máquina possam permitir a geração automática de amostras. Esta automação já está sendo trabalhada no grupo de desenvolvimento do pacote *sits* e, em breve, esta opção poderá ser adicionada ao *TerraCollect*. Onde o usuário fornecerá a área de estudo e o *design*

amostral desejado para gerar amostras e depois refiná-las sem sair da plataforma. Após os testes com uma API servindo as funções do *sits* em um serviço *web*, a classificação de imagens de satélite é algo que também poderá ser adicionado a este conjunto de ferramentas. Além dessas possibilidades, o fato de servir as operações e métodos fornecidos pelo *sits* em um serviço *web*, permite não só a integração nesta plataforma de coleta, mas em outros aplicativos relacionados a visualização de dados geoespaciais, como por exemplo o *software* QGIS.

O contexto para o agrupamento SOM do método de controle de qualidade e redução de ruído de classe carece de mais estudos de casos para o seu *design* na interface, pois como é um método relativamente inovador, ainda não há referências. Neste caso, o *design* para este método na interface gráfica do *TerraCollect* foi estruturado em etapas como primeiro o agrupamento SOM com a inferência *bayesiana* e depois a avaliação de qualidade, com visualizações disponíveis durante todo o processo. Mas são necessárias etapas cíclicas de desenvolvimento de *software* e o estudo da experiência do usuário na plataforma. É preciso disponibilizar uma interface básica para o método a um grupo de usuários-alvo e coletar críticas construtivas.

O *TerraCollect* é uma ferramenta em desenvolvimento, por isso, ainda pode sofrer alterações conforme demandas de usuários. A versão da extensão de análise apresentada nos resultados possui uma instância nos servidores de desenvolvimento do BDC. Esta versão tem os métodos para extração de séries temporais, análise exploratória, controle de qualidade e uma pequena parte da predição de probabilidades e cálculo de métricas com *Active Learning*. O *TerraCollect* em conjunto com a extensão de análise, permite que os usuários visualizem os resultados de forma interativa em um painel de controle dinâmico, identificando amostras com baixa qualidade e pouco representativas que causam problemas na classificação de imagens.

REFERÊNCIAS BIBLIOGRÁFICAS

ALMEIDA, C. A. d.; COUTINHO, A. C.; ESQUERDO, J. C. D. M.; ADAMI, M.; VENTURIERI, A.; DINIZ, C. G.; DESSAY, N.; DURIEUX, L.; GOMES, A. R. High spatial resolution land use and land cover mapping of the Brazilian Legal Amazon in 2008 using Landsat-5 TM and MODIS data. **Acta Amazonia**, v. 46, n. 3, p. 291 – 302, set. 2016. 1

ANANDHI, A.; DOUGLAS-MANKIN, K. R.; SRIVASTAVA, P.; AIKEN, R. M.; SENAY, G.; LEUNG, L. R.; CHAUBEY, I. DPSIR-ESA vulnerability assessment (DEVA) framework: synthesis, foundational overview, and expert case studies. **Transactions of the ASABE**, v. 63, n. 3, p. 741–752, 2020. 7

ANJOS, A. E.; FERREIRA, K. R.; QUEIROZ, G. R.; ZIOTI, F.; SANSIGOLO, G. Integrating analysis methods during land use and land cover sampling. In: BRAZILIAN SYNPOSIUM ON GEOINFORMATICS, 2022. **Proceedings...** São José dos Campos: INPE, 2022. p. 247–252. Disponível em: <<http://mtc-m21d.sid.inpe.br/rep/8JMKD3MGP3W34T/488BPNL>>. 4, 6, 41

ATZBERGER, C.; EILERS, P. H. C. Evaluating the effectiveness of smoothing algorithms in the absence of ground reference measurements. **International Journal of Remote Sensing**, v. 32, n. 13, p. 3689–3709, 2011. Disponível em: <<https://doi.org/10.1080/01431161003762405>>. 62

BELGIU, M.; BIJKER, W.; CSILLIK, O.; STEIN, A. Phenology-based sample generation for supervised crop type classification. **ISPRS Journal of Photogrammetry and Remote Sensing**, v. 95, nov. 2020. 14, 15, 20, 25

BEY, A.; DIAZ, A. S.-P.; PEKKARINEN, A.; PATRIARCA, C.; MANIATIS, D.; WEIL, D.; MOLLICONE, D.; MARCHI, G.; NISKALA, J.; REZENDE, M.; RICCI, S. **Collect Earth user manual**. [s.n.], 2015. Disponível em: <<https://openforis.org/>>. 4, 32, 37

BEY, A.; DíAZ, A. S.-P.; MANIATIS, D.; MARCHI, G.; MOLLICONE, D.; RICCI, S.; BASTIN, J.-F.; MOORE, R.; FEDERICI, S.; REZENDE, M.; PATRIARCA, C.; TURIA, R.; GAMOGA, G.; ABE, H.; KAIDONG, E.; MICELI, G. Collect Earth: land use and land cover assessment through augmented visual interpretation. **MDPI**, v. 8, n. 10, p. 807, 2016. 5, 32, 38

BRASIL, MINISTRO DO MEIO AMBIENTE. **Cerrado**. 2022. Disponível em: <<https://www.gov.br/mma/pt-br/assuntos/ecossistemas-1/biomas/cerrado>>. Acesso em: 1 Nov. 2023. 31

BRAZIL DATA CUBE (BDC). **Brazil Data Cube**. 2022. Disponível em: <<http://brazildatacube.org/>>. Acesso em: 17 Out. 2023. 13

BRIAN, J.; TATEISHI, R.; XIE, X. Using geographically weighted variables for image classification. **Remote Sensing Letters**, v. 3, n. 6, p. 491–499, 2011. 20

CAL-PSE. **Collect Earth online TerraBio public project**. 2023. Disponível em: <<https://app.collect.earth>>. Acesso em: 22 set. 2023. 39

CAMARA, G.; ASSIS, L. F.; RIBEIRO, G.; FERREIRA, K. R.; LLAPA, E.; VINHAS, L. Big earth observation data analytics: matching requirements to system architectures. In: ACM SIGSPATIAL INTERNATIONAL WORKSHOP ON ANALYTICS FOR BIG GEOSPATIAL DATA. **Proceedings...** ACM, 2016. p. 1–6. 16, 17

CHANG, W.; CHENG, J.; ALLAIRE, J.; SIEVERT, C.; SCHLOERKE, B.; XIE, Y.; ALLEN, J.; MCPHERSON, J.; DIPERT, A.; BORGES, B. **shiny : web application framework for R**. [s.n.], 2021. R package version 1.6.0. Disponível em: <<https://CRAN.R-project.org/package=shiny>>. 44, 45

CHAPELLE, O.; SCHöLKOPF, B.; ZIEN, A. **Semi-supervised learning (adaptive computation and machine learning)**. [S.l.]: The MIT Press, 2006. ISBN 0262033585. 25

CHEN, J.; JöNSSON, P.; TAMURA, M.; GU, Z.; MATSUSHITA, B.; EKLUNDH, L. A simple method for reconstructing a high-quality ndvi time-series data set based on the savitzky–golay filter. **Remote Sensing of Environment**, v. 91, n. 3, p. 332–344, 2004. ISSN 0034-4257. Disponível em: <<https://www.sciencedirect.com/science/article/pii/S003442570400080X>>. 62

CUARTERO, A.; PAOLETTI, M. E.; GARCíA-RODRIGUEZ., P.; HAUT, J. M. QCircularStats: a QGIS-plugin for evaluation bidimensional data by circular statistics. **IEEE Access**, p. 1–1, 2023. ISSN 2169-3536. 34

ESTES, L.; CHEN, P.; DEBATS, S.; EVANS, T.; FERREIRA, S.; KUEMMERLE, T.; RAGAZZO, G.; SHEFFIELD, J.; WOLF, A.; WOOD, E.; CAYLOR, K. A large-area, spatially continuous assessment of land cover map error and its impact on downstream analyses. **Global Change Biology**, v. 24, 09 2017. 7

FERREIRA, K. R.; QUEIROZ, G. R.; VINHAS, L.; MARUJO, R. F. B.; SIMOES, R. E. O.; PICOLI, M. C. A.; CAMARA, G.; CARTAXO, R.; GOMES, V. C. F.; SANTOS, L. A.; SANCHEZ, A. H.; ARCANJO, J. S.; FRONZA, J. G.; NORONHA, C. A.; COSTA, R. W.; ZAGLIA, M. C.; ZIOTI, F.; KORTING, T. S.; SOARES, A. R.; CHAVES, M. E. D.; FONSECA, L. M. G. Earth observation data cubes for Brazil: requirements, methodology and products. **MDPI - Remote Sensing**, v. 12, n. 24, 2020. ISSN 2072-4292. Disponível em: <<https://www.mdpi.com/2072-4292/12/24/4033>>. 3, 8, 9, 10, 11, 14, 15

FLENNINKEN, J. M.; STUGLIK, S.; LANNONE, B. V. Quantum GIS (QGIS): an introduction to a free alternative to more costly GIS platforms. **School of Forest Resources and Conservation**, v. 2020, n. 2, 2020. Disponível em: <<https://journals.flvc.org/edis/article/view/108810>>. 32, 33, 35

FOOD AND AGRICULTURE ORGANIZATION (FAO). **The state of the world's land and water resources for food and agriculture - systems at**

breaking point. [s.n.], 2021. Disponível em: <<https://doi.org/10.4060/cb7654en>>. 1, 2

FOWLER, M.; LEWIS, J. 2014. **Microservices**. Disponível em: <<http://martinfowler.com/articles/microservices.html>>. 34

FRITZ, S.; MCCALLUM, I.; SCHILL, C.; PERGER, C.; GRILLMAYER, R.; ACHARD, F.; KRAXNER, F.; OBERSTEINER, M. Geo-wiki.org: the use of crowdsourcing to improve global land cover. **MDPI - Remote Sensing**, v. 1, n. 3, p. 345–354, 2009. ISSN 2072-4292. Disponível em: <<https://www.mdpi.com/2072-4292/1/3/345>>. 20

GAVIN, D.; DHU, T.; SAGAR, S.; MUELLER, N.; DUNN, B.; LEWIS, A.; LYMBURNER, L.; MINCHIN, S.; OLIVER, S.; ROSS, J.; THANKAPPAN, M. Digital earth Australia - from Satellite data to better decisions. In: INTERNATIONAL GEOSCIENCE AND REMOTE SENSING SYMPOISUM, 2018. **Proceedings...** 2018. p. 8633–8635. Disponível em: <<https://ieeexplore.ieee.org/document/8518160>>. 3, 10

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. Disponível em: <<https://doi.org/10.1080/20964471.2017.1398903>>. 3, 8, 10

GIULIANI, G.; CHATENOUX, B.; PILLER, T.; MOSER, F.; LACROIX, P. Data cube on demand (DCoD): generating an earth observation data cube anywhere in the world. **International Journal of Applied Earth Observation and Geo Information**, v. 87, 2020. Disponível em: <<https://www.sciencedirect.com/science/article/pii/S0303243419310372>>. 8, 10

GOMES, V. C. F.; QUEIROZ, G. R.; FERREIRA, K. R. An overview of platforms for Big Earth Observation data management and analysis. **MDPI - Remote Sensing**, v. 12, n. 8, 2020. ISSN 2072-4292. Disponível em: <<https://www.mdpi.com/2072-4292/12/8/1253>>. 9, 31

GOODFELLOW, I.; BENGIO, Y.; COURVILLE, A. **Deep learning**. [S.l.]: MIT Press, 2016. 8, 14, 21, 24, 26, 28

GOOGLE. **Google Earth Engine Guide**. 2023. Disponível em: <<https://earthengine.google.com/>>. Acesso em: 10 Mai. 2023. 32, 35, 36

GORELICK, N.; HANCHER, M.; DIXON, M.; ILYUSHCHENKO, S.; THAU, D.; MOORE, R. Google earth engine: planetary-scale geospatial analysis for everyone. **Remote Sensing of Environment**, v. 202, p. 18–27, 2017. ISSN 0034-4257. Disponível em: <<https://www.sciencedirect.com/science/article/pii/S0034425717302900>>. 35

GREGORIO, A. D.; JANSEN, L. J. **Land cover classification system (LCCS): classification concepts and user manual**. [S.l.: s.n.], 2000. 7

GRIMM, N. B.; FAETH, S. H.; GOLUBIEWSKI, N. E.; REDMAN, C. L.; WU, J.; BAI, X.; BRIGGS, J. M. Global change and the ecology of cities. **Science**, p. 756–760, 2008. 1

HAMA, L.; BEECHAM, R.; LOMAX, N. TGVE: a tool for analysis and visualization of geospatial data. In: HOELLT, T.; AIGNER, W.; WANG, B. (Ed.). **EuroVis 2023 - short papers**. [S.l.]: The Eurographics Association, 2023. ISBN 978-3-03868-219-6. 46

HANSEN, M. C.; LOVELAND, T. R. A review of large area monitoring of land cover change using Landsat data. **Remote Sensing of Environment**, v. 122, p. 66–74, 2012. ISSN 0034-4257. 1, 7

HOSTERT, P.; GRIFFITHS, P.; LINDEN, S. van der; PFLUGMACHER, D. Time series analyses in a new era of optical satellite data. In: KUENZER, C.; DECH, S.; WAGNER, W. (ED.). REMOTE SENSING TIME SERIES: REVEALING LAND SURFACE DYNAMIS. CHAM: SPRINGER, 2015. [S.l.], 2015. v. 22, p. 25–41. ISBN 9783319159669. 5

HU, R.; DELANY, S. J.; NAMEE, B. M. EGAL: Exploration Guided Active Learning for TCBR. In: INTERNATIONAL CONFERENCE ON CASE-BASED REASONING, 2010. **Proceedings...** Dublin Institute of Technology, 2010. p. 156–170. Disponível em: <<https://link.springer.com/chapter/10.1007/978-3-642-14274-1_13>>. 20, 28

HUANG, H.; WANG, J.; LIU, C.; LIANG, L.; LI, C.; GONG, P. The migration of training samples towards dynamic global land cover mapping. **ISPRS Journal of Photogrammetry and Remote Sensing**, v. 161, p. 27–36, 2020. ISSN 0924-2716. Disponível em: <<https://www.sciencedirect.com/science/article/pii/S0924271620300101>>. 4

JAIN, N.; BHANSALI, A.; MEHTA, D. Angularjs: a modern mvc framework in javascript. **Journal of Global Research in Computer Science**, v. 5, n. 12, p. 17–23, 2014. 71

JIA, L.; YAO, W.; JIANG, Y.; LI, Y.; WANG, Z.; LI, H.; HUANG, F.; LI, J.; CHEN, T.; ZHANG, H. Development of interactive biological web applications with R/Shiny. **Briefings in Bioinformatics**, v. 23, n. 1, p. bbab415, 10 2021. ISSN 1477-4054. Disponível em: <<https://doi.org/10.1093/bib/bbab415>>. 45

KOHONEN, T. The self-organizing map. **Proceedings of the IEEE**, v. 78, n. 9, p. 1464–1480, 1990. 29

LI, J.; BIOUCAS-DIAS, J. M.; PLAZA, A. Semisupervised hyperspectral image segmentation using multinomial logistic regression with active learning. **IEEE Transactions on Geoscience and Remote Sensing**, v. 48, n. 11, p. 4085–4098, 2010. 25, 26

LU, M.; PEBESMA, E.; SANCHEZ, A.; VERBESSELT, J. Spatio-temporal change detection from multidimensional arrays: detecting deforestation from MODIS time series. **ISPRS Journal of Photogrammetry and Remote Sensing**, v. 117, p. 227–236, jul. 2016. 4

MACIEL, A. M.; PICOLI, M. C. A.; VINHAS, L.; CAMARA, G. Identifying land use change trajectories in Brazil's agricultural frontier. **Land**, v. 9, n. 12, 2020. ISSN 2073-445X. Disponível em: <<https://www.mdpi.com/2073-445X/9/12/506>>. 3, 9

MACIEL, A. M.; VINHAS, L. Time series classification using features extraction to identification of use land and cover land: a case study in the municipality of Itaqui, South Region of Brazil. In: SIMPÓSIO BRASILEIRO DE SENSORIAMENTO REMOTO, 18., 2017. **Anais...** São José dos Campos: INPE, 2017. ISSN 978-85-17-00088-1. 4

MACKAY, D. J. **Information theory, inference, and learning algorithms**. [S.l.]: Cambridge University Pres, 2003. ISBN 978-0521642989. 26

Malambo, L.; Heatwole, C. D. Automated training sample definition for seasonal burned area mapping. **ISPRS Journal of Photogrammetry and Remote Sensing**, v. 160, p. 107–123, 2020. 20

MATTILA, T.; HELIN, T.; ANTIKAINEN, R.; SOIMAKALLIO, S.; PINGOUD, K.; WESSMAN, H. **Land use in life cycle assessment**. [S.l.]: Finnish Environment Institute, 2011. ISBN 978–952–11–3926–0. 7

MERONI, M.; D'ANDRIMONT, R.; VRIELING, A.; FASBENDER, D.; LEMOINE, G.; REMBOLD, F.; SEGUINI, L.; VERHEGGHEN, A. Comparing land surface phenology of major european crops as derived from sar and multispectral data of sentinel-1 and -2. **Remote Sensing of Environment**, v. 253, p. 112232, 2021. ISSN 0034-4257. Disponível em: <<https://www.sciencedirect.com/science/article/pii/S0034425720306052>>. 5, 7, 26, 32

MILLARD, K.; RICHARDSON, M. On the importance of training data sample selection in random forest image classification: a case study in peatland ecosystem mapping. **MDPI - Remote Sensing**, v. 7, n. 7, p. 8489–8515, 2015. ISSN 2072-4292. Disponível em: <<https://www.mdpi.com/2072-4292/7/7/8489>>. 4

NEDD, R.; LIGHT, K.; OWENS, M.; JAMES, N.; JOHNSON, E.; ANANDHI, A. A synthesis of land use/land cover studies: definitions, classification systems, meta-studies, challenges and knowledge gaps on a global landscape. **Land**, v. 10, n. 9, 2021. ISSN 2073-445X. Disponível em: <<https://www.mdpi.com/2073-445X/10/9/994>>. 7

OLOFSSON, P.; FOODY, G. M.; HEROLD, M.; STEHMAN, S. V.; WOODCOCK, C. E.; WULDER, M. A. Good practices for estimating area and assessing accuracy of land change. **Journal of Remote Sensing of Environment**, v. 148, p. 42–57, fev. 2014. 7, 14, 19

OPEN GEOSPATIAL CONSORTIUM (OGC). **Open Geospatial Consortium**. 2022. Disponível em: <<https://www.ogc.org/>>. Acesso em: 17 Mar. 2022. 32

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. **MDPI - Remote Sensing**, v. 9, n. 2, 2017. ISSN 2072-4292. Disponível em: <<https://www.mdpi.com/2072-4292/9/2/173>>. 4

PELLETIER, C.; WEBB, G. I.; PETITJEAN, F. Temporal convolutional neural network for the classification of satellite image time series. **MDPI - Remote Sensing**, v. 11, n. 5, p. 523, 2019. 4, 16

PENGRA, B. W.; STEHMAN, S. V.; HORTON, J. A.; DOCKTER, D. J.; SCHROEDER, T. A.; YANG, Z.; COHEN, W. B.; HEALEY, S. P.; LOVELAND, T. R. Quality control and assessment of interpreter consistency of annual land cover reference data in an operational national monitoring program. **Remote Sensing of Environment**, v. 238, p. 111261, 2020. ISSN 0034-4257. Disponível em: <<https://www.sciencedirect.com/science/article/pii/S0034425719302809>>. 4

QGIS ASSOCIATION. **PyQGIS developer cookbook**. QGIS Project, 2020. Disponível em: <<https://docs.qgis.org/3.28/en/docs/pyqgis_developer_cookbook/index.html>>. 34

QGIS DEVELOPEMENT TEAM. **QGIS Geographic Information System**. QGIS Project, 2023. Disponível em: <<https://www.qgis.org>>. 33

QUEIROZ, G. R. d.; FERREIRA, K. R.; VINHAS, L.; CAMARA, G.; COSTA, R. W. d.; SOUZA, R. C. M. d.; MAUS, V. W.; SANCHEZ, A. WTSS: um serviço web para extração de séries temporais de imagens de sensoriamento remoto. In: SIMPÓSIO BRASILEIRO DE SENSORIAMENTO REMOTO, 17., 2015. **Anais...** São José dos Campos: INPE, 2015. Disponível em: <<http://mtc-m21d.sid.inpe.br/rep/8JMKD3MGP3W34T/488BPNL>>. 13

R CORE TEAM. **R: a language and environment for statistical computing**. Vienna, Austria, 2013. Disponível em: <<http://www.R-project.org/>>. 44

ROSSONI, R. A.; MORAES, M. L. Agropecuária e desmatamento na Amazônia legal brasileira: uma análise espacial entre 2007 e 2017. **Geografia em Questão**, v. 13, n. 3, 2020. Disponível em: <<https://e-revista.unioeste.br/index.php/geoemquestao/article/view/23536>>. 1

RWANGA, S. S.; M., N. J. Accuracy assessment of land use/land cover classification using remote sensing and GIS. **International Journal of Geosciences**, v. 8, n. 4, p. 611–622, abr. 2017. ISSN 2156-8359. 1, 5, 21

SAAH, D.; JOHNSON, G.; ASHMALL, B.; TONDAPU, G.; TENNESON, K.; PATTERSON, M.; POORTINGA, A.; MARKERT, K.; QUYEN, N. H.; San Aung, K.; SCHLICHTING, L.; MATIN, M.; UDDIN, K.; ARYAL, R. R.; DILGER, J.; Lee Ellenburg, W.; FLORES-ANDERSON, A. I.; WIELL, D.; LINDQUIST, E.; GOLDSTEIN, J.; CLINTON, N.; CHISHTIE, F. Collect earth: an online tool for systematic reference data collection in land cover and use applications. **Environmental Modelling & Software**, v. 118, p. 166–171, 2019. ISSN 1364-8152. Disponível em: <<https://www.sciencedirect.com/science/article/pii/S1364815218312568>>. 38

SANTOS, L. A.; FERREIRA, K.; CAMARA, G.; PICOLI, M. C. A.; 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. 4, 5, 14, 19, 29, 30, 31, 47, 49, 66, 84, 87, 105

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. **MDPI - Remote Sensing**, v. 13, n. 5, p. 974, mar. 2021. 14, 15, 16, 19, 31

SANTOS, L. B. L.; PEREIRA, M. de A. **Proceedings of the 23rd brazilian symposium on geoinformatics (GEOINFO)**. São José dos Campos: INPE: [s.n.], 2022. ISSN 2179–4847. 6

SCHLOERKE, B.; ALLEN, J. **plumber: an API generator for R**. [s.n.], 2022. Disponível em: <<https://www.rplumber.io,https://github.com/rstudio/plumber>>. 45, 46, 68

SEBBAH, B.; ALAOUI, O. Y.; WAHBI, M.; MAâTOUK, M.; Ben Achhab, N. QGIS-landsat indices plugin (Q-LIP): tool for environmental indices computing using Landsat data. **Environmental Modelling & Software**, v. 137, p. 104–972, 2021. ISSN 1364-8152. Disponível em: <<https://www.sciencedirect.com/science/article/pii/S1364815221000153>>. 34

SETTLES, B. **Active learning literature survey**. University of Wisconsin–Madison: [s.n.], 2010. 27

SIMOES, R. **Land use and land cover classification of satellite image time series using machine learning**. Tese (Doutorado em Computação Aplicada) — Instituto Nacional de Pesquisas Espaciais - INPE, São José dos Campos - SP - Brasil, 2021. Disponível em: <<http://urlib.net/8JMKD3MGP3W34R/44KLMUS>>. 17, 18

SIMOES, R.; CAMARA, G.; QUEIROZ, G.; SOUZA, F.; ANDRADE, P. R.; SANTOS, L.; CARVALHO, A.; FERREIRA, K. Satellite image time series analysis for big earth observation data. **MDPI - Remote Sensing**, v. 13, n. 13, p. 2428, jun. 2021. 1, 3, 7, 11, 14, 15, 16, 17, 68, 84

SIMOES, R.; CAMARA, G.; SOUZA, F.; ANDRADE, P.; SANTOS, L.; FERREIRA, K.; QUEIROZ, G.; 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>>. 8, 12, 17, 18, 19, 23

SIMOES, R.; PICOLI, M. C. A.; 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. 34, n. 7, jan. 2020. 15, 17

SIQUEIRA, A.; LEWIS, A.; THANKAPPAN, M.; SZANTOI, Z.; GORYL, P.; LABAHN, S.; ROSS, J.; HOSFORD, S.; MECKLENBURG, S.; TADONO, T.; ROSENQVIST, A.; LACEY, J. CEOS analysis ready data for land – an overview on the current and future work. In: IEEE INTERNATIONAL GEOSCIENCE AND REMOTE SENSING SYMPOSIUM, 2019. **Proceedings...** IEEE, 2019. p. 5536–5537. 3

SOILLE, P.; BURGER, A.; MARCHI, D. 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. Disponível em: <<https://app.dimensions.ai/details/publication/pub.1092957640>>. 8, 9

TUIA, D.; PASOLLI, E.; EMERY, W. J. Dataset shift adaptation with active queries. In: JOINT URBAN REMOTE SENSING EVENT, 2011. **Proceedings...** [S.l.], 2011. p. 121–124. 24

TUIA, D.; RATLE, F.; PACIFICI, F.; KANEVSKI, M. F.; EMERY, W. J. Active learning methods for remote sensing image classification. **IEEE Transactions on Geoscience and Remote Sensing**, v. 47, n. 7, p. 2218–2232, jul. 2009. 4, 5, 26, 27, 47, 61, 105

TUIA, D.; VOLPI, M.; COPA, L.; KANEVSKI, M.; MUñOZ-MARí, J. A survey of active learning algorithms for supervised remote sensing image classification. **IEEE Journal of Selected Topics in Signal Processing**, v. 5, n. 3, p. 606–617, jun. 2011. 5, 19, 20, 24, 28, 106

TUKEY, J. W. **Exploratory data analysis**. [S.l.]: Addison-Wesley, 1977. ISBN 978-0201076165. 21, 22

VINHAS, L.; QUEIROZ, G. R.; FERREIRA, K. R.; CâMARA, G. Web services for big earth observation data. In: BRAZILIAN SYNPOSIUM ON GEOINFORMATICS, 2017. **Proceedings...** São José dos Campos: INPE, 2017. p. 913–922. 12, 13, 17, 32

WALPOLE, R. E.; MYERS, R. H.; MYERS, S. L.; YE, K. **Probability & statistics for engineers & scientists**. 9. ed. [S.l.]: Pearson, 2012. ISBN 978-0-321-62911-1. 28

WICKHAM, H.; GROLEMUND, G. **R for data science: import, tidy, transform, visualize, and model data**. O'Reilly Media, 2017. Paperback. ISBN 1491910399. Disponível em: <<http://r4ds.had.co.nz/>>. 5, 21, 22, 47, 105

WIECZOREK, W.; DELMERICO, A. Geographic information systems. **Computational Statistics**, v. 2, n. 1, p. 167–186, 2009. 31

WULDER, M. A.; MASEK, J. G.; COHEN, W. B.; LOVELAND, T. R.; WOODCOCK, C. E. Opening the archive: how free data has enabled the science and monitoring promise of landsat. **Remote Sensing of Environment**, v. 122, p. 2–10, 2012. ISSN 0034-4257. Disponível em: <<https://www.sciencedirect.com/science/article/pii/S003442571200034X>>. 3

ZIOTI, F.; FERREIRA, K. R.; QUEIROZ, G. R.; NEVES, A. K.; CARLOS, F. M.; SOUZA, F. C.; SANTOS, L. A.; SIMOES, R. E. A platform for land use and land cover data integration and trajectory analysis. **International Journal of Applied Earth Observation and Geoinformation**, v. 106, p. 102655, 2021. ISSN 0303-2434. Disponível em: <<https://www.sciencedirect.com/science/article/pii/S0303243421003627>>. 15, 41