Integrando Métodos de Análise Durante a Coleta de Amostras de Uso e Cobertura da Terra

Abner Ernâni dos Anjos [1], Karine Reis Ferreira [1], Gilberto Ribeiro de Queiroz [1], Fabiana Zioti [1], Gabriel Sansigolo [1]

[1] Instituto Nacional de Pesquisas Espaciais (INPE),
Av. dos Astronautas 1758 – 12227-010, São José dos Campos – SP, Brasil

{abner.anjos, karine.ferreira, gilberto.queiroz,
fabiana.zioti, gabriel.sansigolo}@inpe.br

Abstract. *The Land Use and Land Cover (LULC) mapping based on machine learning requires collecting samples for the training of predictive models. This step is a time-consuming and expensive process, in addition this activity is subject to mislabeling samples. Therefore, several methods for quality assessment have been developed and applied successfully, but attaching the collect process with assessment becomes a challenge on some platforms. This article presents a work in progress that aims to integrate methods for sample analysis during the collection process in the TerraCollect system under development in the Brazil Data Cube.*

Resumo. *A produção de mapas de uso e cobertura da terra com base em aprendizado de máquina requer a coleta de amostras para treinamento de modelos preditivos. Esta etapa é custosa e pode demandar tempo, além de ser uma atividade sujeita a erros na rotulação. Por conta disso, abordagens para análise da qualidade de dados de treinamento têm sido desenvolvidas e empregadas com sucesso, porém unir estas duas atividades ainda é um desafio em diversas plataformas. Este artigo apresenta um trabalho em andamento que busca a integração de métodos para a análise das amostras durante a coleta na plataforma web TerraCollect em desenvolvimento no projeto Brazil Data Cube.*

1. Introdução

Mapas de uso e cobertura da terra, em inglês *Land Use and Land Cover* (LULC), são importantes para representar a ação humana no meio ambiente. Esses mapas são indispensáveis para órgãos governamentais que procuram desenvolver políticas públicas ambientais e territoriais eficazes para prevenir altos impactos no meio-ambiente [Hansen and Loveland 2012]. Esses mapas são produzidos principalmente a partir de imagens de sensoriamento remoto. Atualmente, diferentes missões estão lançando satélites de observação da Terra e produzindo grandes volumes de dados abertos de imagens consistentes ao longo do tempo e do espaço.

A análise de séries temporais provenientes do empilhamento de imagens de satélite combinadas com métodos de aprendizado de máquina é uma técnica amplamente usada para produzir mapas de LULC [Simoes et al. 2020]. A maioria desses métodos são supervisionados, ou seja, necessitam de um conjunto representativo de amostras rotuladas a ser subdividido em treinamento, validação e teste [Santos et al. 2021, Goodfellow et al. 2016]. Portanto, a qualidade das amostras de treinamento é crucial no processo de classificação, pois leva a mapas com melhor acurácia.

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 de LULC é alta e intrínseca em diferentes regiões e períodos devido à heterogeneidade da biodiversidade, bem como condições climáticas e práticas de manejo distintas. Por isso, as amostras, sejam coletadas em campo ou por interpretação visual a partir de imagens de alta resolução, podem conter ruídos e também estão sujeitas a erros na rotulação [Santos et al. 2021]. A coleta pode ser realizada através do uso de Sistemas de Informação Geográfica (SIG) de propósito geral, como o QGIS, ou por sistemas que foram desenvolvidos especificamente para este processo, como o *CollectEarth*.

O uso destes sistemas facilita a avaliação pois disponibilizam métodos de processamento de imagens e geometrias [Rwanga et al. 2017]. Enquanto SIG's de propósito geral possuem diversas ferramentas analíticas que exigem experiência e conhecimento específico para a configuração, os sistemas de coleta possuem funcionalidades mais restritas e objetivas e não integram muitas ferramentas analíticas.

Diversos métodos para melhorar a qualidade do conjunto de amostras têm sido desenvolvidos e empregados com sucesso [Santos et al. 2021, Tuia et al. 2009, Rwanga et al. 2017, Wickham and Grolemund 2017]. No entanto, as ferramentas existentes para coleta não integram esses métodos para auxiliar na produção de amostras representativas [Santos et al. 2021].

Este trabalho apresenta uma aplicação que integra diferentes abordagens para análise de amostras de LULC que auxiliam os especialistas durante e após a atividade de coleta. Essa aplicação está sendo implementada como uma extensão da plataforma *TerraCollect* em desenvolvimento no projeto *Brazil Data Cube* [Ferreira et al. 2020]. As abordagens incluem a Análise Exploratória de Dados [Wickham and Grolemund 2017], Estimativa de Probabilidades, *Active Learning* [Tuia et al. 2009] e *Class Noise Reduction* [Santos et al. 2021].

2. Arquitetura Geral

TerraCollect é uma plataforma *web* para coleta de amostras de LULC que está sendo desenvolvida no projeto Brazil Data Cube (BDC). Esta plataforma utiliza os serviços *web* do BDC para descoberta e acesso a dados de observação da Terra, como por exemplo o *Spatiotemporal Asset Catalog* (STAC), *Web Time Series Service* (WTSS) e *Web Land Trajectory Service* (WLTS).

O *TerraCollect* fornece a visualização de imagens de sensoriamento de cubos e coleções de dados produzidas pelo BDC a partir dos satélites *CBERS-4*, *CBERS-4A*, *Landsat-8*, *Sentinel-2* e *MODIS* catalogados com o STAC. A plataforma também permite o acesso às séries temporais e trajetórias de LULC, respectivamente, com os serviços WTSS e WLTS para auxiliar os especialistas na tomada de decisão durante a coleta.

Após 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). As amostras são representadas por seus atributos temporais ($start\_date$ e end_date), sua localização espacial (longitude e latitude) e a sua classe LULC. Com esses atributos, a extensão para a análise de amostras utiliza o SITS para a extração de séries temporais e para execução dos métodos de análise. A Figura 1 apresenta a arquitetura de *software* com a visão geral da integração desses métodos no *TerraCollect*.

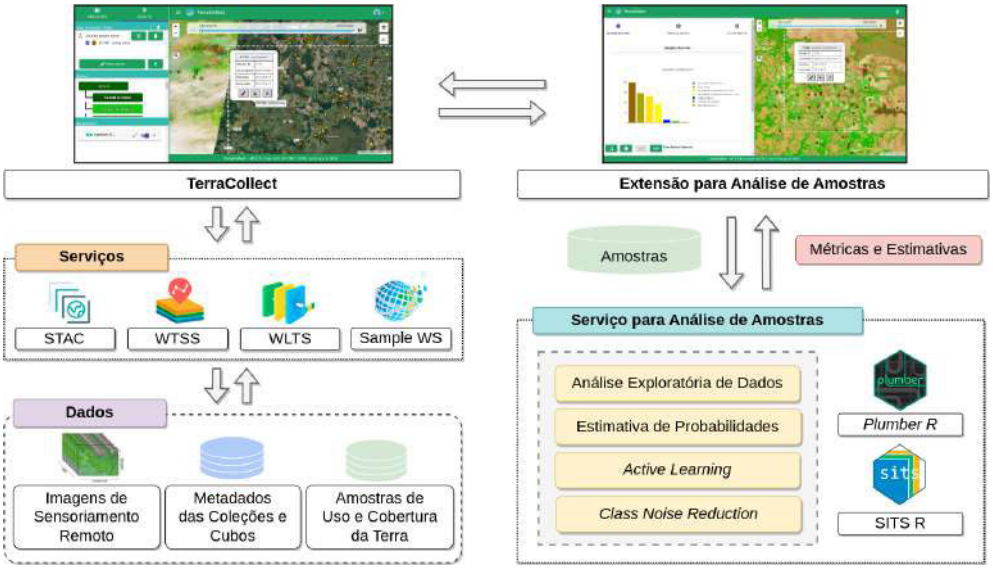

Figura 1. Arquitetura do *TerraCollect* com a extensão para a análise de amostras.

O pacote SITS fornece um conjunto de ferramentas para a extração, análise e classificação de séries temporais com aprendizado de máquina. O SITS adapta algoritmos como o *Random Forest*, *Support Vector Machines* (SVM), *Convolutional Neural Networks* (CNN's), dentre outros para trabalhar com este formato de dados [Simões 2021]. Com estas ferramentas é possível implementar as abordagens para a análise de amostras estendendo a aplicação do *TerraCollect*. A seguir são apresentados os dois principais componentes da extensão de análise de amostras.

2.1. Extensão para a Plataforma *TerraCollect*

Com as amostras disponíveis, os usuários podem acessar a extensão pela aplicação *web* do *TerraCollect*. Esta aplicação foi desenvolvida com o *framework Angular* [Jain et al. 2014] para criar uma interface gráfica com visualizações interativas de resultados. A extensão possui componentes para a execução das abordagens implementadas no SITS: Análise Exploratória de Dados, Estimativa de Probabilidades, *Active Learning* e *Class Noise Reduction*. Para integrar essas abordagens em linguagem de programação R na aplicação *TerraCollect* foi desenvolvido um serviço especializado para análise de amostras.

2.2. Serviço para Análise de Amostras

O serviço para análise de amostras é uma *Application Programming Interface* (API) desenvolvida com o pacote chamado *Plumber R* [Schloerke and Allen 2022]. Uma API permite a comunicação entre diferentes aplicações e serviços *web* [Fowler and Lewis 2014]. Para este trabalho um dos requisitos essenciais é a comunicação entre a aplicação *TerraCollect* e o SITS.

O pacote *Plumber R* converte as funções implementadas em $scripts$ R pré-existente em uma API usando uma coleção de comentários especiais para encapsula-los nos métodos *GET*, *POST* e *DELETE* [Schloerke and Allen 2022]. Desta forma o *TerraCollect* envia as requisições para executar a extração de séries temporais e a análise de amostras no SITS.

3. Abordagens para a Análise de Amostras

Essa seção apresenta a descrição das abordagens para a análise de amostras que serão integradas na arquitetura proposta.

3.1. Análise Exploratória de Amostras

A Análise Exploratória de Dados (AED) é um método usado para resumir as principais características dos dados, geralmente usando técnicas de visualização com gráficos ou tabelas. Desta forma é possível identificar padrões, detectar anomalias e verificar se os dados atendem aos objetivos esperados ou se necessitam de ajustes [Wickham and Grolemund 2017]. Na extensão de análise de amostras, o componente para a AED fornece o gráfico da distribuição de frequência de amostras por classe e a visualização do padrão das séries temporais com as bandas e índices de vegetação para resumir o conjunto de amostras disponíveis. O SITS usa uma *Generalized Additive Models* (GAM) para suavizar a série temporal, e assim, estimar um padrão de comportamento no período. Com essas visualizações o analista consegue distinguir as características de cada classe.

3.2. Estimativa de Probabilidades

As amostras armazenadas no banco de dados do *TerraCollect* podem ser usadas como um conjunto de dados base de treinamento para um modelos de aprendizado de máquina baseados em *RandomForest*, SVM's ou CNN's de forma supervisionada. Desta forma, durante o processo de coleta de novos dados, esses modelos podem ser utilizados para inferir a probabilidade de uma nova amostra pertencer a uma dada classe [E. Walpole et al. 2012]. No entanto, o uso do resultado da predição pode gerar problemas de *overfitting* resultando na especificação do modelo. Por essa razão, estudos usando a incerteza dos resultados são necessários como as técnicas de *Active Learning* apresentadas por Tuia et al. 2009.

3.3. *Active Learning*

Active Learning usa um modelo de aprendizado de máquina pré-treinado com um subconjunto do treinamento para analisar as divergências nas previsões do modelo ao predizer novas amostras [Tuia et al. 2009]. Este método é derivado do mesmo princípio da estimativa de probabilidades, porém é realizado um cálculo com base na predição como a entropia e a margem de confiança que medem a incerteza para avaliar a representatividade de uma amostra.

3.4. *Class Noise Reduction Method*

Santos et al. 2021 propôs um método que avalia a qualidade das amostras de LULC com base na inferência *bayesiana* aplicada no agrupamento com o método *Self-organizing maps* (SOM). Como apresentado no estudo, os vizinhos de cada neurônio em um mapa SOM fornecem informações sobre a variabilidade intraclasse e interclasse, desta forma a inferência é aplicada em cada um dos neurônios para mensurar a probabilidade do rótulo ter sido atribuído corretamente. Assim, é feita a análise com base nos filtros de *threshold* para a probabilidade *a priori* e *posteriori* resultando na classificação das amostras em $status$ de "$clean$"(permanece no conjunto), "$analysis$"(análise necessária) e "$remove$"(remoção aconselhável). O componente para a execução desse método possui a visualização do mapa SOM onde o analista consegue explorar os neurônios e recuperar as amostras de cada grupo. Com a amostra selecionada pode-se analisar a série temporal e compará-la com o padrão da classe.

4. Resultados Parciais

Foi realizado um estudo de caso no Bioma Cerrado com 591 amostras de agricultura e floresta extraídas dos mapas do *TerraClass*. Essas amostras foram fornecidas pela Embrapa para re-classificação com base na interpretação visual das imagens de satélite com o auxílio da plataforma *TerraCollect* para o período de Agosto de 2019 à Agosto de 2020. Para análise foram usadas as séries temporais do cubo de imagens do *CBERS-4* sensor *WFI* com uma resolução espacial de 64 metros e 16 dias de resolução temporal usando os índices de vegetação *Enhanced Vegetation Index* (EVI) e *Normalized Difference Vegetation Index* (NDVI).

A Figura 2.a mostra a tela inicial da ferramenta com as opções disponíveis para realizar a análise exploratória. Esta tela apresenta o gráfico de distribuição de frequência das amostras por classe destacado na Figura 2.b. A Figura 2.c apresenta em destaque a seleção de um padrão de séries temporais para uma dada classe de Agricultura.

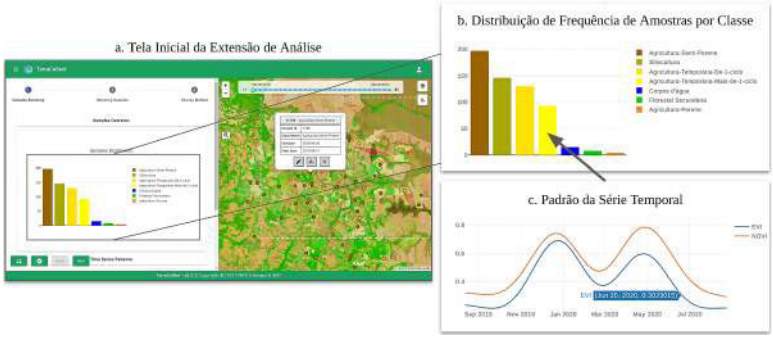

Figura 2. Tela inicial da extensão com a Análise Exploratória de Amostras.

Para fins de teste, foram selecionadas somente as amostras de agricultura das 591 amostras para treinar um modelo baseado no *Random Forest* com 200 árvores. Com o modelo treinado, foi selecionado uma nova amostra para estimar a probabilidade. A Figura 3.a apresenta o gráfico de pizza da predição resultando em 50% de chance de Agricultura semi-perene.

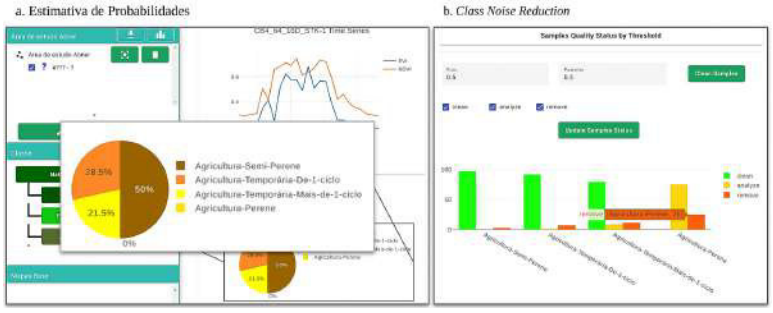

Figura 3. Recorte da tela da extensão com os resultados da análise.

Já a Figura 3.b apresenta o resultado do método *Class Noise Reduction* com gráfico da distribuição da quantidade de amostras por $status$. Para o estudo de caso, a classe de Agricultura Perene apresentou maior número de amostras ruidosas para análise e remoção. O analista pode selecionar as amostras e usar o $status$ em conjunto com as ferramentas do *TerraCollect* (como o WLTS e WTSS) para tomar a decisão sobre quais dados devem permanecer no conjunto.

5. Considerações finais

Este artigo apresenta uma aplicação para integrar métodos de análise de amostras durante a atividade de coleta. A aplicação foi implementada como uma extensão da plataforma *TerraCollect* para a execução das abordagens implementadas no SITS. Os experimentos permitiram um estudo mais elaborado dos métodos de análise de amostras e a organização e implementação desses métodos em uma aplicação *web*. Porém, ainda é necessário realizar estudos e experimentos com as técnicas de *Active Learning* para demonstrar a sua viabilidade no refinamento de amostras.

Para trabalhos futuros, espera-se revisar as ferramentas para a extração e o armazenamento das séries temporais. Devido ao volume de dados usados nos experimentos, conclui-se que ainda há um alto custo de tempo para a extração. Dessa forma é necessário propor um novo modelo de banco de dados para o armazenamento e espera-se que futuramente a extração seja otimizada usando programação paralela ou demais técnicas relacionadas.

Referências

E. Walpole, R., Myers, R. H., Myers, S. L., and Ye, K. (2012). *Probability & Statistics for Engineers & Scientists*. Pearson, 9 edition.

Ferreira, K. R., Queiroz, G. R., Vinhas, L., and et.al. (2020). Earth observation data cubes for brazil: Requirements, methodology and products. *Remote Sensing*, 12(24).

Fowler, M. and Lewis, J. (2014). Microservices.

Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep Learning*. MIT Press.

Hansen, M. C. and Loveland, T. R. (2012). A review of large area monitoring of land cover change using Landsat data. *Remote Sensing of Environment*, 122:66–74.

Jain, N., Bhansali, A., and Mehta, D. (2014). Angularjs: A modern mvc framework in javascript. *Journal of Global Research in Computer Science*, 5(12):17–23.

Rwanga, S. et al. (2017). Accuracy Assessment of Land Use/Land Cover classification Using Remote Sensing and GIS. *International Journal of Geosciences*, 8(4):611–622.

Santos, L. A. et al. (2021). Quality control and class noise reduction of satellite image time series. *ISPRS Journal of Photogrammetry and Remote Sensing*, 177:75–88.

Schloerke, B. and Allen, J. (2022). *plumber: An API Generator for R*. https://www.rplumber.io, https://github.com/rstudio/plumber.

Simoes, R. et al. (2020). Land use and cover maps for Mato Grosso State in Brazil from 2001 to 2017. *Scientific Data*, 34(7).

Simões, R. (2021). *Land use and land cover classification of satellite image time series using machine learning*. PhD thesis, Instituto Nacional de Pesquisas Espaciais - INPE, São José dos Campos - SP - Brasil.

Tuia, D. et al. (2009). Active Learning Methods for Remote Sensing Image Classification. *IEEE Transactions on Geoscience and Remote Sensing*, 47(7):2218–2232.

Wickham, H. and Grolemund, G. (2017). *R for Data Science: Import, Tidy, Transform, Visualize, and Model Data*. O'Reilly Media, 1 edition.