The role of Landsat and Sentinel-2 data harmonization in monitoring agricultural dynamics on smallholder farming regions

Taya Cristo Parreiras[269], Édson Luis Bolfe[270]

Abstract

This study explores the potential of Harmonized Landsat Sentinel-2 (HLS) data for detailed agricultural mapping in diversified farming regions of São Paulo, Brazil. Focusing on Casa Branca and Caconde, the research integrates multitemporal HLS imagery (2021–2024) to perform crop classifications at multiple levels. In Caconde, high-resolution temporal data (2–4 day revisit) enabled strong performance in distinguishing perennial crops, particularly coffee, which achieved an average sensitivity of 0.97 and specificity of 0.91. Phenological stages of coffee, such as *Producing* and *Newly Planted*, were reliably mapped, while *Stumping* and *Skeletoning* showed lower consistency. In Casa Branca, six field campaigns supported the construction of a robust training dataset across up to nine growing seasons. Integrating Landsat 9 into the HLS collection more than doubled temporal resolution over the study period, enhancing model accuracy and phenological tracking. Future work will focus on model transferability across time and space and on evaluating the relative performance of HLS versus individual Landsat and Sentinel data.

Keywords: HLS; machine learning; crop diversity; agriculture; coffee production; irrigation

1. Introduction

Monitoring agricultural dynamics through multitemporal crop mapping in Brazil has advanced considerably in the 21st century, driven by high-quality orbital data from public space agencies and, more recently, by cloud computing and machine learning capabilities. However, most studies remain concentrated in regions dominated by large-scale commodity crops, such as soy, corn, and cotton, in states like Mato Grosso,

[269] State University of Campinas. Institute of Geosciences. ORCID: 0000-0003-2621-7745. Email: t234520@dac.unicamp.br.
[270] Embrapa Digital Agriculture; State University of Campinas. ORCID: 0000-0001-7777-2445. Email: edson.bolfe@embrapa.br

Goiás, and Bahia (MATOPIBA), leaving behind areas characterized by small and medium-sized family farms (Parreiras et al., 2025).

These diversified regions, where annual and perennial crops coexist, are still assessed using generalized data. While major crops cover about 60% of Brazil's cultivated area, the remaining 40%, approximately 38 million hectares, are poorly monitored (Parreiras et al., 2025). This gap limits our understanding of agricultural dynamics, which is essential for improving productivity estimates, policymaking, and agro-environmental modeling.

Mapping such heterogeneous landscapes presents challenges for remote sensing due to crop variety, complex calendars, limited sensor resolution, and the absence of detailed spectral-temporal profiles (Parreiras et al., 2025). This ongoing Ph.D. research investigates the integration of Landsat-8/9 (OLI) and Sentinel-2 (MSI) data through the Harmonized Landsat Sentinel-2 (HLS) product. With 30 m resolution and 2–3 day revisit intervals, HLS offers a promising balance between spatial and temporal detail.

2. Methods

2.1. Study area

The location of Casa Branca and Caconde is exhibited in Figure 1 along with the latest Land Use Land Cover (LULC) map, from 2023, produced by the MapBiomas Project. Casa Branca stands out for its highly diversified agricultural landscape (Parreiras et al., 2024), featuring permanent crops, particularly citrus, and a wide range of annual crops cultivated across more than 300 center pivots and rainfed fields. Although less diversified, Caconde features a highly fragmented landscape dominated by family farming. The region is primarily characterized by coffee production, livestock pasture, and small-scale cultivation of sugarcane and corn.

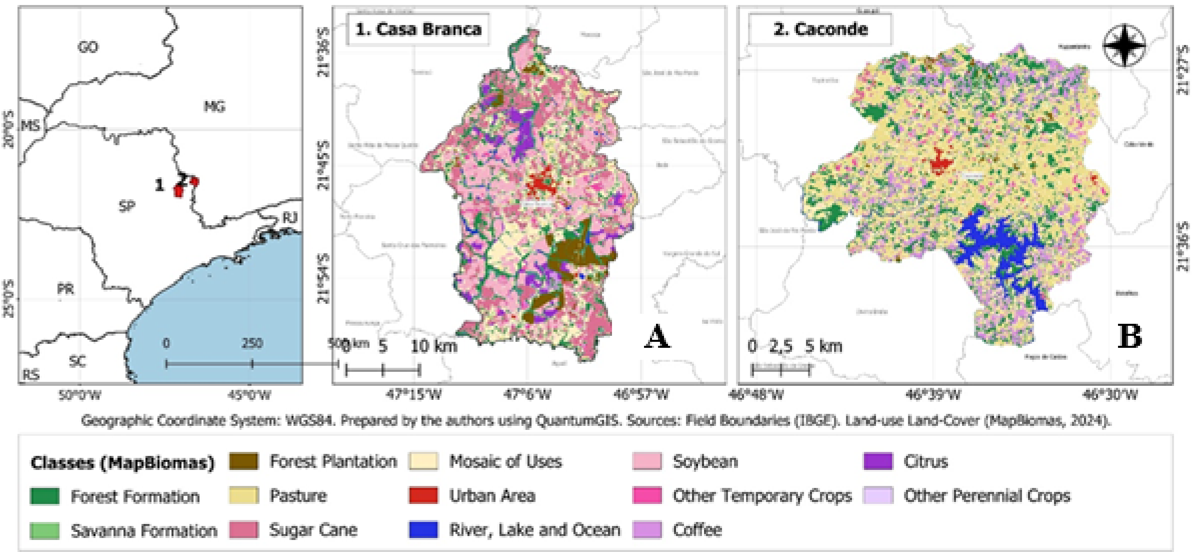

Figure 1. Location and Land Use Land Cover in 2023 of the two study areas: (A) Casa Branca, and (B) Caconde, São Paulo, Brazil.
Source: IBGE (2025), MapBiomas (2025).

2.2. Datasets and methods of analysis

This study generated two datasets: 1) ground sample data collected through field campaigns between 2021 and 2024; and 2) multitemporal multispectral imagery with varying spatial resolutions, as summarized in Table 1. These datasets are used to train and validate machine learning classification algorithms, namely the U-Net and Mask R-CNN for object-based landscape segmentation and pivot detection (at 10m and 2 or 5m*), and Random Forest (RF) and Extreme Gradient Boosting (XGBoost) for pixel-based mapping (at 10m and 30 m).

Table 1. Summary of sample datasets, imagery, and covariates employed in the study.

Study Area	Sample dataset	Imagery dataset	Covariables
Casa Branca	> 500 crop samples across cycles of three crop years	HLS, Landsat 8/9, Sentinel-2, Planet Scope or CBERS 4A*	Spectral Indices, Surface Phenology Indicators, LST, and Slope
Caconde	> 250 samples of four different stages of coffee production (2023-24)	HLS, Landsat 8/9, Sentinel-2, Planet Scope *	Spectral data, Surface Phenology, GLCM, LST, and Slope

The GLCM is the Gray Level Co-Occurrence Matrix; HLS is the Harmonized Landsat Sentinel-2; and LST means Land Surface Temperature.

The selected dimensionality reduction techniques are Recursive Feature Elimination (RFE) and Principal Component Analysis (PCA). Model performances are compared using hypothesis tests, with the null hypothesis stating that HLS outperforms Landsat 8/9 and Sentinel-2 for crop mapping. Hierarchical classification schemes were designed to improve class separation in each case (Figure 2). In Casa Branca, the primary focus is the detailed mapping of all agricultural production, particularly the region's vast diversity and intensity of annual crops. In Caconde, the main objective is to enhance the detection and characterization of coffee plantations, considering four phenological stages: newly planted, producing, skeletonizing, and stumped.

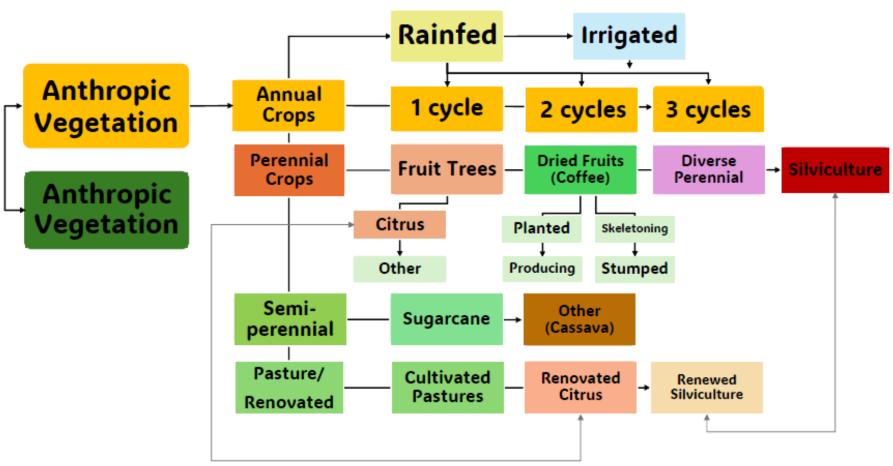

Figure 2. Hierarchical crop classification scheme based on mapped classes in Casa Branca and Caconde, São Paulo.

3. Preliminary Results and Discussion

3.1. Coffee mapping and characterization in Caconde

In an initial analysis phase in Caconde, the multiscale land use classifications using HLS time series and RF for the year 2023, focused on coffee cultivation and the distinction of four production stages, yielded highly promising results. These findings have already been submitted in article format to the *ISPRS Journal of Photogrammetry and Remote Sensing*. The outcomes were made possible by carefully conducted fieldwork in October 2023 and by the high temporal resolution of the HLS data, with revisit intervals of two days during the dry season and four days during the rainy season. This temporal density enabled the reconstruction of phenological signatures for the target classes, significantly enhancing class separability.

The classification models across four levels using HLS data demonstrated strong performance in detecting and characterizing native and anthropic vegetation types. For detecting coffee crops, accuracy remains consistently above 0.90 across varied data splits and algorithm configurations. The average sensitivity for perennial crops reached 0.95, with coffee standing out at 0.97 sensitivity and 0.91 specificity, confirming the high discriminative power of the dataset. When analyzing phenological stages, the *Producing* (PR) and *Newly Planted* (PL) classes were accurately distinguished, with average sensitivity above 0.85 and balanced accuracy often exceeding 0.87. In contrast, the *Stumping* (ST) and *Skeletoning* (SK) stages presented greater variability across models, with average balanced accuracy values of 0.62 and 0.80, respectively, suggesting increased difficulty distinguishing these stages.

These findings highlight the potential of HLS as a reliable data source for detailed monitoring of coffee plantations and their productive stages. As widely reported in the literature, minority classes tend to exhibit lower classification metrics (Bolfe et al., 2023; Parreiras et al., 2025). In this study, annual crops were responsible for the highest volume of misclassifications. These results support the notion that expanding the number of training samples leads to increasingly accurate models (Ramezan et al., 2021). Among spectral variables, the green band and Green-based Normalized Difference Vegetation Index (GNDVI) most effectively distinguished perennial crops from pastures and annuals. Shortwave infrared (SWIR), blue, and green bands accounted for 90% of the importance in separating eucalyptus and coffee. GNDVI was the key variable for differentiating coffee stages. Texture features added little value and increased data volume and processing time. Building on the insights gained from this initial study, a new field campaign was conducted in December 2024. A new coffee mapping phase will be carried out using revised methodological configurations, which will also compare the performance of HLS data with individual Landsat and Sentinel observations.

3.2. Casa Branca

Seven field campaigns have been conducted in Casa Branca to date, during which approximately 500 areas were monitored across up to nine crop seasons (up to three seasons per year) between 2021-2022 and 2023-2024. These areas constitute valuable sampling points, with class distribution illustrated in Figure 3, and will serve as input for classification models yet to be executed.

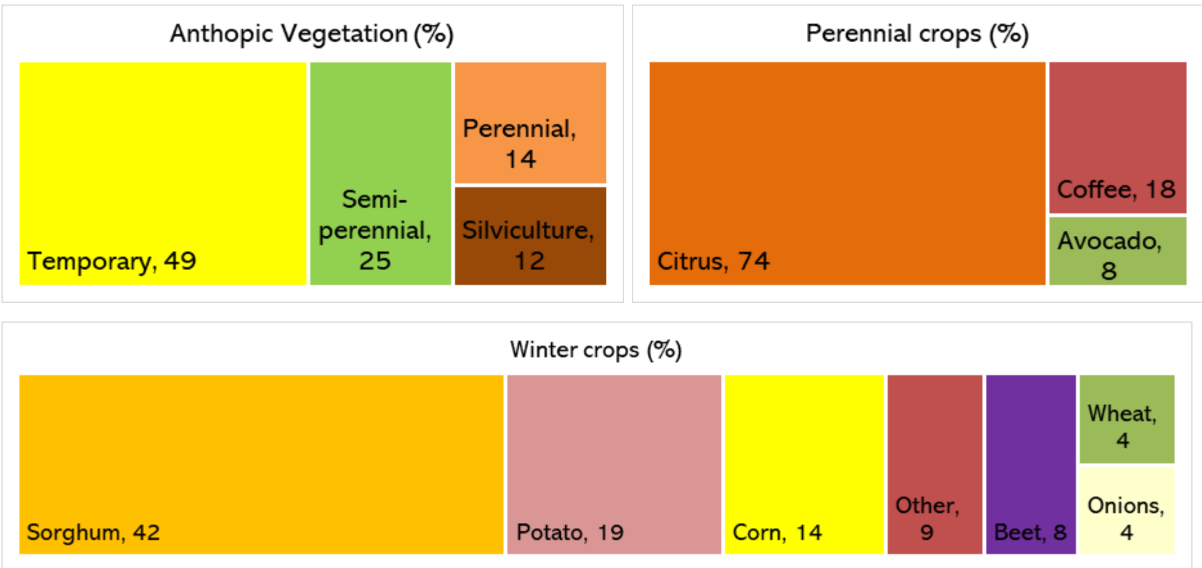

Figure 3. Tree graph with percentage of crop samples collected in Casa Branca across multiple growing seasons from 2022 to 2024.

This study is currently in the dataset construction phase, based on the HLS time series acquired between 2021 and 2024, aimed at extracting phenological indicators and other products derived from dimensionality reduction processes. The temporal resolution of HLS data in Casa Branca more than doubled over the study period, increasing from a 5-day interval in 2021 to just 2 days in 2024. This improvement was made possible by introducing Landsat 9 data into the collection. As observed by Bolfe et al. (2023) and

Parreiras et al. (2025), HLS-based models are expected to provide more detailed and accurate mapping of agricultural diversity in the region. Dimensionality reduction, sample size, and class balance are expected to be the main challenges and will be addressed through different test configurations. The best-performing models will be further adapted to evaluate temporal and spatial transferability to other DATs and municipalities within the microregion.

4. Conclusions

Based on the preliminary results, our main conclusions are: 1) HLS data enabled detailed crop mapping with improved temporal resolution (from 5 to 2 days), supporting phenological reconstruction and class separability, especially for diverse farming systems; 2) Coffee classification showed high accuracy above 90%, with strong sensitivity and specificity; *Producing* and *Newly Planted* stages were well distinguished, while *Stumping* and *Skeletoning* are more challenging; 3) Shortwave infrared (SWIR) and green bands were crucial for distinguishing coffee and their stages; and 4) challenges of dimensionality, sample size, and class imbalance will be addressed in future tests.

Acknowledgements

The authors would like to thank Fapesp (Processes 2019/26222-6, 2022/09319-9, 2024/13150-5) for the funding and other project colleagues for their valuable contributions.

References

BOLFE, E. L.; PARREIRAS, T. C.; SILVA, L. A. P. da; SANO, E. E.; BETTIOL, G. M.; VICTORIA, D. da C.; DEL´ARCO SANCHES, I.; VICENTE, L. E. Mapping agricultural intensification in the Brazilian savanna: a machine learning approach using harmonized data from Landsat Sentinel-2. **ISPRS International Journal of Geo-Information**, v. 12, n. 7, 263, July 2023. DOI: https://doi.org/10.3390/ijgi12070263.

IBGE. **Malhas territoriais**. Rio de Janeiro. Available at: https://www.ibge.gov.br/geociencias/organizacao-do-territorio/malhas-territoriais.html. Accessed on: June 1, 2025.

MAPBIOMAS. Available at: https://plataforma.brasil.mapbiomas.org/. Accessed on: June 1, 2025.

PARREIRAS, T. C.; BOLFE, E. L.; PEREIRA, P. R. M.; SOUZA, A. M. de; ALVES, V. F. Applications, challenges and perspectives for monitoring agricultural dynamics in the Brazilian savanna with multispectral remote sensing. **Remote Sensing Applications**: Society and Environment, v. 37, 101448, Jan. 2025. DOI: https://doi.org/10.1016/j.rsase.2025.101448.

PARREIRAS, T. C.; BOLFE, E. L.; SANO, E. E.; SILVA, L. A. P. da. Dinâmica espaço-temporal da expansão de pivôs centrais e da diversificação agrícola no município de Casa Branca (SP). **Labor & Engenho**, v. 18, e024004, 2024. DOI: https://doi.org/10.20396/labore.v18i00.8674282.

RAMEZAN, C. A.; WARNER, T. A.; MAXWELL, A. E.; PRICE, B. S. Effects of training set size on supervised machine-learning land-cover classification of large-area high-resolution remotely sensed data. **Remote Sensing**, v. 13, n. 3, 368, Feb. 2021. DOI: https://doi.org/10.3390/rs13030368.