



Deposited via The University of Sheffield.

White Rose Research Online URL for this paper:

<https://eprints.whiterose.ac.uk/id/eprint/243149/>

Version: Accepted Version

Article:

Correa, S.P.L.P., Reis, M.S., da Silva, M.P. et al. (2026) Rethinking reference data quality: the role of mixed pixels in remote sensing classification. *International Journal of Remote Sensing*, 47 (12). pp. 5194-5218. ISSN: 0143-1161

<https://doi.org/10.1080/01431161.2026.2664863>

© 2026 The Authors. Except as otherwise noted, this author-accepted version of a journal article published in *International Journal of Remote Sensing* is made available via the University of Sheffield Research Publications and Copyright Policy under the terms of the Creative Commons Attribution 4.0 International License (CC-BY 4.0), which permits unrestricted use, distribution and reproduction in any medium, provided the original work is properly cited. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>

Reuse

This article is distributed under the terms of the Creative Commons Attribution (CC BY) licence. This licence allows you to distribute, remix, tweak, and build upon the work, even commercially, as long as you credit the authors for the original work. More information and the full terms of the licence here: <https://creativecommons.org/licenses/>

Takedown

If you consider content in White Rose Research Online to be in breach of UK law, please notify us by emailing eprints@whiterose.ac.uk including the URL of the record and the reason for the withdrawal request.

Rethinking Reference Data Quality: The Role of Mixed Pixels in Remote Sensing Classification

S. P. L. P. Correa^a, M. S. Reis^b, M. P. da Silva^a, J. A. dos Santos^c, H. N. Oliveira^a, J. C. Oliveira^a, T. S. Körting^b, A. P. dos Santos^a, L. V. Dutra^d

^aUniversidade Federal de Viçosa, Viçosa, Minas Gerais, Brazil; ^bNational Institute for Space Research (INPE), São José dos Campos, São Paulo, Brazil; ^cUniversity of Sheffield, Sheffield, United Kingdom; ^d*In memoriam*.

ARTICLE HISTORY

Compiled July 7, 2026

ABSTRACT

Current recommendations for supervised machine learning classification in Remote Sensing advocate for using only high-quality reference data, which is often associated to pure pixels or endmembers. This focus, however, overlooks a fundamental challenge: real-world environmental images are rarely composed primarily of pure pixels. As a result, using only pure pixels as training data can leave a significant portion of the image unrepresented and hinder adequate classification. To demonstrate this effect, we introduce the Reference Sample Selection (RSS) approach. RSS systematically varies pixel purity in the training dataset to assess its impact on classification results. Pixel purity is determined based on a finer spatial resolution image. In this study, we present a case study within a select region of the Brazilian Amazon Rainforest. We applied RSS to the commonly used medium spatial resolution data from the Sentinel-2 MultiSpectral Imager (MSI) to target four land cover types: forest, water, grassland, and bare soil. The analysis used three common shallow classifiers: K-Nearest Neighbours (KNN), Support Vector Machines (SVM), and Random Forests (RFR). Our results demonstrate that including pixels with varied purity levels can significantly alter classification accuracy, depending on the land cover class. This finding challenges the conventional definition of reference data quality and highlights the need for using training samples that represent the entire image, not just its purest components. This method is readily applicable to a wide range of Remote Sensing studies. The source code and used data are available at https://github.com/paeslemesa/rss_approach/ and <https://www.kaggle.com/datasets/sabrinacorra/uruara-s2msi-samples>.

KEYWORDS

Supervised classification, Reference data collection, Mixed pixels, Pixel heterogeneity, Pattern Recognition, Thematic Accuracy, Classification Accuracy.

1. INTRODUCTION

Land cover analyses have been a central focus of several studies using Remote Sensing. The use of Machine Learning (ML) has increased drastically in the last decades, driven by data availability, hardware development and, consequently, evolving model architectures (Janga et al. 2023). For instance, different projects key for Brazilian governance

currently rely on shallow classifiers (e.g. random forests), such as TerraClass (Santos et al. 2025), Restore+ (INPE RESTORE+ Project 2025) and MapBiomass (Souza et al. 2020; MapBiomass Project 2025). For land cover applications, accuracy depends on image characteristics (Topaloglu et al. 2016), class definitions (Reis et al. 2018), and the chosen classifier (Papoutsis et al. 2023).

Classification results are also directly affected by some well-known factors: (i) number of samples, as studies suggest that a too small sample size ($n \leq 30$) would not be sufficiently representative of the population, whereas a substantially larger amount of samples may not make significant differences (Ouma et al. 2023; Moraes et al. 2024), or introduce biases into the model (Gu et al. 2016; Ustuner and Simsek 2025); (ii) sampling design, which can vary across several situations (Foody 2024; Stehman and Foody 2019; Olofsson et al. 2014); (iii) feature extraction and selection, which can either highlight target information or introduce noise (Bishop and Bishop 2024); and (iv) mixed pixels, often treated as a problem. For instance, there are studies considering unmixing them, as shown in Kavzoglu et al. (2024) and Shimabukuro and Ponzoni (2018). Nonetheless, for training data collection, the standard approach is to select pixels as pure as possible (endmembers), as emphasised in the following statement:

It is very easy to use an image processing system to pull out “training samples” from an image but it is a lot more difficult to ensure that these training samples are not contaminated either by spatial correlation effects or by the inclusion in the training sample of pixels which are not “pure” but “mixed” and therefore atypical of the class which they are supposed to represent. (Mather and Koch 2004, 217).

The existence of mixed pixels, however, leads to a critical and often overlooked factor for supervised classification: as real-world images rarely consist only of pure pixels (Fisher 1997; Kavzoglu 2009), the quality and representativeness of the reference sample data used for both training and validation (Elmes et al. 2020) depend on adequately representing pixels of very different levels of purity. This may be a particularly challenging concept in studies with imagery with lower spatial resolution, as more elements tend to contribute to a pixel’s spectral response, meaning that pixel heterogeneity remains a persistent challenge (Hsieh et al. 2001).

Although persistent, this problem is not new. Ignoring the existence of impure pixels in reference data can create a “performance illusion”, in which accuracy metrics are high but the classification is a poor representation of real phenomena. In modern Remote Sensing, it is increasingly recognised that meaningful information lies not just in individual spectral features, but in the relationships among pixels and spatial contexts. While recent studies emphasise learning these dependencies to capture land-cover structure (Dou et al. 2024a), per-class analyses of feature relations remain essential for robustness in heterogeneous landscapes.

Our study addresses the handling of mixed pixels by asking: *Can classification improve by the inclusion of reference samples with varied pixel purity?* This question aligns with the Data-Centric perspective (Roscher et al. 2024), which stresses improving data understanding before the predictive model. We propose the Reference Sample Selection (RSS) approach, a data-centric method that systematically controls pixel purity in a reference dataset. By using RSS with shallow ML classifiers, we challenge the standard definition of reference data quality and show why collecting more representative samples is critical for Remote Sensing applications.

2. Reference Sample Selection (RSS)

The RSS is a data-centric methodology designed to assess the impact of pixel purity on image classification models. This process allows us to systematically test how models handle variations in training and validation data quality. The approach relies on two primary inputs: a finer-resolution (FR) image and a coarser-resolution (CR) image.

The core concept is that a CR pixel’s purity can be determined by the proportion of classified FR pixels contained within it (entitled herein as *prop*). A CR pixel is considered as “pure”, or an endmember, if all FR pixels covered by it belong to the same class. Otherwise, it is considered as “mixed”, with a quantifiable degree of purity. The RSS approach is structured into five main phases, as shown in the flowchart in Figure 1.

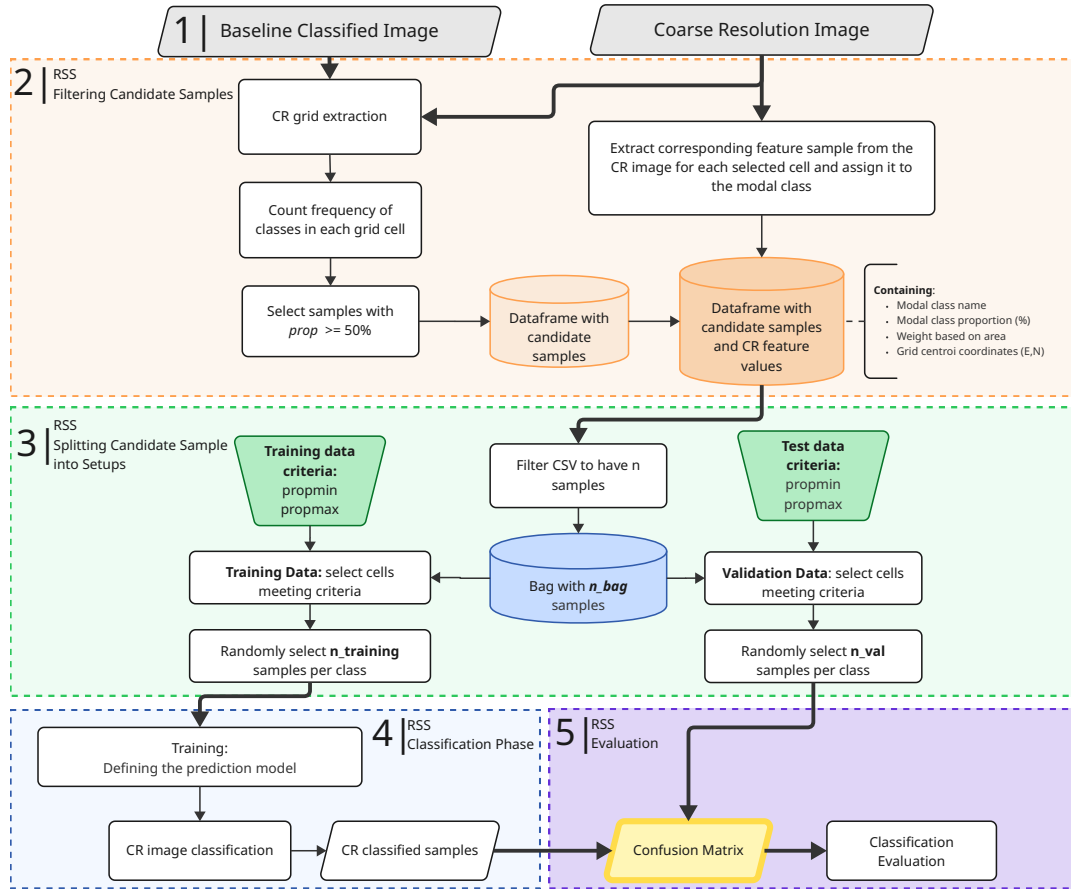


Figure 1.: Flowchart for RSS, using as input FR and CR images and processing with the six main steps: (1) Baseline Classification; (2) Filtering Candidate Samples; (3) Splitting Candidate Samples into Setups; (4) Classification Phase, and (5) Evaluation. In this image, thicker arrows mean output/input to the next phase.

Additionally, in order to properly perform the RSS approach, we defined five criteria:

- I. Both input images, FR and CR, should be well registered (geometric corrected) to avoid inadequate weighting of pixels in later stages of RSS. Inherent errors due to positional discrepancies in image acquisition and processing should be minimised for adequate analysis.
- II. The same land cover (LC) classes must be used in the baseline classification and

- during the RSS approach. This initial study aims to aid interpretation, though this condition may change in future studies.
- III. The thematic accuracy of the Baseline classification is to be $\geq 90\%$, so there is minimal error propagation from the baseline classification to the RSS approach and subsequent classification analysis.
 - IV. A candidate CR sample is chosen if, and only if, its proportion of FR pixels from the same class is equal to or higher than 50% ($prop \geq 50\%$).
 - V. To adjust for registration errors (which should be minimal at this point) and avoid neglecting border pixels, we propose to weigh the CR samples according to the total number of BL pixels (classified fine-resolution pixels). Weight is equal to 1 if the entire CR pixel has BL pixels in it. The weight decreases in proportion to the area covered by BL pixels in a CR pixel, considering the number of BL pixels completely inside a CR pixel, divided by the total number of possible BL pixels inside a CR pixel. The weight that the candidate sample has depends on the number of BL pixels in it.

2.1. Phase 1: Baseline Classification

The first phase involves creating a Baseline (BL) classification over the fine-resolution (FR) image. This classification can be performed by an analyst or derived from a pre-existing, highly accurate land cover product (e.g., MapBiomass from [Souza et al. \(2020\)](#)).

2.2. Phase 2: Filtering Candidate Samples

This is the phase where we identify the candidate samples for the classification experiments. For this, we analyse the superposition between the fine-resolution (FR) classified image and the coarse-resolution image grid. Then, for each CR pixel, we determine the modal class (i.e., the most frequent class among the FR pixels within it). In the eventuality of a tie, the class is chosen by alphabetical order.

A CR pixel is selected as a candidate sample if and only if its modal class proportion is equal to or greater than 50% ($prop \geq 50\%$). For instance, as shown in Figure 2, a CR pixel containing 21 out of 36 FR pixels of a single class ($\approx 58\%$) would be selected, while one with only 17 out of 36 ($\approx 47\%$) would not.

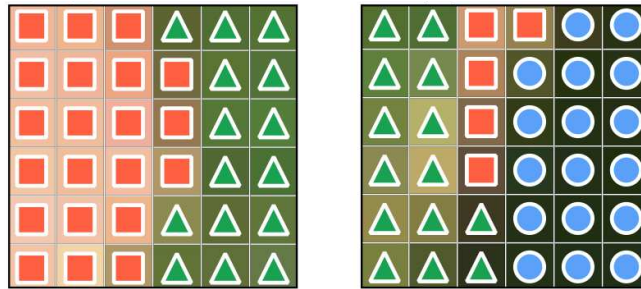


Figure 2.: Two coarse-resolution pixels (left and right) and spatially corresponding classified fine-resolution pixels. In these images, red squares denote areas of bare soil, green triangles denote forests, and blue circles denote water. We emphasise that this is a non-real illustration.

To account for potential registration errors between the two images and to avoid neglecting border pixels, we weight each candidate sample based on the area of classified

FR pixels it fully contains. The weight is calculated as the ratio of the number of FR pixels with more than half inside the CR pixel to the total number of FR pixels that would ideally fit within a CR pixel (i.e., 36 for a 60 m CR pixel with 10 m FR pixels). For this solution, we used Bresenham’s line algorithm, shown in Gillies (2019)¹.

2.3. Phase 3: Splitting Candidate Samples into Setups

The selected samples ($prop \geq 50\%$) are then split into training and validation sets according to six different setups, detailed in Table 1. The default proportion variations range from 50% to 100% in 5% increments, creating strata of samples for each purity range (i.e., $[50\%, 55\%[$, $[55\%, 60\%[$, etc.). Endmembers ($prop = 100\%$) are studied separately.

Table 1.: Summary of different simulation setups with corresponding train and test set compositions.

Setup	Train Sets	Validation Set	Simulation Description
1	$[prop, prop + 5\%[$	$[50\%, 100\%]$	Tests the performance of increasingly purer sets of training samples, though in a limited range, against a generalised test set, which is supposedly representative of typical modal class areas.
2	$[prop, prop + 5\%[$	$[100\%]$	Tests the performance of increasingly purer sets of training samples, though in a limited range, against a pure set of pixels within the modal class. Idealised test sets are the ones supposedly chosen in supervised tasks.
3	$[prop, 100\%]$	$[50\%, 100\%]$	Tests the performance of classifiers estimated with increasingly purer training samples against a generalised test set, which is supposedly representative of typical modal class areas.
4	$[prop, 100\%]$	$[100\%]$	Tests the performance of classifiers estimated with increasingly purer and complete pure sets of training samples against an idealised test set.
5	$[prop, prop + 5\%[$	$[prop, 100\%]$	Tests the performance of increasingly purer sets of training samples in a limited range against a generalised test set, which is supposedly representative of typical modal class areas.
6	$[prop, 100\%]$	$[prop, 100\%]$	This is the scenario of common reference data acquisition, where the analyst acquires both training and test data according to their collecting experience.

In this table, $prop$ refers to the proportion of finer-resolution pixels of the modal class within a coarser-resolution pixel.

For each setup and run, the reference samples are selected using the bagging process. The approach is based on a stratified random sampling approach previously explained in Stehman (2000); Congalton (2007). However, in the bagging, we aim to ensure an equal number of samples per class and per purity stratum. Thus, the total number of samples in the bag, n_{bag} , is calculated using Equation (1), where n_{class} is the number of classes, n_{samples} is the number of samples from the least frequent stratum, and 11 represents the 5% increments from 50% to 100%.

$$n_{\text{bag}} = 11 \times n_{\text{class}} \times n_{\text{samples}} \quad (1)$$

¹Please also refer to: <https://rasterio.readthedocs.io/en/latest/api/rasterio.mask.html>

A point to be addressed is the availability of samples in the *bag*, which will vary according to the Setup (Table 1) and the used *prop*. For example, for the Training Set on Setup 1 ($[prop, prop + 5\%]$), the sample availability will always be n_{samples} because our interest is only in one stratum. On the other hand, for this same Setup, in the Validation set ($[50\%, 100\%]$), the number of available samples will be n_{bag} minus the already selected samples.

2.4. Phase 4: Image Classification Phase

The datasets are split only into train and validation in the proportion of 70%/30% (Hastie et al. 2009), respectively, and then they are subjected to Monte Carlo Cross Validation (MCCV) (Xu and Liang 2001) with 1,000 repetitions. The accuracy metrics (see Section 2.5) are stored, and the average metric and its standard deviation are computed for each iteration.

The prepared datasets are used to train and validate three shallow machine learning classifiers: K-Nearest Neighbours (KNN), Support Vector Machine (SVM), and Random Forests (RFR). These models were chosen for their distinct mechanisms for separating classes, which are expected to behave differently under varied data purity scenarios. No deep learning models were used in this initial study to better understand the fundamental effects of data purity. The used shallow machine learning classifiers are briefly described as follows. They are further described in Du et al. (2020).

2.4.1. K-Nearest Neighbours (KNN)

This is a non-parametric model that classifies samples based on their proximity to the k nearest training samples in the feature space (Kramer 2013; Kavzoglu et al. 2024). The classification is determined by a majority vote among its k neighbours, where the selection of k is crucial as a small value can lead to noise sensitivity (overfitting) and a large value can introduce bias. Because it is a lazy learner that performs no actual training, its computational cost is primarily incurred during the prediction phase, requiring distance metrics like Euclidean distance to define proximity.

2.4.2. Support Vector Machines (SVM)

This classifier finds an optimal hyperplane to separate samples from different classes, a process that can effectively handle non-linear problems (Bishop 2006; Doğan et al. 2016). This is achieved using the kernel trick, which implicitly maps the input features into a higher-dimensional space where a linear separation becomes possible, with common kernels being the Radial Basis Function (RBF) and polynomial kernels. The hyperplane is optimally positioned to maximise the margin between the closest training points of different classes, known as the support vectors, which are the critical elements defining the model.

2.4.3. Random Forests (RFR)

This ensemble method combines multiple decision trees and uses a bootstrap sampling approach to enhance model robustness and avoid overfitting. It is widely used in Remote Sensing due to its high performance and built-in feature importance capabilities (Breiman 2001; Belgiu and Drăguţ 2016; Nguyen and Saha 2024). Each tree is independently grown using a random subset of the training data and a random subset of

features at each split, and the final classification is determined by averaging the outputs or a majority vote across all individual trees. It is widely used in Remote Sensing due to its high performance and built-in feature importance capabilities, which aid in variable selection and model interpretability.

2.5. Phase 5: Evaluation

In all scenarios, for comparing the results, there are two different types of analyses: completeness and thematic accuracy (ISO 2023). Both cases are computed based on the confusion matrix (Congalton 1991, 2007).

Completeness is related to the presence or absence of features in the map (ISO 2023). According to ISO 19157:2023 (ISO 2023), completeness is divided into Commission Error (CE, excess of data, or false positive rate) and Omission Error (OE, absence of data, or false negative rate). Another way to understand these metrics is by the definition of User Accuracy ($UA = 1 - CE$), which is the same in Machine Learning as Precision, and Producer Accuracy ($PA = 1 - OE$), also called Recall. Precision and Recall are presented in Table 2, respectively.

Table 2.: Completeness (or per-class) metrics equations.

Metric name	General equation	Binary equation
Precision / User Accuracy (UA)	$\frac{x_{k,k}}{x_{k,+}}$	$\frac{TP}{TP + FP}$
Recall / Producer Accuracy (PA)	$\frac{x_{k,k}}{x_{+,k}}$	$\frac{TP}{TP + FN}$

Where: $x_{k,k}$ is the main diagonal element of the confusion matrix, $x_{+,k}$ is the sum of the k^{th} column, and $x_{k,+}$ is the sum of the k^{th} row. TP , FP , and FN denote true positives, false positives, and false negatives, respectively.

Note: The *General equation* follows the notation proposed by Congalton (1991), while the *Binary equation* reflects the formulation most commonly adopted in recent literature, presented in Foody (2024).

Thematic accuracy, on the other hand, generally infers how the cover class, at a certain time, is compared to the reference data (Congalton 2007); they can be measured by the Overall Accuracy, OA, the F1-Score (Van Rijsbergen 1979), or the Mean Intersection Over Union (mIoU), also known as Jaccard Index (Jaccard 1912). OA is a generic metric that sums up the amount of correct predictions over mistaken predictions. Differently, F1-score is a metric known for the harmonic average between Precision and Recall, which tends to account for imbalanced data. Finally, mIoU (Jaccard Index) also accounts for unbalanced data (Maxwell et al. 2021). All thematic accuracy equations are presented in Table 3.

3. Materials

This section describes the study area, the data used, and the specific experimental design of the Reference Sample Selection (RSS) approach.

3.1. Study Area

The study area is located in the Brazilian Amazon Rainforest, specifically in the municipalities of Altamira, Placas, and Uruará, in the state of Pará, as shown in Figure 3.

Table 3.: Thematic (or global) metrics.

Metric name	General equation	Binary equation
Overall Accuracy (OA)	$\frac{1}{N} \sum_{k=1}^K x_{k,k}$	$\frac{TP + TN}{TP + TN + FP + FN}$
F1-score	$\frac{1}{K} \sum_{k=1}^K \left(2 \cdot \frac{\text{Precision}_k + \text{Recall}_k}{\text{Precision}_k \cdot \text{Recall}_k} \right)$	$\frac{2 \cdot TP}{(2 \cdot TP) + FN + FP}$
Intersection over Union (IoU)	$\frac{1}{K} \sum_{k=1}^K \left(\frac{x_{k,k}}{x_{k,+} + x_{+,k} - x_{k,k}} \right)$	$\frac{TP}{TP + FP + FN}$

Where: K is the total number of analysed classes, k is the respective class; N is the total number of samples. TN means True Negative.

The site spans approximately 12,028 km² and overlaps with the Iriri River Extractive Reserve (RESEX) in its southwestern portion. The area is georeferenced using the Universal Transverse Mercator (UTM) Zone 21S projection with the SIRGAS2000 datum².

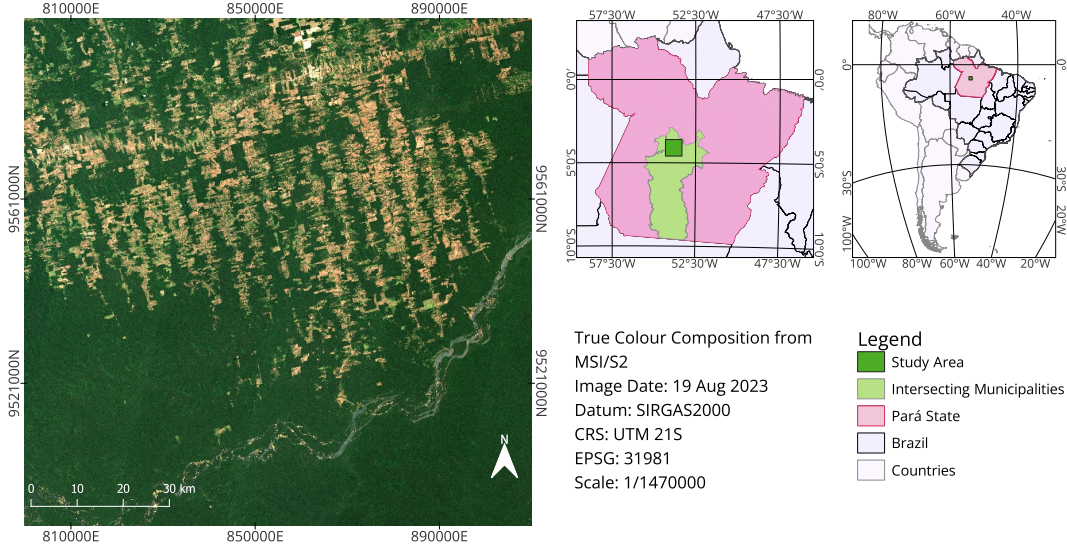


Figure 3.: Study Area location showing (a) Sentinel-2/MSI true colour image for the area; (b) its location in Brazil and (c) its location in the municipalities of Altamira, Placas, and Uruará, in Pará State.

3.2. Data and Preprocessing

In this section, we present the information on the data used and how it was preprocessed for the experimental planning, considering the explanation for land cover classes, the used Remote Sensing sensors, and the respective satellite and feature extraction.

²SIRGAS2000 stands for Geodetic Reference System for the Americas for the year 2000.

Table 4.: Land-cover classes, brief descriptions, and general spectral cues for the study area.

Class	Description	Spectral cues (S2/MSI)	Illustr.
Bare soil	Exposed soil or substrate with little to no vegetation; in this schema, built-up areas are grouped herein.*	<i>Red-NIR</i> : moderate, often increasing with wavelength; <i>Green peak</i> : weak; <i>Indices</i> : NDVI $\lesssim$ 0.1.	
Crops	Cultivated fields, shrubs/litter, and grasslands; strong seasonality/phenology.	<i>Red</i> : low (chlorophyll absorption in B4); <i>NIR</i> : high (B8) when leafy; <i>Index</i> : NDVI $\sim$ 0.3–0.8 (growth-stage dependent).	
Forest	Mature or dense forests with a closed canopy.	<i>Red</i> : very low (strong absorption); <i>NIR</i> : very high plateau (B8); <i>Index</i> : NDVI $\gtrsim$ 0.6.	
Water	Rivers, lakes, reservoirs, and other surface-water bodies.	<i>RGB-NIR</i> : very low (dark in Blue/Green/Red); <i>NIR</i> : near zero (strong absorption); <i>Turbid</i> : brighter in Green/Red; <i>Index</i> : NDVI $\ll$ 0.	

*Urban materials are spectrally heterogeneous: bright roofs/concrete can be bright in RGB-NIR; asphalt/shadows appear very dark.

These NDVI values are qualitative cues, not classification thresholds.

Sources: [Reis et al. \(2020\)](#); [Jensen \(2014\)](#); [Lillesand et al. \(2008\)](#).

3.2.1. Land Cover Classes

The classes used for this experiment are presented in Table 4. These classes were selected as a relatively easy-to-distinguish set, based on *in situ* studies of land use and land cover over the Tapajós reserve – near the study site ([Reis et al. 2020, 2018](#)).

3.2.2. Satellite Data

As this is an experimental study, we opted to use a single satellite sensor to avoid unknown geometric distortions. We used one image from the MultiSpectral Imager (MSI) on board Sentinel-2 (S2) ([European Space Agency 2025](#)), acquired on August 19, 2023, and chosen for having the lowest cloud cover percentage recorded for the area between 2017 and May 2025.

For the FR image, we used four bands from the MSI sensor—blue, green, red, and near-infrared (NIR), all of which have a 10 m spatial resolution. Additionally, the Normalised Difference Vegetation Index (NDVI) was calculated to enhance the feature set, as it is known to improve land cover classification with shallow learning classifiers ([Kavzoglu et al. 2024](#)). The final feature set for the classification consists of five bands: blue, green, red, NIR, and NDVI.

For experimental purposes, we defined a CR pixel to have its spatial resolution at least 6 times greater than the finer resolution. Here, we imagined a scenario compatible with the use of imagery from sensors such as the Wide Field Imaging Camera (WFI) onboard the CBERS family and Amazonia-1. To create the CR image, we resampled the 10 m bands to 60 m using the Nearest Neighbour approach. Additionally, 2 m shift was intentionally applied to the raster coordinates to simulate geometric misalignment between different sensors. This shift ensures that the weighting procedure is applied to account for these real-world errors.

For initial experiments, our recommendation is to use the coarser-resolution image, which is a low-resolution (50 – 250 m) image, while a finer-resolution image would be a

medium-resolution (10 – 50 m) image, following the taxonomies defined in [Ehlers et al. \(2002\)](#). However, we emphasise that this approach is intended to be agnostic to the taxonomy of the spatial resolution, as long as there is this variation between coarser- and finer-resolution images.

4. Experimental Planning

This section explains how the experiment was conducted, organised by each phase of the RSS process (see Section 2).

4.1. Phase 1: Baseline Classification

A baseline classification of the 10 m FR MSI/S2 image was performed using the Random Forest (RFR) classifier with 100 trees, a maximum depth of 2, and the Gini impurity criterion. The random state was set to 42³. This classifier was chosen for its ability to provide feature importance, which can be useful for further analysis.

To ensure a stratified and balanced sample, reference samples were selected through visual interpretation (see Table 4). The number of samples per class was standardised to the number of samples from the least frequent class (Water), which was 31,571. This yielded a total of 126,284 samples, which were then randomly split into 75% for training (94,713 samples) and 25% for validation (31,571 samples).

4.2. Phase 2: Filtering Candidate Samples

For each 60 m CR pixel, we identified the modal class from the underlying 10 m baseline classification. A CR pixel was selected as a candidate sample only if the proportion of its modal class was equal to or greater than 50%.

4.3. Phase 3: Splitting Candidate Samples into Setups

The next step is splitting the candidate samples into Setups, or bagging. The main goal is to guarantee that, for all Setups (see Table 1), there is the same number of candidate samples per stratum, in order to follow a stratified random sampling as sampling design, chosen according to [Stehman \(2009\)](#) and [Olofsson et al. \(2014\)](#). This controlled scenario was designed specifically to avoid known problems in Remote Sensing processing, such as imbalanced data that are known to affect shallow classifiers more drastically than deep classifiers ([Thanh Noi and Kappas 2017](#)).

A stratum, in this case, is the class together with the proportion of pixels from the modal class (*prop*). The set that dictates how the bag will be is Train/Validation. So, for instance, for Setups 1, 2, and 5 (Train/Validation set being $[prop, prop + 5]$), for each $prop = 50 + 5k, k \in [0, 11]$, there will be $n_{\text{classes}} \times n_{\text{samples}}$ samples in the bag (n_{bag}). On the other hand, for Setups 3, 4, and 6, $n_{\text{bag}} = n_{\text{prop}} \times n_{\text{classes}} \times n_{\text{samples}}$.

³In honour of Douglas Adams in his fiction book, *The Hitchhiker's Guide to the Galaxy*.

Table 5.: Confusion matrix for the baseline classification.

Predicted	Reference				Total
	Forest	Water	Crops	Bare Soil	
Forest	7822	0	83	0	7905
Water	10	7963	0	5	7978
Crops	378	0	7328	165	7871
Bare Soil	103	3	308	7403	7817
Total	8313	7966	7719	7573	31571

4.4. Phase 4: Classification Phase

The classification phase consisted of performing shallow learning classification using the following hyperparameters:

- **KNN:** 5 neighbours, leaf-size of 30, Euclidean distance, weighted.
- **SVM:** One Against One (OAO), regularisation parameter of 1.0, Radial Basis Function (RBF), gamma = scale, shrinking heuristic, no iterations limited, weighted.
- **Random Forest:** 100 trees, Gini criterion, maximum depth = 2, OOB-score = True.

For all classifications, there was a Monte Carlo Cross Validation (MCCV) with 1,000 repetitions. Phase 5 (Evaluation) will be discussed in the next section.

5. Results

Our results are presented in three main parts: an analysis of the baseline classification, a characterisation of the prepared candidate samples, and a detailed evaluation of the different RSS setups.

5.1. Phase 1: Baseline Classification

The baseline classification, performed using a Random Forest model on the 10 m S2/MSI image, achieved an OA of 96.68% and an OOB score of 0.965. The confusion matrix in Table 5 confirms the model’s high performance and near-optimal classification for the FR image. The feature importance analysis for the Gini criterion revealed that NDVI (32.00%) and NIR (26.55%) were the most influential bands for classification, followed by Green (19.33%), Red (15.99%), and Blue (6.13%), which is consistent with typical land use and land cover (LULC) studies in forested areas. As for completeness, the results are presented in Table 6, with the lowest accuracy of 93.10% (Crops). This result is compatible with Remote Sensing results for the Amazon LULC classifications over forested areas.

5.2. Phase 2: Filtering Candidate Samples

After creating the baseline image, we generated a dataset of candidate samples from the 60 m CR grid. The number of samples per purity stratum is shown in Figure 4. The

Table 6.: Completeness elements for each land cover class in the baseline classification.

Class	Precision (UA)	Recall (PA)
Forest	94.09%	98.95%
Water	99.96%	99.81%
Crops	94.93%	93.10%
Bare Soil	97.76%	94.70%

lowest number of samples was observed in the Water class for the $[50\%, 55\%[$ proportion interval, which is 1,354. In our study area, the number of samples for the Forest class increased most drastically with higher purity levels, indicating that this class tends to have more pure pixels than others.

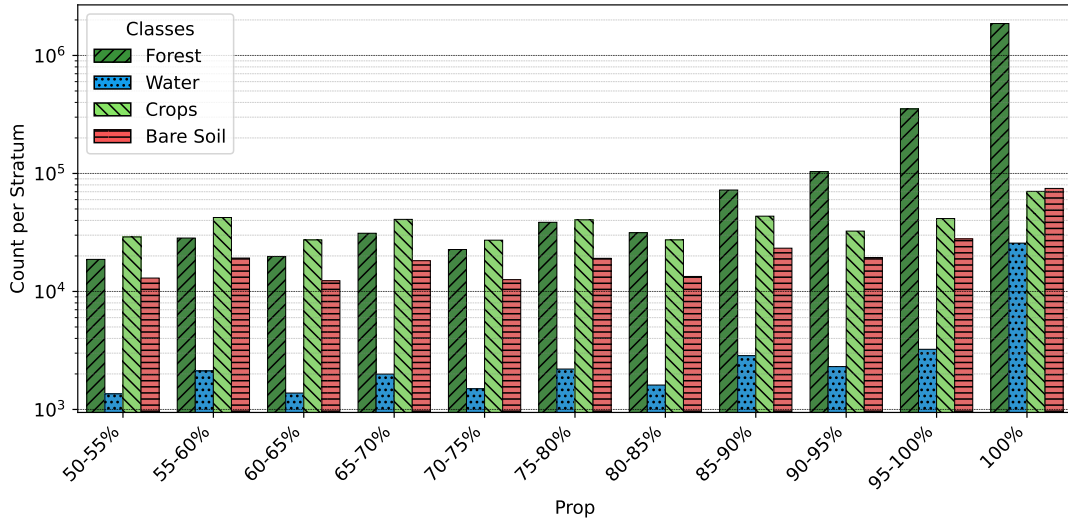


Figure 4.: Number of candidate samples per stratum considering the closed interval on the lowest number and the open interval on the higher number ($[prop, prop + 5\%[$). This figure is on a logarithmic scale on the y -axis.

Section 5.2 illustrates the feature space of selected labelled samples. Only the less correlated channels (Red, NIR, and NDVI) are shown in this figure. To minimise the amount of data in the plot, we selected $prop \geq 50\%$ and 1,350 samples for each class. This is Setup 6, which mimics selection by an analyst. As can be seen in this figure, Water seems to be more separated than the other classes. Bare Soil presents lower NDVI values with little mix with Crops. Crops and Forest have some mix as well. This shows that the Water class is more easily distinguished in the feature space; hence, the classifiers are more likely to better discriminate this class.

5.3. Phases 3 and 4: Splitting Candidate Samples into Setups and Classification

Once all candidate samples are studied, we can observe the lowest number of samples in a stratum, which is 1,354 for Water in $prop = [50\%, 55\%[$. To round the value, we determined the number of samples to be 1,350 per stratum. Consequently, considering the entire possibilities of samples in the bag, its size is 59,400 ($1,350 \times 11_{\text{intervals}} \times 4_{\text{classes}}$).

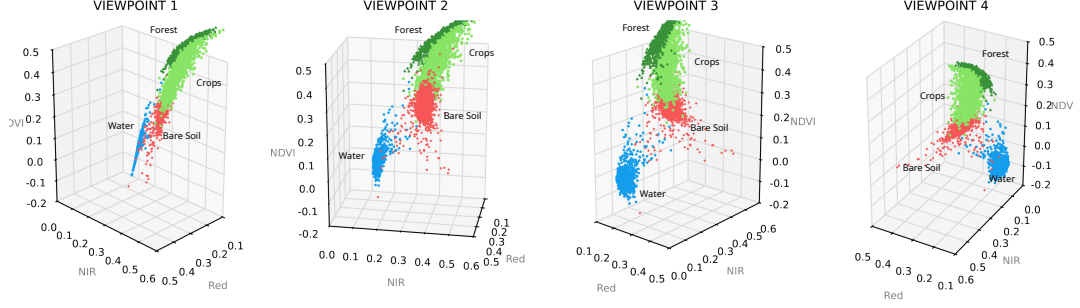


Figure 5.: Feature Space of all classes with $prop \in [50\%, 100\%]$, with 1,350 samples for each class. This feature space shows reflectance values of Red and NIR on the x - and y -axes, respectively, and NDVI values on the z -axis. Water is presented by blue colour with circular markers, Forest is presented by dark green colour with triangular markers, Crops is presented by square markers with light green colour, and Bare Soil is illustrated by red colour in the ‘ x ’ marker. The viewpoints consider elevation and azimuth (in degrees) for each one, respectively: (20, 45), (20, 10), (20, 310), (30, 120). These angles aimed to provide better visualisation of sample distribution in the feature space.

Considering the Classification Phase, we split the reference samples into 70%/30% (see Section 2.4), which results in 945 and 405 samples, respectively, per stratum (i.e., class and $prop$, as detailed in Table 1). We split the analysis into Thematic Accuracy (Section 5.4.1), presenting results for F1-score, and Completeness (Section 5.4.2), presenting Precision and Recall results.

5.4. Phase 5: Evaluation – Impact of Reference Sample Purity on Classification Performance

5.4.1. Thematic Accuracy

Figure 6 shows the F1-score for all setups. We do not present results from overall accuracy and Jaccard Index (mIoU) as their trends did not diverge from the F1-score. Several key trends are apparent:

- **Overestimation from Pure Validation Data:** Setups with a pure validation set (Setups 2 and 4) consistently show overestimated F1-score values, often approaching 1.0. This demonstrates that using only high-quality validation data can lead to a misleadingly optimistic view of a model’s true performance.
- **Performance with Varied Purity:** The classifiers generally perform best when the quality of the training data aligns with the quality of the validation data (Setup 6), where both sets include a wide range of pixel purities.
- **Classifier Sensitivity:** Random Forest (RFR) generally shows a higher F1-score than SVM and KNN, demonstrating its robustness across different data purities. SVM appears to be the most sensitive to data quality, as its performance is often more volatile, particularly in setups where training and validation data purities diverge.
- **Influence of Impure Samples:** Contrary to common practice, using only pure training samples does not guarantee the best performance. In fact, training models with a wider range of purities (e.g., Setup 6) led to the highest overall thematic

accuracy for both RFR and SVM.

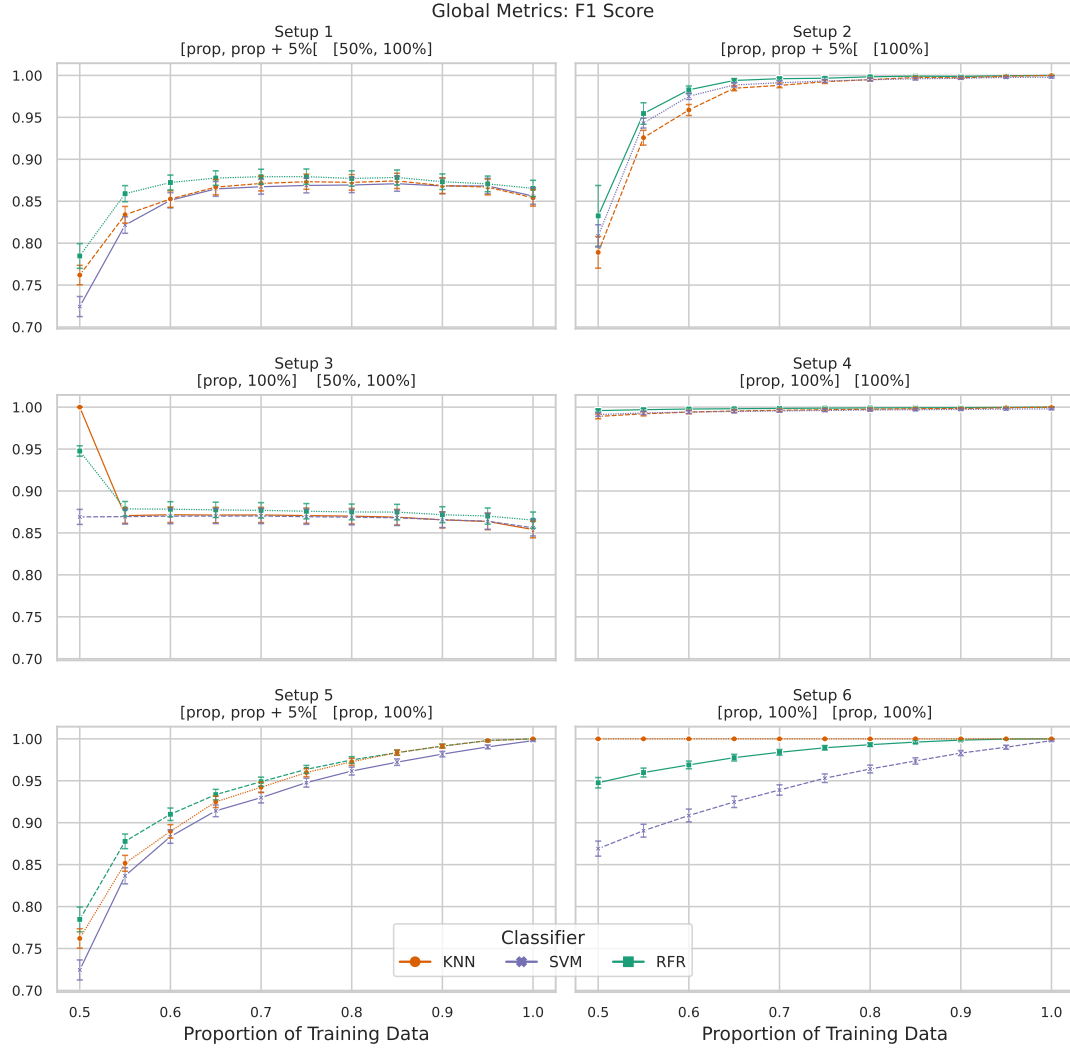


Figure 6.: Thematic Accuracy: F1-score versus *prop* for all Setups considering KNN (blue circle with dotted line), SVM (orange x with dashed line), and Random Forest (RFR, green squares with continuous line). All values represent the average of 1,000 iterations, along with their respective standard deviations. The set on the left represents the training reference data, while the set on the right represents the validation data.

5.4.2. Completeness

The per-class analysis of Precision and Recall is presented in Figure 7 (Random Forest), Figure 8 (SVM), and Figure 9 (KNN). Results are presented per class.

5.4.2.1. Water. As anticipated from the feature space analysis (Section 5.2), the Water class can be easily distinguished from other classes, with Precision and Recall values consistently above 0.9 across all classifiers and setups. Water separability has been previously assessed by Kavzoglu et al. (2024).

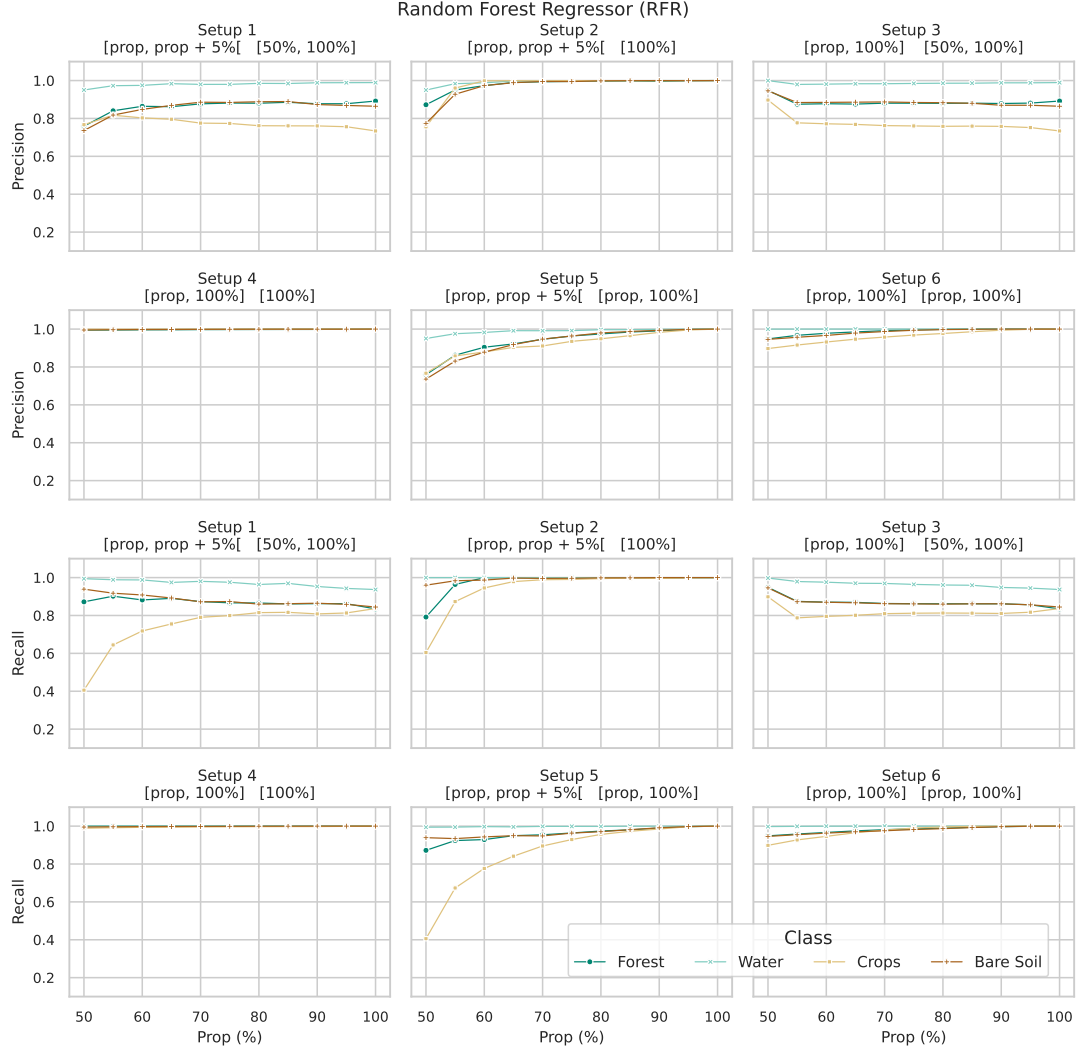


Figure 7.: Completeness: Precision (User Accuracy) on the upper charts and Recall (Producer Accuracy) on the lower charts for each LULC class for Random Forest Classifier for all setups.

5.4.2.2. Bare Soil and Forest. These classes show a similar, counterintuitive pattern. For all classifiers, the Recall (PA) tends to drop as the training data purity increases. This suggests that when models are trained on only pure samples (endmembers), they struggle to correctly identify the full extent of these classes in the real, more mixed landscape, leading to higher omission errors. The inclusion of mixed samples (Setups 5 and 6) helps the models generalise better, resulting in improved Recall.

5.4.2.3. Crops. The Crops class exhibited the most unique and complex behaviour among the land cover types, largely due to its variable spectral characteristics, which place it in a transitional space between Bare Soil and Forest. The analysis of its completeness metrics (Precision and Recall) across different setups revealed that model performance was highly sensitive to the quality of the training data. For instance, all classifiers struggled to correctly identify this class in setups with low-purity training

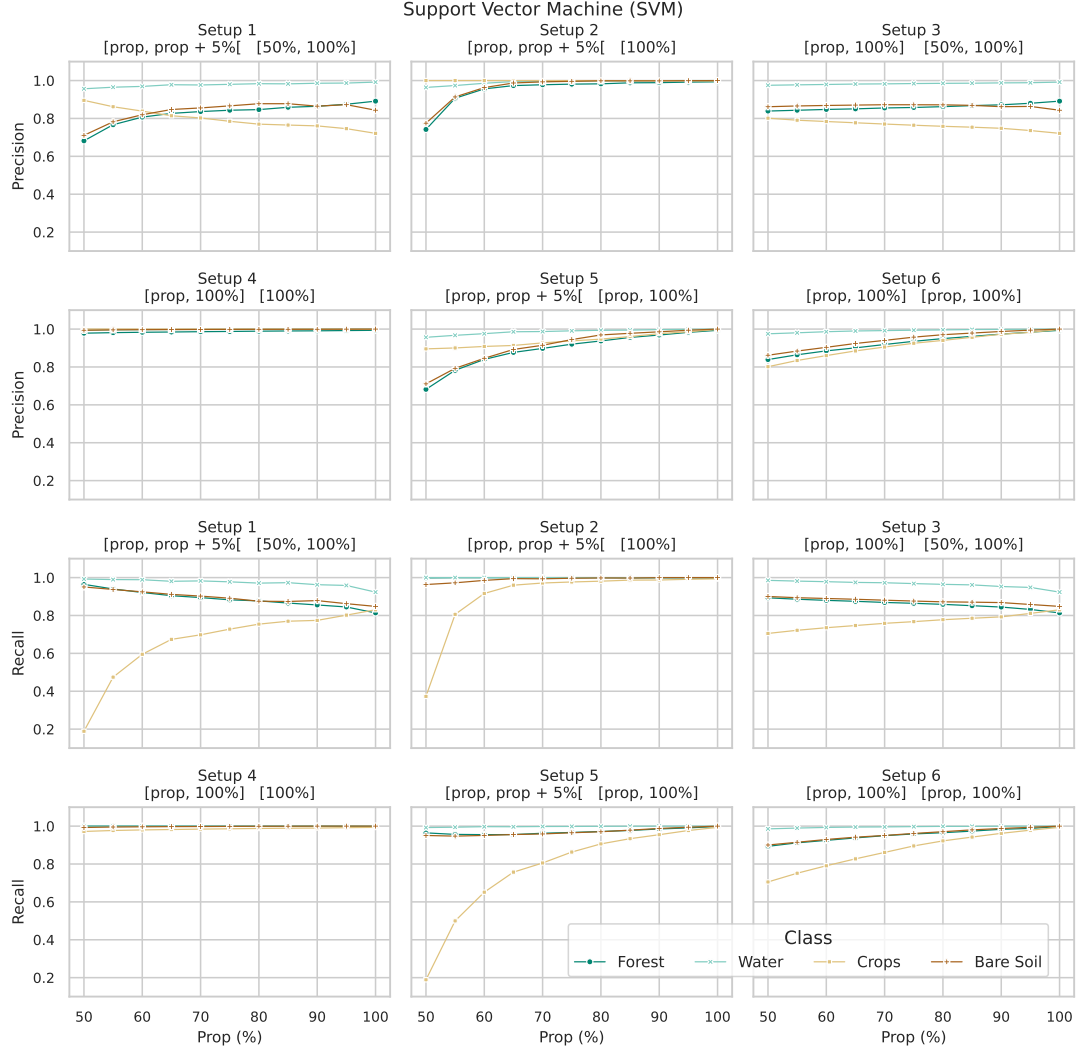


Figure 8.: Completeness: Precision (User Accuracy) on the upper charts and Recall (Producer Accuracy) on the lower charts for each LULC class for Support Vector Machine with Radial Basis Function Classifier for all setups.

data, resulting in high error rates, particularly for Recall. Conversely, in a setup that mimicked real-world data collection (Setup 6), both SVM and Random Forest showed stable and improving completeness metrics. This class performance highlights a critical finding: using a more diverse range of pixel purities in the training data is essential for a model to accurately identify and classify complex, transitional land covers. While some classifiers (like KNN) tended to overfit in ideal scenarios, SVM and Random Forest demonstrated a greater ability to generalise, with their performance improving as the training data became more representative of the varied conditions found in the real world. This suggests that the conventional practice of using only pure pixels may lead to a significant number of false negatives for classes with mixed characteristics.

In summary, the completeness analysis confirms that using only pure samples, while a common recommendation, can lead to increased omission errors for complex and transitional classes. Our results demonstrate that classifiers benefit from training on a more representative dataset that includes a range of pixel purities, as this helps them

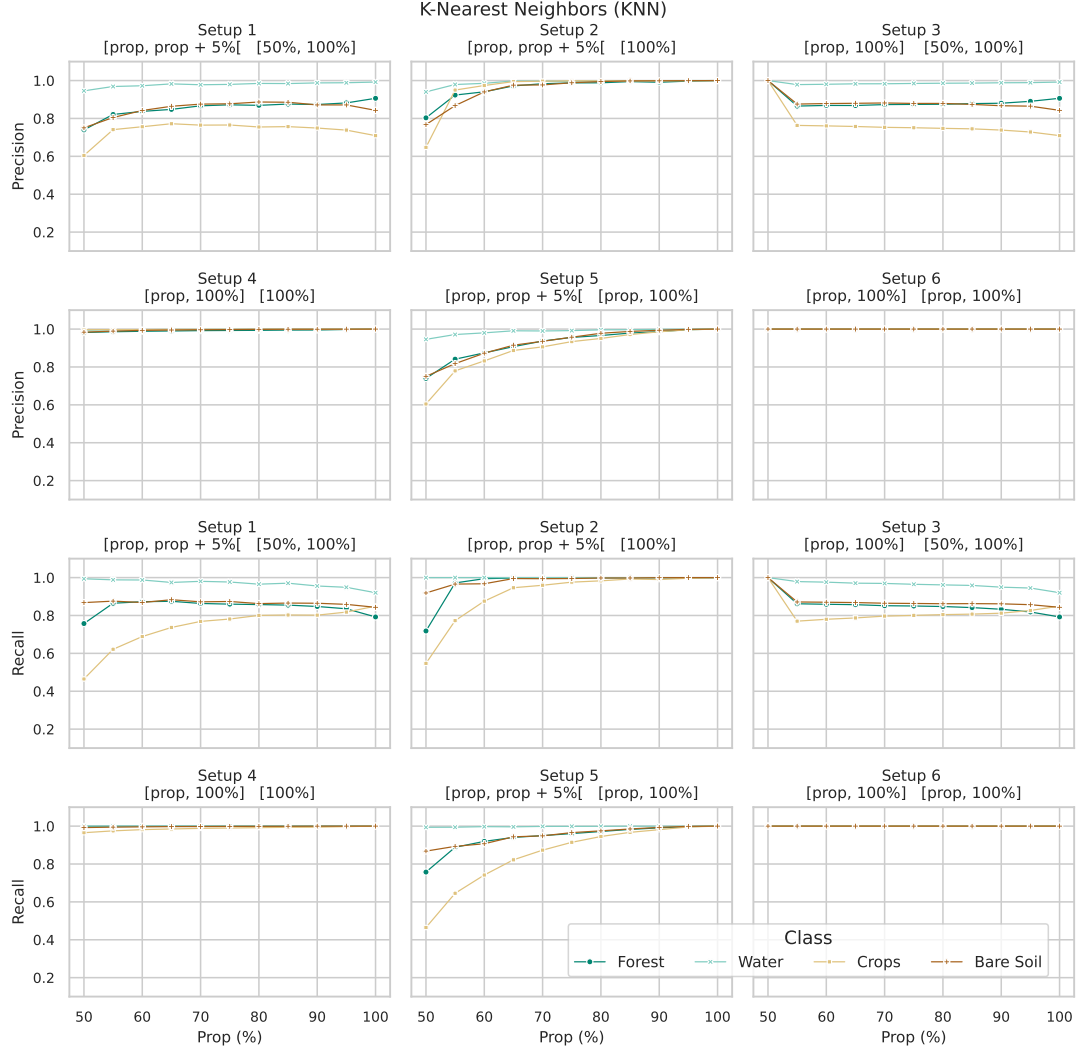


Figure 9.: Completeness: Precision (User Accuracy) on the upper charts and Recall (Producer Accuracy) on the lower charts for each LULC class for K-Nearest Neighbours ($k = 5$) Classifier for all setups.

better generalise to real-world conditions.

6. Discussion

Our study sought to understand how different classifiers, namely Random Forest (RFR), Support Vector Machine (SVM), and K-Nearest Neighbours (KNN), handle variations in the quality of reference samples for training and validation. Through 198 distinct classifications across six setups, we analysed our results using both F1-score (together with Overall Accuracy and Jaccard Index) and completeness (Precision and Recall). A key finding is that thematic accuracy alone, as a single proxy for classification quality, is insufficient. By breaking down the average thematic accuracy into per-class completeness metrics, we were able to clearly observe distinct generalisation patterns among the classes.

Our analysis revealed that the results are best understood by examining the interplay between three perspectives: (i) the classifiers themselves, (ii) the land cover classes, and (iii) the experimental setups.

The Classifiers: Crucially, we cannot analyse classification accuracy based on sample quality alone, as it is intrinsically tied to the classes and classifiers involved. When viewed this way, SVM emerged as more sensitive to changes in pixel purity, particularly when training and validation data qualities diverged. However, SVM was also more stable for Precision and showed a superior ability to classify heterogeneous classes like Crops. Random Forest (RFR), in contrast, was more effective at separating classes like Bare Soil and Forest. KNN, for its part, demonstrated a high dependency on training data quality and struggled with more heterogeneous classes.

Land Cover Classes: Examining the LULC classes themselves showed that their inherent heterogeneity directly affects classifier performance. We observed a clear progression from a spectrally homogeneous class (Water) to two with some variation (Forest and Bare Soil) and finally a highly heterogeneous one (Crops), which lies spectrally between Bare Soil and Forest. The more distinct the class in feature space, the better all classifiers performed. This finding connects to our third point regarding the setups: the quality of the validation data directly influences evaluation metrics.

The Experimental Setups: In scenarios where training and validation data qualities did not align, performance metrics tended to be lower. This confirms that pure pixels do not, on their own, define a class in either spectral or feature space. The results highlight the need to acknowledge and incorporate the inherent heterogeneity of land cover classes into the reference data, moving beyond the limiting practice of using only pure pixels (endmembers).

Therefore, acknowledging the inherent heterogeneity of land cover classes in the reference data (training/test/validation) and the effect of this heterogeneity on these three points may be a game changer in studies of the area. From these results, we can recommend some perspectives for the approach on how it could be best. In sequence, we propose an Operational Guidelines for reproducibility.

6.1. *Perspectives for RSS*

Heterogeneity of Land Cover Transition Classes: The RSS could be helpful to find transitional signatures, once these pixels are highly likely to be more heterogeneous rather than pure. For instance, secondary vegetation may be easier to find compared to primary vegetation by locating different *prop* values.

Relationship with terrain variation: A potential relationship was observed between the Digital Elevation Model (DEM) and the proportion maps, consistent with the findings of [Amin et al. \(2024\)](#). Visual analysis showed clusters of mixed pixels in areas that should theoretically contain only endmembers, specifically where there is significant topographic variation. Although causality cannot be confirmed at this stage, this relationship merits further study.

Large-scale scenarios: Considering the fact the selecting only endmembers can be misleading, further analysis regarding global products, such as ESA WorldCover 2020 ([Zanaga et al. 2022](#)) and ESRI Global Land Cover ([Esri 2024](#); [Karra et al. 2021](#)) are encouraged, not only by assessing best thematic accuracies or using different sampling design; such products and studies may consider pixels heterogeneity with our approach our and/or [Shimabukuro and Ponzoni \(2018\)](#).

Translation of RSS to Deep Learning: The perspective of translating RSS to Deep

Learning is strongly encouraged. We emphasise, though, that the latent space may need to be taken into account, as how each classifier or model architecture deals with different classes does interfere with the results. Shallow-deep models, such as those presented in [Dou et al. \(2024b\)](#), have shown themselves more able to handle large-scale areas.

6.2. Operational Guidelines

To facilitate the reproduction of this study, we sum up the approach into the following operational guidelines. We pinpoint that there is no universal “cake recipe” approach for all geographic areas; therefore, we offer a strategy for analysing class proportions across varying scenarios, using our specific case as a benchmark.

- (1) **Prior knowledge of the study area:** before starting to perform the classification, we suggest that the analyst/researcher understand how the area is designed, the dynamics of the Land Use and Land Cover. Questions such as: what are the primary uses of the land? Is this land mostly urban? How is this land occupation, if any? How is the topography of the terrain?
- (2) **Selection of Imagery:** Ideally, the temporal gap between these images should be minimised to reduce the likelihood of land cover change. Furthermore, prioritise images with low cloud cover to ensure the quality of the subsequent analysis. In this moment, it is also crucial to assess the geometric correction between the imageries (spatial data quality) and minimise such distortions.
- (3) **Initial Visual Pattern Recognition:** Observe the patterns within the study area using a True Colour Imagery (TCI/RGB). In this step, we can observe how the land cover classes are distributed along the imagery. For instance, in [Figure 3](#), it is evident that the majority of the area is forested, interspersed with significant patches of bare soil which, upon closer inspection, reveal traces of crops and sparse vegetation. In this moment, the Land Cover classes shall be defined.
- (4) **Shapes Analysis:** Conduct a preliminary observation of the shapes and boundaries of specific land cover classes to understand their spatial behaviour. If the TCI/RGB imagery proves insufficient, we recommend consulting the Baseline Classification. For unknown regions, [Zanaga et al. \(2022\)](#) serves as a valuable reference; however, a local, up-to-date land cover map is preferable as it more accurately reflects regional vegetation, occupation and/or land-use characteristics. In our example, by observing the classified image (the left-hand panel of [Figure 10](#)), it is evident that Crop areas tend to be smaller and more fragmented, resulting in a higher frequency of boundaries. Conversely, Bare Soil occupies larger contiguous areas, although these are often characterised by numerous internal ‘holes’ or patches of alternative cover. Consequently, we can observe that an abundance of boundaries, coupled with smaller patch sizes, results in lower *prop* values, as illustrated in the right-hand panel of [Figure 10](#).
- (5) **Select of Area of Interest (AOI) / Labelling:** When labelling, prior to selecting only endmembers, the analyst/researcher should select areas that, in fact, represent the land cover class and not aim solely for endmembers, considering the observations made in [Items \(3\) and \(4\)](#). For instance, if bare soils are also presented in roads, which tend to be heterogeneous depending on the spatial resolution of the study, and they are part of the imagery to be studied, then it is worth it to try to label some of them.
- (6) **Fine Resolution Image Classification:** Or Baseline classification, as determined in this study in Phase 1 ([Section 2.1](#)). The researcher/analyst should con-

sider the criteria defined in Section 2.

- (7) **Pixel Purity Distribution Analysis:** Before performing the classification, we strongly recommend that the researcher/analyst understand how the heterogeneity is distributed: spatially and in quantity. Thus, following Phase 2, the generation of charts, such as those in Figure 4 (manuscript), provides a clearer understanding of how proportions (*props*) are distributed across the area in quantity, along some spectral patterns. Another observation is how it is spatially distributed, so we mapped the data in Figure 10 (manuscript). We observe a high variability in *props* for classes other than Water and Forest; this suggests that Bare Soil and Crops may not be adequately represented by endmembers, which we later confirmed in the results.
- (8) **Splitting and Selecting the Candidate Samples:** This is what we call the *bagging* process. The analyst/researcher should take into account the best Setups (Table 1, manuscript) to study. We strongly recommend using all of them for areas in which the researcher/analyst is not yet fully adapted.
- (9) **Classification (Phase 4) and Evaluation (Phase 5):** Perform the coarse resolution (CR) image classification using the selected candidate samples as reference samples and we recommend to analyse (but no limited to) the classifier: (i) the classifiers behaviours; (ii) how the Land Cover classes behave separately; (iii) the variations over the experimental setups. Obviously, these three points intertwine in the RSS approach, but we recommend a focus on each of them.

7. Conclusions and Recommendations

This study’s findings can be summarised in three main observations that challenge conventional approaches to reference data. First, the quality of individual pixels is distinct from their purity; we found that the quality of a pixel is directly related to its representativity in the real world and not only the feature space. This highlights that a dataset of varying pixel purities, which better reflects real-world conditions, can be more valuable for training than a dataset composed solely of “pure” pixels.

Second, a prior understanding of the spectral characteristics of the land cover classes is crucial before entering data into a classification model. This knowledge helps analysts anticipate and mitigate the challenges posed by spectrally heterogeneous classes, which traditional thematic accuracy metrics can easily mask.

Finally, we conclude that completeness metrics like Precision and Recall are a better proxy for classifier assessment than a single, generalised measure like F1-Score or Overall Accuracy, as they provide a more nuanced view of per-class performance and error types.

Our research underscores the importance of a data-centric approach in Remote Sensing, demonstrating its relevance through an experimental framework. The results show that focusing on the class representativity of reference data can lead to more robust and trustworthy models.

We recommend broadening this study to new geographical areas and incorporating different satellite sensors to test the generalisability of our findings. For instance, studying the topographic relationship with pixel heterogeneity or in large-scale datasets.

We also recommend studies considering imbalanced classes, which are more commonly seen in other datasets nowadays.

Additionally, future research should explore how advanced deep learning models would handle these variations in reference data purity, especially in active learning.

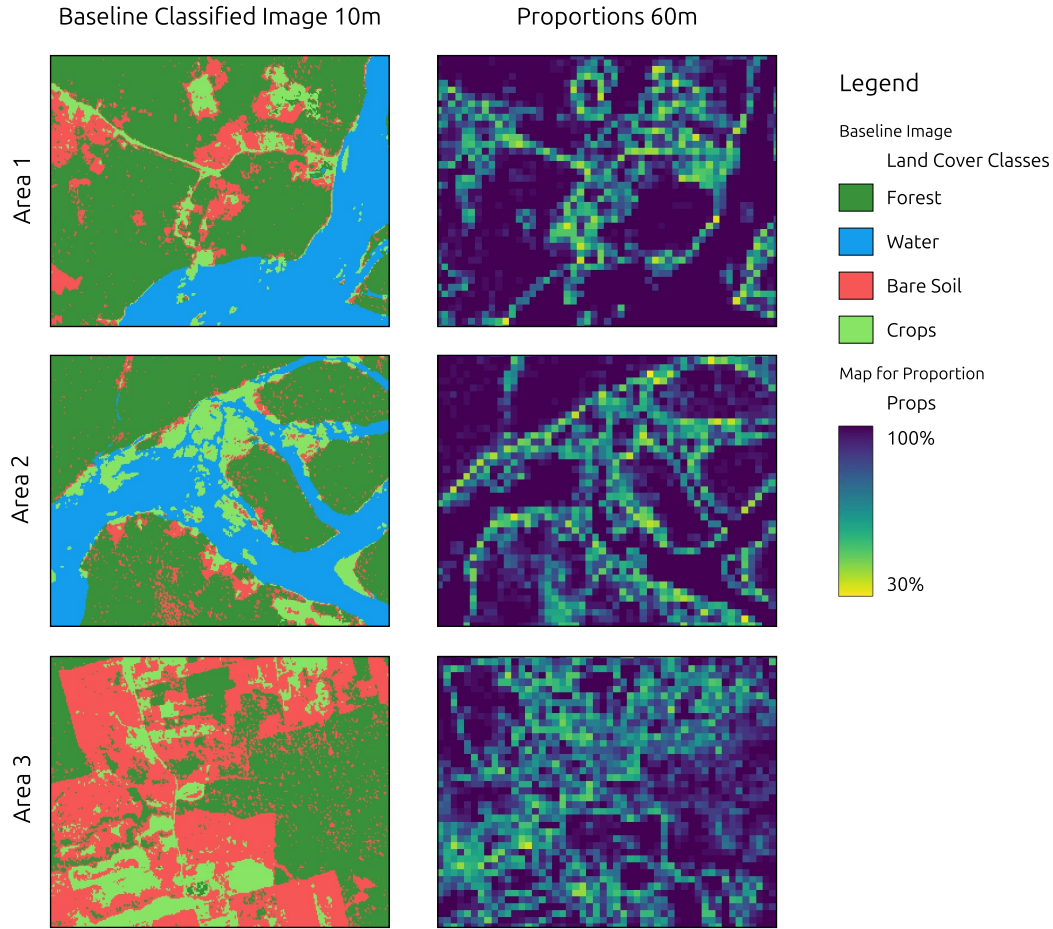


Figure 10.: Visual comparison of three sample areas (Area 1, 2, and 3) as represented by the Baseline Classification (10m resolution) and their corresponding calculated proportions (*prop*) at 60m resolution. **Note:** The greyscale version of this image is available via [Kaggle](#).

8. Data Availability and Accessibility

To encourage researchers to reproduce our study, we share all data and codes in these links⁴:

- Source code: https://github.com/paesleimesa/rss_approach.
- Dataset: <https://www.kaggle.com/datasets/sabrinacorra/uruara-s2msi-samples/>.

The images were adjusted to be colourblind friendly (Brewer 2016), and for those which may present more difficulties to be read, the authors uploaded a Black and White version in the Kaggle repository.

⁴Please contact the corresponding author for questions and/or suggestions.

9. Acknowledgements

The main acknowledgement is to Dr Luciano Vieira Dutra (*in memoriam*), who aided us every step of the way, and we shall never forget his positive impact on the academy and in our lives.

We would also like to express our gratitude to the reviewers who presented very thoughtful insights and suggestions to our study.

This research was financed in part by the Coordination for Higher Education Staff Development – Brazil (CAPES) – Finance Code 001, the National Council of Technological Scientific Development (CNPq) – #140345/2026-4, #141203/2025-0, #302205/2023-3, and #422354/2023-6 (Monitoring and Alerts of Land Cover Changes in Brazilian Biomes – Training and Semi-Automation of the BiomasBR Program), and the São Paulo Research Foundation (FAPESP) #2023/09118-6.

References

- Bhargavi Janga, Gokul Prathin Asamani, Ziheng Sun, and Nicoleta Cristea. A review of practical ai for remote sensing in earth sciences. *Remote Sensing*, 15(16), 2023. ISSN 2072-4292. . URL <https://www.mdpi.com/2072-4292/15/16/4112>.
- Priscilla Azevedo dos Santos, Marcos Adami, Michelle Cristina Araujo Picoli, Victor Hugo Roden Prudente, Júlio César Dalla Mora Esquerdo, Gilberto Ribeiro de Queiroz, Cleverton Tiago Carneiro de Santana, and Michel Eustáquio Dantas Chaves. Land use and land cover products for agricultural mapping applications in brazil: Challenges and limitations. *Remote Sensing*, 17(13), 2025. ISSN 2072-4292. . URL <https://www.mdpi.com/2072-4292/17/13/2324>.
- INPE RESTORE+ Project. Land use and land cover maps for the brazilian amazonia biome: 2000-2024, 2025. URL <https://restore-plus.github.io/amz-maps/>. Part of the LUC-Brazil initiative produced with support from the RESTORE+ project at the request of Brazil’s Ministry of Science and Technology, General Coordination for Climate Science (GCCL). Primary purpose is to support Brazil’s National Inventory of Greenhouse Gases Emissions.
- Carlos M. Souza, Julia Z. Shimbo, Marcos R. Rosa, Leandro L. Parente, Ane A. Alencar, Bernardo F. T. Rudorff, Heinrich Hasenack, Marcelo Matsumoto, Laerte G. Ferreira, Pedro W. M. Souza-Filho, Sergio W. de Oliveira, Washington F. Rocha, Antônio V. Fonseca, Camila B. Marques, Cesar G. Diniz, Diego Costa, Dyeden Monteiro, Eduardo R. Rosa, Eduardo Vélez-Martin, Eliseu J. Weber, Felipe E. B. Lenti, Fernando F. Paternost, Frans G. C. Pareyn, João V. Siqueira, José L. Viera, Luiz C. Ferreira Neto, Marciano M. Saraiva, Marcio H. Sales, Moises P. G. Salgado, Rodrigo Vasconcelos, Soltan Galano, Vinicius V. Mesquita, and Tasso Azevedo. Reconstructing three decades of land use and land cover changes in brazilian biomes with landsat archive and earth engine. *Remote Sensing*, 12(2735), 2020. ISSN 2072-4292. .
- MapBiomas Project. *Algorithm Theoretical Basis Document (ATBD). General Document Collection 10.0*. MapBiomas, 2025. URL <https://brasil.mapbiomas.org/en/atbd/>. Accessed through the link: <https://doi.org/10.58053/MapBiomas/2XSLB3>.
- Raziye Hale Topaloğlu, Elif Sertel, and Nebiye Musaoğlu. ASSESSMENT OF CLASSIFICATION ACCURACIES OF SENTINEL-2 AND LANDSAT-8 DATA FOR LAND COVER / USE MAPPING. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, XLI-B8:1055–1059, June 2016. ISSN 1682-1750. . URL <https://isprs-archives.copernicus.org/articles/XLI-B8/1055/2016/isprs-archives-XLI-B8-1055-2016.html>. Conference Name: XXIII ISPRS Congress, Commission VIII (Volume XLI-B8) - 12–19 July 2016, Prague, Czech Republic Publisher: Copernicus GmbH.

- Mariane S. Reis, Maria Isabel S. Escada, Luciano V. Dutra, Sidnei J. S. Sant'Anna, and Nathan D. Vogt. Towards a Reproducible LULC Hierarchical Class Legend for Use in the Southwest of Pará State, Brazil: A Comparison with Remote Sensing Data-Driven Hierarchies. *Land*, 7(2):65, June 2018. ISSN 2073-445X. .
- Ioannis Papoutsis, Nikolaos Ioannis Bountos, Angelos Zavras, Dimitrios Michail, and Christos Tryfonopoulos. Benchmarking and scaling of deep learning models for land cover image classification. *ISPRS Journal of Photogrammetry and Remote Sensing*, 195:250–268, January 2023. ISSN 0924-2716. . URL <https://www.sciencedirect.com/science/article/pii/S0924271622003057>.
- Yashon O Ouma, Amantle Keitsile, Boipuso Nkwae, Phillimon Odirile, Ditiro Moalafhi, and Jiaguo Qi. Urban land-use classification using machine learning classifiers: comparative evaluation and post-classification multi-feature fusion approach. *European Journal of Remote Sensing*, 56(1):2173659, 2023.
- Daniel Moraes, Manuel L Campagnolo, and Mário Caetano. Training data in satellite image classification for land cover mapping: a review. *European Journal of Remote Sensing*, 57(1): 2341414, 2024.
- Yingxin Gu, Bruce K. Wylie, Stephen P. Boyte, Joshua Picotte, Daniel M. Howard, Kelcy Smith, and Kurtis J. Nelson. An optimal sample data usage strategy to minimize overfitting and underfitting effects in regression tree models based on remotely-sensed data. *Remote Sensing*, 8(11), 2016. ISSN 2072-4292. . URL <https://www.mdpi.com/2072-4292/8/11/943>.
- Mustafa Ustuner and Fatih Fehmi Simsek. An assessment of training data for agricultural land cover classification: a case study of bafra, türkiye. *Earth Science Informatics*, 18(1):7, 2025.
- Giles M. Foody. Ground Truth in Classification Accuracy Assessment: Myth and Reality. *Geomatics*, 4(1):81–90, March 2024. ISSN 2673-7418. . URL <https://www.mdpi.com/2673-7418/4/1/5>. Number: 1 Publisher: Multidisciplinary Digital Publishing Institute.
- Stephen V. Stehman and Giles M. Foody. Key issues in rigorous accuracy assessment of land cover products. *Remote Sensing of Environment*, 231:111199, September 2019. ISSN 0034-4257. . URL <https://www.sciencedirect.com/science/article/pii/S0034425719302111>.
- Pontus Olofsson, Giles M. Foody, Martin Herold, Stephen V. Stehman, Curtis E. Woodcock, and Michael A. Wulder. Good practices for estimating area and assessing accuracy of land change. *Remote Sensing of Environment*, 148:42–57, May 2014. ISSN 00344257. . URL <https://linkinghub.elsevier.com/retrieve/pii/S0034425714000704>.
- Christopher M. Bishop and Hugh Bishop. *Deep Learning: Foundations and Concepts*. Springer International Publishing, Cham, 2024. ISBN 978-3-031-45467-7 978-3-031-45468-4. . URL <https://link.springer.com/10.1007/978-3-031-45468-4>.
- Taskin Kavzoglu, Brandt Tso, and Paul M. Mather. *Classification Methods for Remotely Sensed Data*. CRC Press, Boca Raton, 3 edition, July 2024. ISBN 978-1-003-43917-2. . URL <https://www.taylorfrancis.com/books/9781003439172>.
- Yosio Edemir Shimabukuro and Flávio Jorge Ponzoni. *Spectral mixture for remote sensing: Linear model and applications*. Springer, 2018.
- Paul M. Mather and Magaly Koch. *Computer Processing of Remotely-Sensed Images: An Introduction*. Wiley, 3. ed., repr edition, 2004. ISBN 978-0-470-84918-7 978-0-470-84919-4.
- Peter Fisher. The pixel: a snare and a delusion. *International Journal of Remote Sensing*, 18(3):679–685, 1997.
- Taskin Kavzoglu. Increasing the accuracy of neural network classification using refined training data. *Environmental Modelling & Software*, 24(7):850–858, 2009.
- Arthur Elmes, Hamed Alemohammad, Ryan Avery, Kelly Caylor, J. Ronald Eastman, Lewis Fishgold, Mark A. Friedl, Meha Jain, Divyani Kohli, Juan Carlos Laso Bayas, Dalton Lunga, Jessica L. McCarty, Robert Gilmore Pontius, Andrew B. Reinmann, John Rogan, Lei Song, Hristiana Stoyanova, Su Ye, Zhuang-Fang Yi, and Lyndon Estes. Accounting for Training Data Error in Machine Learning Applied to Earth Observations. *Remote Sensing*, 12(6):

- 1034, January 2020. ISSN 2072-4292. . URL <https://www.mdpi.com/2072-4292/12/6/1034>. Number: 6 Publisher: Multidisciplinary Digital Publishing Institute.
- Pi-Fuei Hsieh, Lou C Lee, and Nai-Yu Chen. Effect of spatial resolution on classification errors of pure and mixed pixels in remote sensing. *IEEE Transactions on Geoscience and Remote Sensing*, 39(12):2657–2663, 2001.
- Peng Dou, Chunlin Huang, Weixiao Han, Jinliang Hou, and Ying Zhang. Time series remote sensing image classification using feature relationship learning. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–13, 2024a.
- Ribana Roscher, Marc Russwurm, Caroline Gevaert, Michael Kampffmeyer, Jefersson A. Dos Santos, Maria Vakalopoulou, Ronny Hänsch, Stine Hansen, Keiller Nogueira, Jonathan Prexl, and Devis Tuia. Better, Not Just More: Data-centric machine learning for Earth observation. *IEEE Geoscience and Remote Sensing Magazine*, 12(4):335–355, December 2024. ISSN 2168-6831, 2473-2397, 2373-7468. .
- Sean Gillies. Rasterio documentation. *MapBox: San Francisco, CA, USA*, 23, 2019.
- Stephen V. Stehman. Practical Implications of Design-Based Sampling Inference for Thematic Map Accuracy Assessment. *Remote Sensing of Environment*, 72(1):35–45, April 2000. ISSN 00344257. . URL <https://linkinghub.elsevier.com/retrieve/pii/S0034425799000905>.
- Russell G Congalton. Thematic and positional accuracy assessment of digital remotely sensed data. In *In: McRoberts, Ronald E.; Reams, Gregory A.; Van Deusen, Paul C.; McWilliams, William H., eds. Proceedings of the seventh annual forest inventory and analysis symposium; October 3-6, 2005; Portland, ME. Gen. Tech. Rep. WO-77. Washington, DC: US Department of Agriculture, Forest Service: 149-154.*, volume 77, 2007.
- Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning*. Springer Series in Statistics. Springer, New York, NY, 2009. ISBN 978-0-387-84857-0 978-0-387-84858-7. . URL <http://link.springer.com/10.1007/978-0-387-84858-7>.
- Qing-Song Xu and Yi-Zeng Liang. Monte Carlo cross validation. *Chemometrics and Intelligent Laboratory Systems*, 56(1):1–11, April 2001. ISSN 0169-7439. . URL <https://www.sciencedirect.com/science/article/pii/S0169743900001222>.
- Peijun Du, Xuyu Bai, Kun Tan, Zhaohui Xue, Alim Samat, Junshi Xia, Erzhu Li, Hongjun Su, and Wei Liu. Advances of four machine learning methods for spatial data handling: A review. *Journal of Geovisualization and Spatial Analysis*, 4(1):13, 2020.
- Oliver Kramer. K-Nearest Neighbors. In Oliver Kramer, editor, *Dimensionality Reduction with Unsupervised Nearest Neighbors*, pages 13–23. Springer, Berlin, Heidelberg, 2013. ISBN 978-3-642-38652-7. . URL https://doi.org/10.1007/978-3-642-38652-7_2.
- Christopher M. Bishop. *Pattern Recognition and Machine Learning*. Information Science and Statistics. Springer, New York, 2006. ISBN 978-0-387-31073-2.
- Ürün Doğan, Tobias Glasmachers, and Christian Igel. A unified view on multi-class support vector classification. *The Journal of Machine Learning Research*, 17(1):1550–1831, 2016.
- Leo Breiman. Random Forests. *Machine Learning*, 45(1):5–32, October 2001. ISSN 1573-0565. . URL <https://doi.org/10.1023/A:1010933404324>.
- Mariana Belgiu and Lucian Drăguț. Random forest in remote sensing: A review of applications and future directions. *ISPRS Journal of Photogrammetry and Remote Sensing*, 114:24–31, 2016. ISSN 0924-2716. .
- Autumn Nguyen and Sulagna Saha. Machine learning and multi-source remote sensing in forest carbon stock estimation: A review. *arXiv preprint arXiv:2411.17624*, 2024.
- ISO. Geographic information — Data quality Part 1: General requirements, 2023.
- Russell G. Congalton. A review of assessing the accuracy of classifications of remotely sensed data. *Remote Sensing of Environment*, 37(1):35–46, July 1991. ISSN 00344257. . URL <https://linkinghub.elsevier.com/retrieve/pii/S003442579190048B>.
- C Van Rijsbergen. Information retrieval: theory and practice. In *Proceedings of the joint IBM/University of Newcastle upon tyne seminar on data base systems*, volume 79, pages 1–14, 1979.
- Paul Jaccard. The Distribution of the Flora in the Alpine Zone. *New Phytologist*, 11(2):37–50,

1912. ISSN 1469-8137. .
- Aaron E Maxwell, Timothy A Warner, and Luis Andres Guillen. Accuracy assessment in convolutional neural network-based deep learning remote sensing studies—part 2: Recommendations and best practices. *Remote Sensing*, 13(13):2591, 2021.
- Mariane Souza Reis, Luciano Vieira Dutra, Maria Isabel Sobral Escada, and Sidnei João Siqueira Sant’anna. Avoiding Invalid Transitions in Land Cover Trajectory Classification With a Compound Maximum a Posteriori Approach. *IEEE Access*, 8:98787–98799, 2020. ISSN 2169-3536. .
- John R. Jensen. *Remote sensing of the environment: an earth resource perspective*. Pearson, Harlow, England, pearson new international, second edition edition, 2014. ISBN 978-1-292-02170-6 978-1-292-03493-5.
- Thomas M. Lillesand, Ralph W. Kiefer, and Jonathan W. Chipman. *Remote sensing and image interpretation*. Wiley, Hoboken, NJ, 6. ed edition, 2008. ISBN 978-0-470-05245-7.
- European Space Agency. S2 Mission. <https://sentiwiki.copernicus.eu/web/s2-mission>, 2025.
- Manfred Ehlers, Ronald Janowsky, and Monika Gaehler. New remote sensing concepts for environmental monitoring. In *Remote Sensing for Environmental Monitoring, GIS Applications, and Geology*, volume 4545, pages 1–12. SPIE, 2002.
- Stephen V. Stehman. Sampling designs for accuracy assessment of land cover. *International Journal of Remote Sensing*, 30(20):5243–5272, September 2009. ISSN 0143-1161, 1366-5901. . URL <https://www.tandfonline.com/doi/full/10.1080/01431160903131000>.
- Phan Thanh Noi and Martin Kappas. Comparison of random forest, k-nearest neighbor, and support vector machine classifiers for land cover classification using sentinel-2 imagery. *Sensors*, 18(1):18, 2017.
- Gomal Amin, Iqra Imtiaz, Ehsan Haroon, Najum us Saqib, Muhammad Imran Shahzad, and Majid Nazeer. Assessment of machine learning algorithms for land cover classification in a complex mountainous landscape. *Journal of Geovisualization and Spatial Analysis*, 8(2):34, 2024.
- Daniele Zanaga, Ruben Van De Kerchove, Dirk Daems, Wanda De Keersmaecker, Carsten Brockmann, Grit Kirches, Jan Wevers, Oliver Cartus, Maurizio Santoro, Steffen Fritz, Myroslava Lesiv, Martin Herold, Nandin-Erdene Tsendbazar, Panpan Xu, Fabrizio Ramoino, and Olivier Arino. Esa worldcover 10 m 2021 v200, October 2022. URL <https://doi.org/10.5281/zenodo.7254221>.
- Esri. Sentinel-2 land cover explorer. <https://livingatlas.arcgis.com/landcover/>, 2024. Accessed: 2025-05-22.
- Krishna Karra, Caitlin Kontgis, Zoe Statman-Weil, Joseph C Mazzariello, Mark Mathis, and Steven P Brumby. Global land use/land cover with sentinel-2 and deep learning. In *2021 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, pages 4704–4707. IEEE, 2021. .
- Peng Dou, Huanfeng Shen, Chunlin Huang, Zhiwei Li, Yujun Mao, and Xinghua Li. Large-scale land use/land cover extraction from landsat imagery using feature relationships matrix based deep-shallow learning. *International Journal of Applied Earth Observation and Geoinformation*, 129:103866, 2024b.
- Cynthia Brewer. *Designing Better Maps: A Guide for GIS users*. ESRI press, 2 edition, 2016. ISBN 978-1589484405.