UCLouvain

AGRO LOUVAIN

Faculté des bioingénieurs

Evaluating Multi-Sensor Remote Sensing for Parcel-Level Mapping of Milpa Mixed-Cropping and Maize Monocrop Fields in Oaxaca, Mexico

Auteur :	Md Huraira Rabbi Shahriar
Promoteur(s) :	Pierre Defourny; Gerald Blasch
Co-superviseur :	Guillaume Jadot
Lecteur(s) :	Sophie Bontemps; Bruno Gérard
Année académique :	2025-2026

Mémoire de fin d'études présenté en vue de l'obtention du diplôme de
Master SAIV - MSc Erasmus+ GEM: Geo-information Science and Earth Observation for Environmental Modelling and Management

Acknowledgements

I would like to express my sincere gratitude to my supervisors, Prof. Dr. Pierre Defourny and Dr. Gerald Blasch, for their guidance, advice and support throughout this thesis. Their comments helped shape the research direction, strengthen the methodological design and improve the interpretation of the results.

I am also deeply grateful to Guillaume Jadot, my co-supervisor, for his continuous support, careful feedback and practical guidance during the data-processing, analysis and writing stages of this work.

I would also like to express my sincere gratitude to the jury for taking the time to read this work. I hope that they will find the study engaging and worthwhile.

I would like to thank the academic staff, researchers and colleagues who contributed directly or indirectly through discussions, technical assistance and constructive suggestions. Their input helped improve the quality and clarity of this thesis.

Finally, I am sincerely thankful to my family and friends for their patience, encouragement and support throughout this demanding period.

Table of Contents

List of Figures

List of Tables

Contributions

Stage	Md. Huraira Rabbi Shahriar	Prof. Pierre Defourny, Guillaume Jadot and Dr. Gerald Blasch	Artificial Intelligence
Conceptualisation	✓	✓	
Methodology	✓	✓	Programming support during implementation, including selected Python code drafting, debugging and consistency checks.
Investigation	✓		
Formal analysis	✓	✓	Python script drafting and debugging support. Analyses were executed, checked and interpreted by the author.
Validation	✓	✓	
Writing – original draft	✓		Language polishing and grammatical refinement after the author prepared the original draft.
Writing – review & editing	✓	✓	Support with language refinement and consistency checks during revision.
Visualisation	✓		Assistance with Python plotting scripts and figure-layout refinement
Supervision		✓	

Explanatory statement

The research objectives and methodological framework were developed by the author through review of scientific literature and discussions with the supervisors. The author conducted the investigation, executed and checked the analyses, interpreted the findings and prepared the original manuscript drafts. Supervisory contributions included guidance on research design, discussion of analytical decisions, scientific validation and manuscript review. Maps and spatial visualisations were produced by the author using ArcGIS and ArcPy, with Python used for selected plots and graphical outputs. Generative artificial intelligence was used selectively for programming support, code debugging, language refinement, consistency checking and limited figure-layout assistance. All scientific decisions, reported results and final thesis content were reviewed and approved by the author.

✓ indicates a substantive contribution to the corresponding stage.

List of Abbreviations and Acronyms

Acronym	Meaning
AOI	Area of interest
BOA	Bottom of atmosphere
CC BY	Creative Commons Attribution
CI	Chlorophyll Index
CNN	Convolutional neural network
CNN-SR	Convolutional neural network super-resolution
DEM	Digital elevation model
EDSR	Enhanced Deep Residual Network for Single Image Super-Resolution
ESA	European Space Agency
EVI2	Enhanced Vegetation Index 2
F1	F1-score
FN	False negative
FP	False positive
IQR	Interquartile range
MAE	Mean absolute error
MSI	Multispectral Instrument
NDMI	Normalised Difference Moisture Index
NDRE	Normalised Difference Red Edge Index
NDVI	Normalised Difference Vegetation Index
NIR	Near infrared
OOF	Out-of-fold
PSRI	Plant Senescence Reflectance Index
RBF	Radial basis function
RF	Random Forest
RGB	Red, green and blue
RMSE	Root mean squared error
S2	Sentinel-2
SR	Super-resolution
SRTM	Shuttle Radar Topography Mission
SRTMGL1	Shuttle Radar Topography Mission Global 1 arc-second
SVM	Support Vector Machine
SWIR	Shortwave infrared
TN	True negative

Acronym	Meaning
TP	True positive
UAV	Unmanned aerial vehicle
UDM2	Usable Data Mask 2
UDM2.1	Usable Data Mask 2.1
UTM	Universal Transverse Mercator
WGS	World Geodetic System

Chapter 1: Introduction

Agricultural maps usually assign each field to its dominant crop, but this is not always sufficient in smallholder landscapes where several crops may be cultivated together. Milpa is a family of maize-centred mixed-cropping systems whose crop composition, planting arrangement and management vary across Mesoamerica. Maize may remain visually dominant even when companion crops are present, so a Milpa field can resemble maize monocrop while still representing a different cropping system. Milpa is also associated with agrobiodiversity, local agricultural knowledge and cultural practices, although these functions vary among communities (Fonteyne et al., 2023; López-Ridaura et al., 2021; Benrey et al., 2024).

Remote sensing can complement field surveys by providing repeated observations over wider areas, but satellite imagery does not directly record crop-system identity. The signal within a parcel can combine maize, companion crops, weeds, exposed soil, residues and boundary features, and the proportions of these components change through the season. Small and irregular fields further increase mixed-pixel effects, while variation among Milpa fields can be substantial. These characteristics make parcel-level discrimination from maize monocrop a problem of both class definition and spatial-spectral representation (Persello et al., 2019; Ibrahim et al., 2021).

Sentinel-2 provides a useful seasonal baseline because its repeated visible, red-edge, near-infrared and shortwave-infrared observations can describe crop development through time. Its value, however, depends on the quality and temporal distribution of the observations: clouds, shadows and other invalid pixels create gaps, and reconstructed values are not equivalent to measurements recorded directly by the sensor. Spatial support is also important because the 10–20 m agricultural bands may contain few interior pixels in narrow fields, and relationships learned in one area may weaken when transferred geographically (Defourny et al., 2019; Karasiak et al., 2022; Orynbaikyzy et al., 2022). SkySat and PlanetScope offer finer spatial sampling that may provide complementary field information, but greater spatial detail does not automatically yield better classification (Rao et al., 2021).

Evidence remains limited for parcel-level remote-sensing discrimination of the Milpa mixed-cropping system from maize monocrop in Oaxaca. This thesis therefore evaluates Sentinel-2 as the principal seasonal baseline and asks whether complementary information from SkySat and PlanetScope improves that distinction. Three routes are examined separately: direct sensor predictors and predictor-level integration, SkySat-guided Sentinel-2 super-resolution, and classification-level integration of independently fitted sensor models. The primary formulation distinguishes Milpa from maize monocrop, while a secondary three-class formulation retains Milpa, landrace maize and hybrid maize. Model development is separated from a geographically distinct secondary benchmark, and class-specific Milpa performance is considered alongside aggregate measures. The overall aim is to assess whether these complementary routes provide additional, geographically transferable discrimination beyond the Sentinel-2 parcel baseline.

Chapter 2: Literature Review

This chapter reviews the literature needed to understand and evaluate the remote-sensing-based mapping of the Milpa mixed-cropping system in Oaxaca. It first defines Milpa as a crop-system mapping target and examines the challenges created by small, heterogeneous agricultural parcels. It then considers Sentinel-2 observation reliability, spectral-temporal and spatial feature representation, machine-learning classification and geographic evaluation. The review subsequently assesses the potential contribution of SkySat and PlanetScope imagery, predictor-level integration, cross-sensor super-resolution and classification-level integration. The chapter concludes by synthesising the remaining evidence gaps and positioning the contribution of the present study.

2.1 Milpa as a crop-system mapping target

2.1.1 Cropping-system terminology

Intercropping is generally defined as the cultivation of two or more crop species in the same field simultaneously during overlapping periods of growth. Mixed cropping is often used to denote crops grown together without any fixed spatial arrangement, though the differentiation between these terms varies across the literature (Brooker et al., 2015; Maitra et al., 2021).

Milpa does not have a fixed composition of crops nor a fixed arrangement of planting. Maize can be cultivated with various companion crops and the combinations of these vary across locations and farming systems (López-Ridaura et al., 2021; Fonteyne et al., 2023). For simplicity, the term “Milpa mixed-cropping system” is used here without implying a fixed set of companion crops or a uniform planting geometry. In the terminology used in this study, landrace and hybrid refer to maize types, whereas maize monocrop refers to the cropping system in which maize is grown without the mixed-cropping structure of Milpa.

2.1.2 Composition and variability of the Milpa mixed-cropping system

Milpa is a traditional Mesoamerican maize-based cropping system. The term originates from the Nahuatl words "milli", meaning a planted field, and "pan", meaning upon (Vazeux-Blumental et al., 2024). Although Milpa is commonly associated with maize, beans and squash, this classic association represents only one of several documented crop combinations (Vazeux-Blumental et al., 2024; López-Ridaura et al., 2021).

The commonly cited maize–bean–squash association is illustrated in Figure 2.1.

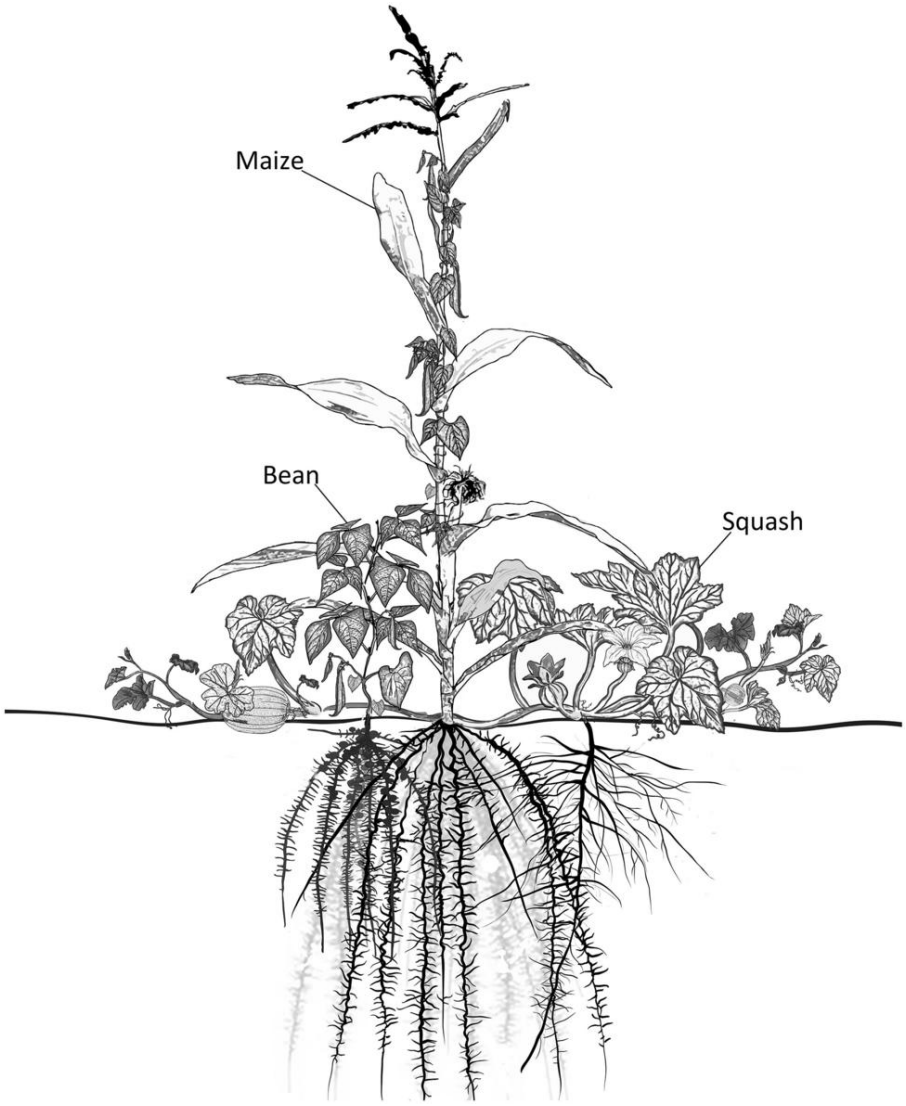


Figure 2.1. The classic Milpa association of maize, bean and squash. Source: López-Ridaura et al. (2021, Fig. 1); illustration by Melinda Ridaura. Reproduced under the Creative Commons Attribution 4.0 International licence (CC BY 4.0).

Across Mesoamerica, Milpa includes a wider range of maize-centred crop associations. Documented companion crops include beans, squash, faba bean, potato, chilli, tomato, amaranth, fruit trees, coffee and other vegetables (López-Ridaura et al., 2021). Their composition varies across locations and has been associated with agroecological conditions such as climate, soils and topography, as well as farmers' needs, knowledge and traditions (López-Ridaura et al., 2021).

In Chinantec agroecosystems in Oaxaca, Mateos-Maces et al. (2016) reported variation in crop combinations, moisture regimes for planting, and the timing of sowing, flowering and harvesting. In the Western Highlands of Guatemala, López-Ridaura et al. (2021) documented several crop combinations across 1,205 maize plots cultivated by 989 households, with maize–bean associations more frequent than the complete maize–bean–squash combination. These studies show that Milpa varies considerably in crop composition and growing conditions rather than representing a single uniform crop mixture.

2.1.3 Agronomic, nutritional and socio-cultural importance

Several crops grown together can occupy the field differently and use light, water, nutrients and soil space in complementary ways (Brooker et al., 2015; Fonteyne et al., 2023; Vazeux-Blumental et al., 2024). Maize can support climbing beans, legumes may contribute nitrogen through biological fixation, and lower-growing crops such as squash can shade the soil surface, suppress some weeds and conserve soil moisture (Fonteyne et al., 2023). These interactions can improve system productivity, although their effects are context-dependent and should not be treated as a guarantee of higher maize yield (Brooker et al., 2015; Fonteyne et al., 2023; Vazeux-Blumental et al., 2024).

Milpa can also diversify household food production and contribute to nutritional self-sufficiency (Novotny et al., 2021). In two highland communities in Oaxaca, Novotny et al. (2021) reported that Milpa produced more food per unit area than sole maize or bean fields, while nutritional self-sufficiency varied between communities. One hectare of Milpa could meet the assessed requirements of at least two people per year in Santa Catarina Tayata and one person in San Cristóbal Amoltepec when limiting nutrients were considered (Novotny et al., 2021). Vitamin B12 was not supplied by crop production, while zinc and vitamins A, B9 and C were among the limiting nutrients (Novotny et al., 2021). The study also found sole maize to be more labour-efficient and reported farmers' concerns about the high labour demand of Milpa (Novotny et al., 2021).

Its importance extends beyond production and nutrition. Milpa is linked to agrobiodiversity and locally maintained crop varieties (Benrey et al., 2024; Fonteyne et al., 2023). It is also associated with traditional agricultural knowledge and, in some contexts, cultural identity (Fonteyne et al., 2023; Romero-Natale et al., 2024). These relationships vary across communities, while socioeconomic conditions and labour requirements can influence farmers' choices (Fonteyne et al., 2023; Novotny et al., 2021). Milpa therefore has importance beyond maize yield alone, reflecting its production, nutritional, ecological and socio-cultural functions.

The associated above- and below-ground ecological relationships are illustrated conceptually in Figure 2.2.

Figure 2.2. Conceptual illustration of the Milpa mixed-cropping system, showing maize, beans and squash together with associated above- and below-ground ecological components. Source: Benrey et al. (2024, Fig. 1). The source notes that the figure was partially created with DALL·E. Reproduced under the Creative Commons Attribution 4.0 International licence (CC BY 4.0).

2.1.4 Implications for remote-sensing class definition

The same diversity that gives Milpa agricultural importance makes it difficult to represent as one homogeneous remote-sensing class (Fonteyne et al., 2023; Ibrahim et al., 2021; Persello et al., 2019). A Milpa signal may contain different contributions from maize, companion crops, exposed soil and weeds, and these contributions can vary with crop development and spatial arrangement (Fonteyne et al., 2023; Bégué et al., 2018).

A maize-dominated part of a parcel may resemble maize monocropping, while another part may contain stronger contributions from companion crops or exposed soil (Fonteyne et al., 2023; Ibrahim et al., 2021). Variation can occur both between Milpa parcels and within a single parcel, so one universal spectral signature should not be expected (Fonteyne et al., 2023; Ibrahim et al., 2021; Bégué et al., 2018). Differences used to distinguish Milpa from maize monocropping therefore need to be evaluated in the local context rather than assumed to apply across regions (Bégué et al., 2018; Ibrahim et al., 2021).

Comparable work provides context without supplying direct evidence for Oaxaca. Ibrahim et al. (2021) used Sentinel-2 time-series imagery to map crop types and cropping systems in

Nigeria, including maize-based mixed-cropping classes. The study showed that mixed systems can be represented as explicit mapping classes, although spectral similarity and temporal heterogeneity can make separation difficult. Because the crops, environment and management differ, the study cannot determine how Milpa will appear in Oaxaca. It nevertheless supports a class definition that separates Milpa from maize monocropping while allowing genuine differences among Milpa parcels (Ibrahim et al., 2021; Fonteyne et al., 2023).

2.2 Remote sensing of smallholder crop systems

2.2.1 Remote sensing and agricultural monitoring

Agricultural monitoring requires information on where crops are grown, how they develop through the season and how cultivated areas change over time (Weiss et al., 2020; Wu et al., 2023). Field observations and other reference data remain important because they provide the information needed to identify crop classes and interpret management conditions (Bégué et al., 2018; Wu et al., 2023). Remote sensing complements these observations by providing repeated coverage over broad areas, making it useful for crop mapping, land-use monitoring, yield forecasting, irrigation assessment and food-security applications (Karthikeyan et al., 2020; Weiss et al., 2020; Wu et al., 2023).

Satellite imagery does not provide crop labels directly. Instead, crop identity is inferred from measurements recorded within image pixels, while the amount of ground detail represented depends on the spatial resolution of the sensor (Bégué et al., 2018; Weiss et al., 2020). Suitable reference data are therefore still needed for crop classification. Cloud cover, irregular temporal coverage and limited spatial detail can also reduce the information available for monitoring and make interpretation more difficult (Bégué et al., 2018; Weiss et al., 2020; Wu et al., 2023).

2.2.2 Crop-type versus crop-system mapping

Crop type is a common target in agricultural remote-sensing studies, but crop identity alone does not describe all aspects of a cropping system (Ibrahim et al., 2021; Tariq et al., 2023). Cropping systems can also differ in the way crops are combined, arranged or grown in sequence, as well as in their management (Bégué et al., 2018). Mapping such systems therefore requires classes that represent these broader characteristics rather than crop type alone.

This distinction is particularly relevant to Milpa, where maize is grown together with other crops rather than being defined simply by the presence of maize (Fonteyne et al., 2023). A maize monocrop and a maize-based Milpa parcel can both contain maize, but they represent different cropping systems. Reference labels for this type of classification therefore need to capture the crop association or arrangement that distinguishes the two classes (Bégué et al., 2018; Ibrahim et al., 2021).

Previous studies show that these broader cropping-system characteristics can be mapped explicitly. Ibrahim et al. (2021) mapped crop types and mixed-cropping systems in Nigeria, including maize-based associations. Tariq et al. (2023) similarly mapped individual crop

types together with cropping patterns in Pakistan. These examples show that crop associations and cropping patterns can themselves form mapping classes rather than being represented only by the identity of an individual crop.

2.2.3 Small fields, fragmentation, mixed pixels and class variability

Here, the term "smallholder landscape" is used descriptively to refer to a mosaic of relatively small fields. Such landscapes often contain small, irregular parcels, while field boundaries may be indistinct or interrupted by roads, ditches or vegetation, making parcel delineation difficult (Persello et al., 2019).

The effect of these boundaries depends partly on field size relative to image resolution. When fields are small, a larger share of pixels lies close to field edges, where the recorded signal may include contributions from neighbouring fields or other boundary features. Pixels farther inside a sufficiently large and uniform field are more likely to represent the field itself. Mixed-pixel effects therefore become more important as field size decreases relative to pixel size and as the landscape becomes more fragmented (Persello et al., 2019; Tariq et al., 2023; Wu et al., 2023).

Boundary mixing should be distinguished from heterogeneity that occurs within the field itself. Mixed-cropping systems may contain more than one crop within the same parcel, and Ibrahim et al. (2021) reported strong spatial heterogeneity in intercropped classes. For Milpa, variation within the parcel should therefore not automatically be interpreted as boundary contamination. A pixel within the parcel may reflect genuine within-field variability, whereas a boundary pixel may also be influenced by adjacent land cover (Ibrahim et al., 2021; Persello et al., 2019).

2.2.4 Parcel- and pixel-level spatial support

Spatial support refers to the geographical unit at which a measurement, feature or label is represented. In pixel-level classification, each pixel is treated as an observation, whereas parcel-level approaches aggregate information within field boundaries. Object-based image analysis instead groups neighbouring pixels into image objects that are used as classification units (Kussul et al., 2016; Xue et al., 2023). No single representation is best in all settings. Pixel-level analysis retains local spatial detail, while parcel-level aggregation can provide more consistent field-level predictions but may conceal genuine variation within a field. Parcel- and object-based approaches have improved map consistency in some studies by incorporating field context and reducing fragmented predictions (Kussul et al., 2016; Xue et al., 2023), although Maolan et al. (2024) found slightly higher overall accuracy for pixel-based than object-oriented classification in their smallholder case study.

Spatial support also affects validation because spatial dependence between training and test samples can lead to optimistic accuracy estimates (Karasiak et al., 2022). The detailed implications for data splitting and geographic validation are discussed in Section 2.5.3. The choice among pixel, object and parcel support should therefore follow the class definition, reference-label scale and intended mapping product.

2.3 Sentinel-2 observations and time-series reliability

2.3.1 Sentinel-2 mission, bands, resolutions and products

Sentinel-2 was developed within the European Union's Copernicus Earth-observation programme to provide repeated multispectral imagery for monitoring land, vegetation and inland and coastal waters. It was designed as a two-satellite constellation with a nominal revisit interval of about five days over land (European Space Agency, 2015). Clouds, shadows and other defects make clear agricultural observations less frequent than the nominal revisit (Bukowiecki et al., 2021; Defourny et al., 2019).

Each satellite carries a Multispectral Instrument that records reflected solar energy in 13 bands at native resolutions of 10, 20 or 60 m. A spectral band represents a defined interval of the electromagnetic spectrum rather than an ordinary photographic colour. The bands span the visible, red-edge, near-infrared and shortwave-infrared regions, with the red-edge bands being particularly relevant to vegetation and chlorophyll-related applications (Clevers and Gitelson, 2013; European Space Agency, 2015).

Table 2.1. Sentinel-2 spectral bands and native resolutions.

Band	Spectral region	Approx. centre	Native resolution
B01	Coastal aerosol	443 nm	60 m
B02	Blue	490 nm	10 m
B03	Green	560 nm	10 m
B04	Red	665 nm	10 m
B05	Red edge	705 nm	20 m
B06	Red edge	740 nm	20 m
B07	Red edge	783 nm	20 m
B08	Near infrared	842 nm	10 m
B8A	Narrow near infrared	865 nm	20 m
B09	Water vapour	945 nm	60 m
B10	Cirrus	1,375 nm	60 m
B11	Shortwave infrared	1,610 nm	20 m
B12	Shortwave infrared	2,190 nm	20 m

Source: European Space Agency (2015).

Sentinel-2 products are distributed at different processing levels. Level-1C provides orthorectified top-of-atmosphere reflectance, whereas Level-2A provides atmospherically corrected bottom-of-atmosphere, or surface, reflectance. Standard Level-2A products contain 12 surface-reflectance bands and omit the cirrus band B10 (Copernicus Sentinel-2, 2021; European Space Agency, 2015). Atmospheric correction does not remove all cloud and shadow contamination, so observation-quality screening is still required. These issues are discussed further in Section 2.3.2.

2.3.2 Optical-data quality problems and analysis-ready observations

An available Sentinel-2 image does not necessarily provide a usable observation for every field. Clouds and cloud shadows can obscure or contaminate surface reflectance, reducing the number of clear observations available for agricultural monitoring (Defourny et al., 2019; Magno et al., 2021). Consequently, the nominal revisit frequency does not necessarily correspond to the temporal density of observations that remain suitable for analysis.

Other inconsistencies can affect comparability among dates even when observations are cloud-free. Sentinel-2 bands are acquired at different native spatial resolutions, so spatial harmonisation is required when several bands are analysed together (Perich et al., 2023). Multi-temporal geometric consistency is also important because residual misregistration can reduce the spatial correspondence between repeated acquisitions (Yan et al., 2018). Surface reflectance is also directional: the same surface can produce different measured reflectance as solar and viewing geometry changes, introducing variation between acquisitions that is unrelated to actual land-surface change (Roy et al., 2017).

Analysis-ready preparation therefore requires quality screening and spatial harmonisation before observations are combined into a time series. Perich et al. (2023), for example, used the Sentinel-2 Scene Classification Layer to exclude clouds, cloud shadows, dark areas and defective pixels, and resampled the categorical 20 m layer to 10 m using nearest-neighbour interpolation, while 20 m spectral bands were resampled using cubic interpolation. This example illustrates that categorical quality information and continuous reflectance data can require different treatment when they are placed on a common spatial grid. After invalid observations are removed, however, the remaining temporal sequence may still be irregular and incomplete, leading to the temporal regularisation and gap-filling problem considered in Section 2.3.3.

2.3.3 Temporal regularisation, gap filling and uncertainty

After invalid observations are removed, optical time series are often irregular and incomplete because clear observations are not available on the same dates for every parcel (Chen et al., 2021; Lira Melo de Oliveira Santos et al., 2019). Temporal regularisation addresses this inconsistency by organising observations or reconstructed values on a common date sequence. Missing periods can then be handled using different approaches, including temporal compositing, interpolation and spline fitting, smoothing methods such as Savitzky–Golay or Whittaker filtering, harmonic or other model-based approaches, and more complex spatio-temporal reconstruction methods (Chen et al., 2021; Liu et al., 2017; Lira Melo de Oliveira Santos et al., 2019). These approaches are not interchangeable: their performance depends on the vegetation trajectory, the amount and timing of missing data, noise in the observations and the intended analytical use of the reconstructed series (Liu et al., 2017; Lira Melo de Oliveira Santos et al., 2019).

Linear interpolation is a comparatively simple approach that has been used to fill missing dates in agricultural image time series and can be sufficient in some applications when temporal change is relatively smooth and gaps are limited (Lira Melo de Oliveira Santos et al., 2019; Liu et al., 2017). Interpolation estimates values within the period bounded by observations, whereas extrapolation extends beyond the first or last available observation and

is therefore less constrained by measured data. Reconstruction becomes less reliable as gaps become longer or when missing observations coincide with periods of rapid change in the vegetation trajectory (Chen et al., 2021; Liu et al., 2017). Reconstruction uncertainty therefore depends not only on the selected method, but also on the duration and temporal position of a gap relative to the underlying seasonal development of the vegetation (Liu et al., 2017).

More flexible smoothing methods can reduce residual noise and produce stable seasonal trajectories (Chen et al., 2021), but this benefit involves a trade-off. Excessive smoothing can weaken inflection points, peaks, valleys or other short-lived features that form part of the original temporal signal (Liu et al., 2017). Gap-filled values should therefore be interpreted as reconstructed estimates rather than observations directly recorded by the sensor (Chen et al., 2021; Liu et al., 2017; Lira Melo de Oliveira Santos et al., 2019). Assessing how much of a time series is supported by original observations, and where longer unsupported periods occur, is consequently important for judging the reliability of the resulting temporal information; this issue is considered in Section 2.3.4.

2.3.4 Observation support and reliability assessment

A complete-looking regular time series can conceal differences in observational support because it may contain both directly observed and reconstructed values (Chen et al., 2021; Liu et al., 2017). The reliability of the resulting temporal information therefore depends not only on whether the series is complete, but also on the amount, timing and continuity of the original observations (Defourny et al., 2019; Liu et al., 2017). Useful indicators include the number or proportion of valid observations (Defourny et al., 2019), their distribution through the season (Defourny et al., 2019; Liu et al., 2017), and the duration of continuous gaps without valid observations (Chen et al., 2021; Liu et al., 2017).

These indicators describe different aspects of temporal support. Two parcels may contain the same number of valid observations, yet one may be observed relatively consistently through the growing season while the other contains a long gap during an important period of crop development (Liu et al., 2017). Because reconstruction performance depends on both the amount and timing of missing data, observation counts alone cannot fully describe the quality of temporal support (Defourny et al., 2019; Liu et al., 2017). Observation support and reconstruction accuracy are therefore complementary aspects of reliability: the first describes how much original information is available and how it is distributed through time, whereas the second concerns how well missing values can be estimated from that information (Chen et al., 2021; Liu et al., 2017).

Having some usable imagery is likewise not equivalent to having sufficient temporal support for interpreting a seasonal trajectory (Defourny et al., 2019; Liu et al., 2017). Regular sampling and the timing of available observations remain important when constructing agricultural time series (Defourny et al., 2019; Liu et al., 2017). Reconstructed values must also remain distinguishable from directly observed reflectance because their reliability depends on the surrounding observations, the duration and timing of the gap, and the underlying vegetation trajectory (Chen et al., 2021; Liu et al., 2017).

Reconstruction accuracy can be assessed separately from observation support through simulated-gap evaluations that test gap-filling performance under different missing-data conditions (Liu et al., 2017). Because the adequacy of a particular observation pattern depends on vegetation dynamics, gap conditions, data availability and the intended analytical use, support thresholds and gap-duration categories are best treated as context-specific operational criteria rather than universal standards (Liu et al., 2017; Perich et al., 2023). Maintaining this distinction between observed and reconstructed information provides a clearer basis for the temporal and spectral feature representations discussed in Section 2.4.

2.4 Spectral-temporal and spatial feature representation

Reliable image time series must be transformed into classification predictors before their spectral, temporal and spatial information can be used by a model. These predictors can be organised into five complementary families: raw-band statistics, spectral-index statistics, phenology and temporal stability, within-parcel heterogeneity, and observed-image texture. Together they describe crop condition and field structure, but they also enlarge the feature pool and require explicit control of dimensionality and redundancy.

2.4.1 Spectral bands and spectral indices

Satellite reflectance provides the starting point for describing agricultural vegetation. Green vegetation generally absorbs visible red radiation and reflects strongly in the near-infrared (Clevers and Gitelson, 2013; Li et al., 2022). The red-edge region captures the rapid transition between red absorption and near-infrared reflectance and is responsive to chlorophyll and canopy development (Qiao et al., 2024; Sun et al., 2020; Xie et al., 2018), while shortwave-infrared reflectance adds information related to vegetation and soil moisture, senescence and dry matter (Li et al., 2022).

Individual bands retain wavelength-specific reflectance information, whereas spectral indices emphasise particular contrasts through normalised differences or ratios. Spectral indices can highlight greenness, moisture, pigment condition or senescence (Sun et al., 2020; Li et al., 2022; Zhao et al., 2025). They remain indirect indicators of crop identity, and different crops may produce similar values at some stages of development (Tufail et al., 2025; Xie et al., 2018).

The six indices considered here represent complementary aspects of vegetation condition. Their exact formulations and interpretations are summarised below.

Table 2.2. Spectral indices used for Sentinel-2 feature construction.

Index	Formula using Sentinel-2 bands	Interpretation and agricultural relevance	Main limitation	Supporting source
Normalised Difference Vegetation Index (NDVI)	(B08−B04)/(B08+B04)	Contrasts red absorption with near-infrared reflectance and is widely used to describe vegetation greenness, cover and canopy development	Can saturate in dense canopies and become less responsive at higher leaf area	Xie et al. (2018); Qiao et al. (2024)

Index	Formula using Sentinel-2 bands	Interpretation and agricultural relevance	Main limitation	Supporting source
Enhanced Vegetation Index 2 (EVI2)	2.5(B08−B04)/(B08 +2.4B04+1)	Uses red and near-infrared reflectance to represent vegetation condition while reducing some soil-background sensitivity relative to simple red–NIR ratios	Remains an indirect descriptor of vegetation condition and does not directly indicate crop type	Jiang et al. (2008); Tufail et al. (2025)
Normalised Difference Red Edge Index (NDRE)	(B06−B05)/(B06+B 05)	Measures contrast between two red-edge bands and represents chlorophyll- and canopy-related variation within the red-edge region	Interpretation depends on the selected red-edge bands, canopy density and pigment condition	Sun et al. (2020); Zhao et al. (2025)
Chlorophyll Index Red Edge (CI red edge)	B08/B07−1	Contrasts near-infrared and red-edge reflectance and is used as an indicator of chlorophyll-related canopy variation	Can be influenced by leaf area and may become less responsive at high vegetation density	Xie et al. (2018); Sun et al. (2020)
Normalised Difference Moisture Index (NDMI)	(B08−B11)/(B08+B1 1)	Combines near-infrared and shortwave-infrared reflectance to describe moisture-related differences in vegetation and the land surface	Moisture-related responses can also vary with dry matter and soil-background conditions	Griffiths et al. (2019); Cheng et al. (2023)
Plant Senescence Reflectance Index (PSRI)	(B04−B02)/B06	Represents changes in the balance among visible and red-edge reflectance associated with pigment loss and crop senescence	Responses may also reflect vegetation stress and phenological transition	Zhao et al. (2025)

Source: Exact formulations used in this study; see Appendix Table A.1. Agricultural interpretations are supported by the sources shown in the table.

Red-edge indices can add information beyond visible and broad near-infrared bands, particularly where conventional greenness indices become less responsive (Qiao et al., 2024; Sun et al., 2020; Xie et al., 2018). Shortwave-infrared information provides complementary sensitivity to vegetation and soil moisture, senescence and dry matter (Li et al., 2022).

2.4.2 Temporal summaries, phenology and stability

A single image captures only one moment in crop development. Vegetation changes through emergence, canopy expansion, peak development, senescence and harvest, so crops that appear similar on one date may follow different seasonal trajectories or become separable

only at particular stages. Multi-temporal observations are therefore often more informative than isolated acquisitions (Hu et al., 2019; Liu et al., 2023; Xu et al., 2021).

An ordered stack of band or index values preserves detailed seasonal information and lets a model respond to particular dates, but predictor numbers increase rapidly with bands and observations, and full stacks remain sensitive to missing observations and irregular temporal support. Liu et al. (2023) nevertheless found full time-series stacks to outperform statistical and phenological summaries in one comparison, showing that detailed temporal information can remain valuable despite its dimensionality.

Seasonal statistics such as the mean, median, minimum, maximum, standard deviation, interquartile range, percentiles and amplitude provide a more compact representation of typical conditions and overall variation. This compression can hide brief discriminative events or trajectories with similar ranges but different timing (Liu et al., 2023). Phenological features, including peak value and timing, green-up and decline rates and season length, preserve more explicit information about when and how a canopy changes (Hu et al., 2019; Woźniak et al., 2022; Zhao et al., 2025).

Temporal stability measures summarise the degree of fluctuation in a signal through time. Such variability may reflect crop development and management, but it can also arise from residual cloud effects, irregular observational support or reconstructed values. Features derived from interpolated series must therefore be interpreted alongside the distinction between observed and reconstructed information established in Section 2.3.

2.4.3 Within-parcel heterogeneity and observed-image texture

Parcel-level averages may conceal spatial differences in crop cover, vigour, moisture, exposed soil, weeds, residues or planting arrangement. Such within-parcel heterogeneity can affect field-level spectral representation and retain information lost through averaging (Chen and Liu, 2024; Squeri et al., 2021; Wang et al., 2025). In Milpa it may reflect companion crops, unequal planting density or soil exposure, but it is not a unique class signature because maize monocrops can also contain weeds, missing plants, residues, uneven emergence or stress.

Parcel-wide statistics and texture describe different aspects of this variation. A field-level standard deviation summarises dispersion among all parcel pixels without considering their arrangement. Texture examines neighbouring pixels and can represent local contrast, continuity or patchiness. In a moving window, a statistic is calculated from the neighbourhood around each central location (Wang et al., 2025). Local variance measures dispersion within that neighbourhood, while local standard deviation expresses the same variability in the units of the original reflectance or index.

At 10 m grid spacing, 3 × 3 and 5 × 5 windows cover approximately 30 × 30 m and 50 × 50 m. Smaller windows respond to fine contrasts; larger windows capture broader patterns but are more likely to cross parcel boundaries or combine several structures in small fields. No window size is universally optimal because usefulness depends on parcel dimensions, spatial resolution and the scale of the pattern (Fang et al., 2025; Wang et al., 2025).

Texture has improved crop discrimination in some settings by representing canopy discontinuity and soil–vegetation mixtures not captured by spectral averages (Fang et al., 2025; Tang et al., 2025). Its value may decline when parcels contain few pixels or neighbourhoods extend outside the field. As discussed in Section 2.3, reconstructed values are estimates rather than observations; texture calculated from reconstructed imagery may therefore partly reflect the reconstruction process rather than only recorded spatial structure. Observed imagery provides a more direct basis for interpreting texture as a recorded spatial pattern.

2.4.4 High dimensionality, redundancy and feature control

Combining dates, bands, indices, temporal summaries and texture can produce a very large feature set. The curse of dimensionality describes the increasing difficulty of analysis as predictor dimensions grow; in supervised classification, the related Hughes phenomenon occurs when accuracy initially improves with informative variables but later declines because the feature space becomes too large for the available labelled sample (Gómez-Chova et al., 2003; Jia et al., 2022).

Remote-sensing predictors are also often redundant: adjacent bands may respond to similar vegetation properties, neighbouring dates can capture the same crop stage, and several indices reuse the same bands. Additional variables may therefore increase computation and overfitting without providing equivalent new information. Crop-mapping studies confirm that more predictors do not always improve classification. Tang et al. (2025) reported performance rising and then declining as variables were added, while Orynbaikyzy et al. (2020) found that feature selection clarified a large correlated feature set without necessarily increasing overall accuracy.

Common controls include correlation filtering, sequential selection, model-based importance rankings and compact panels representing several feature families. Correlation filtering reduces strong redundancy but may discard useful interactions; sequential methods can be computationally demanding and sensitive to the development sample; and Random Forest importance can efficiently rank nonlinear predictors, although correlation may divide or distort apparent importance (Gómez-Chova et al., 2003; Orynbaikyzy et al., 2020; Tang et al., 2025).

Feature importance may vary across seasons or geographical areas because crop differences are not spatially or temporally constant (Fang et al., 2025; Xu et al., 2021). Repeated development-only evaluation can therefore test whether variables remain influential across different samples. A compact panel should be judged not only by its highest score but also by redundancy, stability, interpretability and representation of spectral, temporal and spatial processes.

Feature selection must be confined to model-development data. The same rule applies to imputation and any preprocessing parameter estimated from the data: each operation must be fitted on development data and then applied unchanged to evaluation data (Ambroise and McLachlan, 2002; Varma and Simon, 2006; Cawley and Talbot, 2010). If holdout observations influence variable selection, imputation or repeated panel comparison,

information enters model development and the performance estimate is no longer independent. Feature control and final geographic evaluation must therefore remain separate. Once a defensible representation is established, the selected predictors can be supplied to the classification and validation approaches reviewed in Section 2.5.

2.5 Machine-learning classification and geographic evaluation

2.5.1 Supervised machine learning for crop classification

Once the spectral, temporal and spatial information described in Section 2.4 has been reduced to a controlled set of predictors, supervised machine learning can be used to test whether those predictors distinguish the reference classes. A supervised classifier learns from observations for which both predictor values and correct class labels are known. These labelled observations form the training data, while observations kept outside model development are needed to assess whether the learned relationship applies beyond the cases already seen.

The model does not understand the agronomic meaning of a Milpa mixed-cropping system. It identifies statistical relationships between the reference labels and combinations of predictors. This distinction matters because the classes may overlap. A maize-dominant Milpa parcel may resemble a maize monocrop, while variation among Milpa parcels may broaden the range of patterns assigned to the same class (Persello et al., 2019; Ibrahim et al., 2021). Performance therefore depends not only on the algorithm but also on the quality, representativeness and balance of the reference data (Orynbaikyzy et al., 2019; Tariq et al., 2023).

Crop-mapping studies use several algorithm families. Random Forest and support vector machines are established methods for nonlinear, high-dimensional remote-sensing data (Orynbaikyzy et al., 2019; Mountrakis et al., 2011). Deep-learning models can learn complex representations from multi-temporal inputs and have outperformed conventional methods in some studies (Kussul et al., 2017; Zhong et al., 2019). Other studies have reported strong performance from support vector machines, Random Forest or hybrid approaches (Mountrakis et al., 2011; Orynbaikyzy et al., 2019; Yao et al., 2022). No classifier is universally superior: performance varies with the sensor, feature representation, class structure, sample size and geographical domain (Zhong et al., 2019; Yao et al., 2022).

Model fitting must also remain separate from feature selection, hyperparameter tuning and final evaluation. Using evaluation observations to guide these development decisions can produce an overly favourable estimate of performance, particularly when labelled data are limited (Varma and Simon, 2006; Cawley and Talbot, 2010).

2.5.2 Random Forest, class imbalance and probability outputs

Random Forest combines many decision trees instead of relying on a single tree. Each tree is fitted to a bootstrap sample of the development data, and only a random subset of predictors is considered at each split. The trees then vote collectively for the predicted class. This design reduces dependence on any one tree and allows nonlinear relationships and interactions

among spectral, temporal and spatial variables to be represented (Breiman, 2001; Akbari et al., 2020).

The method is widely used in crop mapping because it accommodates mixed predictor types, usually requires little predictor scaling and provides measures of variable importance. These advantages do not remove the need for careful design. Hyperparameters influence tree complexity, correlated predictors can divide or distort apparent importance, and a large ensemble can reproduce biases in an unrepresentative training sample. Performance may also decline when a model is applied in areas with different crop calendars, environmental conditions or feature distributions (Orynbaikyzy et al., 2022; Mahmood et al., 2025).

Random Forest can also produce continuous class-support outputs that are often treated as estimated probabilities. These values summarise ensemble support for each class, but their calculation is implementation-dependent and they are not automatically calibrated probabilities (Boström, 2007; Niculescu-Mizil and Caruana, 2005). A support value of 0.80 does not, by itself, show that eight out of ten comparable parcels will belong to that class. Probability calibration and the use of such outputs in classification-level probability fusion, where outputs from separately fitted models are combined, are considered separately in Section 2.8.

Class imbalance creates another challenge when one class contains substantially more reference observations than another. A model may increase the total number of correct predictions by favouring the majority class while failing to detect the minority class. Overall accuracy may then appear satisfactory even when the main mapping target is poorly recognised. Responses include class weighting, over-sampling and synthetic-sample generation (Douzas et al., 2019; Cheng et al., 2025). These approaches involve trade-offs because resampling changes the representation of the original training distribution, while synthetic samples are modelled approximations rather than new field observations. Any balancing operation must therefore be restricted to the development data and must not alter the evaluation set.

2.5.3 Parcel-controlled and geographic evaluation

Validation design determines what kind of generalisation a reported score represents. In pixel-level evaluation, pixels from the same parcel must remain within a single split because they share the same field label and local conditions; otherwise, within-parcel similarity can produce an overly favourable estimate of transfer to new parcels. Parcel-controlled splitting prevents this leakage, but it does not remove spatial dependence among neighbouring fields. Random parcel-level evaluation can therefore remain less demanding than evaluation across separated locations (Karasiak et al., 2022).

Spatial-block validation, leave-one-area-out designs and geographically separated holdouts provide stronger tests of spatial transfer. Their purpose is not simply to produce a lower score, but to examine whether relationships learned in one place remain useful where environmental conditions, management, crop prevalence and observation quality differ. Crop-classification studies often report reduced performance in unseen regions, showing that strong

within-area discrimination does not guarantee geographic transferability (Orynbaikyzy et al., 2022; Hoppe et al., 2024).

Geographic out-of-fold predictions can support development-stage comparisons when each region is predicted by a model trained on other regions, but they remain part of model development rather than an untouched final benchmark. A geographically separated holdout can provide a secondary benchmark, but it should not be described as independent external validation when it derives from the same study programme. It must not influence feature selection, data-derived preprocessing, hyperparameter tuning, model choice or decision-threshold selection (Ambroise and McLachlan, 2002; Varma and Simon, 2006; Cawley and Talbot, 2010). These operations must be fitted using development data and then applied unchanged to the holdout; repeatedly consulting the holdout while revising the workflow gradually turns it into development data.

2.5.4 Classification metrics and interpretation

A confusion matrix compares reference classes with predicted classes. In the binary notation used below, the Milpa mixed-cropping system is the positive class. A true positive (TP) is a Milpa parcel correctly predicted as Milpa, while a false negative (FN) is a Milpa parcel incorrectly predicted as maize monocrop. A true negative (TN) is a maize monocrop parcel correctly predicted as maize monocrop, while a false positive (FP) is a maize monocrop parcel incorrectly predicted as Milpa.

Table 2.3. Classification metrics and interpretation.

Metric	Formula	Interpretation
Overall accuracy	(TP+TN)/(TP+TN+FP+FN)	Proportion of all observations classified correctly
Precision	TP/(TP+FP)	Reliability of positive-class predictions
Recall or sensitivity	TP/(TP+FN)	Proportion of actual positive observations detected
Specificity	TN/(TN+FP)	Proportion of actual negative observations correctly rejected
F1-score	2TP/(2TP+FP+FN)	Harmonic balance between precision and recall
Balanced accuracy	½[TP/(TP+FN) + TN/(TN+FP)]	Equal-weight average of positive- and negative-class recall

Sources: Congalton (1991), Foody (2002), Olofsson et al. (2014) and Farhadpour et al. (2024).

In remote-sensing terminology, producer's accuracy corresponds to recall for the class being assessed, while user's accuracy corresponds to precision. In multiclass problems, macro averaging calculates a metric separately for each class and gives every class equal influence. Weighted averaging gives more influence to classes with more observations. This choice can materially change interpretation when class frequencies are unequal (Farhadpour et al., 2024).

No single metric describes every type of error. Overall accuracy may be dominated by the more frequent class. Precision can remain high while many actual positive cases are missed,

and recall can be high at the cost of many false positives. F1 balances precision and recall but does not include true negatives. Balanced accuracy gives equal weight to sensitivity and specificity, although it still compresses two error processes into one value. Cohen's kappa is also widely reported, but its dependence on class prevalence and chance-agreement formulation means that it should not replace the confusion matrix and class-specific metrics (Foody, 2002; Farhadpour et al., 2024).

Performance should therefore be interpreted using several complementary metrics together with confusion-matrix counts, class prevalence and variation across validation folds or geographical areas. This allows performance to be assessed across classes and validation settings rather than relying on a favourable aggregate score within one sample. These principles also provide the evaluation basis for the approaches considered in Section 2.6.

2.6 High-resolution optical imagery and predictor-level integration

2.6.1 Finer spatial resolution in smallholder crop-system mapping

Sentinel-2 offers a valuable combination of spectral coverage and repeated observations, but its 10–20 m agricultural bands cannot always represent small or irregular fields in sufficient spatial detail. Pixels near parcel boundaries may contain contributions from neighbouring fields or other land-surface components. As field dimensions approach the pixel size, such mixed pixels become more common and can reduce field-level spectral homogeneity and complicate crop classification (Rao et al., 2021; Ibrahim et al., 2021).

Finer-resolution imagery can reduce some of this boundary mixing and represent field boundaries and within-field variation more clearly. This additional spatial detail can be particularly useful in heterogeneous smallholder landscapes, where several crops or different levels of vegetation cover may occur within relatively small fields (Rao et al., 2021; Ibrahim et al., 2021; Wang et al., 2022).

Greater spatial detail does not, however, guarantee better crop classification. Spatial resolution is only one component of the information provided by an optical sensor, and some finer-resolution commercial imagery may offer fewer spectral bands or less regular seasonal coverage than Sentinel-2. Its usefulness therefore depends on the mapping problem and on the balance between spatial, spectral and temporal information available from the sensor (Rao et al., 2021).

2.6.2 SkySat imagery for agricultural monitoring

SkySat is a Planet-operated very-high-resolution optical constellation that acquires blue (450–515 nm), green (515–595 nm), red (605–695 nm) and near-infrared (740–900 nm) multispectral bands together with a broad panchromatic band (450–900 nm). Planet distributes orthorectified products sampled at 0.50 m pixel spacing and provides tasking with multiple revisit opportunities per day (Planet Labs PBC, 2026a, 2026c). This product sampling should not, however, be interpreted as the effective spatial resolution of the measurement. Independent characterisation of SkySat has assessed its spatial response using full width at half maximum and modulation transfer function measurements, which provide

information on image sharpness beyond nominal ground sampling distance (Kim et al., 2022).

The visible and near-infrared bands permit vegetation-index calculation and can provide detailed information on variation within agricultural fields. Ibrahim et al. (2021) used in-season SkySat imagery to augment reference data and delineate field boundaries for field-level assessment of Sentinel-2-based crop and intercropping-system mapping; SkySat imagery itself was not used as a predictor in the classifier. Li et al. (2025) used SkySat together with Sentinel-2 and PlanetScope for in-season assessment of Japanese squash and reported stronger relationships between SkySat-derived vegetation indices and yield in that case study. These studies show the agricultural value of fine spatial observations without implying that SkySat will improve every crop-classification problem.

SkySat also differs from Sentinel-2 in its observation strategy and spatial coverage. Its imagery is task-driven rather than a systematic seasonal monitoring record, and individual scenes or imaging swaths cover relatively limited areas (Planet Labs PBC, 2026a). Accurate geolocation and co-registration are particularly important when fine-resolution imagery is compared across acquisitions or summarised within small parcels (Kim et al., 2022). SkySat is therefore better understood as a source of detailed complementary spatial and spectral information than as a direct replacement for a dense seasonal Sentinel-2 time series.

2.6.3 PlanetScope imagery for agricultural monitoring

PlanetScope provides finer product sampling than Sentinel-2 together with frequent observations from a large satellite constellation. Current SuperDove (PSB.SD) imagery contains eight spectral bands: coastal blue, blue, green I, green, yellow, red, red edge and near-infrared, centred approximately at 443, 490, 531, 565, 610, 665, 705 and 865 nm, respectively. Earlier Dove Classic (PS2) and Dove-R (PS2.SD) generations provided four-band blue, green, red and near-infrared imagery. Orthorectified PlanetScope products are distributed at approximately 3 m pixel spacing, but this sampling interval should not be interpreted as the effective spatial resolution or measurement footprint. The constellation is designed for daily or near-daily land imaging, although usable observations remain less frequent because of cloud and image-quality constraints (Planet Labs PBC, 2026b, 2026c).

This combination of fine sampling and frequent acquisition can be useful in smallholder agriculture. Rao et al. (2021) reported improved classification of the smallest farms with Planet imagery relative to Sentinel-2, associated with reduced mixed-pixel effects at field boundaries. PlanetScope time series have also been used to characterise seasonal crop-growth patterns in small fields (Sakuma and Yamano, 2020), while high-resolution PlanetScope imagery has supported field-boundary delineation at large spatial scales (Estes et al., 2022). Sadeh et al. (2021) further demonstrated its value for crop and vegetation monitoring through daily 3 m surface-reflectance reconstruction and wheat leaf-area-index assessment.

These benefits involve trade-offs. PlanetScope generations differ in spectral configuration, and some configurations contain fewer spectral bands than Sentinel-2. Radiometric differences among PlanetScope sensors and acquisitions can also require harmonisation for consistent time-series analysis (Sadeh et al., 2021). Moreover, nominal daily acquisition

capability does not guarantee a cloud-free, analysis-ready observation each day (Sakuma and Yamano, 2020; Estes et al., 2022). Finer product sampling should therefore be considered together with spectral content, temporal availability and data consistency rather than assumed to provide consistently superior crop classification.

2.6.4 Predictor-level multi-sensor integration

Predictor-level integration combines variables derived from different sensors in a shared feature set before fitting a single classifier (Orynbaikyzy et al., 2020). In parcel-based applications, this requires sensor-derived variables to be aligned to the same analytical units and reference labels. Sentinel-2 predictors may provide broad spectral and seasonal information, while PlanetScope or SkySat variables may contribute finer-scale information on vegetation condition and within-parcel spatial variation.

This strategy differs from integration at other analytical levels. Pixel- or image-level integration combines or transforms image measurements before classification, whereas classification- or decision-level integration combines outputs from separately fitted models (Orynbaikyzy et al., 2020; Useya and Chen, 2018). Super-resolution, reviewed in Section 2.7, is a distinct image-level strategy that reconstructs an image representation at a finer spatial scale. Predictor-level integration instead changes the feature representation supplied to a classifier without reconstructing the image or combining outputs from separately fitted models.

Multi-sensor integration has improved crop classification in several studies combining optical and radar observations, although the integration architectures differ. Orynbaikyzy et al. (2020) directly evaluated feature stacking of Sentinel-1 and Sentinel-2 variables, while Ibrahim et al. (2023) integrated information from the two sensors for crop classification and Zhang et al. (2023) combined them through a learned feature-fusion architecture. Within the literature reviewed here, direct comparative evidence for parcel-level predictor integration specifically combining Sentinel-2 with SkySat or PlanetScope remains limited. The optical–radar literature therefore provides methodological context rather than direct evidence that these optical-sensor combinations will improve classification.

Improvement is not guaranteed. Feature stacking can increase dimensionality and include correlated or redundant information, while unequal data availability among sensors can complicate temporal correspondence (Orynbaikyzy et al., 2020). Combined representations should therefore be compared with corresponding single-sensor baselines to determine whether integration provides incremental classification value.

As discussed in Section 2.5.3, any data-dependent imputation, scaling, feature selection or model tuning must remain restricted to development data and then be applied unchanged during evaluation. Predictor-level comparisons are most interpretable when spatial support, temporal correspondence, missing data and redundancy are controlled consistently across sensor representations.

2.7 Cross-sensor super-resolution

2.7.1 Concept and purpose of super-resolution

Super-resolution aims to reconstruct an image at a finer spatial scale than the original input. In cross-sensor super-resolution, imagery from a coarser sensor is paired with observations from a finer-resolution sensor so that a model can learn the relationship between the two spatial representations (Latte and Lejeune, 2020). This can be useful in smallholder landscapes, where individual Sentinel-2 pixels may be large relative to parcel dimensions and may conceal field boundaries or variation within fields (Rao et al., 2021; Ibrahim et al., 2021).

Super-resolution must be distinguished from conventional resampling. Resampling places values on a finer grid without providing observations measured directly at that finer scale, whereas learned super-resolution estimates finer-scale structure from modelled relationships. Unlike predictor-level integration described in Section 2.6.4 and classification-level integration reviewed in Section 2.8, super-resolution reconstructs the image representation itself. The resulting fine-scale detail nevertheless remains estimated rather than directly observed and may include hallucinated spatial patterns or radiometric and spatial inconsistencies relative to the true scene (Donike et al., 2025; Aybar et al., 2026).

2.7.2 Conventional and learning-based approaches

Conventional interpolation provides a useful non-learning baseline for evaluating super-resolution. Bicubic interpolation is commonly used as a baseline because it requires no model training and provides a reproducible finer-grid representation; unlike learning-based super-resolution, however, it does not estimate a mapping from training examples (Lim et al., 2017).

Super-resolution methods can use either single or multiple low-resolution observations. Single-image super-resolution reconstructs a finer-resolution image from one low-resolution input, whereas multi-image super-resolution exploits repeated observations of the same area to recover complementary information across acquisitions. Razzak et al. (2023), for example, used multiple Sentinel-2 revisits together with PlanetScope reference imagery and showed that multi-image reconstruction could outperform corresponding single-image approaches on image-fidelity measures. Pansharpening represents a related but distinct form of spatial enhancement in which a higher-resolution panchromatic image is combined with lower-resolution multispectral imagery to produce a multispectral image with finer spatial detail (Palsson et al., 2014).

Learning-based approaches instead estimate fine-scale spatial relationships from training examples. Convolutional neural networks learn transformations from coarse image patches to finer-resolution outputs, while residual architectures use skip connections to preserve information and learn residual corrections across network layers. Enhanced Deep Residual Networks for Single Image Super-Resolution (EDSR) improved reconstruction by removing batch-normalisation layers from conventional residual blocks, increasing model capacity and performing upsampling near the end of the network (Lim et al., 2017). Related CNN-based

approaches have since been adapted specifically to multispectral Sentinel-2 imagery (Qian et al., 2023; Vasilescu et al., 2023).

More complex generative approaches can emphasise perceptual sharpness. GAN-based satellite super-resolution can produce visually sharper reconstructions (Moustafa and Sayed, 2021), while more recent diffusion-based work has highlighted the need to avoid hallucinated structures and spectral distortions (Donike et al., 2025). A visually sharper reconstruction should therefore not automatically be interpreted as a more radiometrically or spatially faithful representation of the observed scene.

2.7.3 Cross-sensor pairing and data preparation

Reliable cross-sensor training depends on overlapping coverage, accurate geometric co-registration, spectrally corresponding bands and closely matched acquisition dates. Cloud, shadow and invalid pixels should be excluded from paired supervision, while differences in spectral response and sensor-specific spatial response need to be considered. Misregistration can cause nominally corresponding pixels to represent different ground locations and can therefore distort reconstruction errors (Latte and Lejeune, 2020; Okabayashi et al., 2024; Aybar et al., 2024, 2026).

Temporal matching is especially important in agriculture because vegetation conditions can change between acquisitions. If coarse- and fine-resolution observations are collected on different dates, part of their difference may reflect temporal change rather than spatial resolution. A cross-sensor model may then learn temporal variation as though it represented fine-scale spatial information (Okabayashi et al., 2024).

Synthetic training pairs are often generated by degrading a fine-resolution image to create an artificial coarse-resolution input. This provides precise spatial correspondence, but the simulated degradation may not reproduce the blur, noise, atmospheric effects and spectral response of a real coarse-resolution sensor. Training with genuine Sentinel-2 and higher-resolution observations better represents real cross-sensor differences, although such datasets require more careful spectral harmonisation and geometric co-registration (Aybar et al., 2024, 2026).

Cross-sensor training also requires leakage-controlled data preparation. Following the parcel-control principles described in Section 2.5.3, spatially related patches derived from the same parcel should remain within a single data partition. Normalisation and other data-derived preprocessing should be estimated from training or development data and then applied unchanged to evaluation imagery.

2.7.4 Validation of reconstructed imagery

Super-resolution is commonly evaluated using complementary measures of reconstruction quality. Pixel-wise errors such as mean absolute error and root mean squared error quantify the magnitude of differences between reconstructed and reference imagery. Signed bias retains the direction of those differences and can reveal a systematic tendency to over- or under-estimate the reference signal. Peak signal-to-noise ratio relates reconstruction error to the available signal range, while the structural similarity index evaluates local similarity in luminance, contrast and spatial structure. For multispectral imagery, spectral-angle measures

compare the orientation of spectral vectors and can reveal spectral distortion that spatial or pixel-wise measures alone may not capture (Vasilescu et al., 2023; Major et al., 2025).

No single metric provides a complete assessment of reconstruction quality. Low numerical error does not by itself demonstrate that local spatial structures are faithfully represented, while strong structural similarity does not guarantee preservation of spectral relationships important for quantitative remote sensing. Pixel- and structure-based comparisons can also be affected by geometric misregistration. Validation is therefore stronger when radiometric or band-wise errors are interpreted together with spectral and spatial evidence rather than as a single quality score (Vasilescu et al., 2023; Major et al., 2025). A transparent non-learning comparator such as bicubic interpolation is also useful for determining whether a learned reconstruction provides improvement beyond conventional upsampling.

2.7.5 Downstream agricultural and classification value

The value of super-resolution ultimately depends on the application for which the reconstructed imagery is used. Finer reconstructed imagery may support applications such as field delineation, vegetation monitoring, within-parcel analysis and crop classification, but improvement in reconstruction metrics does not necessarily translate into better downstream performance. A model may reproduce or sharpen spatial structure while altering spectral relationships that are important for quantitative remote sensing or class discrimination (Vasilescu et al., 2023; Major et al., 2025; Donike et al., 2025).

Downstream evaluation should therefore compare learned super-resolution with transparent alternatives, including the original coarse imagery and conventional interpolation, and with native finer-resolution observations where an appropriate reference is available. Comparisons should use consistent reference data, analytical units and evaluation procedures so that differences in performance can be attributed as clearly as possible to the image representation rather than to changes in the classification design. Data-dependent preprocessing and model selection must remain restricted to development data, as discussed in Section 2.5.3.

Within the literature reviewed here, direct evidence that cross-sensor super-resolution consistently improves classification of heterogeneous crop systems remains limited. Any downstream improvement should therefore be interpreted as evidence that the reconstructed representation was useful for the tested task, rather than as proof that all reconstructed details were spatially or radiometrically faithful. Where information from different sensors is instead retained in separate predictive models, classification-level integration provides an alternative means of combining their outputs without reconstructing a new image. Section 2.8 reviews this approach.

2.8 Classification-level integration and uncertainty

2.8.1 Classification-level integration

Classification-level integration combines the outputs of separately trained models after each model has processed information from its own sensor. Here, classification-level integration refers to decision-level fusion, in which model predictions or class-support outputs are combined after classification (Orynbaikyzy et al., 2020; Useya and Chen, 2018). This differs

from image-level integration, which combines or reconstructs image information before classification, and from predictor-level integration, which combines sensor-derived variables before fitting a single classifier. Post-classification agreement analysis is also distinct because it can describe similarities and differences among model outputs without necessarily combining them into a new prediction.

This approach allows sensor-specific models to retain separate processing pathways while their outputs are compared or combined at a common prediction unit, such as a parcel. It can therefore accommodate sensors that differ in spatial resolution, spectral content, temporal coverage or data availability. Crop-mapping studies have applied decision-level integration to both optical-only and optical–radar combinations. Reported benefits vary: fusion has improved classification in some settings, but decision-level approaches have not consistently outperformed the strongest single-sensor model or alternative fusion strategies (Useya and Chen, 2018; Orynbaikyzy et al., 2020; Ofori-Ampofo et al., 2021; Valero et al., 2021).

2.8.2 Fusion rules and weighting

Hard voting combines final class labels, commonly by selecting the class supported by the majority of contributing models. Soft voting instead combines continuous class-support or probability outputs before a final class decision is made, thereby retaining information that would be lost if each model were first reduced to a single label. Equal-weight soft voting gives the same influence to each model, whereas weighted soft voting assigns different contributions to their outputs (Delgado, 2022). Stacking provides a more flexible alternative by fitting an additional model to combine predictions from the base models (Wolpert, 1992).

These approaches differ in the amount of additional estimation required at the fusion stage. Equal weighting does not require learned fusion weights, whereas weighted schemes may require weights to be specified or estimated and stacking introduces a second-stage model. When such parameters are estimated from labelled data, their selection must remain within development data to avoid information leakage, consistent with the validation principles described in Section 2.5.3. Regardless of the fusion rule, the outputs being combined must refer to the same target classes and prediction units.

2.8.3 Complementarity and disagreement

Classification-level fusion is most useful when the contributing models capture complementary information and make different, informative errors. Optical and radar sensors, for example, respond to different properties of crops and the land surface and can therefore produce different classification strengths and error patterns (Orynbaikyzy et al., 2020; Valero et al., 2021). Models that repeatedly make the same errors provide less opportunity for mutual correction, while adding a consistently weak model may provide little benefit to a stronger partner.

Disagreement should not, however, be interpreted automatically as useful complementarity. Differences among model predictions may also reflect ambiguous classes, unequal sensor support or changes in the domain in which the models are applied (Ofori-Ampofo et al., 2021; Valero et al., 2021). Agreement and disagreement analysis is therefore most informative when examined by class and prediction unit, because it can show where sensor-

specific models provide different decisions and where they instead share the same successes or failures.

Cohen's kappa provides a chance-adjusted summary of agreement between two sets of categorical predictions (Cohen, 1960). It should be interpreted alongside raw agreement and the underlying confusion patterns, because agreement alone does not show whether shared predictions are correct or whether disagreement is useful for fusion.

2.8.4 Probability outputs and uncertainty

Class-support outputs, estimated probabilities and calibrated probabilities are not equivalent. Random Forest outputs are commonly interpreted as class probabilities, but they should not automatically be assumed to correspond to observed event frequencies (Boström, 2007; Niculescu-Mizil and Caruana, 2005). A probabilistic classifier is calibrated when predictions assigned a given probability occur at approximately that frequency; for example, events assigned a probability of 0.80 should occur in approximately 80% of comparable cases. Reliability diagrams and proper scoring rules such as the Brier score provide complementary ways of assessing probabilistic predictions and their calibration (Dimitriadis et al., 2021). The Brier score is the mean squared difference between a predicted probability and the corresponding binary outcome, so lower values indicate better probabilistic accuracy (Brier, 1950).

This distinction is relevant when continuous class-support outputs are combined through soft voting. Averaging uncalibrated outputs does not make the resulting score a calibrated probability, so fused support values should be interpreted accordingly (Niculescu-Mizil and Caruana, 2005; Dimitriadis et al., 2021). Predictive uncertainty can also be summarised from the distribution of class support, for example through low maximum class support or high predictive entropy (Ji and Tang, 2025). Such indicators describe uncertainty in model outputs but do not by themselves establish why a prediction is uncertain.

When two classifiers are evaluated on the same parcels, uncertainty in their performance difference should preserve that pairing rather than resample the models independently. A paired non-parametric bootstrap repeatedly resamples the analytical units and applies the same sampled units to each compared model, producing an empirical distribution of the paired difference from which percentile confidence intervals can be obtained (Efron and Tibshirani, 1986). Where class imbalance or geographic structure is important, the resampling design can preserve relevant strata. Such intervals describe sampling uncertainty conditional on the observed analytical units and should be interpreted cautiously when samples are small or geographically structured.

2.8.5 Evaluation and interpretation

Classification-level integration should be evaluated against each contributing single-sensor model using the same reference labels, prediction units and validation design. Where more complex weighting or stacking strategies are considered, equal-weight soft voting provides a transparent fusion comparator. Performance should be interpreted using balanced and class-specific measures together with confusion matrices, because a small aggregate improvement

can coincide with poorer recognition of an individual class (Foody, 2002; Farhadpour et al., 2024).

Geographically controlled evaluation is also important when the objective is to assess whether fusion remains useful across different areas. Sensor-specific model performance and the relationships between classes may change geographically, so gains observed within one domain should not automatically be assumed to transfer elsewhere (Karasiak et al., 2022; Orynbaikyzy et al., 2022). Within the literature reviewed here, direct comparative evidence for classification-level integration specifically combining Sentinel-2 with SkySat or PlanetScope remains limited. Evidence from other sensor combinations therefore provides methodological context rather than direct proof that these optical-sensor combinations will improve classification.

These considerations provide the basis for Section 2.9, which synthesises the remaining evidence gaps and positions the contribution of the present study.

2.9 Synthesis, research gap and positioning of the present study

2.9.1 Synthesis of the reviewed evidence

The literature reviewed in Sections 2.1–2.8 shows that mapping the Milpa mixed-cropping system is not equivalent to identifying a single crop species. Milpa is maize-centred, but its associated crops, management, spatial arrangement and seasonal expression vary among places and households. It is therefore better understood as a crop-system class whose remote-sensing expression may overlap with maize monocropping (Fonteyne et al., 2023; Ibrahim et al., 2021). The challenge is intensified by small, irregular and internally heterogeneous parcels. Where field dimensions approach sensor resolution, mixed pixels can contain information from multiple crops or neighbouring land-surface components, while parcel averages may conceal within-field variation (Ibrahim et al., 2021; Rao et al., 2021).

Sentinel-2 provides a strong basis for seasonal crop observation, but interpretation depends on the timing and continuity of valid observations. Temporal gaps and unequal observation density can affect confidence in seasonal trajectories, and interpolated values should not be treated as equivalent to directly observed measurements (Defourny et al., 2019; Perich et al., 2023). Spectral, temporal and spatial predictors can describe complementary crop properties, but larger feature spaces also introduce redundancy and increase the importance of leakage-controlled preprocessing and feature selection. Data-derived preparation must therefore remain restricted to development data (Orynbaikyzy et al., 2020; Ambroise and McLachlan, 2002; Varma and Simon, 2006; Cawley and Talbot, 2010).

Classification quality consequently depends not only on algorithm choice but also on class representation and validation design. Parcel-controlled splitting prevents observations from the same field entering both sides of an evaluation, whereas geographically controlled designs provide a stronger test of transfer across spatially separated areas (Karasiak et al., 2022). Confusion matrices, class-specific measures and balanced metrics are also important because aggregate accuracy can conceal poor recognition of a less frequent or more difficult class (Foody, 2002).

Finer-resolution imagery can reduce boundary mixing and represent small fields in greater spatial detail, but finer sampling does not automatically provide more discriminative information (Rao et al., 2021; Ibrahim et al., 2021). The value of SkySat and PlanetScope also depends on temporal availability, valid coverage, spectral compatibility and geometric alignment. Predictor-level integration, super-resolution and classification-level integration address multi-sensor information at different analytical stages: the first combines variables before model fitting, the second reconstructs a finer image representation, and the third combines outputs from separately trained models (Orynbaikyzy et al., 2020; Latte and Lejeune, 2020; Useya and Chen, 2018). These strategies are not interchangeable. Super-resolved detail remains reconstructed and may not be fully spatially or radiometrically faithful (Donike et al., 2025; Aybar et al., 2026), while disagreement and class-support outputs should not automatically be interpreted as evidence of useful complementarity or calibrated probability.

2.9.2 Remaining research gap

Within the literature reviewed here, direct remote-sensing evidence for treating the Milpa mixed-cropping system as an explicit parcel-level mapping class remains limited. Existing studies establish the diversity and agricultural importance of Milpa, while research in other smallholder settings shows that mixed-cropping systems can be represented as distinct remote-sensing classes. However, these studies do not establish whether heterogeneous Milpa parcels can be distinguished consistently from maize monocropping in Oaxaca (Fonteyne et al., 2023; Novotny et al., 2021; Ibrahim et al., 2021). This is a demanding distinction because both classes can be maize-dominated, while companion crops and within-parcel heterogeneity may modify the observed signal.

The reviewed evidence also leaves uncertainty about the incremental value of complementary sensors and integration strategies for this specific problem. Finer-resolution imagery has supported small-field mapping, but direct evidence that SkySat or PlanetScope improves Milpa–maize monocrop discrimination beyond a seasonal Sentinel-2 baseline is lacking within the literature reviewed here. The same applies to parcel-level predictor integration involving these sensors. Super-resolution studies demonstrate that finer image representations can be reconstructed, but improved reconstruction does not establish improved crop-system classification, while most evidence for classification-level integration comes from other optical or optical–radar sensor combinations.

The resulting gap is therefore broader than the absence of another classification model. What remains insufficiently established is how seasonal Sentinel-2 information, finer-resolution observations, reconstructed imagery and model-level integration contribute to the same Milpa–maize monocrop mapping problem when evaluated under consistent parcel definitions, leakage-controlled model development, class-specific assessment and geographic transfer. An integrated evaluation is needed to determine not only whether these information sources improve classification, but also where they provide limited or inconsistent additional value.

2.9.3 Positioning and contribution of the present study

The present study responds to these gaps by treating the Milpa mixed-cropping system as an explicit parcel-level class separate from maize monocropping. It establishes a Sentinel-2 baseline in which seasonal information is interpreted together with observation support and the distinction between directly observed and temporally reconstructed values. Spectral, temporal and spatial predictors are used to represent complementary aspects of crop development and within-parcel variation without assuming that any single feature family uniquely defines Milpa.

SkySat and PlanetScope are evaluated as potential complements rather than inherently superior substitutes for Sentinel-2. The study separates predictor-level integration, cross-sensor super-resolution and classification-level integration and compares these routes with corresponding single-sensor or conventional baselines. This design tests their incremental value without assuming that finer spatial sampling, additional predictors, reconstructed imagery or fusion must improve classification.

The evaluation framework uses parcel-controlled data preparation and geographically separated assessment to examine transfer beyond the fitting domain. The geographically separated holdout is treated as a secondary benchmark rather than as independent or external validation. Confusion matrices, class-specific precision and recall, balanced measures, geographic variation, and uncertainty and disagreement diagnostics are used to interpret aggregate performance together with class-specific failure. The contribution of the study therefore lies in applying a common analytical and evaluative framework to determine where complementary information sources improve, fail to improve or alter Milpa–maize monocrop discrimination.

The subsequent chapters operationalise this framework for Oaxaca through the reference data, sensor-specific preparation, feature construction, model development and controlled comparisons required to evaluate the mapping problem.

Chapter 3: Objectives

3.1 General context

The Milpa mixed-cropping system is a maize-centred agricultural system whose composition and spatial arrangement can vary among parcels and locations. This variability complicates remote-sensing discrimination from maize monocrop because maize may dominate both systems, while companion crops and other within-field components occur unevenly.

Sentinel-2 provides the seasonal multispectral baseline for this study. SkySat and PlanetScope are evaluated as complementary sources of finer-spatial information, while the analysis also tests image-level reconstruction and post-classification integration. The central question is whether these additional information routes improve Milpa–maize monocrop discrimination beyond the Sentinel-2 baseline under geographically separated evaluation.

3.2 Objectives

The overall objective of this thesis is to assess whether complementary information from SkySat and PlanetScope improves parcel-level discrimination between the Milpa mixed-cropping system and maize monocrop beyond a seasonal Sentinel-2 baseline in Oaxaca. A secondary three-class formulation is retained to examine the additional separation of Milpa, landrace maize and hybrid maize.

The specific objectives are:

1. To develop and evaluate a Sentinel-2 parcel-level baseline for Milpa versus maize monocrop, with a secondary three-class assessment of Milpa, landrace maize and hybrid maize.

2. To determine whether finer-spatial information from SkySat and PlanetScope provides additional classification value beyond Sentinel-2, using both native-sensor predictors and predictor-level integration with Sentinel-2.

3. To assess SkySat-guided Sentinel-2 super-resolution relative to bicubic interpolation and determine whether reconstruction improvements translate into better downstream Milpa–maize classification.

4. To determine whether classification-level integration of independently fitted Sentinel-2, SkySat and PlanetScope models improves discrimination relative to the individual sensor models.

All objectives are assessed using geographically controlled development analyses and a separated secondary benchmark, with class-specific Milpa performance considered alongside aggregate measures.

Chapter 4: Methodology

This chapter describes the data and analytical workflow used to map the Milpa mixed-cropping system and maize monocrop across the four study areas in Oaxaca. It first introduces the geographical and topographic setting of the study and describes the field-reference parcels, target-class definitions and sensor-specific parcel preparation. The Sentinel-2 workflow is then presented, including image harmonisation, cloud and invalid-pixel masking, temporal interpolation, observation-support assessment, feature construction and development of the primary classification baseline.

The later sections describe the complementary use of imagery with finer spatial sampling. SkySat was evaluated both as a source of direct parcel-level predictors and as a reference for experimental Sentinel-2 super-resolution, while PlanetScope provided an additional set of spectral and spatial parcel predictors. The chapter then distinguishes predictor-level, image-level and classification-level integration and describes the geographically separated development and secondary benchmark framework used across the comparative analyses. Particular attention is given to maintaining consistent parcel identities, preventing information leakage and distinguishing primary, sensitivity, diagnostic and corroborative analyses.

4.1 Study area

4.1.1 Geographical setting and AOI organisation

The study was conducted in four agricultural areas of interest (AOIs) in the state of Oaxaca, Mexico: Monte Verde (AOI1), Nochixtlan (AOI2), Tlaxiaco (AOI3), and Valles Centrales (AOI4). These AOIs formed the spatial framework for organising the field reference parcels and the satellite datasets used in the study.

All spatial data were processed in WGS 84 / UTM Zone 14N (EPSG:32614), providing a common metric coordinate system for raster–vector alignment, parcel-based spatial operations, and area calculations. Figure 4.1 shows the location of Oaxaca within Mexico, the distribution of the four AOIs within the state, and the reference parcels available in each study area.

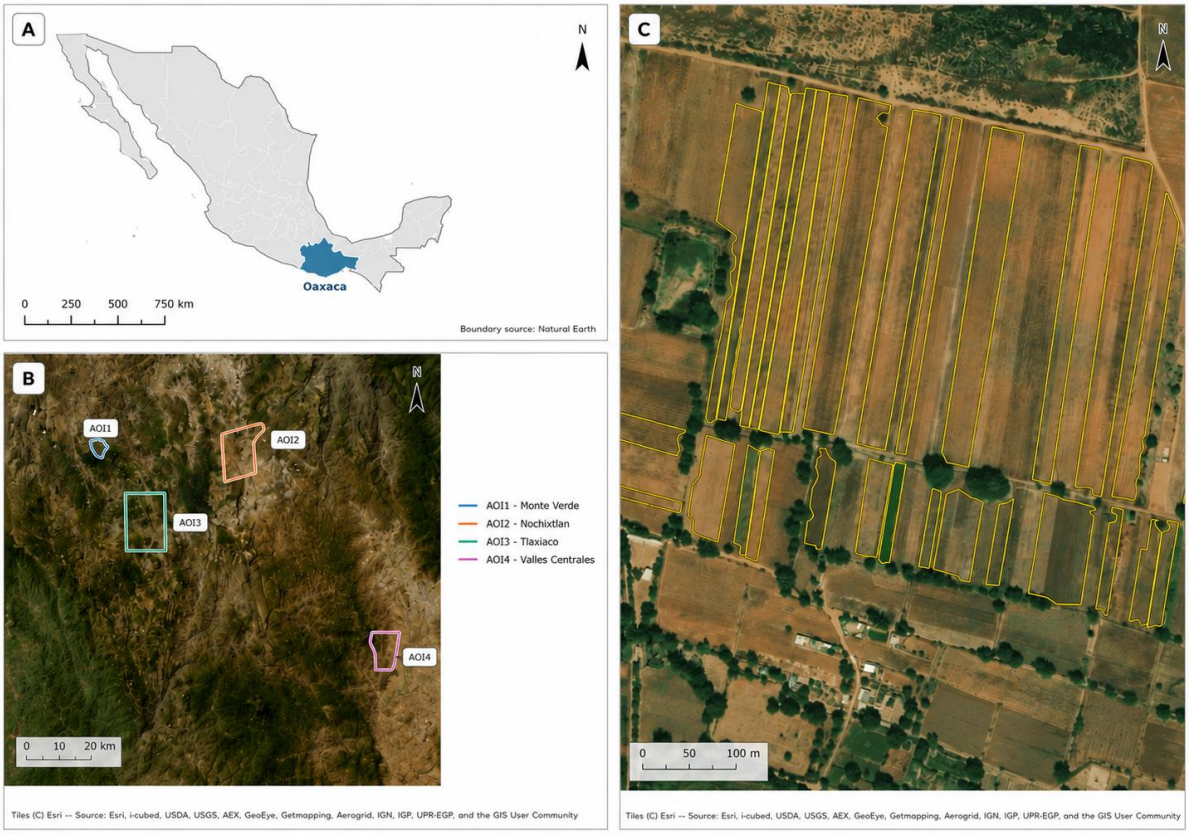


Figure 4.1. Study-area context and reference-parcel setting. (A) Oaxaca within Mexico; (B) the four AOIs within Oaxaca; and (C) an illustrative parcel context showing the field polygons used for satellite extraction.

4.1.2 Topographic setting of the reference-parcel footprints

Topographic conditions across the field reference dataset were characterised using the NASA Shuttle Radar Topography Mission Global 1 arc-second digital elevation model (SRTMGL1 DEM; NASA JPL, 2013). The DEM was reprojected from geographic coordinates to WGS 84 / UTM Zone 14N (EPSG:32614) and resampled to a 30 m grid using bilinear interpolation, which is suitable for continuous elevation data. Slope was calculated in degrees from the horizontal gradients of the projected elevation surface. Before topographic extraction, the parcel datasets were checked against the expected AOI-specific parcel counts and combined footprint areas. For each area of interest (AOI), elevation and slope values were extracted from DEM cells whose centres fell within the combined footprint of the original reference-parcel polygons. Minimum, maximum and mean elevation, together with mean slope, were then calculated for each AOI. Table 4.1 therefore describes the terrain represented by the sampled agricultural parcels rather than the complete geographic extent of each AOI. These variables were used solely to characterise the study setting and were not included as predictors in any classification model.

Table 4.1. Topographic characteristics of the original reference-parcel footprints in the four AOIs, derived from the 30 m SRTMGL1 DEM.

AOI	Area name	Elevation min–max (m)	Mean elevation (m)	Mean slope (°)
AOI1	Monte Verde	2144–2480	2286	12.08

AOI	Area name	Elevation min–max (m)	Mean elevation (m)	Mean slope (°)
AOI2	Nochixtlan	2045–2310	2090	2.93
AOI3	Tlaxiaco	1977–2318	2148	4.58
AOI4	Valles Centrales	1497–1555	1521	3.23

4.2 Field-reference data and parcel preparation

This section describes the field-reference dataset and the parcel preparation steps used to establish a consistent reference population for the satellite analyses. It first presents the original crop labels and target-class formulations, followed by the geometry checks and sensor-specific inward buffers applied before data extraction.

4.2.1 In-situ reference dataset and target-class formulation

The reference dataset was provided by the supervisory team and was reported to originate from a field survey conducted in 2024. It comprised 457 georeferenced agricultural parcel polygons distributed across Monte Verde (AOI1), Nochixtlan (AOI2), Tlaxiaco (AOI3), and Valles Centrales (AOI4). The available reference documentation did not specify the collecting organisation, survey protocol or parcel-boundary delineation procedure. Crop labels used in the analysis were derived from the source attribute Crop_cl_l2. The source values were harmonised for analysis, and their AOI-wise composition after grouping into the three target crop categories and non-target crops is presented in Table 4.2.

Of the 457 parcels, 328 belonged to the target categories: 118 Milpa, 164 landrace maize, and 46 hybrid maize parcels. The remaining 129 non-target parcels were excluded from the Milpa–maize classification targets. Two target formulations were used according to the analysis. The three-class formulation retained Milpa, landrace maize, and hybrid maize as separate labels. For the binary formulation, landrace and hybrid maize were combined before model fitting because the analytical question concerned cropping-system type rather than maize-subtype identification. This retained the original maize-type distinction where relevant while providing the Milpa mixed-cropping system versus maize monocrop comparison required by the principal research objective.

Table 4.2. Composition of the original 2024 field-reference dataset before parcel preparation and satellite-data screening

AOI	Study area	Total parcels	Milpa	Landrace maize	Hybrid maize	Non-target crops
AOI1	Monte Verde	115	50	59	0	6
AOI2	Nochixtlan	146	8	42	27	69
AOI3	Tlaxiaco	47	9	29	3	6
AOI4	Valles Centrales	149	51	34	16	48
Total		457	118	164	46	129

4.2.2 Parcel geometry quality control and inward buffering

The original parcel polygons were checked for missing, empty, and invalid geometries, and the AOI-specific parcel counts were verified. All 457 parcels passed these checks. Original parcel identifiers, AOI assignments, and crop labels were retained to maintain the link between the field records and all derived geometries.

Inward buffering was used to reduce the influence of neighbouring fields, paths, and other land-cover types along parcel boundaries. For Sentinel-2, a 10 m inward buffer, equivalent to one pixel on the analysis grid, was applied to the original polygons. Parcels that did not retain valid interior geometry were removed, leaving 270 buffered parcels. Of these, 226 satisfied the Sentinel-2 eligibility criteria described in Section 4.3.3. Restricting this set to the three target crop categories produced the final 149-parcel cohort.

The same 149 parcel identifiers were linked back to their original polygons before the sensor-specific geometries were prepared. A 1 m inward buffer was used for SkySat predictor extraction and the super-resolution workflow, while a 3 m inward buffer was applied for the PlanetScope workflow. All buffers were generated from the original polygons rather than from the 10 m buffered layer, thereby avoiding cumulative reduction of parcel area.

Figure 4.2 summarises the successive screening stages and shows how the common parcel identifiers were linked to the sensor-specific buffered geometries.

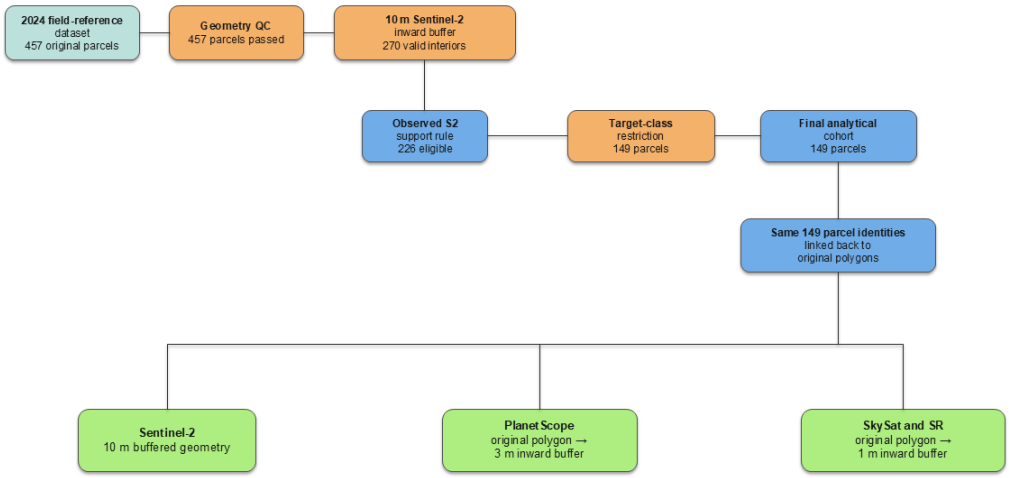


Figure 4.2. Reference-parcel preparation, screening and sensor-specific geometry lineage. The 457 original parcels were checked, buffered and screened to obtain the 149 target-class parcels, after which the common identifiers were linked back to the original polygons for the Sentinel-2, PlanetScope, SkySat and super-resolution geometries.

4.3 Sentinel-2 data preparation

Sentinel-2 Level-2A imagery was prepared to generate spatially and temporally consistent inputs for analysis across the four AOIs. The preparation included tile-wise organisation, spatial harmonisation, cloud and invalid-pixel masking, temporal resampling and gap filling,

AOI clipping, and parcel eligibility screening. Two complementary products were retained: observed cloud-masked images for data-support assessment and texture analysis, and regular 10-day gap-filled stacks for spectral-temporal feature extraction.

4.3.1 Time-series composition, spatial harmonisation and cloud masking

Sentinel-2 Level-2A images acquired throughout 2024 were used to capture seasonal variation in crop development. The Level-2A products provided surface reflectance generated through atmospheric correction with the Sen2Cor processor (Copernicus Sentinel-2, 2021). The images were organised by acquisition date for tiles T14QPE, covering AOI1, AOI2 and AOI3, and T14QQD, covering AOI4.

The preprocessing inventory retained eleven selected Sentinel-2 bands. B02, B03, B04 and B08 were available at 10 m, whereas B05, B06, B07, B8A, B11 and B12, together with the available 20 m representation of the natively 60 m B01 band, were resampled to the common 10 m analysis grid using bilinear interpolation. Binary mask layers were resampled using nearest-neighbour interpolation to preserve their categorical codes. B01 was retained in the prepared products and subsequent temporal harmonisation for preprocessing quality control but was excluded before feature construction. Parcel- and pixel-level modelling therefore used ten bands: B02, B03, B04, B05, B06, B07, B08, B8A, B11 and B12. The native spatial resolutions of the Sentinel-2 spectral bands are documented by the European Space Agency (2015).

Cloud and cloud-shadow contamination was identified using FMask-derived validity masks supplied through the Sen4Stat processing chain (Bontemps et al., 2024). Sen4Stat is an ESA-funded processing system developed to support the use of Earth-observation information in agricultural statistics (Bontemps et al., 2024). In the supplied FMask classification, cloud-shadow and cloud pixels corresponded to classes 2 and 4, respectively; these classes were flagged as invalid in the corresponding binary masks used in the analysis. Before masking, each binary mask was checked against its corresponding reflectance image for coordinate reference system, raster dimensions and spatial transform. The masks were aligned to the common 10 m grid using nearest-neighbour resampling. Pixels flagged as invalid were assigned as no-data, while unflagged pixels retained their reflectance values. The resulting cloud-masked images were organised by acquisition date, and the operational temporal stack used for gap filling contained 69 common input dates between 3 January and 23 December 2024. These observed products were retained for support assessment and texture calculation and provided the valid-pixel information used during temporal gap filling.

Figure 4.3 illustrates the relationship between the source imagery, the categorical quality mask and the observed cloud-masked product retained for support assessment and subsequent temporal processing.

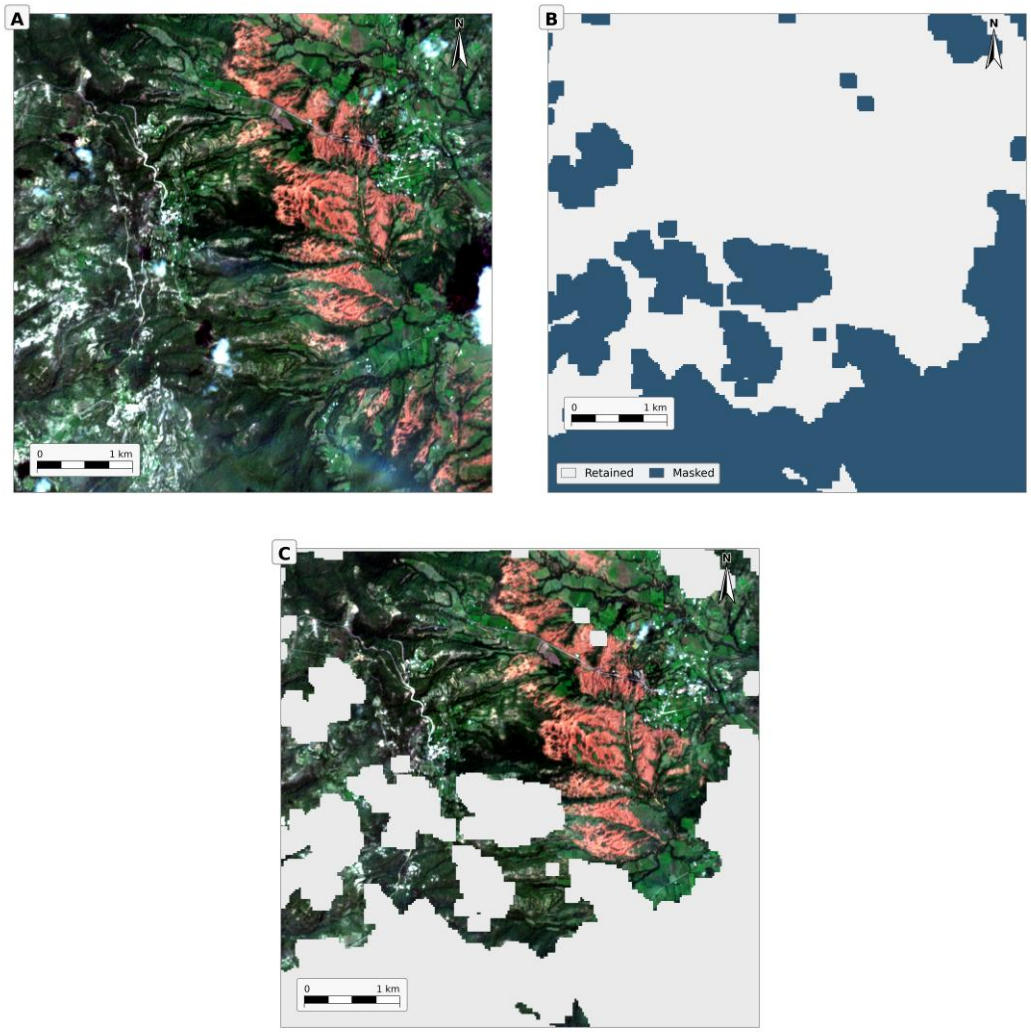


Figure 4.3. Sentinel-2 cloud and invalid-pixel masking. (A) Source Sentinel-2 image before masking; (B) categorical mask distinguishing retained and masked pixels; and (C) the observed cloud-masked product retained for support assessment, texture calculation and gap filling.

4.3.2 Temporal resampling and gap filling

After cloud masking, valid Sentinel-2 observations were unevenly distributed through time and differed among pixels. A regular temporal sequence was therefore required to compare seasonal reflectance patterns across parcels and AOIs. The 69 common input dates were resampled to a 10-day sequence containing 36 dates.

Temporal gap filling was performed for each band and pixel using linear interpolation implemented with Orfeo ToolBox. Where valid observations were available, they were retained; otherwise, values at target dates bound by valid observations were estimated from the nearest valid observations before and after the target date. Linear interpolation was used as a simple and reproducible method for representing gradual changes in surface reflectance between observed dates.

The contribution of original cloud-free observations was recorded for each AOI and output date to distinguish well-supported periods from those relying more heavily on interpolation. The resulting 36-date stacks were used for spectral and temporal feature extraction. The original cloud-masked observations were retained separately for data-support assessment and texture analysis.

4.3.3 AOI-specific products and parcel eligibility

Following temporal preprocessing, the observed cloud-masked images and the 36-date gap-filled stacks were clipped to the four areas of interest. Parcel eligibility was then assessed using the 10 m inward-buffered geometries described in Section 4.2.2. A parcel-date was considered supported when at least one pixel whose centre fell within the 10 m inward-buffered parcel contained jointly valid B04 and B08 values. A parcel was classified as Sentinel-2 eligible when this condition was met on at least one observed date. Of the 270 parcels that retained valid buffered geometries, 226 met this minimum support requirement. Restricting the eligible parcels to Milpa, landrace maize and hybrid maize produced the final 149-parcel Sentinel-2 cohort used for subsequent feature engineering and classification. This eligibility criterion established only whether a parcel had any usable observed support; it did not describe the amount, continuity or temporal distribution of that support, which were evaluated separately.

4.3.4 Observed-data support and interpolation quality assessment

Annual observation support was assessed for the buffered parcels using the number and proportion of supported Sentinel-2 acquisition dates, the longest consecutive sequence of unsupported dates, and the number of jointly valid parcel-interior B04–B08 pixels. Pixel support was summarised as the total accumulated across the observed acquisitions and as the maximum and mean support per acquisition date. Both coverage and continuity were considered because a parcel could have valid observations on several dates while still containing a long uninterrupted period without usable data. These support measures were summarised using the study-specific operational classes shown in Table 4.3.

Table 4.3. Study-specific operational classes used to summarise annual Sentinel-2 observation support.

Support class	Supported acquisition dates	Maximum consecutive unsupported acquisition dates
Strong	≥60%	≤4
Moderate	≥30%	≤8
Weak	Did not meet the moderate criteria	—

A parcel was classified as strong only when both strong-support conditions were met. Parcels that did not meet these conditions were classified as moderate when both moderate-support conditions were satisfied; all remaining parcels were classified as weak. These classes were study-specific quality-assurance strata used for reporting, interpretation and sensitivity analysis. They were not universal reliability standards and did not determine primary parcel eligibility.

Interpolation quality was assessed according to the calendar-day interval between the valid observations surrounding a missing value. These intervals were grouped relative to the regular 10-day temporal grid, as shown in Table 4.4.

Table 4.4. Operational interpolation-gap categories defined relative to the 10-day temporal grid.

Gap category	Interval between bracketing valid observations	Approximate temporal-grid span
Short	≤20 days	Up to 2 intervals
Medium	21–40 days	3–4 intervals
Long	41–60 days	5–6 intervals
Very long	>60 days	More than 6 intervals

Interpolation reliability was tested through leave-one-valid-observation-out validation of the Normalised Difference Vegetation Index (NDVI). Each interior valid observation was temporarily withheld and reconstructed from the immediately preceding and following valid observations. The first and last valid observations were not withheld because they were not bounded on both sides, and extrapolation beyond the available observations was not permitted.

Saved gap-filled NDVI values were also checked against observed values at matching dates and, separately, against off-grid observations within ±5 days, with exact matches excluded. The latter comparison was treated as supporting context rather than same-day validation because its differences could combine interpolation discrepancy with genuine phenological change. August–October was examined independently because persistent gaps during this important crop-development period required closer assessment. The seasonal checks included B02, B03, B04, B05, B06, B08 and B11, together with NDVI, the B08–B05 normalised difference red-edge index, the Plant Senescence Reflectance Index (PSRI), and the B08–B11 Normalised Difference Moisture Index (NDMI).

Finally, the records were screened for residual dark-shadow or cloud-like behaviour, physically implausible index values, and isolated temporal spikes or drops. Retention of parcels was compared across the original, buffered and Sentinel-2-usable cohorts by AOI, crop group and original parcel-size class. These diagnostics were used to assess how strongly each parcel time series was supported by original observations, whether successive screening stages altered the composition of the reference cohort, and whether the main classification conclusions remained consistent within a stricter, better-supported subset. They were not included as classification predictors because they described data availability and reliability rather than crop characteristics, and could therefore introduce acquisition- or AOI-related bias. They did not trigger automatic parcel removal beyond the primary eligibility rule.

4.4 Sentinel-2 feature engineering and classification

The prepared Sentinel-2 data were converted into spectral, temporal and spatial predictors for crop-system classification. The primary analysis used one feature record per reference parcel, consistent with the field-level reference labels. A separate pixel-level workflow examined an alternative spatial representation. In both workflows, development and evaluation domains were separated geographically to prevent information leakage.

4.4.1 Parcel-level feature construction

The feature design represented two complementary properties of the crop systems: seasonal spectral behaviour and spatial variation within fields. Features were constructed for the final 149 parcels using ten modelling bands, B02, B03, B04, B05, B06, B07, B08, B8A, B11 and B12, across the 36 regular Sentinel-2 dates.

For the main temporal branches, valid pixels were first averaged within each 10 m buffered parcel to produce one reflectance trajectory for each band. NDVI, EVI2, NDRE, CI red edge, NDMI and PSRI were then calculated from these parcel-mean trajectories. This calculation order produced one consistent spectral and index record for each field. Finite-value and near-zero-denominator safeguards were applied during index calculation. The definitions, band inputs and formulae for the six indices are provided in Appendix Table A.1.

Before phenological and stability metrics were calculated, the parcel-level index trajectories were smoothed using a centred three-date rolling median. This reduced the influence of isolated fluctuations between consecutive dates while retaining the broader seasonal trajectory. Phenological metrics were calculated from NDVI, EVI2, NDRE, NDMI and PSRI, whereas stability metrics were calculated from NDVI, EVI2, NDRE and NDMI.

The spatial branches followed a different calculation order because averaging the pixels first would remove the variation that these branches were intended to measure. Within-parcel heterogeneity variables and the indices used for texture calculation were therefore derived at pixel level before being summarised within each parcel.

Six temporal windows represented different parts of the annual crop cycle. January to March described the early season context, June to August represented development before the seasonal peak, September to October covered peak and late season conditions, October to November described the transition towards senescence, November to December represented the end of the growing season, and the complete annual sequence described overall seasonal behaviour.

Raw band summaries used the median, mean, standard deviation, interquartile range, coefficient of variation, P10, P90 and the difference between P90 and P10. Index summaries used the median, standard deviation, interquartile range, P10, P90 and the difference between P90 and P10.

Within-parcel heterogeneity was calculated spatially across valid pixels within each 10 m buffered parcel at each date using P10, P25, P50, P75, P90, interquartile range, standard deviation, standard deviation divided by the absolute mean, and the difference between P90 and P10. The resulting date-specific statistics were then aggregated using the median across dates within each temporal window. Observed-image texture was calculated from pixel-level B07, B8A, B11, NDRE and NDMI using the available cloud-masked observations. Local standard deviation was computed on the observed raster using full 3 × 3 and 5 × 5 neighbourhoods; parcel summaries were then calculated from neighbourhood centres falling within the 10 m buffered parcel. For the September–October and October–November windows, the parcel-level texture measures were the mean local standard deviation for the 3 × 3 and 5 × 5 neighbourhoods and the 90th percentile of local standard deviation for the 5 × 5

neighbourhood. These date-specific values were aggregated using the median across the available observed dates within each temporal window.

Figure 4.4 illustrates the centred rolling-median smoothing applied to an index trajectory and the neighbourhood-based construction of observed-image texture within a parcel.

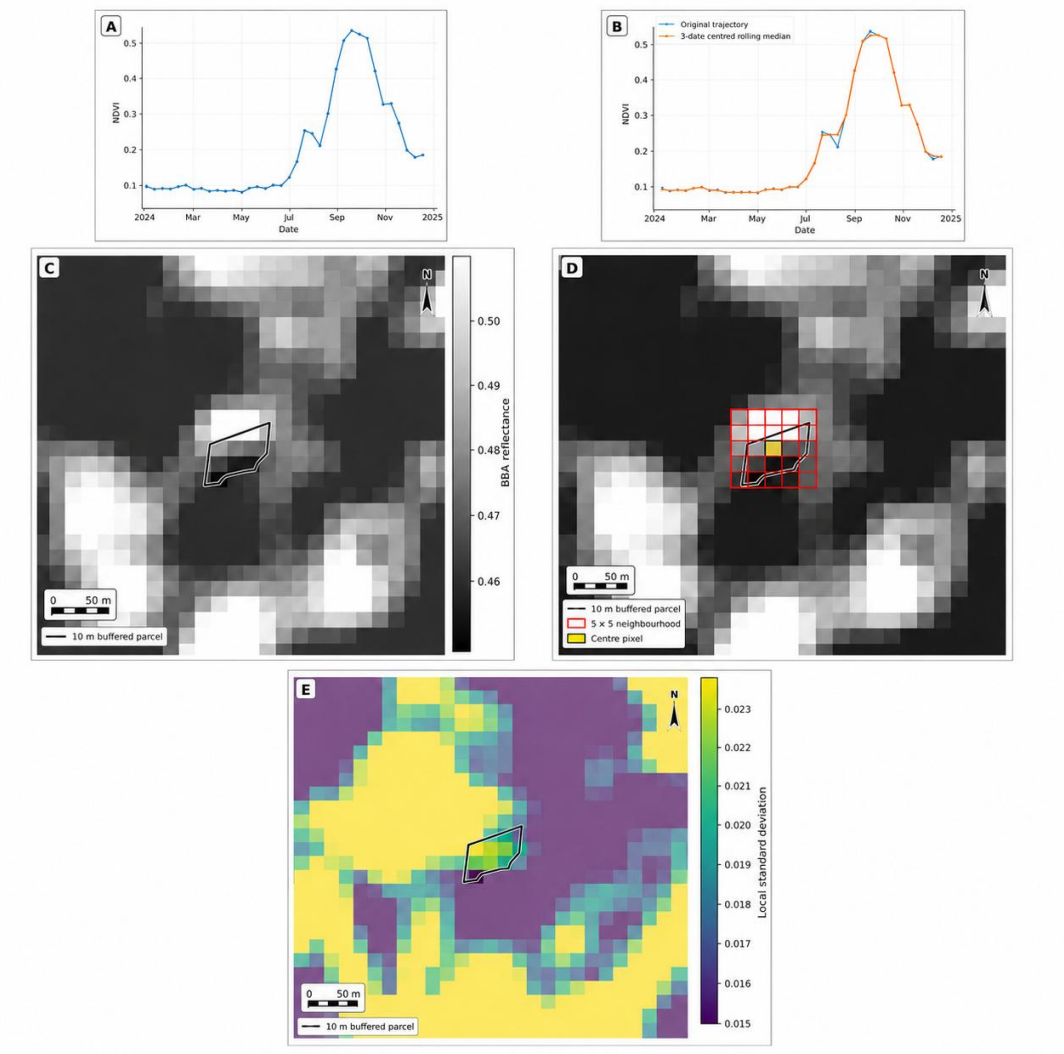


Figure 4.4. Sentinel-2 feature-construction examples. (A) NDVI trajectory before centred three-date rolling-median smoothing; (B) a comparison of the original NDVI trajectory and the centred three-date rolling median; (C) an observed B8A parcel patch; (D) the 5 × 5 neighbourhood used for local texture calculation; and (E) the local-standard-deviation surface derived from the neighbourhood analysis.

Table 4.5. Sentinel-2 parcel-level feature families and their representation in the final 25-predictor panel.

Feature branch	Construction	Temporal basis	Candidates	Final 25
Raw-band statistics	Ten parcel-mean band trajectories summarised using eight statistics	Six temporal windows	480	8
Index statistics	Six indices calculated from parcel mean bands and summarised using six statistics	Six temporal windows	216	8
Phenology and stability	Five indices with eight phenology metrics and four indices with five stability metrics	Complete 36-date sequence	60	6
Within-parcel heterogeneity	Eight pixel-level variables summarised using nine spatial statistics, followed by the median across dates	Three temporal windows	216	2
Observed-image texture	B07, B8A, B11, NDRE and NDMI represented by mean local standard deviation in 3 × 3 and 5 × 5 neighbourhoods and P90 local standard deviation in a 5 × 5 neighbourhood, followed by median aggregation across available observed dates	Two temporal windows September–October; October–November	30	1

The raw-band, index, phenology and stability, and heterogeneity branches used the relevant gap-filled products so that the same temporal positions were available for all parcels. Texture was calculated from the observed cloud-masked imagery. This prevented spatial variation created by temporal interpolation from being interpreted as observed field structure.

The five branches produced a pool of 1,002 candidate features. Missing values in the observed texture branch were replaced using medians estimated only from the AOI3 and AOI4 development parcels. The same development-derived medians were then applied to the evaluation parcels. The merged matrix was checked for remaining missing or infinite values, duplicate feature names and duplicate-valued columns. The resulting matrix contained one complete feature record for each of the 149 parcels and was passed to development-only dimensionality and redundancy control.

4.4.2 Development-only feature control and final predictor panel

The 1,002 candidate predictors were too numerous relative to the 71 parcels available for model development, and many represented closely related spectral or temporal properties. Feature control was therefore performed entirely within AOI3 and AOI4. The 78 parcels from AOI1 and AOI2 remained outside this process as the separated secondary benchmark; they were not used for imputation, feature scoring, redundancy control, panel comparison or final predictor selection.

Redundancy was first assessed within each feature family using an absolute Spearman correlation threshold of 0.90, retaining representatives from highly correlated groups. The retained predictors were then compared across feature families using a threshold of 0.95. These stages reduced the feature pool from 1,002 predictors to 558 within-family representatives and then to 442 cross-family representatives.

The remaining predictors were scored using complementary development-only evidence: eta-squared (η^2), an effect-size measure expressing the proportion of total feature variance

associated with class grouping, for the three-class and binary targets; a robust Milpa–maize effect measure; repeated Random Forest importance; and Top 50 selection frequency. The robust effect measure was the absolute difference between the Milpa and maize monocrop medians divided by the interquartile range of the features across the development parcels. Repeated importance was estimated from 300-tree Random Forests using three stratified folds and eight repeats, giving 24 fitted forests.

Candidate panel sizes were then assessed entirely within the development data using repeated stratified three-fold cross-validation with 20 repeats, giving 60 folds. Feature selection was repeated within each training fold before the selected panel was evaluated on the corresponding validation fold. Panels of 15, 20, 25 and 30 predictors were used for the principal panel-size comparison, with 40 predictors retained as an additional diagnostic. A panel was considered within tolerance when its development performance was no more than 0.03 below the best macro-F1 and no more than 0.05 below the best Milpa recall and Milpa F1. The smallest panel meeting these criteria contained 15 predictors.

This result showed that a compact predictor set was sufficient but did not automatically determine the final panel. Repeated-fold selection frequency was subsequently used as the main stability criterion, with full-development feature scores and representation across the five feature branches used as supporting controls. Predictors with zero selection frequency were excluded. The resulting candidates were checked for residual redundancy and branch representation, producing a ranked 40-feature representative list. Ranks 1–25 were retained for the primary models and ranks 26–40 as reserve features. This was a development-only stability and diversity adjudication rather than an automatic optimum from the panel-size comparison, and no AOI1 or AOI2 performance was used. The final 25 predictors and reserve ranks 26–40 are reported in Appendix Table A.2.

4.4.3 Parcel-level classification and sensitivity analyses

The primary classification was conducted at parcel level, with one predictor vector and one class label for each parcel. The primary classifier was a Random Forest using the Gini criterion, unrestricted tree depth and bootstrap sampling. The model used 800 trees, considered the square root of the available predictors at each split, used a minimum split size of two and required at least two parcels in each terminal leaf. Balanced class weights were used, and the random state was fixed at 42. The number of trees represented the size of the Random Forest ensemble rather than the number of training parcels. A development-only sensitivity check across 100–1,000 trees, with all other settings fixed, showed only small performance differences once several hundred trees were used; the 800-tree configuration was therefore retained. Models were fitted for two target formulations: a three-class formulation distinguishing Milpa, landrace maize and hybrid maize, and a binary formulation contrasting Milpa with maize monocrop.

Development assessment was conducted only within AOI3 and AOI4 using repeated stratified three-fold cross-validation with 20 repeats. After the predictor panel and model configuration had been fixed, a final model for each target formulation was fitted using all 71 development parcels and evaluated once on the 78 parcels from AOI1 and AOI2. No predictor, model parameter or class definition was changed after the secondary-benchmark predictions were

obtained. Performance was assessed using accuracy, balanced accuracy, macro-F1, weighted F1, class-specific precision, recall and F1, and confusion matrices, allowing performance for the less frequent classes to be examined separately from overall accuracy.

A better-supported subset was analysed to assess sensitivity to weaker observation support. The subset contained 132 parcels, comprising all 71 development parcels and 61 benchmark parcels. It retained parcels with strong or moderate annual observation support and excluded parcels classified as unreliable for the August–October period, unsuitable for fine phenological interpretation, or affected by a hard residual artefact flag. The same 25-predictor panel and Random Forest configuration were retained so that the analysis differed only in observational support. Quality-control variables were used only to define the subset and were not included as model predictors.

Support Vector Machine models were evaluated as a diagnostic comparison with the Random Forest. Predictors were standardised before fitting a Support Vector Classifier with balanced class weights. Linear and radial basis function kernels were examined. For both kernels, C was tested at 0.1, 1 and 10, while for the radial basis function kernel, gamma was tested as scale, 0.01 and 0.1. Parameter selection used three-fold cross-validation within the development data, with macro-F1 as the selection criterion. The selected three-class model used a linear kernel with C = 0.1, while the selected binary model used a radial basis function kernel with C = 1 and gamma = scale.

4.4.4 Pixel-level diagnostic

A pixel-level workflow was implemented as a secondary diagnostic to examine an alternative spatial representation while retaining the same geographic separation as the primary parcel-level analysis. Sentinel-2 pixels whose centres fell inside the buffered parcel geometries were extracted, producing 3,367 pixels from the 149 parcels. The parcel-level geographic split was applied before feature preparation: AOI3 and AOI4 provided 1,465 development pixels, while AOI1 and AOI2 provided 1,902 benchmark pixels. Pixels from the same parcel therefore could not occur in both domains.

Because larger parcels contained more pixels, a maximum of 20 pixels was sampled from each development parcel using the fixed random seed 20260702. This produced 811 training pixels. Benchmark pixels were retained without capping or resampling. The initial pixel representation contained 1,764 predictors, calculated as 18 x [(6 x 16) + 2]. The 18 band and index variables were summarised using 16 statistics across six temporal windows together with two seasonal contrasts. The complete variable, statistic and contrast definitions are provided in Appendix Table A.3.

Feature filtering and imputation were based only on the development data. A total of 216 predictors that were entirely missing or constant in the development matrix were removed, leaving 1,548 predictors. Missing values in both the development and benchmark matrices were then replaced using medians estimated from the development data. The final matrices contained 811 x 1,548 development records and 1,902 x 1,548 benchmark records.

Pixel classification used a Random Forest with the Gini criterion, unrestricted tree depth, bootstrap sampling and a minimum split size of two. The model used 500 trees, considered

the square root of the predictors at each split, required at least two pixels in each terminal leaf and used balanced-subsample class weights. The random state was fixed at 20260702. Both three-class and binary diagnostics were performed. Predictions were assessed first at pixel level and then aggregated to parcel level by averaging the predicted class probabilities across all pixels belonging to each parcel and assigning the class with the highest mean probability.

4.5 SkySat preprocessing and predictor-level classification

Section 4.4 established the Sentinel-2 baseline for the final 149 parcels. SkySat was then used to represent the same fields at finer spatial sampling, allowing the predictor sets to be compared while keeping the reference parcels and geographic evaluation framework unchanged.

4.5.1 Scene preparation, masking and mosaic construction

The SkySat archive comprised 33 analytic surface-reflectance scenes and their corresponding UDM2 quality products. Each analytic scene was paired with its quality product and checked for spatial agreement. Pixels were retained when the UDM2 clear layer equalled 1 and the cloud, shadow, light-haze, heavy-haze and snow layers equalled 0. Cleaned scenes were combined only when they represented the same AOI, acquisition date and location, producing 14 AOI/date/location mosaics.

The temporal distribution of the 14 mosaics is summarised in Table 4.6. SkySat coverage was concentrated between 10 and 25 October 2024 and differed among the four AOIs. Multiple mosaics could occur for the same AOI and date where separate locations were represented; mosaics were not combined across AOIs.

Table 4.6. Temporal distribution of the SkySat mosaics used in the analysis.

AOI	Number of mosaics	Acquisition dates in 2024
AOI1	2	16 and 25 October
AOI2	5	10, 13 and 16 October
AOI3	4	15, 16 and 23 October
AOI4	3	12, 15 and 16 October
Total	14	10–25 October 2024

The four AOI-specific crop masks available in the processing archive were aligned to the corresponding SkySat grids using nearest-neighbour resampling. The aligned masks were intersected with valid SkySat image support to produce 14 crop-only mosaics. The exact producer and derivation procedure of these crop masks could not be established from the verified processing archive. They were therefore used only to restrict extraction to mapped crop support and were not treated as an independent source of reference labels.

The spectral products retained Blue, Green, Red and NIR bands. NDVI was calculated from valid Red and NIR observations where both values were finite and the denominator was non-zero. The crop-only spectral mosaics and corresponding NDVI products formed the SkySat inputs for parcel-level extraction.

Figure 4.5 shows how the aligned crop-support mask and valid-pixel screening produced the crop-only RGB and NDVI inputs used for parcel extraction.

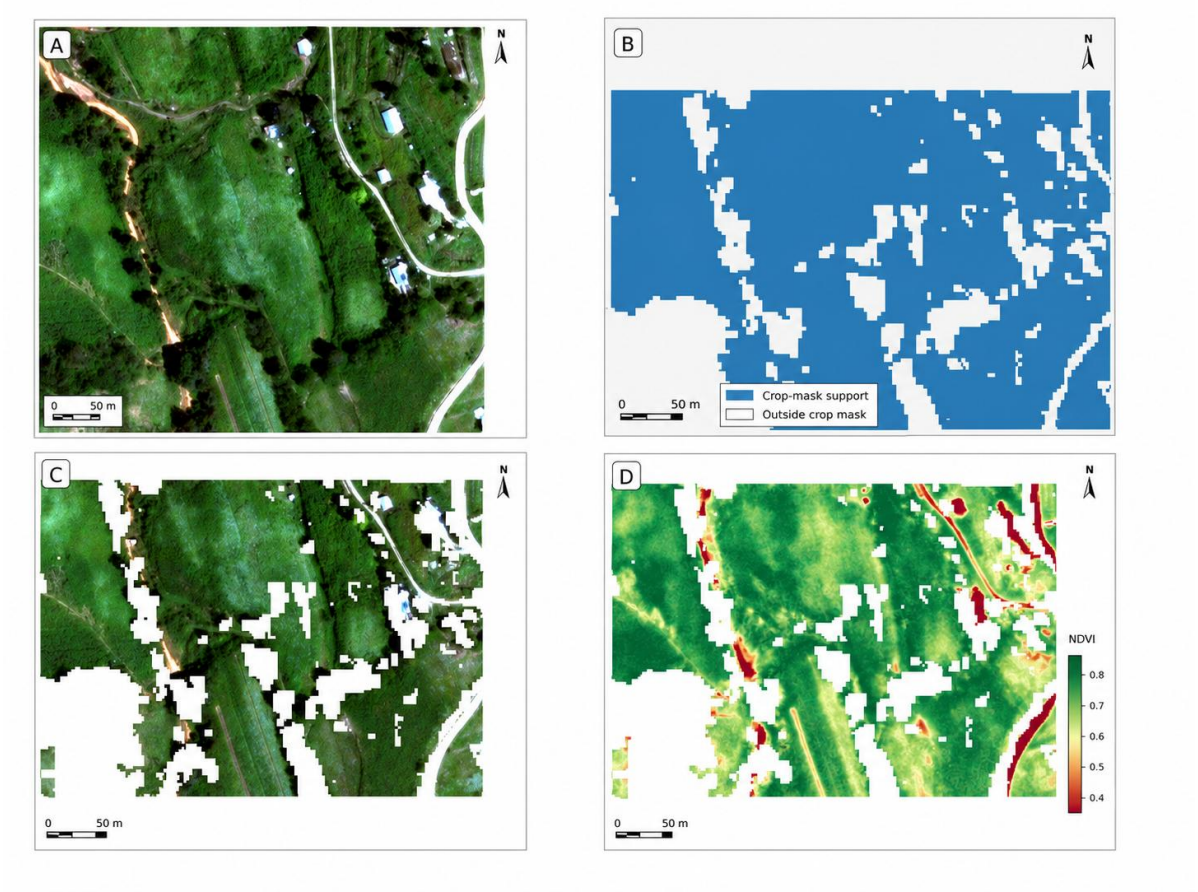


Figure 4.5. SkySat preprocessing and parcel-level predictor inputs. (A) Cleaned SkySat RGB mosaic; (B) aligned crop-support mask; (C) crop-only RGB product; and (D) crop-only NDVI product used for parcel summarisation.

4.5.2 Original-geometry parcel extraction and predictor construction

The SkySat workflow retained the same 149 parcel identities used for Sentinel-2, but parcel extraction was based on the original field geometries rather than the 10 m Sentinel-2 buffers. The Sentinel-2 geometry layer was used only to link each parcel identifier back to its original polygon. Candidate polygons were restricted to the same AOI, and the polygon with the greatest intersection relative to the linkage geometry was retained when the overlap ratio was at least 0.95. A one-to-one match was enforced so that each parcel identifier was linked to only one original polygon.

A 1 m inward buffer was applied directly to each matched original polygon, and all 149 buffered parcels remained non-empty. Pixels were retained when their centres fell within the buffered polygon and the aligned crop mask indicated valid crop support. A parcel-product record required at least one pixel with finite values in all four spectral bands, with at least one of the four values differing from zero. NDVI observations were additionally required to be finite and within the range -1 to 1.

For each valid parcel-product combination, the mean and median were calculated for Blue, Green, Red, NIR and NDVI, giving ten variables and 316 parcel-scene records. Where a

parcel intersected more than one location product from the same acquisition date, the corresponding predictor values were averaged within the parcel-date group. Consolidating 71 such duplicate groups produced 245 unique parcel-date records.

The ten variables were then averaged across the distinct valid dates available for each parcel, with each date contributing equally. All 149 parcels retained valid SkySat support, and no imputation was required. The resulting ten SkySat predictors were used alone and together with the 25 Sentinel-2 predictors defined in Section 4.4.2, giving Sentinel-2-only, SkySat-only and 35-predictor combined representations.

4.5.3 SkySat and Sentinel-2–SkySat predictor-level classification design

The predictor-level comparison addressed the binary distinction between Milpa and maize monocrop using three representations: the 25 Sentinel-2 predictors, the ten SkySat predictors, and the combined set of 35 predictors. The same 149 parcels, target definition, classifier specification and spatial evaluation design were used for all three representations so that the comparison differed only in predictor information.

Each representation was evaluated using the same Random Forest specification with 1,000 trees, the Gini criterion, unrestricted tree depth, bootstrap sampling, square-root feature sampling, a minimum split size of two, a minimum terminal-leaf size of two, balanced-subsample class weights and a random state of 42. The Sentinel-2-only model used in this comparison was therefore fitted under the same 1,000-tree specification as the SkySat and combined models and was distinct from the primary 800-tree Sentinel-2 model described in Section 4.4.3.

Geographic transfer within the development domain was assessed through two directional folds: AOI3-to-AOI4 and AOI4-to-AOI3. Together, these produced out-of-fold predictions for all 71 development parcels, with each AOI used once for model fitting and once for geographic validation.

After the predictor definitions and classifier specification had been fixed, each representation was fitted using all 71 development parcels and evaluated once on the 78 parcels from AOI1 and AOI2. This assessment was treated as a secondary, non-confirmatory benchmark because these AOIs had been examined earlier in the thesis. They remained outside model development, so benchmark outcomes were not used to revise the predictor definitions or classifier settings. Evaluation followed the measures defined in Section 4.4.3.

4.6 SkySat-guided Sentinel-2 super-resolution

Section 4.5 used the prepared SkySat observations as direct parcel-level predictors. In a separate experiment, the same observations were used as finer-spatial-sampling targets to reconstruct four observed Sentinel-2 bands from the 10 m grid to an experimental 5 m grid, defining a controlled 2× super-resolution task.

4.6.1 Cross-sensor pairing and paired-patch construction

Pairing was restricted to observed, cloud-masked Sentinel-2 B02, B03, B04 and B08 and the corresponding SkySat Blue, Green, Red and NIR bands. For each AOI, a predefined observed Sentinel-2–SkySat pair was attempted first; where this did not produce an acceptable patch,

another observed Sentinel-2 acquisition within ±5 days of the SkySat observation was considered. Gap-filled Sentinel-2 imagery was excluded. Sentinel-2 was retained as the reference grid, and AOI-specific alignment corrections were applied only to the SkySat target during paired extraction. The x/y corrections were 0/0 m for AOI1, 0/+10 m for AOI2, +10/+10 m for AOI3 and 0/+10 m for AOI4. On the 5 m target grid, these ground translations corresponded to twice as many pixels.

The same 149 parcel identities used in Sections 4.4 and 4.5 were linked to their original parcel polygons, and a 1 m inward buffer was applied before patch extraction. Each low-resolution sample consisted of a four-band 16 × 16 array on the 10 m Sentinel-2 grid. The corresponding SkySat target was resampled to a four-band 32 × 32 array on a 5 m grid over the same 160 m × 160 m ground footprint. Common-valid masks identified pixels supported by both sensors, and accepted patches required at least 80% common-valid support. Pixels outside the common-valid mask were excluded from both the training loss and reconstruction metrics. A maximum of five patches was retained per parcel to limit unequal parcel representation.

A deterministic sample of 12 paired patches was rendered for visual checking of alignment, spectral correspondence and mask behaviour; this inspection did not alter the dataset or exclude parcels. The final paired dataset contained 515 accepted patches from 112 parcels, while 37 of the 149 starting parcels had no patch satisfying the temporal, alignment and common-valid-support requirements.

4.6.2 Parcel-controlled split and mask-aware EDSR-lite training

Because several patches could originate from the same parcel, all patches from a parcel were assigned to the same partition. AOI3 and AOI4 formed the development domain, while AOI1 and AOI2 were retained as the secondary, non-confirmatory evaluation domain and were not used for super-resolution training. The training, validation and test sets contained 222, 55 and 238 patches from 47, 12 and 53 parcels, respectively, with no parcel overlap between partitions. Per-band normalisation statistics were calculated only from valid training pixels and then applied unchanged to the validation and test sets.

Reconstruction used a lightweight convolutional neural network based on Enhanced Deep Super-Resolution, referred to as EDSR-lite. The four-band network used 32 feature channels, four residual blocks and PixelShuffle 2x upsampling, with 122,564 trainable parameters. Training used a masked Charbonnier loss with epsilon = 0.001, calculated only over common-valid target pixels. The model was optimised using AdamW with an initial learning rate of 0.001, a batch size of four and a maximum of 100 epochs. Model checkpoints were selected using validation performance, with early stopping after 20 epochs without improvement. The remaining optimisation settings are reported in Appendix Table A.4.

4.6.3 Reconstruction and downstream classification assessment

Reconstruction quality was assessed against bicubic interpolation as a non-learned comparator at the same 2x scale. Bicubic interpolation and the frozen CNN were applied to the held-out 238-patch AOI1+AOI2 test set. Mean absolute error, root mean squared error and signed bias were calculated over identical common-valid support in the normalised image

space defined by the training-only SkySat target statistics. Metrics were summarised for the complete test set, by spectral band and by AOI. An equal-parcel assessment was also calculated by first summarising errors within each parcel and then averaging across the 53 represented test parcels.

Figure 4.6 provides a matched visual comparison of the observed Sentinel-2 input, bicubic baseline, CNN reconstruction and SkySat reference used in the super-resolution assessment.

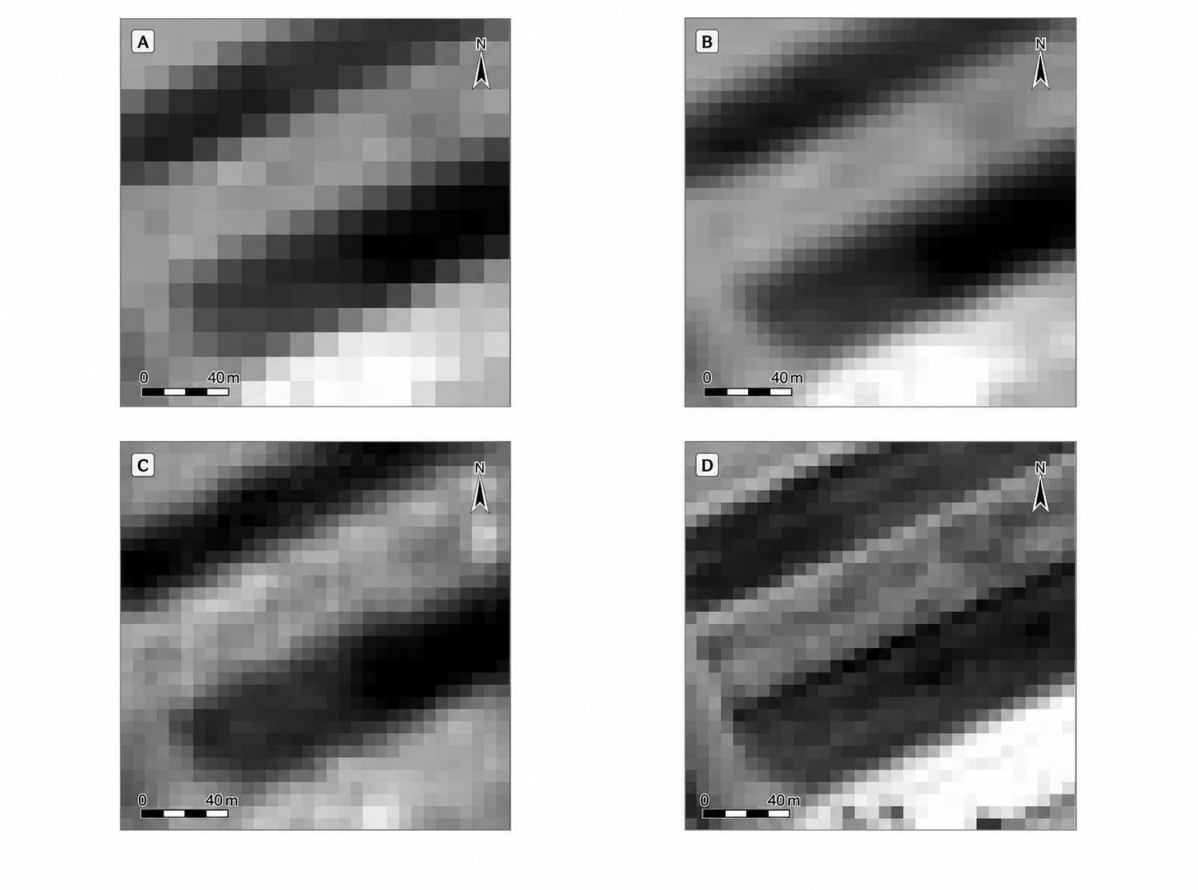


Figure 4.6. Matched comparison of the SkySat-guided Sentinel-2 super-resolution route. (A) Observed Sentinel-2 NIR at 10 m; (B) bicubic interpolation at 5 m; (C) CNN-based super-resolved NIR at 5 m; and (D) the matched SkySat reference NIR at 5 m.

A separate downstream diagnostic compared parcel-level discrimination between the Milpa mixed-cropping system and maize monocrop using native observed Sentinel-2 at 10 m, bicubic Sentinel-2 at 5 m and CNN super-resolution at 5 m. For Blue, Green, Red, NIR and NDVI, each patch was summarised using the mean, standard deviation, P10, P50 and P90. The mean and standard deviation of these 25 patch-level summaries were then calculated for each parcel, producing 50 predictors for each representation. Missing predictor values were replaced using medians estimated from the development data.

The same Random Forest specification was used for all three representations, with 300 trees, square-root feature sampling, a minimum terminal-leaf size of two, balanced class weights and a random state of 42. Model development used the 59 represented parcels from AOI3 and AOI4. Evaluation used 18 mixed-class parcels from AOI1 as the primary geographic-transfer diagnostic, while the 35 represented AOI2 parcels, all maize monocrop, were used only as a

maize false-positive control. The reconstruction network remained frozen during classification. The remaining Random Forest settings are reported in Appendix Table A.4.

4.7 PlanetScope preprocessing and predictor-level classification

Section 4.6 examined whether finer spatial information could be reconstructed from Sentinel-2 using SkySat guidance. The PlanetScope analysis provided a separate source of observed multispectral predictors while retaining the same 149-parcel population and spatial evaluation framework established in Sections 4.4 and 4.5.

4.7.1 Scene preparation and original-geometry parcel extraction

PlanetScope was used as a compact October observation set rather than as a dense daily time series. The October 2024 window was selected to maintain temporal consistency with the SkySat analyses while focusing on the main crop period. Candidate observations were screened according to acquisition timing, cloud cover, usable-data proportion and AOI coverage. Two suitable acquisitions were retained for each AOI, giving eight PSScene AnalyticMS Surface Reflectance scenes in total. Each eight-band scene was paired with its UDM2.1 quality product, and stored reflectance values were divided by 10,000. Pixels were retained only where all eight spectral bands contained finite, non-no-data values and the corresponding UDM2.1 information indicated clear, usable conditions.

The same 149 parcel identities were retained, but PlanetScope extraction used the original parcel polygons rather than the 10 m Sentinel-2 geometries. A 3 m inward buffer was applied directly to each original polygon, and raster cells were assigned when their centres fell within the buffered parcel. A parcel-scene record was retained when the buffered parcel had a scene-coverage ratio of at least 0.999 and contained at least one valid interior cell. The resulting parcel-scene observations were used to construct the PlanetScope predictors described in Section 4.7.2.

Figure 4.7 illustrates the joint UDM2.1 validity mask and direct 3 m parcel extraction used to establish scene-level PlanetScope support.

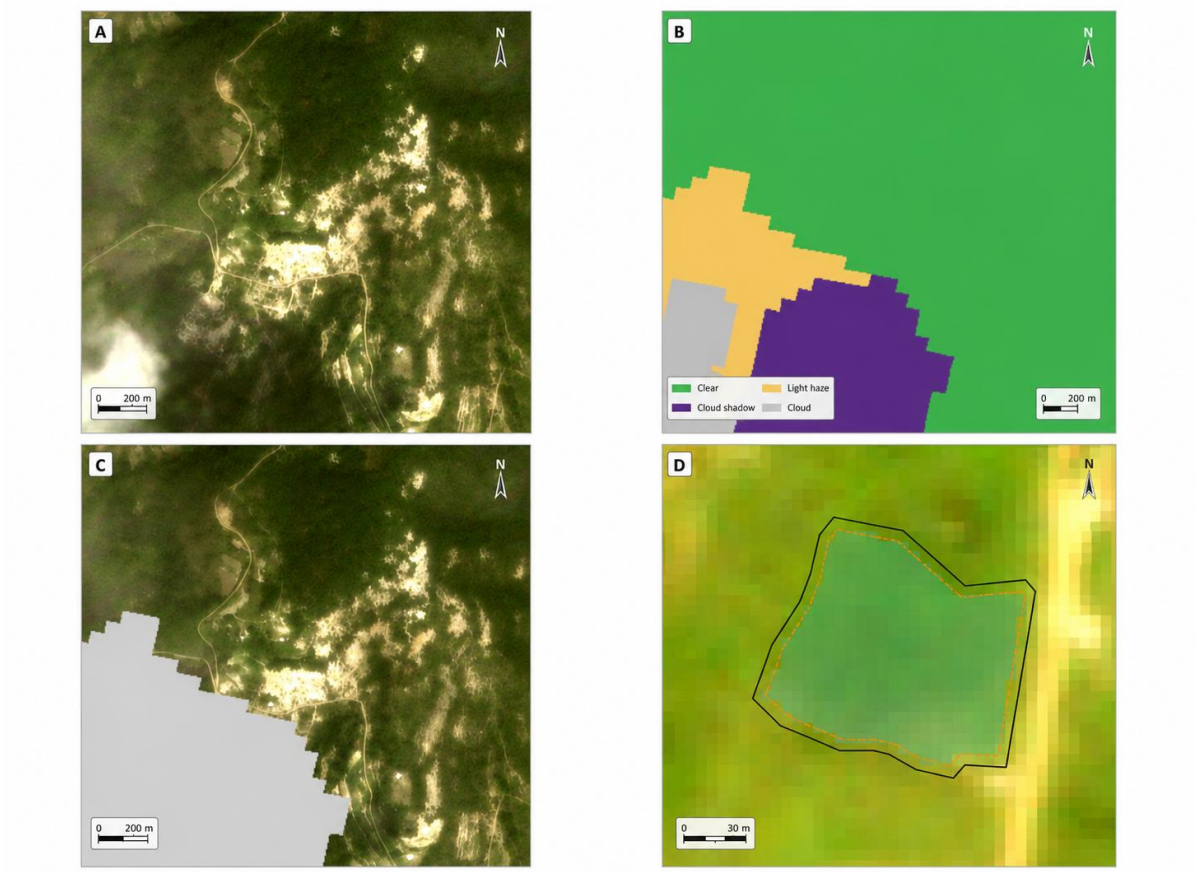


Figure 4.7. PlanetScope validity masking and direct parcel extraction. (A) Raw PlanetScope RGB image; (B) UDM2.1 validity classes; (C) the masked RGB product; and (D) the authoritative parcel boundary, direct 3 m inward buffer and retained interior raster cells used for extraction.

4.7.2 Spectral and spatial predictor construction

For each valid parcel-scene combination, the mean and median were calculated for the eight PlanetScope reflectance bands. Mean and median NDVI and NDRE were also calculated using denominator safeguards, giving 20 spectral variables per scene. Where two valid acquisitions were available for a parcel, the corresponding scene-level values were averaged equally. Parcels with only one valid acquisition used that observation directly; missing acquisitions were not copied or imputed.

Three pre-specified spatial predictors were calculated from valid pixels within the parcel interiors: NIR interquartile range, NDRE interquartile range and the median local standard deviation of NDRE from 3 × 3 neighbourhoods. A neighbourhood was retained only when all nine cells were valid and lay within the parcel. Scene-level spatial values were averaged across the valid acquisitions available for each parcel.

4.7.3 PlanetScope and Sentinel-2–PlanetScope classification design

The classification experiment addressed the binary distinction between the Milpa mixed-cropping system and maize monocrop. Five fixed predictor representations were compared: Sentinel-2 only with 25 predictors; PlanetScope spectral only with 20; Sentinel-2 plus

PlanetScope spectral with 45; PlanetScope spectral plus the three spatial predictors with 23; and Sentinel-2 plus PlanetScope spectral and spatial predictors with 48. All representations used the same 149 parcels, target definition, geographic evaluation design and 1,000-tree Random Forest specification described in Section 4.5.3.

Geographic transfer within the development domain was assessed through the directional AOI3-to-AOI4 and AOI4-to-AOI3 folds, producing out-of-fold predictions for all 71 development parcels. Each representation was then fitted using all 71 development parcels and evaluated once on the 78 parcels from AOI1 and AOI2. This remained a secondary, non-confirmatory benchmark because these AOIs had been examined earlier in the research workflow, and benchmark outcomes were not used to revise the predictor representations or classifier specification. Evaluation followed the measures defined in Section 4.4.3.

4.8 Multi-sensor classification-level integration

Sections 4.5 and 4.7 examined predictor-level integration, while Section 4.6 evaluated image-level enhancement through SkySat-guided Sentinel-2 super-resolution. The final integration experiment combined outputs from independently fitted native-sensor classifiers at the classification level, allowing the three integration routes to be evaluated separately.

4.8.1 Matched native predictions and cross-sensor corroboration

Classification-level integration used the Milpa probabilities produced by independently fitted Sentinel-2-only, SkySat-only and PlanetScope spectral-only models. The PlanetScope spectral-plus-spatial and Sentinel-2–PlanetScope combined-predictor models were not used in this stage. Geographic out-of-fold predictions for the 71 AOI3+AOI4 development parcels and secondary benchmark predictions for the 78 AOI1+AOI2 parcels were matched by parcel identifier, AOI and reference label, providing common prediction support for all 149 parcels while retaining the separation between the development and benchmark domains.

The SkySat and PlanetScope comparison branches each used their corresponding 1,000-tree Sentinel-2-only comparator with the same 25-predictor Sentinel-2 representation and geographic evaluation design described in Sections 4.5.3 and 4.7.3. These models were separate from the primary 800-tree Sentinel-2 baseline in Section 4.4.3. Before fusion, native-model predictions were compared using correlations between Milpa probabilities, binary prediction agreement, Cohen's kappa, shared and sensor-specific errors, and the Brier score. These diagnostics were calculated separately from the fusion procedure and were not used to modify the fusion weights or decision threshold.

4.8.2 Pairwise classification-level integration

Unlike the predictor-level analyses in Sections 4.5 and 4.7, the pairwise experiments combined probabilities only after the native sensor models had produced their predictions. For reporting, Sentinel-2–SkySat was treated as the principal pairwise comparison within the SkySat-centred thesis framework, while Sentinel-2–PlanetScope was presented as a corroborative extension; both used the same fixed integration rule.

For Sentinel-2 and SkySat, the fused Milpa probability was calculated as

$$p_{S2+SkySat} = 0.5p_{S2} + 0.5p_{SkySat}.$$

The Sentinel-2 and PlanetScope combination used

$$p_{S2+PlanetScope} = 0.5p_{S2} + 0.5p_{PlanetScope}.$$

All component probabilities represented the Milpa class. A fused probability greater than 0.50 was assigned to Milpa, whereas a probability equal to or below 0.50 was assigned to maize monocrop.

No classifier was fitted during probability integration, and the native models were not retrained. Sensor predictors were not concatenated, probabilities were not calibrated, and neither the weights nor the decision threshold were optimised. The procedure was fixed without using benchmark outcomes. Assessment was conducted separately for the AOI3-to-AOI4 and AOI4-to-AOI3 development directions, the pooled 71-parcel geographic out-of-fold development set, the individual AOI1 and AOI2 benchmark domains, and the pooled 78-parcel secondary benchmark.

4.8.3 Three-sensor diagnostic fusion and uncertainty assessment

A three-sensor diagnostic combined the Milpa probabilities from the independently fitted Sentinel-2, SkySat and PlanetScope spectral-only models using a fixed equal-weight rule. The fused Milpa probability was calculated as

$$p_{three} = \frac{p_{S2} + p_{SkySat} + p_{PlanetScope}}{3}.$$

Each sensor received a one-third weight, and the same decision rule assigned probabilities greater than 0.50 to Milpa and those equal to or below 0.50 to maize monocrop. The native models were not retrained, probabilities were not calibrated, and neither the fusion weights nor the decision threshold were optimised. The procedure was fixed without using benchmark outcomes.

Uncertainty in fused-minus-component performance differences was assessed using a paired parcel-level bootstrap with 10,000 replicates and a fixed seed of 20260714. Parcels were sampled with replacement within the defined strata, with the same sampled parcel indices applied to all models in each comparison. Directional AOI analyses were stratified by reference class, while the pooled development and pooled benchmark analyses were stratified jointly by AOI and reference class. Two-sided 95% percentile intervals were obtained from the 2.5th and 97.5th percentiles of the bootstrap distributions for differences in accuracy, balanced accuracy, macro-F1, Milpa precision, Milpa recall and Milpa F1. Post hoc diagnostics examined the stored probabilities, class support, shared errors and feature shifts without modifying the fitted models or integration procedure.

Figure 4.8 summarises how the common parcel cohort, Sentinel-2 baseline, complementary sensor routes and classification-level integration were connected while remaining analytically distinct.

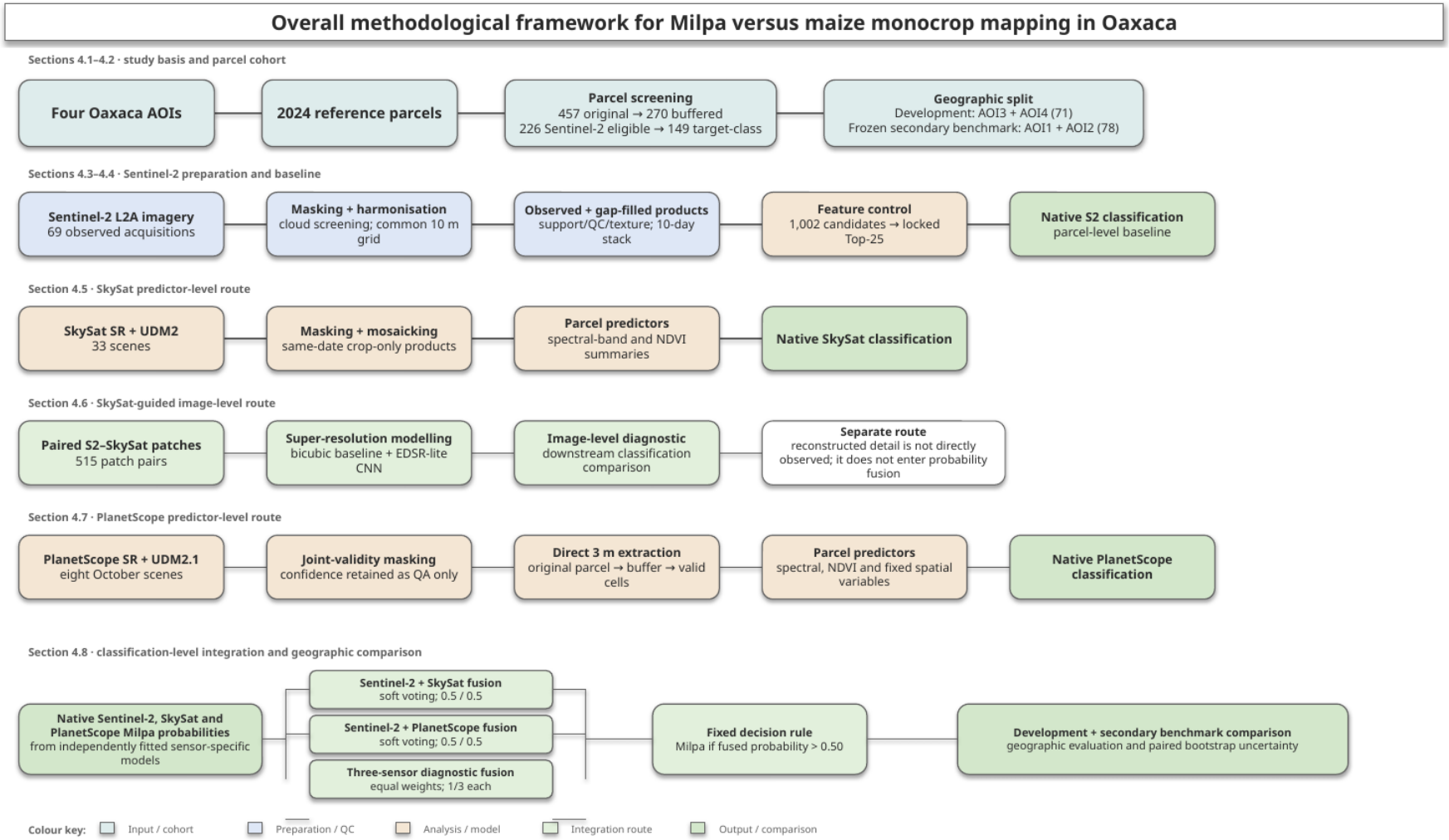


Figure 4.8. Overall methodological framework across Sections 4.1–4.8. The schematic links the common parcel cohort to the Sentinel-2 baseline, the SkySat and PlanetScope predictor-level routes, the separate SkySat-guided image-level route, and the fixed classification-level integration and geographic evaluation.

Chapter 5: Results

5.1 Reference-cohort composition and Sentinel-2 data quality

This section characterises the final analytical cohort and examines the distribution and continuity of its observed Sentinel-2 support. It then evaluates temporal interpolation and reports the quality-control decision applied before the classification analyses.

5.1.1 Analytical-cohort composition

The final analytical cohort contained 149 parcels distributed across four areas of interest (AOIs) and two evaluation domains. Of the 457 original reference parcels, 270 retained valid buffered geometry, 226 were Sentinel-2 eligible and 149 belonged to the three target crop classes. The complete parcel-retention sequence is provided in Appendix Table A5.1. AOI3 and AOI4 formed the 71-parcel development domain used for feature control, model development and development assessment. AOI1 and AOI2 formed the separate 78-parcel secondary, non-confirmatory benchmark.

The cohort was unevenly distributed geographically and by crop class. AOI2 contributed the largest individual group, with 58 parcels, whereas AOI1 contributed 20. The three-class cohort comprised 43 Milpa mixed-cropping system parcels, 73 landrace maize parcels and 33 hybrid maize parcels. Landrace and hybrid maize were combined in the binary analysis, producing 43 Milpa and 106 maize monocrop parcels. Overall, the final cohort represented 32.60% of the original 457 reference parcels. Table 5.1 summarises the domain-level crop composition and observed-support context. The complete AOI-by-class distribution is provided in Appendix Table A5.2.

The subsequent classification analyses therefore used a geographically and class-wise uneven reference cohort.

5.1.2 Distribution of observed Sentinel-2 support

Sentinel-2 eligibility did not mean that all retained parcels had the same amount or continuity of observed data. The study-specific strong, moderate and weak support classes jointly represented the proportion of the 69 acquisition dates with usable parcel observations and the longest consecutive sequence of unsupported acquisitions.

Most parcels in the final cohort had strong or moderate full-year support, with 93 classified as strong and 39 as moderate. However, all 17 weak-support parcels occurred in the benchmark domain, and 16 of these were in AOI2. Neither development AOI contained a weak-support parcel.

Table 5.1. Crop composition and full-year observed Sentinel-2 support in the development and benchmark domains.

Evaluation domain	Parcels	Milpa	Landrace maize	Hybrid maize	Strong / Moderate / Weak	Mean supported acquisition dates (%)	Maximum consecutive unsupported acquisitions
Development	71	29	32	10	58 / 13 / 0	72.05	7
Benchmark	78	14	41	23	35 / 26 / 17	63.47	21

The domain-level difference was also evident in the continuity of the observations. Development parcels had a higher mean proportion of supported acquisition dates and a maximum consecutive unsupported run of seven acquisitions, compared with 21 in the benchmark (Table 5.1). Figure 5.1a shows how the three support classes were distributed across the individual AOIs, while the complete AOI-level statistics are reported in Appendix Table A5.3.

Observed support was therefore uneven across the analytical cohort, with the weakest full-year observed support concentrated in the benchmark domain, particularly AOI2.

5.1.3 Interpolation reliability and seasonal limitations

Across the 226 Sentinel-2-eligible parcels, interpolated Normalised Difference Vegetation Index (NDVI) values closely matched the withheld observations. The 9,989 validation records produced a mean absolute error (MAE) of 0.02 and a root mean squared error (RMSE) of 0.03. The signed bias was negative but less than 0.01 in magnitude, while Pearson r and R^2 were 0.97 and 0.95, respectively.

The clearest limitation occurred when longer gaps coincided with the August–October crop period. During these months, MAE increased from 0.04 for gaps of no more than 20 days to 0.10 for gaps exceeding 60 days. Errors outside August–October were generally lower. The small outside-season long- and very-long-gap groups contained 35 and 18 validation records, respectively, and did not follow a monotonic sequence.

Figure 5.1b presents the complete season-by-gap pattern together with the number of validation records in each group. The results were available only as combined season-by-gap strata; separate pooled estimates for season alone or gap duration alone are therefore not reported. The full numerical matrix and AOI-level interpolation results are provided in Appendix Tables A5.4 and A5.5.

Full-year support also concealed limited observations during the principal crop period. Mean observed support during August–October was 36.68% across the final cohort. Four parcels had no observed support during this period; all four occurred in the AOI2 benchmark, where the longest seasonal gap reached 91 days. The complete seasonal-support classification is retained in Appendix Table A5.6.

Temporal interpolation was reliable in aggregate, but its performance was weaker where long gaps coincided with the principal August–October crop period. The subsequent classification results should therefore be read against this spatially and seasonally uneven support context.

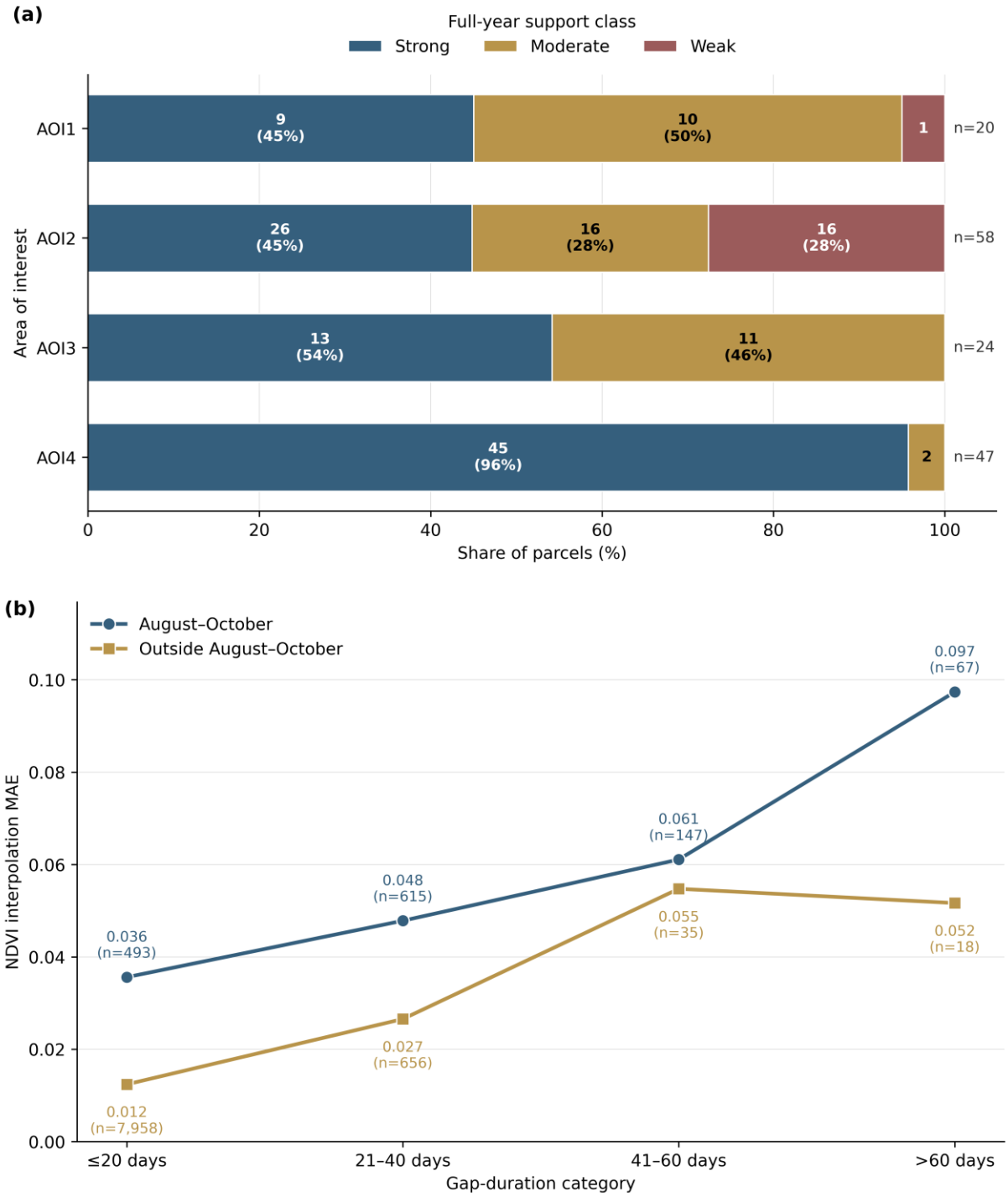


Figure 5.1. Observed Sentinel-2 support and interpolation reliability across the Oaxaca reference parcels. Panel (a) shows the distribution of strong, moderate and weak full-year support classes across the four AOIs in the final 149-parcel cohort. Panel (b) shows withheld-observation NDVI interpolation MAE by season and gap-duration category across the 226 Sentinel-2-eligible parcels; labels indicate the number of validation records in each group.

5.1.4 Quality-control outcome and sensitivity cohort

Soft quality-control warnings remained within the Sentinel-2 records and affected 110 of the final 149 parcels. These warnings were retained as cautionary indicators, but no confirmed hard artefacts were identified and no parcel was removed on this basis.

The primary 149-parcel Sentinel-2 dataset was therefore retained for the downstream classification analyses. A separate, better-supported sensitivity cohort contained 132 parcels. It retained all 71 development parcels but reduced the benchmark from 78 to 61 parcels. This subset did not replace the primary cohort; it was used only to test whether the classification findings changed under stricter observed-support conditions. Detailed warning counts and the composition of the sensitivity cohort are reported in Appendix Tables A5.7 and A5.8.

The classification results for the primary and better-supported cohorts are compared in Section 5.2.2.

5.2 Sentinel-2 baseline classification

The Sentinel-2 baseline was evaluated using the final 149-parcel cohort and the predefined Top-25 predictor panel. A parcel-level Random Forest provided the primary assessment for the three-class formulation comprising the Milpa mixed-cropping system, landrace maize and hybrid maize, and the binary distinction between Milpa and maize monocrop, with development and secondary benchmark results reported separately.

Across the 71 development parcels in AOI3 and AOI4, median gap-filled NDVI trajectories followed broadly similar seasonal patterns across the three crop classes. The largest recorded class medians occurred on 9 September 2024, reaching 0.59 for hybrid maize, 0.51 for landrace maize and 0.48 for Milpa. Although differences in magnitude were visible around this recorded maximum, the trajectories overlapped across much of the observation period (Figure 5.2). These profiles provide descriptive development-domain evidence rather than an independent assessment of secondary benchmark performance. Complementary AOI-specific observed-versus-gap-filled NDVI profiles and date-specific parcel-support counts for the final 149-parcel cohort are provided as descriptive evidence in Appendix 5.2-H.

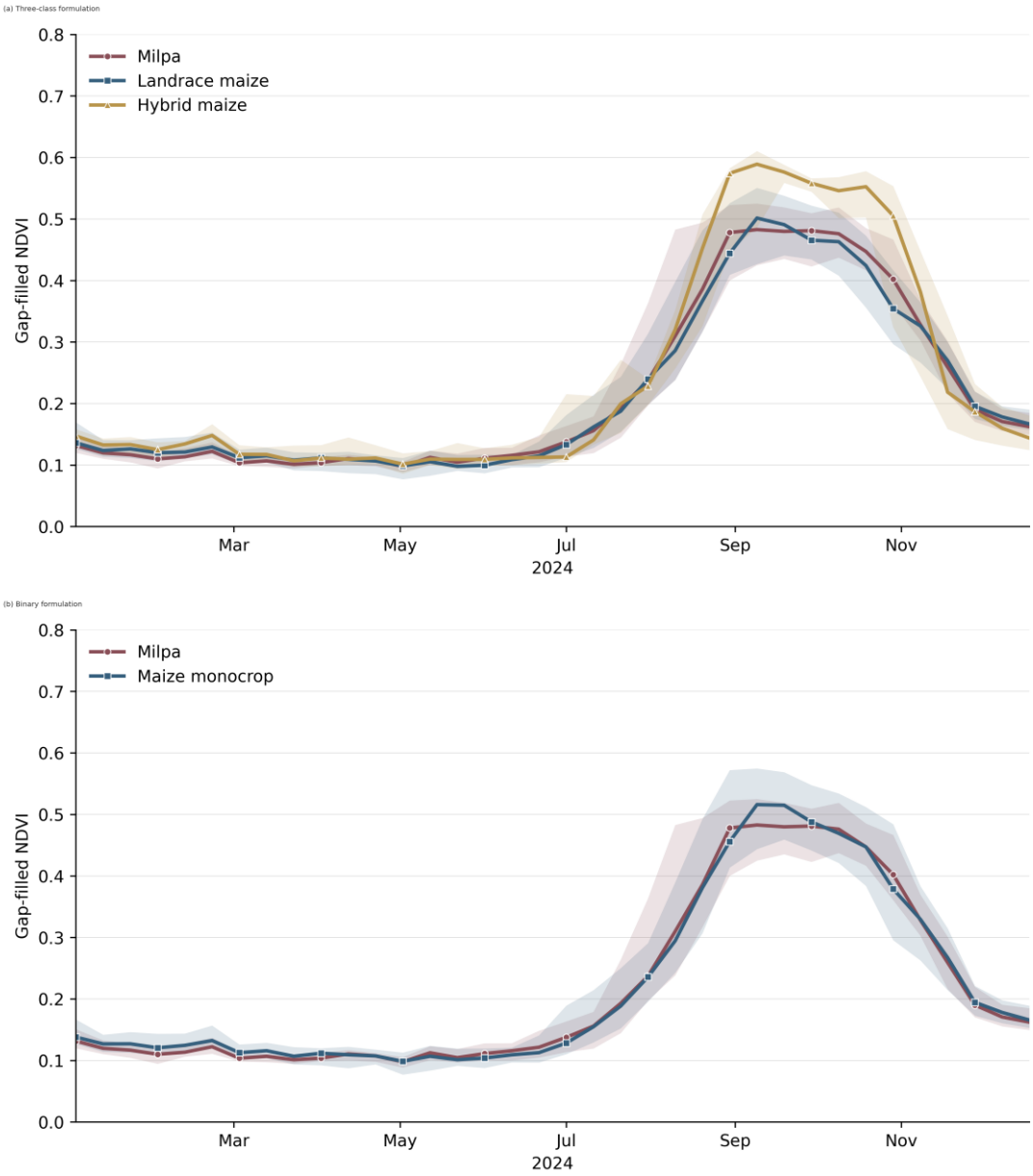


Figure 5.2. Development-domain gap-filled Sentinel-2 NDVI temporal profiles for the three-class and binary crop-system formulations. Lines represent class medians derived from the gap-filled series across the 71 development parcels in AOI3 and AOI4, and shaded envelopes represent the interquartile range. The profiles provide descriptive development-domain evidence and do not represent secondary benchmark performance.

5.2.1 Parcel-level classification

The binary formulation produced stronger development results than the three-class formulation: mean balanced accuracy and macro-F1 increased from 0.62 to 0.78. Development results in Table 5.2 are means across 60 folds, with standard deviations shown in parentheses; benchmark results are single estimates for 78 parcels. Development-fold summaries did not report class-specific Milpa recall or Milpa F1.

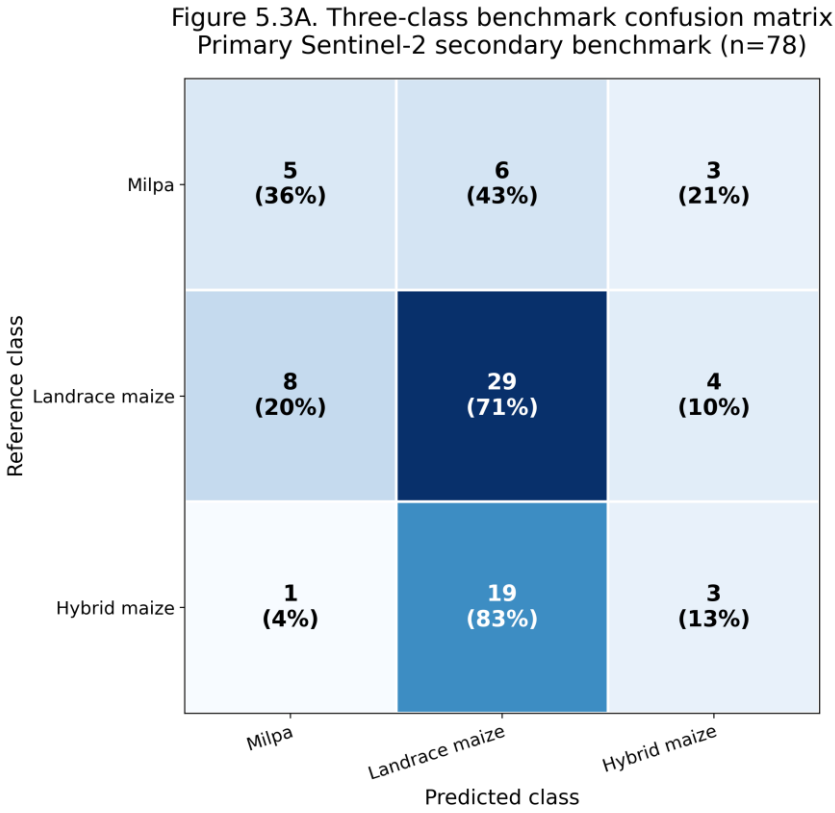
Table 5.2. Development and benchmark performance of the primary Sentinel-2 parcel-level Random Forest.

Formulation	Analytical role	Assessment basis	Accuracy	Balanced accuracy	Macro-F1	Milpa recall	Milpa F1
Three-class	Development assessment	Mean (SD), 60 folds	0.68 (0.08)	0.62 (0.10)	0.62 (0.11)	Not reported	Not reported
Three-class	Secondary benchmark	Single estimate, 78 parcels	0.47	0.40	0.38	0.36	0.36
Binary	Development assessment	Mean (SD), 60 folds	0.79 (0.07)	0.78 (0.08)	0.78 (0.08)	Not reported	Not reported
Binary	Secondary benchmark	Single estimate, 78 parcels	0.78	0.62	0.62	0.36	0.37

Three-class benchmark performance was lower than the corresponding development performance, with an accuracy of 0.47 and macro-F1 of 0.38. Landrace maize was the best-recognised class, with a recall of 0.71, whereas hybrid maize recall was 0.13; 19 of the 23 hybrid maize parcels were predicted as landrace maize. Milpa recall was 0.36, corresponding to five correct classifications among 14 Milpa parcels.

Binary benchmark accuracy was higher at 0.78, but Milpa recall remained 0.36: nine of the 14 Milpa parcels were classified as maize monocrop, while maize monocrop recall reached 0.88. Binary performance also varied between the benchmark AOIs, with balanced accuracy of 0.72 in AOI2 and 0.61 in AOI1, and Milpa recall of 0.60 and 0.22, respectively. AOI1 contained no hybrid maize reference parcels and therefore did not provide a complete AOI-specific three-class assessment. The binary formulation thus provided stronger overall discrimination, although benchmark recognition of Milpa remained limited.

The corresponding three-class and binary benchmark confusion matrices are presented in Figure 5.3.



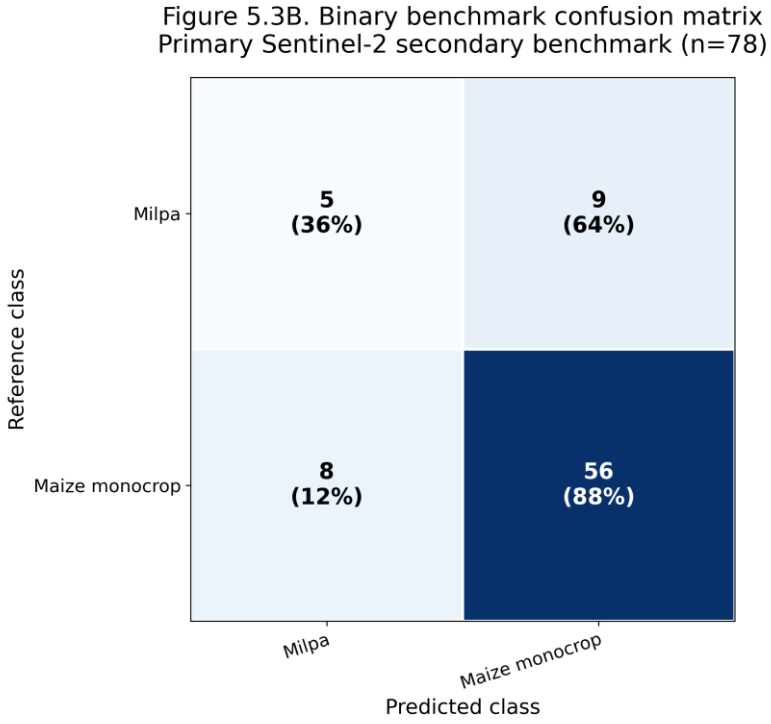


Figure 5.3. Primary Sentinel-2 parcel-level Random Forest confusion matrices for the secondary benchmark. Panel (a) presents the three-class formulation and panel (b) the binary formulation. Rows represent observed classes, columns represent predicted classes and cell values show parcel counts.

Complete class-specific metrics, confusion matrices, AOI-specific results and development-fold ranges are provided in Appendices 5.2-A–5.2-D.

5.2.2 Sensitivity and model-family diagnostics

Restricting the benchmark to better-supported parcels reduced it from 78 to 61 parcels without improving either formulation. Three-class macro-F1 decreased from 0.38 to 0.34 and Milpa recall from 0.36 to 0.27; in the binary formulation, macro-F1 decreased from 0.62 to 0.53 and Milpa recall from 0.36 to 0.18. The Support Vector Machine (SVM) also remained below the primary Random Forest: three-class macro-F1 was 0.36 with Milpa recall of 0.29, while binary macro-F1 was 0.45 and Milpa recall only 0.07.

Table 5.3. Better-supported sensitivity and SVM benchmark diagnostics relative to the primary Random Forest.

Formulation	Model or cohort	Benchmark parcels	Accuracy	Balanced accuracy	Macro-F1	Milpa recall	Milpa F1
Three-class	Primary Random Forest	78	0.47	0.40	0.38	0.36	0.36
Three-class	Better-supported Random Forest	61	0.44	0.37	0.34	0.27	0.29
Three-class	SVM	78	0.40	0.37	0.36	0.29	0.27
Binary	Primary Random Forest	78	0.78	0.62	0.62	0.36	0.37
Binary	Better-supported Random Forest	61	0.75	0.53	0.53	0.18	0.21
Binary	SVM	78	0.71	0.46	0.45	0.07	0.08

Neither the better-supported restriction nor the alternative model-family surpassed the primary parcel-level Random Forest on the principal benchmark measures.

Extended class-specific results and confusion matrices for the better-supported sensitivity and Support Vector Machine diagnostics are provided in Appendices 5.2-E and 5.2-F.

5.2.3 Pixel-level diagnostic

The pixel-level benchmark comprised 1,902 pixels representing all 78 benchmark parcels. In the three-class formulation, pixel-level macro-F1 was 0.29 and increased to 0.32 after mean-probability parcel aggregation, remaining below the primary parcel-level Random Forest value of 0.38; Milpa and hybrid maize recognition also remained weak. Binary pixel accuracy reached 0.87, but balanced accuracy was 0.50 and Milpa recall was 0.08. Aggregation increased macro-F1 to 0.54 and Milpa recall to 0.14, still below the primary parcel model values of 0.62 and 0.36.

Table 5.4. Pixel-level, pixel-aggregated and primary parcel-level Sentinel-2 benchmark performance.

Target formulation	Evaluation level	Evaluation records	Accuracy	Balanced accuracy	Macro-F1	Milpa recall	Milpa F1
Three-class	Benchmark pixels	1,902 pixels	0.50	0.36	0.29	0.10	0.08
Three-class	Mean-probability parcel aggregation	78 parcels	0.53	0.37	0.32	0.14	0.20
Three-class	Primary parcel-level Random Forest	78 parcels	0.47	0.40	0.38	0.36	0.36
Binary	Benchmark pixels	1,902 pixels	0.87	0.50	0.50	0.08	0.07
Binary	Mean-probability parcel aggregation	78 parcels	0.79	0.54	0.54	0.14	0.20
Binary	Primary parcel-level Random Forest	78 parcels	0.78	0.62	0.62	0.36	0.37

Parcel aggregation improved several pixel diagnostic measures but did not surpass the primary parcel model on balanced or Milpa-sensitive measures.

Extended pixel-level and parcel-aggregated class metrics and confusion matrices are provided in Appendix 5.2-G.

Across the Sentinel-2 analyses reported here, the parcel-level Random Forest provided the strongest balance between overall and class-sensitive benchmark performance. Binary separation was stronger overall than three-class classification, although recognition of the Milpa mixed-cropping system remained limited. Neither stricter support filtering, the SVM nor the pixel-level diagnostic surpassed the primary parcel model on the principal benchmark measures. The Sentinel-2 Top-25 parcel representation therefore serves as the reference baseline for the predictor-level comparisons in Section 5.3.

5.3 Predictor-level multi-sensor integration

This section evaluates whether direct SkySat and PlanetScope predictor representations, or their combination with Sentinel-2 predictors, improved the binary distinction between the Milpa mixed-cropping system and maize monocrop. All comparisons used the common final cohort of 149 parcels, comprising 71 development parcels and the same 78-parcel secondary benchmark used in Section 5.2, with Sentinel-2 retained as the reference representation. To provide a common comparison across sensors, development results in this section were derived from pooled geographic AOI3-to-AOI4 and AOI4-to-AOI3 out-of-fold predictions and therefore differ from the repeated-fold development summaries reported for the primary Sentinel-2 model in Section 5.2. Because AOI3 contained only three Milpa parcels, the AOI4-to-AOI3 direction represented a small-minority-class stress test.

5.3.1 Sentinel-2 and SkySat

All 149 parcels retained valid SkySat predictors, with no parcel imputed or excluded. The comparison included the 25-predictor Sentinel-2 representation, the 10-predictor SkySat representation and the combined 35-predictor Sentinel-2 plus SkySat representation.

In the development domain, Sentinel-2 and the combined representation produced identical pooled classifications. Both achieved an accuracy of 0.55, balanced accuracy of 0.46 and macro-F1 of 0.35, with Milpa recall and Milpa F1 of 0.00. The SkySat representation

achieved lower accuracy, balanced accuracy and macro-F1 of 0.41, 0.35 and 0.31, respectively. It correctly identified one of the 29 development Milpa parcels, producing a Milpa recall of 0.03 and Milpa F1 of 0.05. The full development and benchmark results are reported in Table 5.5.

Table 5.5. Pooled development and secondary benchmark performance of the Sentinel-2 and SkySat predictor representations.

Representation	Predictors	Domain	Accuracy	Balanced accuracy	Macro-F1	Milpa recall	Milpa F1
Sentinel-2	25	Development	0.55	0.46	0.35	0.00	0.00
Sentinel-2	25	Benchmark	0.78	0.62	0.62	0.36	0.37
SkySat	10	Development	0.41	0.35	0.31	0.03	0.05
SkySat	10	Benchmark	0.77	0.66	0.65	0.50	0.44
Sentinel-2 + SkySat	35	Development	0.55	0.46	0.35	0.00	0.00
Sentinel-2 + SkySat	35	Benchmark	0.78	0.59	0.60	0.29	0.32

The SkySat representation produced the strongest class-balanced and Milpa-sensitive results on the secondary benchmark. Relative to Sentinel-2, balanced accuracy increased from 0.62 to 0.66, macro-F1 from 0.62 to 0.65, Milpa recall from 0.36 to 0.50 and Milpa F1 from 0.37 to 0.44, while overall accuracy decreased slightly from 0.78 to 0.77. SkySat correctly identified seven Milpa parcels compared with five for Sentinel-2, but the number of maize parcels incorrectly assigned to Milpa increased from eight to eleven. Greater Milpa recall therefore coincided with more maize false positives. The combined 35-predictor representation did not improve the pooled benchmark, as its balanced accuracy, macro-F1, Milpa recall and Milpa F1 were all below the Sentinel-2 reference.

The AOI-specific results were not consistent. In AOI1, the SkySat representation increased Milpa recall from 0.22 to 0.67, but balanced accuracy decreased from 0.61 to 0.47. In AOI2, balanced accuracy decreased from 0.72 to 0.57 and Milpa recall from 0.60 to 0.20. The combined representation produced a small increase in AOI2 macro-F1 but no pooled benchmark advantage. SkySat therefore provided favourable pooled benchmark Milpa-sensitive results, but this advantage was not reproduced in development, and combining its predictors with Sentinel-2 did not improve pooled performance.

5.3.2 Sentinel-2 and PlanetScope

All 149 parcels remained in the final PlanetScope predictor tables. Of these, 144 had two valid scenes. Five development parcels in AOI3, all belonging to the maize monocrop class, had one valid scene and each used its single observed acquisition without imputation, copying or zero filling. The comparison included Sentinel-2 with 25 predictors, PlanetScope spectral with 20, Sentinel-2 plus PlanetScope spectral with 45, PlanetScope spectral plus spatial with 23, and Sentinel-2 plus PlanetScope spectral plus spatial with 48 predictors.

Several PlanetScope representations identified one or two Milpa development parcels, whereas the Sentinel-2 representation identified none. The PlanetScope spectral representation produced a Milpa recall of 0.07 and Milpa F1 of 0.08, but its balanced accuracy and macro-F1 were lower than the Sentinel-2 values. Combining PlanetScope spectral predictors with Sentinel-2 increased macro-F1 from 0.35 to 0.37 and Milpa F1 from

0.00 to 0.06, while accuracy and balanced accuracy decreased. Adding the three spatial predictors did not improve development performance, and the 45-predictor and 48-predictor combined representations produced identical pooled classifications. Table 5.6 presents the complete development and benchmark comparison.

Table 5.6. Pooled development and secondary benchmark performance of the Sentinel-2 and PlanetScope predictor representations.

Representation	Predictors	Domain	Accuracy	Balanced accuracy	Macro-F1	Milpa recall	Milpa F1
Sentinel-2	25	Development	0.55	0.46	0.35	0.00	0.00
Sentinel-2	25	Benchmark	0.78	0.62	0.62	0.36	0.37
PlanetScope spectral	20	Development	0.38	0.33	0.31	0.07	0.08
PlanetScope spectral	20	Benchmark	0.73	0.64	0.61	0.50	0.40
Sentinel-2 + PlanetScope spectral	45	Development	0.52	0.45	0.37	0.03	0.06
Sentinel-2 + PlanetScope spectral	45	Benchmark	0.81	0.60	0.62	0.29	0.35
PlanetScope spectral + spatial	23	Development	0.38	0.33	0.29	0.03	0.04
PlanetScope spectral + spatial	23	Benchmark	0.69	0.59	0.57	0.43	0.33
Sentinel-2 + PlanetScope spectral + spatial	48	Development	0.52	0.45	0.37	0.03	0.06
Sentinel-2 + PlanetScope spectral + spatial	48	Benchmark	0.79	0.60	0.61	0.29	0.33

The PlanetScope spectral benchmark result was mixed. Relative to Sentinel-2, balanced accuracy increased from 0.62 to 0.64, Milpa recall from 0.36 to 0.50 and Milpa F1 from 0.37 to 0.40. However, accuracy decreased from 0.78 to 0.73 and macro-F1 from 0.62 to 0.61. PlanetScope spectral correctly identified seven Milpa parcels rather than five, while maize false positives increased from eight to fourteen. Greater Milpa recognition therefore coincided with more maize misclassification.

The Sentinel-2 plus PlanetScope spectral representation increased benchmark accuracy to 0.81 but did not improve balanced accuracy, macro-F1, Milpa recall or Milpa F1. The full 48-predictor representation showed the same principal non-improving pattern. Adding the three spatial predictors did not improve any benchmark measure. For the PlanetScope representation, all reported measures declined; for the Sentinel-2 combined representation, balanced accuracy and Milpa recall were unchanged while the other reported measures declined.

AOI-specific outcomes were also heterogeneous. PlanetScope spectral produced stronger Milpa-sensitive results in AOI2, while its class-balanced result in AOI1 remained below Sentinel-2. The Sentinel-2 plus PlanetScope spectral representation also performed more favourably in AOI2 than in AOI1, but these differences did not produce a consistent pooled advantage from the combined representation. PlanetScope spectral therefore provided mixed pooled benchmark evidence, predictor concatenation did not improve the principal pooled benchmark measures, and the three spatial predictors were non-improving.

Across the two sensor comparisons, the direct SkySat and PlanetScope representations produced some favourable Milpa-sensitive benchmark results, but these gains were not consistently reproduced in the geographic development evidence. Combining either sensor's predictors with Sentinel-2 did not improve the principal pooled benchmark measures, and the PlanetScope spatial additions were also non-improving. Complete class-specific, confusion-count and AOI-specific diagnostics for both predictor-level sensor branches are provided in Appendix 5.3. Section 5.4 next evaluates a different integration stage at image level through SkySat-guided Sentinel-2 super-resolution.

5.4 Image-level integration through SkySat-guided super-resolution

The final super-resolution branch, restricted to the four observed Sentinel-2 bands shared with SkySat (Blue, Green, Red and NIR), began with all 149 target parcels, of which 112 produced accepted paired patches and 37 produced no accepted patch. The represented parcels contributed 515 paired patches and were divided into 47 training, 12 validation and 53 untouched test parcels. Parcel overlap among the three splits was zero. This section first assesses image reconstruction relative to bicubic interpolation and then examines whether the reconstructed representations improved downstream Milpa and maize monocrop classification.

5.4.1 Reconstruction performance

The primary patch-weighted assessment comprised 238 untouched test patches and 225,615 valid target pixels evaluated on common-valid support in normalised image space. CNN-SR reduced mean absolute error (MAE) from 0.79 under bicubic interpolation to 0.66, corresponding to a reduction of 17.21%. Root mean squared error (RMSE) decreased from 1.09 to 0.97, a reduction of 11.13%, while absolute bias decreased by 63.93%. Pooled MAE, RMSE and absolute bias were therefore all lower for CNN-SR than for bicubic interpolation. The reconstruction comparison is presented in Figure 5.4.

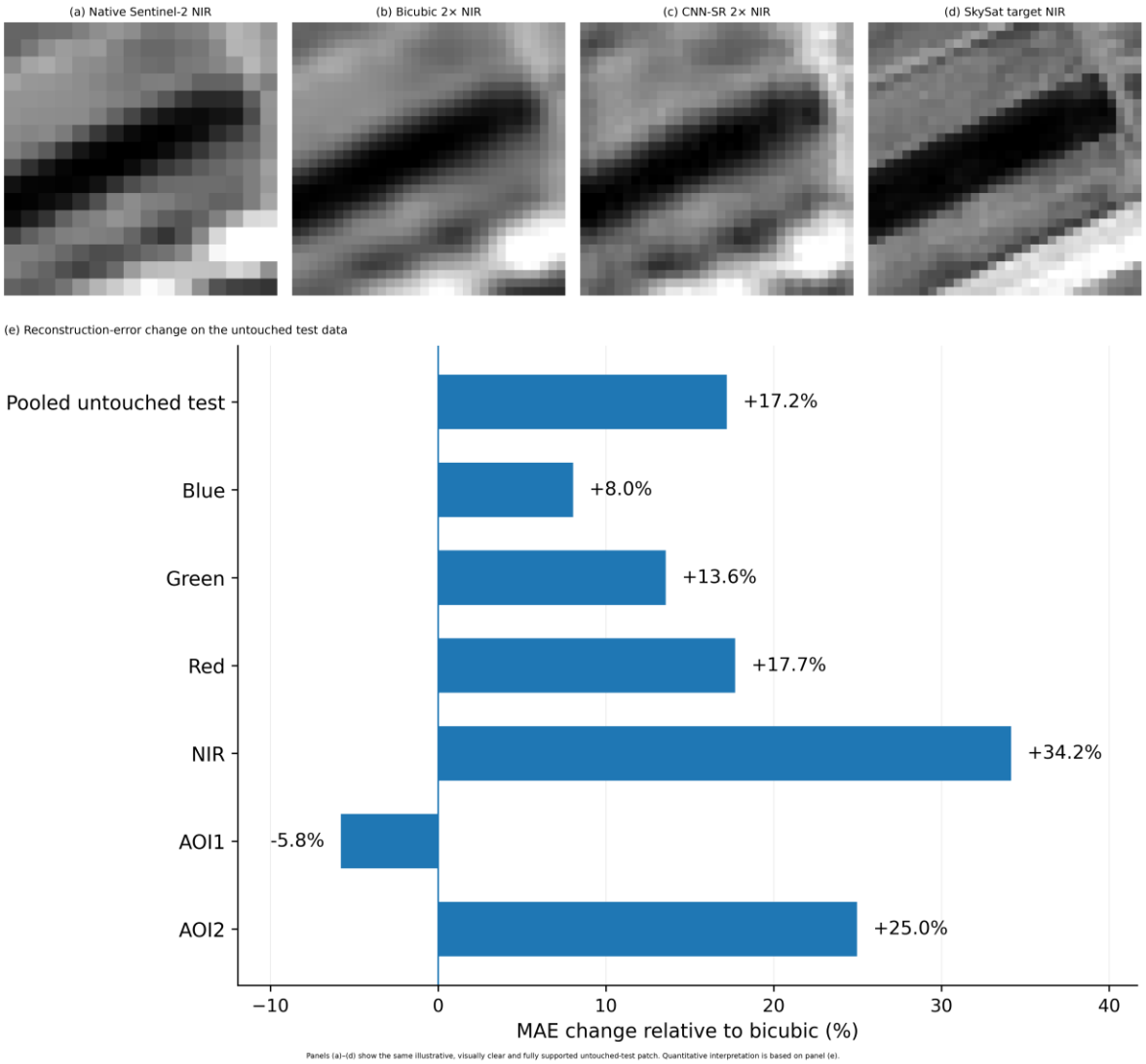


Figure 5.4. Reconstruction performance of the frozen CNN super-resolution model relative to bicubic interpolation on the untouched test data. Panel (a) compares representative native observed Sentinel-2, bicubic, CNN-SR and corresponding SkySat-target patches from the 1 m inward-buffer branch. Panel (b) summarises the percentage change in MAE for the pooled, band-specific and AOI-specific comparisons; positive values indicate lower CNN-SR error and negative values indicate higher CNN-SR error.

5.4.2 Band-, AOI- and parcel-balanced assessment

All four bands showed lower pooled MAE and RMSE under CNN-SR, although the magnitude of improvement varied. NIR produced the largest MAE reduction at 34.16%, followed by Red at 17.71% and Green at 13.57%. Blue produced the smallest reduction at 8.05%.

The AOI-specific results were not uniform. AOI1 contributed 63 test patches, for which CNN-SR increased MAE by 5.80% and RMSE by 7.73%. Within AOI1, NIR improved, whereas Blue, Green and Red did not. AOI2 contributed 175 test patches and showed reductions of 24.97% in MAE and 22.27% in RMSE, with all four bands improving on both

measures. The pooled reconstruction result therefore combined contrasting outcomes between the two test AOIs.

The overall improvement persisted when the 53 test parcels contributed equally. Equal-parcel MAE decreased from 0.80 to 0.69, a reduction of 13.90%, while RMSE decreased from 1.05 to 0.94, a reduction of 10.38%. However, AOI1 equal-parcel MAE increased by 5.71%, whereas AOI2 equal-parcel MAE decreased by 24.94%. Maize monocrop MAE decreased by 17.26%, while the represented Milpa subset showed an increase of 3.71%. CNN-SR therefore retained an overall reconstruction advantage under equal-parcel weighting, but AOI1 and the represented Milpa subset did not improve. An independent crop-class effect cannot be established because all eight test Milpa parcels occurred in AOI1 and the represented AOI2 test cohort was maize-only; consequently, the strong AOI2 reconstruction gain provided no evidence of improved Milpa reconstruction.

5.4.3 Downstream classification diagnostic

The downstream diagnostic used 59 represented development parcels from AOI3 and AOI4, 18 represented AOI1 parcels for the primary mixed-class assessment and 35 represented AOI2 parcels for the maize-only false-positive control. Native observed Sentinel-2 at 10 m, bicubic Sentinel-2 at 5 m and CNN-SR at 5 m were evaluated using the same 50-predictor definition and fixed classification design.

This branch-specific comparison was restricted to the represented super-resolution parcels and was not a rerun of the full-cohort Top-25 Sentinel-2 baseline reported in Section 5.2.

Bicubic interpolation produced the strongest AOI1 classification results across accuracy, balanced accuracy, macro-F1, Milpa recall and Milpa F1 (Table 5.7). It correctly identified four of the eight Milpa parcels, compared with two for native Sentinel-2 and one for CNN-SR. Native Sentinel-2 and bicubic interpolation each assigned one maize parcel to Milpa, whereas CNN-SR assigned no maize parcel to Milpa. CNN-SR therefore eliminated AOI1 maize false positives but detected only one of eight Milpa parcels. The pooled image-fidelity improvement did not transfer to improved AOI1 thematic discrimination.

Table 5.7. AOI1 parcel-level classification performance of the native Sentinel-2, bicubic and CNN super-resolution representations.

Representation	Accuracy	Balanced accuracy	Macro-F1	Milpa recall	Milpa F1	Maize false-positive rate
Native Sentinel-2 10 m	0.61	0.58	0.54	0.25	0.36	0.10
Bicubic Sentinel-2 5 m	0.72	0.70	0.70	0.50	0.62	0.10
CNN-SR 5 m	0.61	0.56	0.48	0.13	0.22	0.00

The represented AOI2 test cohort contained no Milpa parcel; its 35 parcels therefore provided no basis for assessing Milpa recall or Milpa–maize discrimination and were restricted to a maize-only false-positive control. Native Sentinel-2 correctly classified 18 of the 35 maize

parcels, bicubic interpolation classified 19 correctly and CNN-SR classified 21 correctly. The corresponding maize-to-Milpa false-positive rates were 0.49, 0.46 and 0.40, respectively. Balanced accuracy, macro-F1, Milpa recall and Milpa F1 were not defined for this control.

CNN-SR reduced pooled reconstruction error, but the gain was heterogeneous and did not extend to AOI1 or the represented Milpa reconstruction summary. Because the represented AOI2 test cohort was maize-only, its reconstruction gain and lower maize false-positive rate provided no evidence of improved Milpa detection. In the only mixed-class test, AOI1, bicubic interpolation produced the strongest classification metrics, while CNN-SR detected only one of eight Milpa parcels. Detailed represented-cohort, reconstruction and downstream confusion evidence is provided in Appendix 5.4. Section 5.5 next evaluates post-prediction classification-level integration using native sensor probabilities.

5.5 Classification-level integration

Classification-level integration was evaluated using independently produced native Milpa probabilities from the Sentinel-2 only, SkySat only and PlanetScope spectral only models. The common cohort comprised 71 development parcels evaluated through directional geographic out-of-fold predictions for AOI3 and AOI4, together with the frozen 78-parcel secondary benchmark from AOI1 and AOI2, giving 149 parcels overall. The development values reported here therefore differ from the repeated stratified development summaries presented in Section 5.2.

5.5.1 Native-model agreement and complementarity

The three native models showed both shared success and shared failure. In the development geographic-OOF cohort, all three models correctly classified 21 parcels and all three incorrectly classified 28. Exactly one native model was correct for 12 parcels, while exactly two were correct for 10. In the frozen secondary benchmark, all three models were correct for 41 parcels and all three were incorrect for four. Exactly one model was correct for 11 benchmark parcels and exactly two were correct for 22.

Native agreement was strongly class and AOI dependent. All three models unanimously misclassified all 26 Milpa parcels in AOI4. Native-model disagreement therefore did not automatically provide useful complementarity, as substantial shared failure remained within the development Milpa class. Figure 5.5 summarises the native agreement patterns and the numbers of Sentinel-2 predictions corrected or harmed by the subsequent fixed fusions.

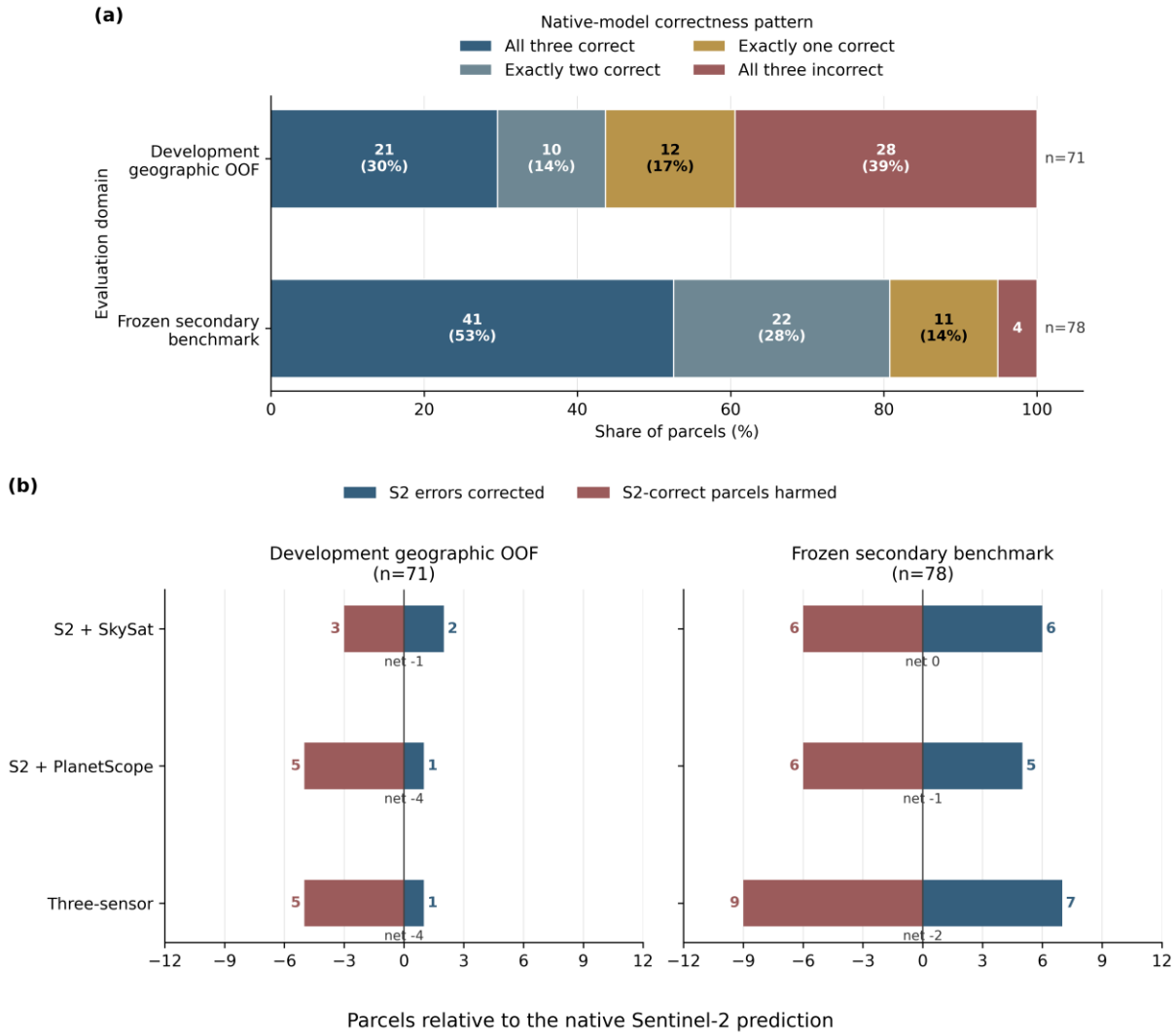


Figure 5.5. Native-model agreement and parcel-level correction or harm under fixed classification-level fusion. Panel (a) summarises the mutually exclusive native-model correctness patterns for the 71-parcel development geographic-OOF cohort and the 78-parcel frozen secondary benchmark: all three native models correct, all three incorrect, exactly one correct and exactly two correct. Panel (b) shows the numbers of Sentinel-2 errors corrected and Sentinel-2-correct parcels harmed by the fixed Sentinel-2 + SkySat, Sentinel-2 + PlanetScope and three-sensor probability fusions. All fusions used predefined equal weights and a 0.50 decision threshold; no exact threshold ties occurred.

5.5.2 Pairwise classification-level fusion

The pairwise models used fixed equal-weight averages of the native Milpa probabilities and a decision threshold of 0.50. No calibration, retraining, weight optimisation or threshold optimisation was applied, and no exact threshold ties occurred.

For Sentinel-2 + SkySat, development accuracy was 0.54, balanced accuracy was 0.45 and macro-F1 was 0.35. Milpa recall and Milpa F1 remained 0.00, and the fusion did not improve on the Sentinel-2 development result. On the frozen secondary benchmark, the fusion achieved an accuracy of 0.78, balanced accuracy of 0.67, macro-F1 of 0.66, Milpa recall of 0.50 and Milpa F1 of 0.45. This was the strongest pooled fixed-fusion benchmark result. It matched the Sentinel-2 accuracy of 0.78, while balanced accuracy increased from 0.62 to 0.67, macro-F1 from 0.62 to 0.66 and Milpa recall from 0.36 to 0.50. Relative to Sentinel-2,

six errors were corrected, while six parcels that Sentinel-2 had classified correctly were harmed.

The Sentinel-2 + PlanetScope fusion produced development accuracy of 0.49, balanced accuracy of 0.42 and macro-F1 of 0.33, with Milpa recall and Milpa F1 of 0.00. It corrected one Sentinel-2 error but harmed five correct Sentinel-2 predictions. On the benchmark, accuracy was 0.77, balanced accuracy was 0.61, macro-F1 was 0.61, Milpa recall was 0.36 and Milpa F1 was 0.36. Five Sentinel-2 errors were corrected and six correct Sentinel-2 predictions were harmed. The pooled benchmark therefore did not improve over Sentinel-2. AOI-specific performance differed markedly: in AOI1, balanced accuracy was 0.51 and Milpa recall was 0.11, whereas in AOI2 they were 0.82 and 0.80, respectively. This contrast did not translate into a pooled benchmark advantage over Sentinel-2.

Table 5.8 summarises the native and fixed-fusion performance across the development geographic-OOF and frozen secondary benchmark domains.

Table 5.8. Classification-level performance of native and fixed-fusion models in the development geographic-OOF and frozen secondary benchmark domains.

Domain	Role	Model	n	Accuracy	Balanced accuracy	Macro-F1	Milpa precision	Milpa recall	Milpa F1
Development geographic OOF	Native reference	Sentinel-2 only	71	0.55	0.46	0.35	0.00	0.00	0.00
Development geographic OOF	Native complementary-sensor model	SkySat only	71	0.41	0.35	0.31	0.07	0.03	0.05
Development geographic OOF	Native complementary-sensor model	PlanetScope spectral only	71	0.38	0.33	0.31	0.11	0.07	0.08
Development geographic OOF	Pairwise fixed fusion	Sentinel-2 + SkySat fusion	71	0.54	0.45	0.35	0.00	0.00	0.00
Development geographic OOF	Pairwise fixed fusion	Sentinel-2 + PlanetScope fusion	71	0.49	0.42	0.33	0.00	0.00	0.00
Development geographic OOF	Three-sensor fixed fusion	Three-sensor fusion	71	0.49	0.42	0.33	0.00	0.00	0.00
Frozen secondary benchmark	Native reference	Sentinel-2 only	78	0.78	0.62	0.62	0.38	0.36	0.37
Frozen secondary benchmark	Native complementary-sensor model	SkySat only	78	0.77	0.66	0.65	0.39	0.50	0.44
Frozen secondary benchmark	Native complementary-sensor model	PlanetScope spectral only	78	0.73	0.64	0.61	0.33	0.50	0.40
Frozen secondary benchmark	Pairwise fixed fusion	Sentinel-2 + SkySat fusion	78	0.78	0.67	0.66	0.41	0.50	0.45
Frozen secondary benchmark	Pairwise fixed fusion	Sentinel-2 + PlanetScope fusion	78	0.77	0.61	0.61	0.36	0.36	0.36
Frozen secondary benchmark	Three-sensor fixed fusion	Three-sensor fusion	78	0.76	0.63	0.62	0.35	0.43	0.39

5.5.3 Three-sensor fusion and bootstrap uncertainty

The three-sensor fusion used the fixed mean of the three native Milpa probabilities:

$$p_{three} = \frac{p_{S2} + p_{SkySat} + p_{PlanetScope}}{3}.$$

In the development geographic-OOF cohort, the three-sensor fusion produced an accuracy of 0.49, balanced accuracy of 0.42 and macro-F1 of 0.33. Milpa recall and Milpa F1 were both 0.00, and no development Milpa parcel was correctly classified.

On the frozen secondary benchmark, the three-sensor fusion achieved an accuracy of 0.76, balanced accuracy of 0.63, macro-F1 of 0.62, Milpa recall of 0.43 and Milpa F1 of 0.39. Relative to Sentinel-2, seven errors were corrected, but nine parcels that Sentinel-2 had classified correctly were harmed. The fusion assigned 11 maize monocrop parcels to Milpa and remained below both SkySat only and the Sentinel-2 + SkySat fusion in balanced accuracy, macro-F1, Milpa recall and Milpa F1.

The descriptive bootstrap assessment placed development balanced accuracy at 0.42 [0.37, 0.46], with a difference from Sentinel-2 of −0.05 [−0.10, 0.00]. Development Milpa recall remained 0.00 [0.00, 0.00]. On the benchmark, balanced accuracy was 0.63 [0.50, 0.76] and Milpa recall was 0.43 [0.21, 0.71]. The balanced-accuracy difference from Sentinel-2 was 0.01 [−0.11, 0.15], while the Milpa-recall difference was 0.07 [−0.14, 0.29].

Complete native-model agreement, fusion correction and harm, AOI-specific performance and bootstrap summaries are provided in Appendix 5.5.

Across the fixed classification-level combinations, Sentinel-2 + SkySat produced the strongest pooled benchmark fusion result, but no fusion improved development Milpa recognition. Sentinel-2 + PlanetScope did not improve the pooled benchmark over Sentinel-2, while the three-sensor fusion did not surpass the strongest pairwise fusion.

Chapter 6: Discussion

6.1 Overall interpretation of the findings

This thesis examined whether parcel-level discrimination of the Milpa mixed-cropping system from maize monocrop could be strengthened by combining a seasonal Sentinel-2 baseline with direct finer-spatial predictor representations, SkySat-guided super-resolution and classification-level integration. The central finding is that the binary distinction was more tractable than separating Milpa, landrace maize and hybrid maize as three classes, but Milpa recognition remained limited when the models were transferred to the secondary benchmark. Sentinel-2 provided the most defensible common reference because it represented the complete analytical cohort through a consistent annual spectral-temporal framework. Even so, the benchmark results showed that the apparent strength of the binary formulation did not amount to reliable recognition of Milpa across locations.

The complementary routes added useful information in specific settings rather than a stable improvement across the study. Direct SkySat and PlanetScope representations increased Milpa recall on the pooled secondary benchmark, but also increased maize false positives and did not reproduce the same advantage in the geographically separated development evidence. Convolutional neural network super-resolution reduced pooled reconstruction error relative to bicubic interpolation, yet it did not improve thematic discrimination in the mixed-class downstream test. Fixed probability fusion produced its strongest benchmark result for Sentinel-2 plus SkySat, but this gain was not reproduced for development Milpa recognition, and the three-sensor combination did not improve on the strongest pairwise fusion.

These outcomes are consistent with the underlying difficulty of treating Milpa as one remotely sensed class. Milpa is a family of maize-centred mixed-cropping systems whose composition, density and arrangement vary among fields, so within-class variability can coexist with strong similarity to maize monocrop (Fonteyne et al., 2023; Mota-Cruz et al., 2025). The contribution of this thesis is therefore not a claim that one additional sensor or integration method provided a complete solution to the mapping problem. It is a controlled account of where additional spatial, reconstructed and model-level information helped, where it failed to help, and which trade-offs became visible when class-sensitive and geographically separated evidence were considered together.

6.2 Sentinel-2 baseline, class separability and geographic transfer

The Sentinel-2 analysis showed that the principal difficulty was not simply obtaining a complete feature table, but learning a crop-system distinction that remained stable across geographical domains. Development-domain Normalised Difference Vegetation Index trajectories overlapped through much of the season. This is consistent with the class definition: a maize-dominated Milpa parcel can resemble maize monocrop, while variation in companion crops, planting geometry and soil exposure broadens the patterns assigned to Milpa. Satellite classifiers learn statistical relationships rather than agronomic meaning, and their performance depends on how consistently the reference classes are represented (Orynbaikyzy et al., 2019; Tariq et al., 2023). The observed NDVI trajectory overlap

therefore indicates that NDVI alone did not provide a clear and consistent seasonal separation between Milpa and maize monocrop in the development domain. The profiles also did not support the simple expectation that Milpa would necessarily appear greener: at the largest recorded class medians on 9 September, Milpa NDVI was 0.48, compared with 0.51 for landrace maize and 0.59 for hybrid maize, while the trajectories overlapped across much of the season. The predictor set also included EVI2, NDRE, CI red edge, NDMI and PSRI; with a larger reference dataset, future work could test whether LAI-related information or more narrowly targeted phenological windows provide additional separation.

The binary formulation was nevertheless more defensible than the three-class formulation. On the secondary benchmark, three-class macro-F1 was 0.38, with hybrid maize recall of 0.13 and Milpa recall of 0.36. Most hybrid parcels were assigned to landrace maize, indicating that the subdivision of maize type was especially weak under geographic transfer. Combining landrace and hybrid maize increased binary balanced accuracy and macro-F1 to 0.62, but Milpa recall remained 0.36. Thus, the higher binary performance was consistent with removing the additional landrace-versus-hybrid distinction, but it did not resolve the Milpa target-class problem. Overall accuracy alone would have obscured this limitation because the more numerous maize monocrop parcels were recognised much more often than Milpa. This illustrates why balanced accuracy, macro-F1, recall and confusion counts must be interpreted together in an imbalanced crop-mapping problem (Foody, 2002; Farhadpour et al., 2024).

The difference between development and benchmark performance further indicates limited geographic transfer. Repeated stratified development summaries supported stronger average discrimination, while the geographically separated benchmark exposed lower three-class performance and restricted Milpa recognition. Crop-classification relationships commonly weaken when models are applied to areas with different feature distributions, crop prevalence or observation conditions, so within-domain performance cannot be assumed to transfer spatially (Orynbaikyzy et al., 2022; Hoppe et al., 2024). The present pattern is consistent with geographic heterogeneity or domain shift, but the analysis did not isolate a single cause. The uneven composition of the areas of interest (AOIs) and classes also narrows interpretation: Milpa prevalence differed substantially among areas, and AOI1 contained no hybrid maize parcel, preventing a complete three-class assessment there.

Observation support was an important context but did not, by itself, explain the benchmark failure. Sentinel-2 interpolation was reliable in aggregate, although errors increased when long gaps coincided with August to October, and the weakest full-year support was concentrated in the benchmark. Yet restricting evaluation to the better-supported subset reduced rather than improved the main class-sensitive measures. This does not show that observation quality was unimportant. It shows only that weak support alone did not account for the principal failure under the tested restriction, which also changed the composition of the evaluated subset. A complete-looking regular series can conceal differences in the timing and continuity of measured observations, so support remains a necessary reliability consideration rather than a sufficient explanation of classification error (Defourny et al., 2019; Perich et al., 2023).

The alternative diagnostics reinforced the choice of the parcel-level Random Forest as the primary baseline. The Support Vector Machine did not surpass it on balanced or Milpa-sensitive benchmark measures. Pixel-level classification was valid because pixels from each parcel remained within a single split, but it performed poorly for Milpa; probability aggregation improved several pixel diagnostic measures without reaching the parcel model on the principal balanced and Milpa-sensitive measures. Parcel-level representation was therefore appropriate for the field-level reference labels, while the controlled pixel diagnostic showed that a pixel-level representation of the same Sentinel-2 observations did not recover a more transferable distinction. This is consistent with the wider principle that parcel-controlled evaluation prevents within-field leakage, whereas geographically separated assessment is still required to test broader transfer (Karasiak et al., 2022).

6.3 Contribution of direct finer-spatial predictor representations

SkySat and PlanetScope supplied finer-spatial observations of the same parcels, and both native representations increased Milpa recall on the pooled secondary benchmark relative to Sentinel-2. These changes showed that the finer-spatial representations altered the decisions for some Milpa parcels that Sentinel-2 missed. They were not unqualified improvements. SkySat and PlanetScope also assigned more maize parcels to Milpa, so increased recall was accompanied by lower precision or weaker overall balance. Moreover, the advantage varied by AOI and was not consistently reproduced in the geographic development predictions. The evidence therefore indicates different decision information, not universal superiority.

This interpretation aligns with literature showing that finer pixels can reduce boundary mixing and represent small fields more directly (Rao et al., 2021; Ibrahim et al., 2021), while greater spatial detail does not by itself guarantee better crop classification; its value also depends on the spectral and temporal information available (Rao et al., 2021). The present comparisons cannot isolate spatial resolution as the cause of the differences. SkySat, PlanetScope and Sentinel-2 also differed in spectral configuration, acquisition timing, temporal coverage, valid support and preprocessing. PlanetScope represented a small October observation set, whereas Sentinel-2 summarised annual spectral-temporal behaviour. SkySat likewise provided detailed but irregular observations rather than a comparable seasonal record. The finer-spatial datasets should therefore be regarded as complementary observations, not as ground truth or automatically superior replacements for Sentinel-2.

Predictor concatenation did not convert these conditional gains into consistent improvement. Combining Sentinel-2 with SkySat or PlanetScope did not improve the principal pooled benchmark measures, and the three pre-specified PlanetScope spatial variables were non-improving. Although multi-sensor predictors have improved crop classification in some optical-radar settings (Orynbaikyzy et al., 2020; Zhang et al., 2023), additional predictors can increase redundancy and dimensionality without providing equivalent new information (Gómez-Chova et al., 2003; Orynbaikyzy et al., 2020; Tang et al., 2025). More variables therefore do not necessarily provide more discriminating information. In this study, the direct sensor models showed that finer-spatial observations could alter the recall-versus-false-positive trade-off, but the combined models did not establish stable incremental value beyond the seasonal Sentinel-2 baseline.

6.4 Super-resolution: reconstruction fidelity and thematic value

The super-resolution experiment showed why image reconstruction and crop classification must be evaluated separately. Across the pooled untouched test patches, the convolutional neural network reduced mean absolute error and root mean squared error relative to bicubic interpolation. The improvement remained under equal-parcel weighting, indicating that it was not produced only by parcels contributing many patches. Nevertheless, the effect was heterogeneous by band, AOI and class representation. AOI2 improved strongly, whereas AOI1 did not, and the represented Milpa subset showed no pooled reconstruction advantage. Because all test Milpa parcels occurred in AOI1 and the represented AOI2 test cohort was maize-only, the analysis could not separate a crop-class effect from the geographical composition of the test data.

The downstream classification evidence was more restrictive. In the only mixed-class test, bicubic interpolation produced the strongest balanced and Milpa-sensitive result, while the convolutional neural network reconstruction detected only one of eight Milpa parcels. The network reduced pooled pixel-level reconstruction error, but this did not preserve or strengthen the distinctions needed for Milpa recognition. The finding supports the broader warning that lower radiometric error or greater apparent detail does not guarantee improved thematic utility (Vasilescu et al., 2023; Major et al., 2025). Image-level fidelity and thematic utility must therefore be assessed separately. Bicubic interpolation, although unable to create newly observed detail, remained an essential baseline and performed better for the tested downstream task.

Interpretation is also bounded by the represented subset. Of the 149 parcels, 112 produced accepted paired patches and 37 did not, so the branch did not evaluate the full cohort. The AOI2 downstream control contained only maize parcels and could assess maize-to-Milpa false positives, not Milpa detection. SkySat was a finer-spatial target observation rather than error-free ground truth. Temporal mismatch within the permitted pairing window, AOI-specific alignment corrections, sensor-response differences and common-valid masking also bounded the comparison. Cross-sensor studies must control these factors because a model can otherwise learn temporal or geometric differences as apparent spatial detail (Okabayashi et al., 2024; Aybar et al., 2024, 2026).

The experiment should therefore be understood as a controlled cross-sensor reconstruction and transfer test, not an operational super-resolution product. Only the Sentinel-2 Blue, Green, Red and near-infrared bands with matched SkySat targets were reconstructed. The analysis did not establish the utility of super-resolving the complete Sentinel-2 spectral system, particularly the red-edge and shortwave-infrared bands that were neither reconstructed nor validated. The strongest conclusion is specific: the tested network improved pooled reconstruction fidelity, but that improvement was geographically uneven and did not translate into better Milpa discrimination in the available mixed-class downstream evidence.

6.5 Classification-level integration and model complementarity

Classification-level integration tested whether separate native models could correct one another after prediction. Disagreement created opportunities for correction, but it was not

equivalent to useful complementarity. In the development geographic out-of-fold predictions, all three native models unanimously misclassified all 26 AOI4 Milpa parcels. This shared failure meant that averaging their outputs could not recover target-class information that none of the component models expressed. The pattern is consistent with the general principle that fusion helps most when models make different, informative errors, whereas correlated or systematically weak models may reinforce the same mistake (Orynbaikyzy et al., 2020; Valero et al., 2021).

Sentinel-2 plus SkySat produced the strongest pooled fixed-fusion benchmark result, increasing balanced accuracy, macro-F1 and Milpa recall relative to Sentinel-2. This was an observed benchmark gain under the predefined rule, but it was conditional. Development Milpa recall remained zero, and the fusion corrected six Sentinel-2 errors while harming six parcels that Sentinel-2 had classified correctly. Sentinel-2 plus PlanetScope did not improve the pooled benchmark, and the three-sensor fusion did not surpass Sentinel-2 plus SkySat. Across the tested combinations, correction and harm occurred together; a gain in one subset or class could be offset by new errors elsewhere.

The three-sensor bootstrap assessment further limited the claim. The paired intervals for differences from Sentinel-2 were broad and included no clear improvement in balanced accuracy or Milpa recall. These intervals were descriptive rather than evidence of a formal superiority test, but they showed that the observed benchmark differences were uncertain within the available sample. The benchmark must not be used retrospectively to tune weights or thresholds, because doing so would weaken the validation hierarchy and turn a fixed comparison into post hoc optimisation (Varma and Simon, 2006; Cawley and Talbot, 2010).

Equal weighting and the fixed threshold provided a transparent diagnostic, although the native class-support outputs were not calibrated to a common probability scale. Random Forest support values do not automatically correspond to empirical probabilities, and averaging differently calibrated outputs can give disproportionate influence to one model (Boström, 2007; Niculescu-Mizil and Caruana, 2005; Dimitriadis et al., 2021). The fusion results therefore support cautious interpretation: SkySat added useful benchmark decision information in the strongest pairwise combination, but model disagreement did not provide consistent or geographically transferable complementarity.

6.6 Limitations, implications and future directions

The most important limitation concerns reference and geographic representation. The final cohort of 149 parcels was small, geographically uneven and class-imbalanced, with large differences in Milpa availability among AOIs and no hybrid maize parcel in AOI1. The available reference documentation did not fully describe the field-survey organisation, label protocol or parcel-boundary delineation. These constraints affect how confidently model errors can be attributed to remotely sensed class overlap rather than reference uncertainty. In addition, AOI1 and AOI2 formed a frozen secondary benchmark drawn from the same study programme. They provided a geographically separated assessment, but not independent, external or confirmatory validation.

Observation and sensor comparability also bounded the comparisons. Sentinel-2 support was uneven, and long gaps during August to October were reconstructed less reliably. SkySat and PlanetScope did not offer equivalent spectral or temporal sampling: the direct finer-spatial representations contained fewer and differently timed observations, while Sentinel-2 represented the annual cycle. Consequently, differences among sensors cannot be assigned uniquely to spatial resolution. Larger multi-year datasets with better coverage of the principal crop period would test whether seasonal relationships persist and would reduce dependence on one year of cloud and acquisition conditions. Temporally matched and radiometrically harmonised observations would also make future cross-sensor comparisons more diagnostic of genuinely complementary information (Okabayashi et al., 2024; Aybar et al., 2024, 2026).

The super-resolution branch had a narrower evidence base than the full classification analysis. Only parcels producing accepted pairs entered reconstruction and downstream testing, the finer-spatial reference contained temporal, alignment and cross-sensor differences, and only Blue, Green, Red and near-infrared transfer was evaluated. Future extension to Sentinel-2 red-edge and shortwave-infrared bands would be scientifically useful only where defensible finer-spatial reference or independent validation evidence exists. More importantly, future work should retain separate reconstruction and thematic tests, because the present results show that improved image fidelity does not establish improved class discrimination.

The integration analysis was deliberately diagnostic. Fixed, uncalibrated equal-weight fusion protected the secondary benchmark from tuning, but it could not account for differences in probability scale or learn when one sensor should be trusted more than another. Calibration, trainable weighting or stacking would be more defensible within a larger development domain, with all calibration, model-selection and threshold decisions completed before independent geographic evaluation (Varma and Simon, 2006; Cawley and Talbot, 2010). Given the small and uneven development sample, a more flexible fusion model would increase the risk of overfitting.

The practical implication is therefore bounded but useful. The evidence does not support an operational claim that finer-spatial imagery, super-resolution or fusion reliably resolves Milpa mapping across Oaxaca. It does support a leakage-controlled parcel-level framework for identifying where discrimination is stronger, where Milpa-specific failure persists and where additional information produces trade-offs rather than stable gains. The most valuable next step is not simply to add more sensors. It is to improve the reference basis through a larger, more balanced and geographically broader sample, with parcel-level records of crop composition, planting arrangement and management. Genuinely independent geographic evaluation should then test whether any improvement persists across AOIs and Milpa-sensitive measures. These changes directly address the principal uncertainty revealed by the thesis: whether the learned distinction reflects transferable crop-system information rather than the particular composition and observation conditions of the current cohort.

Chapter 7: Conclusion

This thesis evaluated whether parcel-level mapping of the Milpa mixed-cropping system in Oaxaca could be improved beyond a seasonal Sentinel-2 baseline through direct SkySat and PlanetScope predictor representations, SkySat-guided super-resolution and classification-level integration. The study treated Milpa as a crop-system class separate from maize monocrop and used a leakage-controlled development framework with a geographically separated secondary benchmark to test not only overall discrimination, but also target-class recognition and spatial transfer.

Sentinel-2 provided the most defensible common baseline across the full 149-parcel cohort. The binary Milpa versus maize monocrop formulation was stronger than the three-class separation of Milpa, landrace maize and hybrid maize. Nevertheless, Milpa recall remained limited on the secondary benchmark, and the difference between development and benchmark evidence showed that the learned distinction did not transfer consistently across the study areas. Stricter observation-support filtering, a Support Vector Machine and a parcel-controlled pixel diagnostic did not surpass the primary parcel-level Random Forest on the principal balanced and Milpa-sensitive measures.

Direct SkySat and PlanetScope representations added some useful decision information. Both increased pooled benchmark Milpa recall relative to Sentinel-2, but these gains involved more maize false positives and were not reproduced consistently in development or across AOIs. Combining either finer-spatial predictor set with Sentinel-2 did not improve the principal pooled benchmark measures, and the PlanetScope spatial additions were non-improving. Finer-spatial observations were therefore complementary in specific comparisons, but they did not establish a stable additional source of geographically transferable class discrimination.

The image-level experiment reached a similarly conditional conclusion. Convolutional neural network super-resolution reduced pooled reconstruction error relative to bicubic interpolation, but the benefit varied by AOI, band and parcel representation. It did not improve the represented Milpa reconstruction summary, and in the only mixed-class downstream assessment it detected only one of eight Milpa parcels while bicubic interpolation produced the strongest classification result. Improved reconstruction fidelity therefore did not translate into improved Milpa discrimination, and the experiment did not establish the value of super-resolving the complete Sentinel-2 spectral system.

At classification level, fixed equal-weight fusion exposed both potential complementarity and substantial shared failure. Sentinel-2 plus SkySat was the strongest tested fusion on the pooled secondary benchmark, but no fusion improved development Milpa recognition. Sentinel-2 plus PlanetScope did not improve the pooled benchmark over Sentinel-2, and the three-sensor fusion did not surpass the strongest pairwise combination. Model disagreement therefore created some correctable errors, but it did not constitute consistent transferable complementarity.

The overall answer to the thesis question is that complementary information from SkySat, PlanetScope, super-resolution and classification-level integration did not provide a consistent

geographically transferable improvement over the Sentinel-2 parcel baseline for Milpa recognition. Their value was conditional, route-specific and often accompanied by false-positive or geographic trade-offs. This conclusion does not negate the usefulness of the additional data. Instead, it defines the conditions under which apparent gains should be accepted or rejected.

The methodological contribution lies in the integrated evaluation framework: an explicit parcel-level Milpa class definition; assessment of observed support and temporal interpolation; separation between development and the secondary benchmark; and controlled comparison of predictor-level, image-level and classification-level integration. Together, the results show that finer spatial sampling, lower reconstruction error and model disagreement do not automatically create transferable crop-system discrimination. Progress towards reliable Milpa mapping will require larger and more representative field references, clearer records of within-parcel crop composition and management, temporally harmonised multi-sensor observations, and genuinely independent geographic evaluation of improvements developed entirely within an adequate development domain.

References

Akbari, E., Boloorani, A. D., Samany, N. N., Hamzeh, S., Soufizadeh, S., and Pignatti, S. (2020). Crop mapping using Random Forest and particle swarm optimization based on multi-temporal Sentinel-2. Remote Sensing, 12(9), 1449. https://doi.org/10.3390/rs12091449

Ambroise, C., and McLachlan, G. J. (2002). Selection bias in gene extraction on the basis of microarray gene-expression data. Proceedings of the National Academy of Sciences of the United States of America, 99(10), 6562–6566. https://doi.org/10.1073/pnas.102102699

Aybar, C., Contreras, J., Donike, S., Portalés-Julià, E., Mateo-García, G., and Gómez-Chova, L. (2026). A radiometrically and spatially consistent super-resolution framework for Sentinel-2. Remote Sensing of Environment, 334, 115222. https://doi.org/10.1016/j.rse.2025.115222

Aybar, C., Montero, D., Contreras, J., Donike, S., Kalaitzis, F., and Gómez-Chova, L. (2024). SEN2NAIP: A large-scale dataset for Sentinel-2 image super-resolution. Scientific Data, 11, 1389. https://doi.org/10.1038/s41597-024-04214-y

Bégué, A., Arvor, D., Bellón, B., Betbeder, J., de Abelleyra, D., Ferraz, R. P. D., Lebourgeois, V., Lelong, C., Simões, M., and Verón, S. R. (2018). Remote sensing and cropping practices: A review. Remote Sensing, 10(1), 99. https://doi.org/10.3390/rs10010099

Benrey, B., Bustos-Segura, C., and Grof-Tisza, P. (2024). The Mesoamerican Milpa system: Traditional practices, sustainability, biodiversity, and pest control. Biological Control, 198, 105637. https://doi.org/10.1016/j.biocontrol.2024.105637

Bontemps, S., Cara, C., Deffense, N., Nicola, L., Nørgaard, B., and Udroiu, C. (2024). ESA Sen4Stat: Sentinels for Agricultural Statistics, D10.1 – Software User Manual (Issue/Rev. 1.3). Submitted to the European Space Agency. https://www.esa-sen4stat.org/wp-content/uploads/content/Sen4Stat_SUM_v1.3.pdf

Boström, H. (2007). Estimating class probabilities in random forests. In Sixth International Conference on Machine Learning and Applications (ICMLA 2007), 211–216. IEEE Computer Society. https://doi.org/10.1109/ICMLA.2007.64

Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5–32. https://doi.org/10.1023/A:1010933404324

Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. Monthly Weather Review, 78(1), 1–3. https://doi.org/10.1175/1520-0493(1950)078<0001:VOFEIT>2.0.CO;2

Brooker, R. W., Bennett, A. E., Cong, W.-F., Daniell, T. J., George, T. S., Hallett, P. D., Hawes, C., Iannetta, P. P. M., Jones, H. G., Karley, A. J., Li, L., McKenzie, B. M., Pakeman, R. J., Paterson, E., Schöb, C., Shen, J., Squire, G., Watson, C. A., Zhang, C., Zhang, F., Zhang, J., and White, P. J. (2015). Improving intercropping: A synthesis of research in agronomy, plant physiology and ecology. New Phytologist, 206(1), 107–117. https://doi.org/10.1111/nph.13132

Bukowiecki, J., Rose, T., and Kage, H. (2021). Sentinel-2 data for precision agriculture? A UAV-based assessment. Sensors, 21(8), 2861. https://doi.org/10.3390/s21082861

Cawley, G. C., and Talbot, N. L. C. (2010). On over-fitting in model selection and subsequent selection bias in performance evaluation. Journal of Machine Learning Research, 11, 2079–2107.

Chen, W., and Liu, G. (2024). A novel method for identifying crops in parcels constrained by environmental factors through the integration of a Gaofen-2 high-resolution remote-sensing image and Sentinel-2 time series. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 17, 450–463. https://doi.org/10.1109/JSTARS.2023.3329987

Chen, Y., Cao, R., Chen, J., Liu, L., and Matsushita, B. (2021). A practical approach to reconstruct high-quality Landsat NDVI time-series data by gap filling and the Savitzky–Golay filter. ISPRS Journal of Photogrammetry and Remote Sensing, 180, 174–190. https://doi.org/10.1016/j.isprsjprs.2021.08.015

Cheng, G., Ding, H., Yang, J., and Cheng, Y. (2023). Crop type classification with combined spectral, texture, and radar features of time-series Sentinel-1 and Sentinel-2 data. International Journal of Remote Sensing, 44(4), 1215–1237. https://doi.org/10.1080/01431161.2023.2176723

Cheng, J., Huang, L., Tang, B.-H., Wu, Q., Wang, M., and Zhang, Z. (2025). A minority-sample-enhanced sampler for crop classification in unmanned aerial vehicle remote-sensing images with class imbalance. Agriculture, 15(4), 388. https://doi.org/10.3390/agriculture15040388

Clevers, J. G. P. W., and Gitelson, A. A. (2013). Remote estimation of crop and grass chlorophyll and nitrogen content using red-edge bands on Sentinel-2 and Sentinel-3. International Journal of Applied Earth Observation and Geoinformation, 23, 344–351. https://doi.org/10.1016/j.jag.2012.10.008

Cohen, J. (1960). A coefficient of agreement for nominal scales. Educational and Psychological Measurement, 20(1), 37–46. https://doi.org/10.1177/001316446002000104

Congalton, R. G. (1991). A review of assessing the accuracy of classifications of remotely sensed data. Remote Sensing of Environment, 37(1), 35–46. https://doi.org/10.1016/0034-4257(91)90048-B

Copernicus Sentinel-2 (2021). MSI Level-2A BOA Reflectance Product. Collection 1. Processed by European Space Agency. https://doi.org/10.5270/S2_-znk9xsj

Defourny, P., Bontemps, S., Bellemans, N., Cara, C., Dedieu, G., Guzzonato, E., Hagolle, O., Inglada, J., Nicola, L., Rabaute, T., Savinaud, M., Udroiu, C., Valero, S., Bégué, A., Dejoux, J.-F., El Harti, A., Ezzahar, J., Kussul, N., Labbassi, K., and Koetz, B. (2019). Near real-time agriculture monitoring at national scale at parcel resolution: Performance assessment of the Sen2-Agri automated system in various cropping systems around the world. Remote Sensing of Environment, 221, 551–568. https://doi.org/10.1016/j.rse.2018.11.007

Delgado, R. (2022). A semi-hard voting combiner scheme to ensemble multi-class probabilistic classifiers. Applied Intelligence, 52(4), 3653–3677. https://doi.org/10.1007/s10489-021-02447-7

Dimitriadis, T., Gneiting, T., and Jordan, A. I. (2021). Stable reliability diagrams for probabilistic classifiers. Proceedings of the National Academy of Sciences of the United States of America, 118(8), e2016191118. https://doi.org/10.1073/pnas.2016191118

Donike, S., Aybar, C., Gómez-Chova, L., and Kalaitzis, F. (2025). Trustworthy super-resolution of multispectral Sentinel-2 imagery with latent diffusion. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 18, 6940–6952. https://doi.org/10.1109/JSTARS.2025.3542220

Douzas, G., Bação, F., Fonseca, J., and Khudinyan, M. (2019). Imbalanced learning in land-cover classification: Improving minority classes' prediction accuracy using the Geometric SMOTE algorithm. Remote Sensing, 11(24), 3040. https://doi.org/10.3390/rs11243040

Efron, B., and Tibshirani, R. (1986). Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. Statistical Science, 1(1), 54–75. https://doi.org/10.1214/ss/1177013815

Estes, L., Ye, S., Song, L., Luo, B., Eastman, J. R., Meng, Z.-Z., Zhang, Q., McRitchie, D., Debats, S., Muhando, J., Kaloo, B. W., Makuru, J., Mbatia, B. K., Muasa, I. M., Mugami, A. M., Mugami, J., and Caylor, K. K. (2022). High-resolution annual maps of field boundaries for smallholder-dominated croplands at national scales. Frontiers in Artificial Intelligence, 4, 744863. https://doi.org/10.3389/frai.2021.744863

European Space Agency (2015). Sentinel-2 user handbook. ESA Standard Document, Issue 1, Revision 2, 24 July 2015. European Space Agency.

Fang, G., Wang, C., Dong, T., Wang, Z., Cai, C., Chen, J., Liu, M., and Zhang, H. (2025). A landscape-clustering zoning strategy to map multi-crops in fragmented cropland regions using Sentinel-2 and Sentinel-1 imagery with feature selection. Agriculture, 15(2), 186. https://doi.org/10.3390/agriculture15020186

Farhadpour, S., Warner, T. A., and Maxwell, A. E. (2024). Selecting and interpreting multiclass loss and accuracy-assessment metrics for classifications with class imbalance: Guidance and best practices. Remote Sensing, 16(3), 533. https://doi.org/10.3390/rs16030533

Fonteyne, S., Castillo Caamal, J. B., López-Ridaura, S., Van Loon, J., Espidio Balbuena, J., Osorio Alcalá, L., Martínez Hernández, F., Odjo, S., and Verhulst, N. (2023). Review of agronomic research on the milpa, the traditional polyculture system of Mesoamerica. Frontiers in Agronomy, 5, 1115490. https://doi.org/10.3389/fagro.2023.1115490

Foody, G. M. (2002). Status of land-cover classification accuracy assessment. Remote Sensing of Environment, 80(1), 185–201. https://doi.org/10.1016/S0034-4257(01)00295-4

Gómez-Chova, L., Calpe-Maravilla, J., Camps-Valls, G., Martín-Guerrero, J. D., Soria-Olivas, E., Vila-Francés, J., Alonso, L., and Moreno, J. (2003). Feature selection of hyperspectral data through local correlation and sequential floating forward selection for crop

classification. In Proceedings of the IEEE International Geoscience and Remote Sensing Symposium 2003, 1, 555–557. https://doi.org/10.1109/IGARSS.2003.1293840

Griffiths, P., Nendel, C., and Hostert, P. (2019). Intra-annual reflectance composites from Sentinel-2 and Landsat for national-scale crop and land cover mapping. Remote Sensing of Environment, 220, 135–151. https://doi.org/10.1016/j.rse.2018.10.031

Hoppe, H., Dietrich, P., Marzahn, P., Weiß, T., Nitzsche, C., von Lukas, U. F., Wengerek, T., and Borg, E. (2024). Transferability of machine-learning models for crop classification in remote-sensing imagery using a new test methodology: A study on phenological, temporal and spatial influences. Remote Sensing, 16(9), 1493. https://doi.org/10.3390/rs16091493

Hu, Q., Sulla-Menashe, D., Xu, B., Yin, H., Tang, H., Yang, P., and Wu, W. (2019). A phenology-based spectral and temporal feature-selection method for crop mapping from satellite time series. International Journal of Applied Earth Observation and Geoinformation, 80, 218–229. https://doi.org/10.1016/j.jag.2019.04.014

Ibrahim, E. S., Rufin, P., Nill, L., Kamali, B., Nendel, C., and Hostert, P. (2021). Mapping crop types and cropping systems in Nigeria with Sentinel-2 imagery. Remote Sensing, 13(17), 3523. https://doi.org/10.3390/rs13173523

Ibrahim, G. F. R., Rasul, A., and Abdullah, H. (2023). Improving crop-classification accuracy with integrated Sentinel-1 and Sentinel-2 data: A case study of barley and wheat. Journal of Geovisualization and Spatial Analysis, 7, 22. https://doi.org/10.1007/s41651-023-00152-2

Ji, C., and Tang, H. (2025). Towards reliable land-cover mapping under domain shift: An overview and comprehensive comparative study on uncertainty estimation. Earth-Science Reviews, 263, 105070. https://doi.org/10.1016/j.earscirev.2025.105070

Jia, W., Sun, M., Lian, J., and Hou, S. (2022). Feature dimensionality reduction: A review. Complex & Intelligent Systems, 8, 2663–2693. https://doi.org/10.1007/s40747-021-00637-x

Jiang, Z., Huete, A. R., Didan, K., and Miura, T. (2008). Development of a two-band enhanced vegetation index without a blue band. Remote Sensing of Environment, 112(10), 3833–3845. https://doi.org/10.1016/j.rse.2008.06.006

Karasiak, N., Dejoux, J.-F., Monteil, C., and Sheeren, D. (2022). Spatial dependence between training and test sets: Another pitfall of classification accuracy assessment in remote sensing. Machine Learning, 111(7), 2715–2740. https://doi.org/10.1007/s10994-021-05972-1

Karthikeyan, L., Chawla, I., and Mishra, A. K. (2020). A review of remote sensing applications in agriculture for food security: Crop growth and yield, irrigation, and crop losses. Journal of Hydrology, 586, 124905. https://doi.org/10.1016/j.jhydrol.2020.124905

Kim, M., Park, S., Sampath, A., Anderson, C., and Stensaas, G. L. (2022). System characterization report on Planet SkySat. U.S. Geological Survey Open-File Report 2021–1030–E. https://doi.org/10.3133/ofr20211030e

Kussul, N., Lavreniuk, M., Skakun, S., and Shelestov, A. (2017). Deep-learning classification of land cover and crop types using remote-sensing data. IEEE Geoscience and Remote Sensing Letters, 14(5), 778–782. https://doi.org/10.1109/LGRS.2017.2681128

Kussul, N., Lemoine, G., Gallego, F. J., Skakun, S. V., Lavrenyuk, M., and Shelestov, A. Y. (2016). Parcel-based crop classification in Ukraine using Landsat-8 data and Sentinel-1A data. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 9(6), 2500–2508. https://doi.org/10.1109/JSTARS.2016.2560141

Latte, N., and Lejeune, P. (2020). PlanetScope radiometric normalization and Sentinel-2 super-resolution (2.5 m): A straightforward spectral-spatial fusion of multi-satellite multi-sensor images using residual convolutional neural networks. Remote Sensing, 12(15), 2366. https://doi.org/10.3390/rs12152366

Li, M., Shamshiri, R. R., Weltzien, C., and Schirrmann, M. (2022). Crop monitoring using Sentinel-2 and UAV multispectral imagery: A comparison case study in northeastern Germany. Remote Sensing, 14(17), 4426. https://doi.org/10.3390/rs14174426

Li, N., Skaggs, T. H., and Scudiero, E. (2025). In-season estimation of Japanese squash using high-spatial-resolution time-series satellite imagery. Sensors, 25(7), 1999. https://doi.org/10.3390/s25071999

Lim, B., Son, S., Kim, H., Nah, S., and Lee, K. M. (2017). Enhanced deep residual networks for single image super-resolution. In 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops, 1132–1140. https://doi.org/10.1109/CVPRW.2017.151

Lira Melo de Oliveira Santos, C., Lamparelli, R. A. C., Figueiredo, G. K. D. A., Dupuy, S., Boury, J., Luciano, A. C. D. S., Torres, R. D. S., and le Maire, G. (2019). Classification of crops, pastures, and tree plantations along the season with multi-sensor image time series in a subtropical agricultural region. Remote Sensing, 11(3), 334. https://doi.org/10.3390/rs11030334

Liu, R., Shang, R., Liu, Y., and Lu, X. (2017). Global evaluation of gap-filling approaches for seasonal NDVI with considering vegetation growth trajectory, protection of key point, noise resistance and curve stability. Remote Sensing of Environment, 189, 164–179. https://doi.org/10.1016/j.rse.2016.11.023

Liu, X., Xie, S., Yang, J., Sun, L., Liu, L., Zhang, Q., and Yang, C. (2023). Comparisons between temporal statistical metrics, time-series stacks and phenological features derived from NASA Harmonized Landsat Sentinel-2 data for crop-type mapping. Computers and Electronics in Agriculture, 211, 108015. https://doi.org/10.1016/j.compag.2023.108015

López-Ridaura, S., Barba-Escoto, L., Reyna-Ramírez, C. A., Sum, C., Palacios-Rojas, N., and Gérard, B. (2021). Maize intercropping in the Milpa system: Diversity, extent and importance for nutritional security in the Western Highlands of Guatemala. Scientific Reports, 11(1), 3696. https://doi.org/10.1038/s41598-021-82784-2

Magno, R., Rocchi, L., Dainelli, R., Matese, A., Di Gennaro, S. F., Chen, C.-F., Son, N.-T., and Toscano, P. (2021). AgroShadow: A new Sentinel-2 cloud-shadow detection tool for precision agriculture. Remote Sensing, 13(6), 1219. https://doi.org/10.3390/rs13061219

Mahmood, T., Usman, M., and Conrad, C. (2025). Selecting relevant features for Random-Forest-based crop-type classifications by spatial assessments of backward feature reduction.

PFG – Journal of Photogrammetry, Remote Sensing and Geoinformation Science, 93(2), 173–196. https://doi.org/10.1007/s41064-024-00329-4

Maitra, S., Hossain, A., Brestič, M., Skalický, M., Ondrisik, P., Gitari, H. I., Brahmachari, K., Shankar, T., Bhadra, P., Palai, J. B., Jena, J., Bhattacharya, U., Duvvada, S. K., Lalichetti, S., and Sairam, M. (2021). Intercropping: A low-input agricultural strategy for food and environmental security. Agronomy, 11(2), 343. https://doi.org/10.3390/agronomy11020343

Major, D., Horváth, Z., Kröber, F., Augustin, H., Sudmanns, M., Ševčík, P., Baraldi, A., Berg, A., Cornel, D., and Tiede, D. (2025). A holistic approach for multi-spectral Sentinel-2 super-resolution and spectral evaluation. International Journal of Remote Sensing, 46(20), 7437–7464. https://doi.org/10.1080/01431161.2025.2549132

Maolan, K., Rusuli, Y., Zhang, X., and Kuluwan, Y. (2024). Sentinel-2 image based smallholder crops classification and accuracy assessment by UAV data. Geocarto International, 39(1), 2361733. https://doi.org/10.1080/10106049.2024.2361733

Mateos-Maces, L., Castillo-González, F., Chávez Servia, J. L., Estrada-Gómez, J. A., and Livera-Muñoz, M. (2016). Manejo y aprovechamiento de la agrobiodiversidad en el sistema Milpa del sureste de México. Acta Agronómica, 65(4), 413–421. https://doi.org/10.15446/acag.v65n4.50984

Mota-Cruz, C., Casas, A., Ortega-Paczka, R., Perales, H., Vega-Peña, E., and Bye, R. (2025). Milpa, a long-standing polyculture for sustainable agriculture. Agriculture, 15(16), 1737. https://doi.org/10.3390/agriculture15161737

Mountrakis, G., Im, J., and Ogole, C. (2011). Support vector machines in remote sensing: A review. ISPRS Journal of Photogrammetry and Remote Sensing, 66(3), 247–259. https://doi.org/10.1016/j.isprsjprs.2010.11.001

Moustafa, M., and Sayed, S. A. (2021). Satellite imagery super-resolution using squeeze-and-excitation-based GAN. International Journal of Aeronautical and Space Sciences, 22, 1481–1492. https://doi.org/10.1007/s42405-021-00396-6

NASA JPL. (2013). NASA Shuttle Radar Topography Mission Global 1 arc second [Data set]. NASA EOSDIS Land Processes Distributed Active Archive Center. https://doi.org/10.5067/MEaSUREs/SRTM/SRTMGL1.003

Niculescu-Mizil, A., and Caruana, R. (2005). Predicting good probabilities with supervised learning. In Proceedings of the 22nd International Conference on Machine Learning (ICML ’05), 625–632. ACM. https://doi.org/10.1145/1102351.1102430

Novotny, I. P., Tittonell, P., Fuentes-Ponce, M. H., López-Ridaura, S., and Rossing, W. A. H. (2021). The importance of the traditional milpa in food security and nutritional self-sufficiency in the highlands of Oaxaca, Mexico. PLOS ONE, 16(2), e0246281. https://doi.org/10.1371/journal.pone.0246281

Ofori-Ampofo, S., Pelletier, C., and Lang, S. (2021). Crop-type mapping from optical and radar time series using attention-based deep learning. Remote Sensing, 13(22), 4668. https://doi.org/10.3390/rs13224668

Okabayashi, A., Audebert, N., Donike, S., and Pelletier, C. (2024). Cross-sensor super-resolution of irregularly sampled Sentinel-2 time series. In 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, 502–511. https://doi.org/10.1109/CVPRW63382.2024.00055

Olofsson, P., Foody, G. M., Herold, M., Stehman, S. V., Woodcock, C. E., and Wulder, M. A. (2014). Good practices for estimating area and assessing accuracy of land change. Remote Sensing of Environment, 148, 42–57. https://doi.org/10.1016/j.rse.2014.02.015

Orynbaikyzy, A., Gessner, U., and Conrad, C. (2019). Crop-type classification using a combination of optical and radar remote-sensing data: A review. International Journal of Remote Sensing, 40(17), 6553–6595. https://doi.org/10.1080/01431161.2019.1569791

Orynbaikyzy, A., Gessner, U., and Conrad, C. (2022). Spatial transferability of Random Forest models for crop-type classification using Sentinel-1 and Sentinel-2. Remote Sensing, 14(6), 1493. https://doi.org/10.3390/rs14061493

Orynbaikyzy, A., Gessner, U., Mack, B., and Conrad, C. (2020). Crop-type classification using fusion of Sentinel-1 and Sentinel-2 data: Assessing the impact of feature selection, optical-data availability and parcel sizes on the accuracies. Remote Sensing, 12(17), 2779. https://doi.org/10.3390/rs12172779

Palsson, F., Sveinsson, J. R., and Ulfarsson, M. O. (2014). A new pansharpening algorithm based on total variation. IEEE Geoscience and Remote Sensing Letters, 11(1), 318–322. https://doi.org/10.1109/LGRS.2013.2257669

Perich, G., Turkoglu, M. O., Graf, L. V., Wegner, J. D., Aasen, H., Walter, A., and Liebisch, F. (2023). Pixel-based yield mapping and prediction from Sentinel-2 using spectral indices and neural networks. Field Crops Research, 292, 108824. https://doi.org/10.1016/j.fcr.2023.108824

Persello, C., Tolpekin, V. A., Bergado, J. R., and de By, R. A. (2019). Delineation of agricultural fields in smallholder farms from satellite images using fully convolutional networks and combinatorial grouping. Remote Sensing of Environment, 231, 111253. https://doi.org/10.1016/j.rse.2019.111253

Planet Labs PBC (2026a). SkySat. Planet Documentation. Updated 8 July 2026. Available at: https://docs.planet.com/data/imagery/skysat/ (Accessed: 27 July 2026).

Planet Labs PBC (2026b). PlanetScope. Planet Documentation. Updated 18 June 2026. Available at: https://docs.planet.com/data/imagery/planetscope/ (Accessed: 27 July 2026).

Planet Labs PBC (2026c). Planet imagery product specifications. June 2026. Planet Labs PBC. Available at: https://assets.planet.com/docs/Planet_Combined_Imagery_Product_Specs_letter_screen.pdf (Accessed: 27 July 2026).

Qian, X., Jiang, T.-X., and Zhao, X. (2023). SelfS2: Self-supervised transfer learning for Sentinel-2 multispectral image super-resolution. IEEE Journal of Selected Topics in Applied

Earth Observations and Remote Sensing, 16, 215–227. https://doi.org/10.1109/JSTARS.2022.3224987

Qiao, K., Zhu, W., Xie, Z., Wu, S., and Li, S. (2024). New three red-edge vegetation index for crop seasonal leaf-area-index prediction using Sentinel-2 data. International Journal of Applied Earth Observation and Geoinformation, 130, 103894. https://doi.org/10.1016/j.jag.2024.103894

Rao, P., Zhou, W., Bhattarai, N., Srivastava, A. K., Singh, B., Poonia, S., Lobell, D. B., and Jain, M. (2021). Using Sentinel-1, Sentinel-2 and Planet imagery to map crop type of smallholder farms. Remote Sensing, 13(10), 1870. https://doi.org/10.3390/rs13101870

Razzak, M. T., Mateo-García, G., Lecuyer, G., Gómez-Chova, L., Gal, Y., and Kalaitzis, F. (2023). Multi-spectral multi-image super-resolution of Sentinel-2 with radiometric consistency losses and its effect on building delineation. ISPRS Journal of Photogrammetry and Remote Sensing, 195, 1–13. https://doi.org/10.1016/j.isprsjprs.2022.10.019

Romero-Natale, A., Acevedo-Sandoval, O. A., and Sánchez-Porras, A. (2024). Ecosystem services in the Milpa system: A systematic review. One Ecosystem, 9, e131969. https://doi.org/10.3897/oneeco.9.e131969

Roy, D. P., Li, J., Zhang, H. K., Yan, L., Huang, H., and Li, Z. (2017). Examination of Sentinel-2A multi-spectral instrument (MSI) reflectance anisotropy and the suitability of a general method to normalize MSI reflectance to nadir BRDF adjusted reflectance. Remote Sensing of Environment, 199, 25–38. https://doi.org/10.1016/j.rse.2017.06.019

Sadeh, Y., Zhu, X., Dunkerley, D., Walker, J. P., Zhang, Y., Rozenstein, O., Manivasagam, V. S., and Chenu, K. (2021). Fusion of Sentinel-2 and PlanetScope time-series data into daily 3 m surface reflectance and wheat LAI monitoring. International Journal of Applied Earth Observation and Geoinformation, 96, 102260. https://doi.org/10.1016/j.jag.2020.102260

Sakuma, A., and Yamano, H. (2020). Satellite constellation reveals crop-growth patterns and improves mapping accuracy of cropping practices for subtropical small-scale fields in Japan. Remote Sensing, 12(15), 2419. https://doi.org/10.3390/rs12152419

Squeri, C., Poni, S., Di Gennaro, S. F., Matese, A., and Gatti, M. (2021). Comparison and ground truthing of different remote and proximal sensing platforms to characterise variability in a hedgerow-trained vineyard. Remote Sensing, 13(11), 2056. https://doi.org/10.3390/rs13112056

Sun, Y., Qin, Q., Ren, H., Zhang, T., and Chen, S. (2020). Red-edge band vegetation indices for leaf-area-index estimation from Sentinel-2 MSI imagery. IEEE Transactions on Geoscience and Remote Sensing, 58(2), 826–840. https://doi.org/10.1109/TGRS.2019.2940826

Tang, Z., Wang, X., Jiang, Q., Pan, H., Deng, G., Chen, H., You, Y., Li, S., and Hou, H. (2025). Parcel-scale crop planting-structure extraction combining time series of Sentinel-1 and Sentinel-2 data based on a semantic edge-aware multi-task neural network. International Journal of Digital Earth, 18, 2497487. https://doi.org/10.1080/17538947.2025.2497487

Tariq, A., Yan, J., Gagnon, A. S., Khan, M. R., and Mumtaz, F. (2023). Mapping of cropland, cropping patterns and crop types by combining optical remote-sensing images with decision tree classifier and random forest. Geo-spatial Information Science, 26(3), 302–320. https://doi.org/10.1080/10095020.2022.2100287

Tufail, R., Tassinari, P., and Torreggiani, D. (2025). Deep-learning applications for crop mapping using multi-temporal Sentinel-2 data and red-edge vegetation indices: Integrating convolutional and recurrent neural networks. Remote Sensing, 17(18), 3207. https://doi.org/10.3390/rs17183207

Useya, J., and Chen, S. (2018). Comparative performance evaluation of pixel-level and decision-level data fusion of Landsat 8 OLI, Landsat 7 ETM+ and Sentinel-2 MSI for crop ensemble classification. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 11(11), 4441–4451. https://doi.org/10.1109/JSTARS.2018.2870650

Valero, S., Arnaud, L., Planells, M., and Ceschia, E. (2021). Synergy of Sentinel-1 and Sentinel-2 imagery for early seasonal agricultural crop mapping. Remote Sensing, 13(23), 4891. https://doi.org/10.3390/rs13234891

Varma, S., and Simon, R. (2006). Bias in error estimation when using cross-validation for model selection. BMC Bioinformatics, 7, 91. https://doi.org/10.1186/1471-2105-7-91

Vasilescu, V., Datcu, M., and Faur, D. (2023). A CNN-based Sentinel-2 image super-resolution method using multiobjective training. IEEE Transactions on Geoscience and Remote Sensing, 61, 1–14. https://doi.org/10.1109/TGRS.2023.3240296

Vazeux-Blumental, N., Manicacci, D., and Tenaillon, M. (2024). The Milpa, from Mesoamerica to present days, a multicropping traditional agricultural system serving agroecology. Comptes Rendus. Biologies, 347, 159–173. https://doi.org/10.5802/crbiol.164

Wang, C., Chen, J., Cai, C., Liu, M., Wang, Z., Zhang, H., and Zhang, H. (2025). Comprehensive analysis of a parcel-level crop-mapping framework from the perspective of spatial heterogeneity and spatiotemporal transferability. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 18, 7290–7303. https://doi.org/10.1109/JSTARS.2025.3532283

Wang, M., Wang, J., Cui, Y., Liu, J., and Chen, L. (2022). Agricultural field-boundary delineation with satellite-image segmentation for high-resolution crop mapping: A case study of rice paddy. Agronomy, 12(10), 2342. https://doi.org/10.3390/agronomy12102342

Weiss, M., Jacob, F., and Duveiller, G. (2020). Remote sensing for agricultural applications: A meta-review. Remote Sensing of Environment, 236, 111402. https://doi.org/10.1016/j.rse.2019.111402

Wolpert, D. H. (1992). Stacked generalization. Neural Networks, 5(2), 241–259. https://doi.org/10.1016/S0893-6080(05)80023-1

Woźniak, E., Rybicki, M., Kofman, W., Aleksandrowicz, S., Wojtkowski, C., Lewiński, S., Bojanowski, J., Musiał, J., Milewski, T., Ślesiński, P., and Łączyński, A. (2022). Multi-temporal phenological indices derived from time-series Sentinel-1 images for country-wide

crop classification. International Journal of Applied Earth Observation and Geoinformation, 107, 102683. https://doi.org/10.1016/j.jag.2022.102683

Wu, B., Zhang, M., Zeng, H., Tian, F., Potgieter, A. B., Qin, X., Yan, N., Chang, S., Zhao, Y., Dong, Q., Boken, V., Plotnikov, D., Guo, H., Wu, F., Zhao, H., Deronde, B., Tits, L., and Loupian, E. (2023). Challenges and opportunities in remote sensing-based crop monitoring: A review. National Science Review, 10(4), nwac290. https://doi.org/10.1093/nsr/nwac290

Xie, Q., Dash, J., Huang, W., Peng, D., Qin, Q., Mortimer, H., Casa, R., Pignatti, S., Laneve, G., Pascucci, S., Dong, Y., and Ye, H. (2018). Vegetation indices combining red and red-edge spectral information for leaf-area-index retrieval. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 11(5), 1482–1493. https://doi.org/10.1109/JSTARS.2018.2813281

Xu, J., Yang, J., Xiong, X., Li, H., Huang, J., Ting, K. C., Ying, Y., and Lin, T. (2021). Towards interpreting multi-temporal deep-learning models in crop mapping. Remote Sensing of Environment, 264, 112599. https://doi.org/10.1016/j.rse.2021.112599

Xue, H., Xu, X., Zhu, Q., Yang, G., Long, H., Li, H., Yang, X., Zhang, J., Yang, Y., Xu, S., Yang, M., and Li, Y. (2023). Object-oriented crop classification using time-series Sentinel images from Google Earth Engine. Remote Sensing, 15(5), 1353. https://doi.org/10.3390/rs15051353

Yan, L., Roy, D. P., Li, Z., Zhang, H. K., and Huang, H. (2018). Sentinel-2A multi-temporal misregistration characterization and an orbit-based sub-pixel registration methodology. Remote Sensing of Environment, 215, 495–506. https://doi.org/10.1016/j.rse.2018.04.021

Yao, J., Wu, J., Xiao, C., Zhang, Z., and Li, J. (2022). The classification-method study of crop remote sensing with deep learning, machine learning and Google Earth Engine. Remote Sensing, 14(12), 2758. https://doi.org/10.3390/rs14122758

Zhang, K., Yuan, D., Yang, H., Zhao, J., and Li, N. (2023). Synergy of Sentinel-1 and Sentinel-2 imagery for crop classification based on DC-CNN. Remote Sensing, 15(11), 2727. https://doi.org/10.3390/rs15112727

Zhao, H., Chang, C., Wang, Z., and Zhao, G. (2025). A large-scale agricultural-land classification method based on synergistic integration of time-series red-edge vegetation index and phenological features. Sensors, 25(2), 503. https://doi.org/10.3390/s25020503

Zhong, L., Hu, L., and Zhou, H. (2019). Deep-learning-based multi-temporal crop classification. Remote Sensing of Environment, 221, 430–443. https://doi.org/10.1016/j.rse.2018.11.032

Appendices

Appendix A. Methodological specifications

This appendix provides the compact specifications referenced from Sections 4.4-4.8. It contains methodological definitions and fixed analytical settings only; no performance results are reported.

Appendix Table A.1. Sentinel-2 spectral indices used in predictor construction

Index	Exact implemented formula	Bands	Denominator safeguard	Methodological representation	Workflow use
NDVI	(B08 - B04) / (B08 + B04)	B08, B04	Finite numerator and denominator; |denominator| > 1 x 10^-6	Greenness contrast	Parcel and pixel workflows
EVI2	2.5(B08 - B04) / (B08 + 2.4B04 + 1)	B08, B04	Finite numerator and denominator; |denominator| > 1 x 10^-6	Greenness with reduced soil-background influence	Parcel and pixel workflows
NDRE (B06/B05)	(B06 - B05) / (B06 + B05)	B06, B05	Finite numerator and denominator; |denominator| > 1 x 10^-6	Red-edge contrast	Parcel and pixel workflows
NDRE (B07/B05)	(B07 - B05) / (B07 + B05)	B07, B05	Finite numerator and denominator; |denominator| > 1 x 10^-6	Alternative red-edge contrast	Pixel diagnostic only
CI red-edge (B08/B07)	B08 / B07 - 1	B08, B07	Finite bands; |B07| > 1 x 10^-6	Relative red-edge response	Parcel workflow
CI red-edge (B06/B05)	B06 / B05 - 1	B06, B05	Finite bands; |B05| > 1 x 10^-6	Alternative relative red-edge response	Pixel diagnostic only
CI red-edge (B07/B05)	B07 / B05 - 1	B07, B05	Finite bands; |B05| > 1 x 10^-6	Alternative relative red-edge response	Pixel diagnostic only
NDMI	(B08 - B11) / (B08 + B11)	B08, B11	Finite numerator and denominator; |denominator| > 1 x 10^-6	NIR-SWIR moisture contrast	Parcel and pixel workflows
PSRI	(B04 - B02) / B06	B04, B02, B06	Finite numerator and denominator; |B06| > 1 x 10^-6	Pigment/senescence contrast	Parcel and pixel workflows

Note: reflectance inputs were scaled to surface-reflectance units before index calculation. Invalid divisions returned missing values rather than infinite values.

Appendix Table A.2. Final and reserve Sentinel-2 predictor ranking

Rank	Exact machine-readable predictor	Panel status	Feature family	Reader-readable construction	Property represented

Rank	Exact machine-readable predictor	Panel status	Feature family	Reader-readable construction	Property represented
1	B11_refl_oct_nov_transition_p10	Final Top-25	Raw-band statistics	October-November, P10	SWIR moisture / late transition
2	NDRE_B06_B05_sep_oct_peak_late_iqr	Final Top-25	Index statistics	September-October, IQR	Red-edge peak variability
3	NDMI_B08_B11_jan_mar_early_context_p10	Final Top-25	Index statistics	January-March, P10	Early/background moisture context
4	B02_refl_full_year_cv	Final Top-25	Raw-band statistics	Full year, coefficient of variation	Visible/background variability
5	B11_refl_nov_dec_late_p10	Final Top-25	Raw-band statistics	November-December, P10	Late SWIR dry-down/residue
6	NDMI_B08_B11_full_year_p10	Final Top-25	Index statistics	Full year, P10	Full-year moisture low-tail
7	CIrededge_B08_B07_jan_mar_early_context_median	Final Top-25	Index statistics	January-March, median	Red-edge contextual condition
8	NDMI_B08_B11_phenology_season_integral	Final Top-25	Phenology	Full 36-date sequence, season integral	Moisture phenology integral
9	NDRE_B06_B05_sep_oct_peak_late_p90_p10_range	Final Top-25	Index statistics	September-October, P90-P10 range	Peak red-edge range
10	PSRI_B04_B02_B06_phenology_minimum_doy	Final Top-25	Phenology	Full 36-date sequence, minimum day of year	Senescence/stress timing
11	B11_withinparcel_oct_nov_transition_p10	Final Top-25	Within-parcel heterogeneity	October-November, P10	Within-parcel SWIR heterogeneity
12	B04_refl_sep_oct_peak_late_cv	Final Top-25	Raw-band statistics	September-October, coefficient of variation	Peak red/visible variability
13	NDMI_B08_B11_stability_full_year_signal_stability_absmean	Final Top-25	Stability	Full year, |mean|/SD stability	Moisture temporal stability
14	B12_refl_sep_oct_peak_late_p10	Final Top-25	Raw-band statistics	September-October, P10	SWIR2 peak moisture
15	CIrededge_B08_B07_full_year_p90	Final Top-25	Index statistics	Full year, P90	Full-year red-edge high response
16	B03_refl_full_year_cv	Final Top-25	Raw-band statistics	Full year, coefficient of variation	Green-band background variability
17	B8A_texture_oct_nov_transition_localstd5x5_p90	Final Top-25	Observed texture	October-November, P90 local SD, 5 x 5	Observed spatial texture
18	NDRE_B06_B05_stability_full_year_iqr	Final Top-25	Stability	Full year, IQR	Red-edge temporal stability
19	NDVI_B08_B04_phenology_minimum_doy	Final Top-25	Phenology	Full 36-date sequence, minimum day of year	Greenness phenology timing

Rank	Exact machine-readable predictor	Panel status	Feature family	Reader-readable construction	Property represented
20	NDRE_B06_B05_phenology_duration_above_70pct_amplitude	Final Top-25	Phenology	Full 36-date sequence, duration above 70% amplitude	Red-edge phenology duration
21	PSRI_B04_B02_B06_oct_nov_transition_p10	Final Top-25	Index statistics	October-November, P10	Stress/senescence transition
22	NDVI_B08_B04_sep_oct_peak_late_p90_p10_range	Final Top-25	Index statistics	September-October, P90-P10 range	Peak greenness range
23	B08_refl_sep_oct_peak_late_iqr	Final Top-25	Raw-band statistics	September-October, IQR	NIR canopy structure
24	B06_withinparcel_oct_nov_transition_p10	Final Top-25	Within-parcel heterogeneity	October-November, P10	Red-edge within-parcel heterogeneity
25	B05_refl_nov_dec_late_p90_p10_range	Final Top-25	Raw-band statistics	November-December, P90-P10 range	Late red-edge range
26	B11_refl_nov_dec_late_iqr	Reserve only	Raw-band statistics	November-December, IQR	Late SWIR spread
27	NDMI_B08_B11_sep_oct_peak_late_p90	Reserve only	Index statistics	September-October, P90	Peak moisture high-tail
28	CIrededge_B08_B07_full_year_median	Reserve only	Index statistics	Full year, median	Full-year red-edge median
29	B04_refl_nov_dec_late_cv	Reserve only	Raw-band statistics	November-December, coefficient of variation	Late visible variability
30	EVI2_B08_B04_phenology_duration_above_70pct_amplitude	Reserve only	Phenology	Full 36-date sequence, duration above 70% amplitude	EVI2 canopy phenology duration
31	PSRI_B04_B02_B06_phenology_peak_to_late_decline	Reserve only	Phenology	Full 36-date sequence, peak-to-late decline	Senescence decline
32	NDMI_withinparcel_jun_aug_pre_peak_cv_absmean	Reserve only	Within-parcel heterogeneity	June-August, SD/|mean|	Pre-peak moisture heterogeneity
33	B05_refl_jun_aug_pre_peak_iqr	Reserve only	Raw-band statistics	June-August, IQR	Pre-peak red-edge spread
34	B08_refl_full_year_p90	Reserve only	Raw-band statistics	Full year, P90	NIR high canopy response
35	CIrededge_B08_B07_full_year_p10	Reserve only	Index statistics	Full year, P10	Red-edge low-tail
36	PSRI_B04_B02_B06_full_year_std	Reserve only	Index statistics	Full year, SD	Full-year stress variability
37	B12_refl_oct_nov_transition_p10	Reserve only	Raw-band statistics	October-November, P10	Transition SWIR2 moisture
38	NDRE_withinparcel_oct_nov_transition_p90	Reserve only	Within-parcel heterogeneity	October-November, P90	Red-edge heterogeneity high-tail

Rank	Exact machine-readable predictor	Panel status	Feature family	Reader-readable construction	Property represented
39	B11_texture_oct_nov_transition_localstd5x5_mean	Reserve only	Observed texture	October-November, mean local SD, 5 x 5	SWIR texture
40	B05_refl_oct_nov_transition_p90_p10_range	Reserve only	Raw-band statistics	October-November, P90-P10 range	Transition red-edge range

Note: ranks 1-25 formed the fitted primary predictor panel. Ranks 26-40 were retained only as reserve predictors and were not included in the primary fitted models.

Appendix Table A.3. Construction of the Sentinel-2 pixel-level diagnostic predictor set

Component	Exact definition	Count or control
Raw-band temporal variables	B02, B03, B04, B05, B06, B07, B08, B8A, B11 and B12 surface reflectance	10 variables
Spectral-index temporal variables	NDVI B08/B04; EVI2 B08/B04; NDRE B06/B05; NDRE B07/B05; CI-red-edge B06/B05; CI-red-edge B07/B05; NDMI B08/B11; PSRI B04/B02/B06	8 variables
Temporal windows	Full year; January-March; June-August; September-October; October-November; November-December	6 windows
Per-window statistics	Mean; median; standard deviation; IQR; P10; P90; P90-P10; SD/|mean|; valid count; valid fraction; peak value; peak day of year; trough value; trough day of year; amplitude; trapezoidal integral by day of year	16 statistics per variable and window
Seasonal contrasts	September-October mean minus January-March mean; September-October mean minus November-December mean	2 contrasts per variable
Initial feature library	18 x [(6 x 16) + 2]	1,764 predictors
Development-only filtering	Remove predictors that were entirely missing or constant in the AOI3+AOI4 development matrix	216 removed; 1,548 retained
Imputation	Estimate medians from the development matrix and apply them unchanged to development and benchmark records	No benchmark-derived imputation
Analytical matrices	Development pixels were capped at 20 per parcel; benchmark pixels were retained without capping	811 x 1,548 development; 1,902 x 1,548 benchmark
Parcel aggregation	Average predicted class probabilities across all pixels belonging to each parcel and assign the class with the greatest mean probability	Diagnostic return to parcel scale

Appendix Table A.4. Consolidated model, training and integration settings

Analytical branch	Cohort / input	Fixed model or operation	Development / training control	Evaluation domain	Analytical role
Sentinel-2 parcel Random Forest	149 parcels; final Top-25; three-class and binary targets	800 trees; Gini; unrestricted depth; min split 2; min leaf 2; sqrt feature sampling; bootstrap; balanced class weights; seed 42; parallel execution	Repeated stratified 3-fold CV x 20 within 71 AOI3+AOI4 parcels; final fit on all 71 development parcels	One evaluation on 78 AOI1+AOI2 parcels	Primary
Better-supported sensitivity	132 parcels: all 71 development parcels plus 61 benchmark parcels; same Top-25	Same 800-tree Random Forest and target formulations as the primary parcel analysis	No feature or parameter revision; quality variables used only to define the subset	61 better-supported benchmark parcels	Sensitivity
Sentinel-2 SVM	149 parcels; final Top-25; StandardScaler + probabilistic SVC; balanced class weights; seed 42	Training grid: linear C = 0.1, 1, 10; RBF C = 0.1, 1, 10 and gamma = scale, 0.01, 0.1. Selected: three-class linear C = 0.1; binary RBF C = 1, gamma = scale	Training-only stratified 3-fold grid search on 71 development parcels; macro-F1 selection	One evaluation on 78 benchmark parcels	Model-family diagnostic
Sentinel-2 pixel Random Forest	3,367 pixels from 149 parcels; 811 capped development pixels; 1,902 benchmark pixels; 1,548 retained predictors	500 trees; Gini; unrestricted depth; min split 2; min leaf 2; sqrt feature sampling; bootstrap; balanced-subsample weights; seed 20260702	Geographic parcel split applied before preparation; development capped at 20 pixels per parcel	Pixel-level and mean-probability parcel-level assessment	Spatial-representation diagnostic
SkySat predictor-level Random Forest	149 parcels; S2 Top-25, SkySat 10, and combined 35 predictors; binary target	1,000 trees; Gini; unrestricted depth; min split 2; min leaf 2; sqrt feature sampling; bootstrap; balanced-subsample weights; seed 42	Directional AOI3->AOI4 and AOI4->AOI3 geographic out-of-fold assessment; final fit on 71 development parcels	AOI1, AOI2 and pooled 78-parcel benchmark	Predictor-level comparison
PlanetScope predictor-level Random Forest	149 parcels; fixed representations of 25, 20, 45, 23 and 48 predictors; binary target	Same fixed 1,000-tree Random Forest as the SkySat comparison; seed 42	Directional AOI3->AOI4 and AOI4->AOI3 geographic out-of-fold assessment; final fit on 71 development parcels	AOI1, AOI2 and pooled 78-parcel benchmark	Predictor-level comparison
EDSR-lite reconstruction	Observed B02, B03, B04 and B08; 16 x 16 at 10 m to 32 x 32 at 5 m; 515 patches from 112 parcels	4 input/output channels; 32 features; 4 residual blocks; PixelShuffle 2x; 122,564 parameters; masked Charbonnier epsilon = 0.001; AdamW LR = 0.001, weight decay = 0.00001; batch 4; gradient clip 1.0; seed 42	222/55 training/validation patches from 47/12 parcels; maximum 100 epochs; early stopping patience 20; ReduceLROnPlateau factor 0.5, patience 8; train-only normalisation	238 held-out patches from 53 AOI1+AOI2 parcels; bicubic comparator on identical valid support	Image-level experiment
Super-resolution downstream Random Forest	Native 10 m, bicubic 5 m and CNN 5 m; 50 predictors per representation; 59 development parcels	300 trees; sqrt feature sampling; min leaf 2; balanced class weights; seed 42; remaining RF settings at scikit-learn defaults	Same classifier for all representations; development-median imputation; frozen reconstruction network	18 mixed-class AOI1 parcels and 35 AOI2 maize parcels	Downstream diagnostic
Pairwise probability fusion	Matched native probabilities for 71 geographic OOF development parcels and 78 benchmark parcels	S2+SkySat: 0.5/0.5; S2+PlanetScope: 0.5/0.5; Milpa when fused probability > 0.50, maize monocrop when <= 0.50	No model fitting, retraining, predictor concatenation, calibration, weight optimisation or threshold tuning	Directional development, pooled development, AOI1, AOI2 and pooled benchmark	Classification-level integration

Analytical branch	Cohort / input	Fixed model or operation	Development / training control	Evaluation domain	Analytical role
Three-sensor probability fusion	Matched S2, SkySat and PlanetScope spectral-only native probabilities for all 149 parcels	Equal one-third weights; same fixed > 0.50 Milpa decision rule	No retraining, concatenation, calibration, weight optimisation, threshold tuning or benchmark-guided revision	Directional development and frozen benchmark domains	Three-sensor diagnostic
Paired bootstrap uncertainty	Fused-minus-component metric differences using the same sampled parcels for every compared model	10,000 replicates; seed 20260714; two-sided 95% percentile intervals (2.5th and 97.5th percentiles)	Directional analyses stratified by class; pooled development and benchmark analyses stratified by AOI x class	Accuracy, balanced accuracy, macro-F1 and Milpa precision, recall and F1 differences	Uncertainty assessment

Note: AOI1+AOI2 are described as a frozen secondary, non-confirmatory benchmark. Predictor-level, image-level and classification-level integration remain separate analytical branches.

Appendix 5A

Supporting evidence for reference-cohort composition and Sentinel-2 data quality

This appendix provides the detailed numerical evidence supporting Section 5.1. The main Results chapter retains only the principal domain-level contrasts needed to explain the final cohort, the distribution of observed Sentinel-2 support, the reliability of temporal interpolation and the final quality-control decision. The tables below preserve the complete parcel-retention, AOI, support, interpolation, seasonal-reliability and residual-quality-control results.

The evidence refers to three distinct populations. The primary analytical cohort contained 149 target parcels. Annual interpolation validation was conducted across the broader cohort of 226 Sentinel-2-eligible parcels. A separate 132-parcel cohort was retained only for the later sensitivity analysis.

A5.1 Parcel retention before the principal analyses

The original reference population was reduced through the successive geometry, Sentinel-2 eligibility and target-class restrictions reported in the Methodology. Table A5.1 records the number retained at each stage and distinguishes stage-specific retention from retention relative to the original population.

Appendix Table A5.1. Retention of reference parcels across the principal analytical-cohort stages.

Cohort stage	Parcels retained	Parcels removed at stage	Retention from preceding stage (%)	Retention from original population (%)
Original reference population	457	—	100.00	100.00
Valid 10 m inward-buffered geometry	270	187	59.08	59.08
Sentinel-2 eligible	226	44	83.70	49.45
Final target-class cohort	149	77	65.93	32.60

A5.2 Complete composition of the final analytical cohort

The final 149 parcels were divided into a 71-parcel development domain comprising AOI3 and AOI4 and a separate 78-parcel benchmark domain comprising AOI1 and AOI2. Landrace and hybrid maize remained distinct in the three-class formulation and were combined as maize monocrop in the binary formulation.

Appendix Table A5.2. Complete composition of the final 149-parcel analytical cohort by evaluation domain, area of interest and crop class.

Domain / AOI	Total parcels	Milpa mixed-cropping	Landrace maize	Hybrid maize	Maize monocrop
Development	71	29	32	10	42
Benchmark	78	14	41	23	64
AOI1	20	9	11	0	11
AOI2	58	5	30	23	53
AOI3	24	3	19	2	21
AOI4	47	26	13	8	21
Final analytical cohort	149	43	73	33	106

A5.3 Full-year observed Sentinel-2 support

The study-specific strong, moderate and weak classes summarised both the proportion of the 69 acquisition dates with usable observed parcel support and the longest consecutive period without such support. The first two rows distinguish the complete eligible cohort from the final target-class cohort. The remaining rows show the domain- and AOI-specific distribution within the final 149 parcels.

Appendix Table A5.3. Full-year observed Sentinel-2 support across the eligible and final analytical cohorts.

Cohort / domain / AOI	Parcels	Strong	Moderate	Weak	Mean supported acquisition dates (%)	Mean longest unsupported run (acquisitions)	Maximum consecutive unsupported acquisitions
Sentinel-2-eligible cohort	226	139	57	30	—	—	—
Final analytical cohort	149	93	39	17	—	—	—
Development	71	58	13	0	72.05	4.18	7
Benchmark	78	35	26	17	63.47	7.28	21
AOI1	20	9	10	1	65.58	6.25	18
AOI2	58	26	16	16	62.75	7.64	21

Cohort / domain / AOI	Parcels	Strong	Moderate	Weak	Mean supported acquisition dates (%)	Mean longest unsupported run (acquisitions)	Maximum consecutive unsupported acquisitions
AOI3	24	13	11	0	71.26	4.54	7
AOI4	47	45	2	0	72.46	4.00	7

A5.4 Overall and AOI-specific interpolation performance

Withheld-observation validation was conducted across the 226 Sentinel-2-eligible parcels. The complete validation contained 9,989 NDVI records and showed low overall error, minimal signed bias and strong correspondence between observed and interpolated values. AOI-specific saved outputs reported MAE, RMSE and (R^2), allowing geographical differences in reconstruction error to be examined separately from the observed-support classes.

Appendix Table A5.4. Overall and AOI-specific withheld-observation NDVI interpolation performance.

Validation scope	Validation records	MAE	RMSE	Signed bias	Pearson (r)	(R^2)
All eligible parcels	9,989	0.02	0.03	−0.00	0.97	0.95
AOI1	—	0.02	0.03	—	—	0.95
AOI2	—	0.02	0.03	—	—	0.96
AOI3	—	0.02	0.03	—	—	0.96
AOI4	—	0.02	0.04	—	—	0.93

The dashes indicate that the corresponding AOI-specific values were not retained in the accepted evidence rather than values of zero.

A5.5 Interpolation performance by season and gap duration

The saved validation evidence was available only for combined season-by-gap groups. Consequently, the rows in Table A5.5 preserve the original joint strata and are not pooled into separate season-only or gap-only estimates. The number of validation records is retained because the outside-season long- and very-long-gap groups were substantially smaller than the other strata.

Appendix Table A5.5. Withheld-observation NDVI interpolation error by season and gap duration.

Period	Gap category	Interval between bracketing observations	Validation records	MAE
August–October	Short	≤20 days	493	0.04
August–October	Medium	21–40 days	615	0.05
August–October	Long	41–60 days	147	0.06

Period	Gap category	Interval between bracketing observations	Validation records	MAE
August–October	Very long	>60 days	67	0.10
Outside August–October	Short	≤20 days	7,958	0.01
Outside August–October	Medium	21–40 days	656	0.03
Outside August–October	Long	41–60 days	35	0.05
Outside August–October	Very long	>60 days	18	0.05

A5.6 August–October observed-support outcomes

The August–October assessment described observed data availability within the final 149-parcel cohort rather than interpolation accuracy. Mean observed support during this period was 36.68%. The four parcels classified as not reliable for August–October had no observed support during the period, occurred in the AOI2 benchmark and experienced a maximum gap of 91 days.

Appendix Table A5.6. August–October observed-support outcomes within the final analytical cohort.

August–October outcome	Parcels / value
Final analytical cohort	149 parcels
Mean observed support	36.68%
Caution	79 parcels
Weak	55 parcels
Not reliable for fine phenological assessment	11 parcels
Not reliable for August–October	4 parcels
Parcels with no observed support	4 parcels
Location of all zero-support parcels	AOI2 benchmark
Maximum observed-support gap	91 days

A5.7 Residual quality-control outcomes

Soft-warning events were retained as data-quality cautions and were not treated as confirmed cloud contamination or grounds for automatic parcel exclusion. For the eligible cohort, a valid-record denominator was available and therefore supported calculation of the warning-event rate. No equivalent denominator was saved after restriction to the final 149 parcels, so no final-cohort percentage is reported. No confirmed hard-artifact event was identified in either cohort.

Appendix Table A5.7. Residual Sentinel-2 quality-control outcomes for the eligible and final analytical cohorts.

Quality-control measure	Eligible 226 parcels	Final 149 parcels
Parcel-date records assessed	15,594	—
RGB–NIR-valid parcel-date records	10,441	—
Parcel-date records with at least one soft warning	3,264	2,211
Warning-event rate among valid records	31.26%	Not calculated
Parcels with at least one warning event	166	110
Total individual soft-warning activations	4,217	2,838
Low-pixel warning activations	1,566	1,311
Bright–low-NDVI warning activations	708	416
Blue/haze-like warning activations	1,943	1,111
Dark-shadow-like hard flags	0	0
Physically implausible-index flags	0	0
Isolated NDVI-drop flags	0	0
Isolated NDVI-spike flags	0	0
Confirmed hard-artifact events	0	0
Parcels with a confirmed hard-artifact flag	0	0

A5.8 Composition and role of the better-supported sensitivity cohort

The stricter cohort was retained only to test whether the later classification findings changed after excluding parcels with weaker support or seasonal reliability. It did not replace the primary 149-parcel cohort. All development parcels remained, while the benchmark was reduced by 17 parcels.

Appendix Table A5.8. Composition of the better-supported sensitivity cohort.

Domain / AOI	Primary cohort	Better-supported cohort	Parcels excluded
Development	71	71	0
Benchmark	78	61	17
AOI1	20	19	1
AOI2	58	42	16
AOI3	24	24	0
AOI4	47	47	0
Total	149	132	17

The primary cohort remained the basis for the main classification analyses, while the 132-parcel cohort supported the sensitivity comparison reported in Section 5.2.2.

Appendix 5.2: Supporting evidence for Sentinel-2 baseline classification

Appendix 5.2-A

Primary parcel level Random Forest class specific performance

Table 5.2-A1 reports the complete class-specific results for the 78 parcel secondary benchmark. Landrace maize and hybrid maize are reported separately in the three-class formulation and combined as maize monocrop in the binary formulation.

Table 5.2-A1. Class-specific benchmark performance of the primary Sentinel-2 parcel-level Random Forest.

Formulation	Class	Precision	Recall	F1	Support
Three class	Milpa mixed cropping system	0.36	0.36	0.36	14
Three class	Landrace maize	0.54	0.71	0.61	41
Three class	Hybrid maize	0.30	0.13	0.18	23
Binary	Milpa mixed cropping system	0.38	0.36	0.37	14
Binary	Maize monocrop	0.86	0.88	0.87	64

Appendix 5.2-B

Primary parcel level Random Forest confusion matrices

The following matrices report observed classes as rows and predicted classes as columns for the complete 78 parcel secondary benchmark.

Table 5.2-B1. Three-class benchmark confusion matrix for the primary Sentinel-2 parcel-level Random Forest.

Observed class	Predicted Milpa	Predicted landrace maize	Predicted hybrid maize
Milpa mixed cropping system	5	6	3
Landrace maize	8	29	4
Hybrid maize	1	19	3

Table 5.2-B2. Binary benchmark confusion matrix for the primary Sentinel-2 parcel-level Random Forest.

Observed class	Predicted Milpa	Predicted maize monocrop
Milpa mixed cropping system	5	9
Maize monocrop	8	56

Appendix 5.2-C

AOI specific benchmark performance

AOI1 and AOI2 are reported separately to show geographic variation within the secondary benchmark. AOI1 contained no hybrid maize reference parcels and therefore did not provide a complete AOI-specific assessment of the three-class formulation.

Table 5.2-C1. AOI-specific benchmark performance of the primary Sentinel-2 parcel-level Random Forest.

Formulation	AOI	Parcels	Accuracy	Balanced accuracy	Macro F1	Weighted F1	Milpa precision	Milpa recall	Milpa F1	Milpa support
Three class	AOI1	20	0.50	0.46	0.30	0.47	1.00	0.11	0.20	9
Three class	AOI2	58	0.47	0.53	0.41	0.42	0.31	0.80	0.44	5
Binary	AOI1	20	0.65	0.61	0.56	0.58	1.00	0.22	0.36	9
Binary	AOI2	58	0.83	0.72	0.64	0.85	0.27	0.60	0.38	5

Appendix 5.2-D

Development domain fold variability

The development assessment comprised 60 folds for each formulation. Table 5.2-D1 reports the accepted aggregate distributions; development fold summaries did not report class specific Milpa recall or Milpa F1.

Table 5.2-D1. Distribution of development fold performance for the primary Sentinel-2 parcel-level Random Forest.

Formulation	Metric	Mean	Standard deviation	Minimum	Maximum
Three class	Accuracy	0.68	0.08	0.50	0.83
Three class	Balanced accuracy	0.62	0.10	0.40	0.86
Three class	Macro F1	0.62	0.11	0.36	0.82
Three class	Weighted F1	0.67	0.08	0.45	0.83
Binary	Accuracy	0.79	0.07	0.57	0.92
Binary	Balanced accuracy	0.78	0.08	0.58	0.91
Binary	Macro F1	0.78	0.08	0.56	0.91
Binary	Weighted F1	0.79	0.08	0.57	0.91

Appendix 5.2-E

Better supported benchmark sensitivity

The better supported sensitivity analysis retained all 71 development parcels but reduced the secondary benchmark from 78 to 61 parcels. The complete benchmark class metrics and confusion matrices are reported in Appendices 5.2-A and 5.2-B and are not repeated here. Restricting the benchmark did not improve the recorded classification metrics.

Table 5.2-E1. Complete and better-supported benchmark performance of the primary Sentinel-2 parcel-level Random Forest.

Formulation	Benchmark cohort	Parcels	Accuracy	Balanced accuracy	Macro F1	Weighted F1	Milpa recall	Milpa F1
Three class	Complete benchmark	78	0.47	0.40	0.38	0.44	0.36	0.36
Three class	Better supported benchmark	61	0.44	0.37	0.34	0.39	0.27	0.29
Binary	Complete benchmark	78	0.78	0.62	0.62	0.78	0.36	0.37
Binary	Better supported benchmark	61	0.75	0.53	0.53	0.74	0.18	0.21

Table 5.2-E2. Class-specific performance for the better-supported benchmark.

Formulation	Class	Precision	Recall	F1	Support
Three class	Milpa mixed cropping system	0.30	0.27	0.29	11
Three class	Landrace maize	0.49	0.73	0.59	30
Three class	Hybrid maize	0.33	0.10	0.15	20
Binary	Milpa mixed cropping system	0.25	0.18	0.21	11
Binary	Maize monocrop	0.83	0.88	0.85	50

The following sensitivity matrices report observed classes as rows and predicted classes as columns.

Table 5.2-E3. Three-class confusion matrix for the 61 parcel better-supported benchmark.

Observed class	Predicted Milpa	Predicted landrace maize	Predicted hybrid maize
Milpa mixed cropping system	3	6	2
Landrace maize	6	22	2
Hybrid maize	1	17	2

Table 5.2-E4. Binary confusion matrix for the 61 parcel better-supported benchmark.

Observed class	Predicted Milpa	Predicted maize monocrop
Milpa mixed cropping system	2	9
Maize monocrop	6	44

Appendix 5.2-F

Support Vector Machine diagnostic

The Support Vector Machine was evaluated on the same 78 parcel secondary benchmark as the primary Random Forest. Its three class and binary results did not improve the principal benchmark measures.

Table 5.2-F1. Overall secondary benchmark performance of the Sentinel-2 Support Vector Machine.

Formulation	Parcels	Accuracy	Balanced accuracy	Macro F1	Weighted F1
Three class	78	0.40	0.37	0.36	0.41
Binary	78	0.71	0.46	0.45	0.69

Table 5.2-F2. Class-specific secondary benchmark performance of the Sentinel-2 Support Vector Machine.

Formulation	Class	Precision	Recall	F1	Support
Three class	Milpa mixed cropping system	0.25	0.29	0.27	14
Three class	Landrace maize	0.56	0.46	0.51	41
Three class	Hybrid maize	0.29	0.35	0.31	23
Binary	Milpa mixed cropping system	0.09	0.07	0.08	14
Binary	Maize monocrop	0.81	0.84	0.82	64

The following matrices report observed classes as rows and predicted classes as columns.

Table 5.2-F3. Three-class secondary benchmark confusion matrix for the Sentinel-2 Support Vector Machine.

Observed class	Predicted Milpa	Predicted landrace maize	Predicted hybrid maize
Milpa mixed cropping system	4	4	6
Landrace maize	8	19	14
Hybrid maize	4	11	8

Table 5.2-F4. Binary secondary benchmark confusion matrix for the Sentinel-2 Support Vector Machine.

Observed class	Predicted Milpa	Predicted maize monocrop
Milpa mixed cropping system	1	13
Maize monocrop	10	54

Appendix 5.2-G

Pixel level and parcel aggregated diagnostic

Pixel results and parcel aggregated results retain separate evaluation units. The pixel benchmark contained 1,902 records, while mean probability aggregation produced one result for each of the 78 benchmark parcels. The primary parcel Random Forest comparison remains in Table 5.4 and is not duplicated here.

Table 5.2-G1. Overall pixel-level and mean probability parcel-aggregated benchmark performance.

Formulation	Evaluation level	Records	Accuracy	Balanced accuracy	Macro precision	Macro recall	Macro F1	Weighted F1
Three class	Benchmark pixels	1,902	0.50	0.36	0.41	0.36	0.29	0.40
Three class	Parcel aggregation	78	0.53	0.37	0.46	0.37	0.32	0.42
Binary	Benchmark pixels	1,902	0.87	0.50	0.50	0.50	0.50	0.88
Binary	Parcel aggregation	78	0.79	0.54	0.58	0.54	0.54	0.76

Table 5.2-G2. Class-specific pixel-level and mean probability parcel-aggregated benchmark performance.

Formulation	Evaluation level	Class	Precision	Recall	F1	Support
Three class	Benchmark pixels	Milpa mixed cropping system	0.06	0.10	0.08	115
Three class	Benchmark pixels	Landrace maize	0.55	0.91	0.68	993
Three class	Benchmark pixels	Hybrid maize	0.63	0.05	0.09	794
Three class	Parcel aggregation	Milpa mixed cropping system	0.33	0.14	0.20	14
Three class	Parcel aggregation	Landrace maize	0.54	0.93	0.68	41
Three class	Parcel aggregation	Hybrid maize	0.50	0.04	0.08	23
Binary	Benchmark pixels	Milpa mixed cropping system	0.06	0.08	0.07	115
Binary	Benchmark pixels	Maize monocrop	0.94	0.92	0.93	1,787
Binary	Parcel aggregation	Milpa mixed cropping system	0.33	0.14	0.20	14
Binary	Parcel aggregation	Maize monocrop	0.83	0.94	0.88	64

The following matrices report observed classes as rows and predicted classes as columns. Pixel and parcel counts are presented separately.

Table 5.2-G3. Three-class confusion matrix for the 1,902 benchmark pixels.

Observed class	Predicted Milpa	Predicted landrace maize	Predicted hybrid maize
Milpa mixed cropping system	12	85	18
Landrace maize	82	905	6
Hybrid maize	91	663	40

Table 5.2-G4. Three-class confusion matrix after mean probability aggregation to 78 benchmark parcels.

Observed class	Predicted Milpa	Predicted landrace maize	Predicted hybrid maize
Milpa mixed cropping system	2	12	0
Landrace maize	2	38	1
Hybrid maize	2	20	1

Table 5.2-G5. Binary confusion matrix for the 1,902 benchmark pixels.

Observed class	Predicted Milpa	Predicted maize monocrop
Milpa mixed cropping system	9	106
Maize monocrop	147	1,640

Table 5.2-G6. Binary confusion matrix after mean probability aggregation to 78 benchmark parcels.

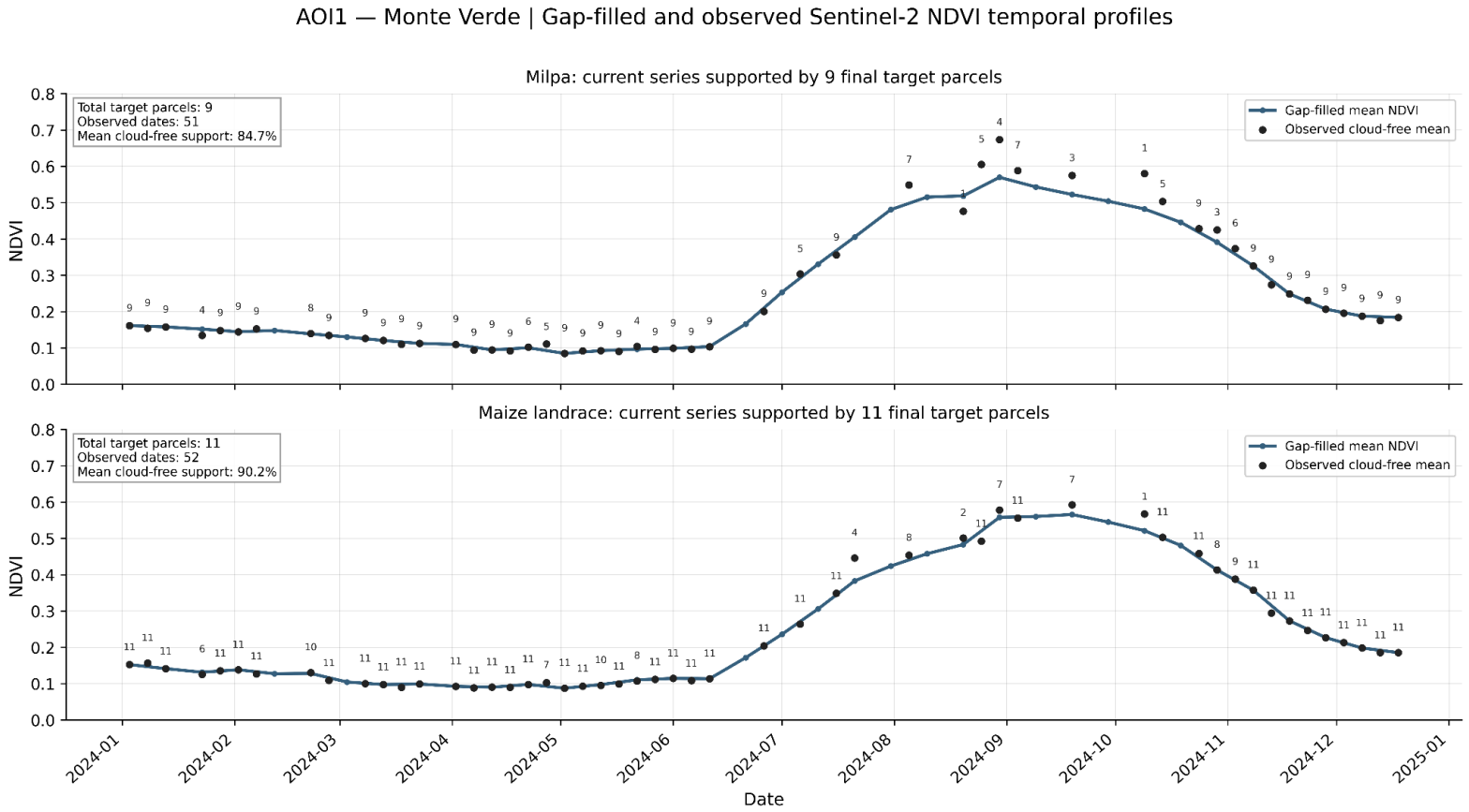
Observed class	Predicted Milpa	Predicted maize monocrop
Milpa mixed cropping system	2	12
Maize monocrop	4	60

Parcel aggregation improved several pixel diagnostic measures but remained below the primary parcel Random Forest on the principal balanced and Milpa sensitive benchmark measures.

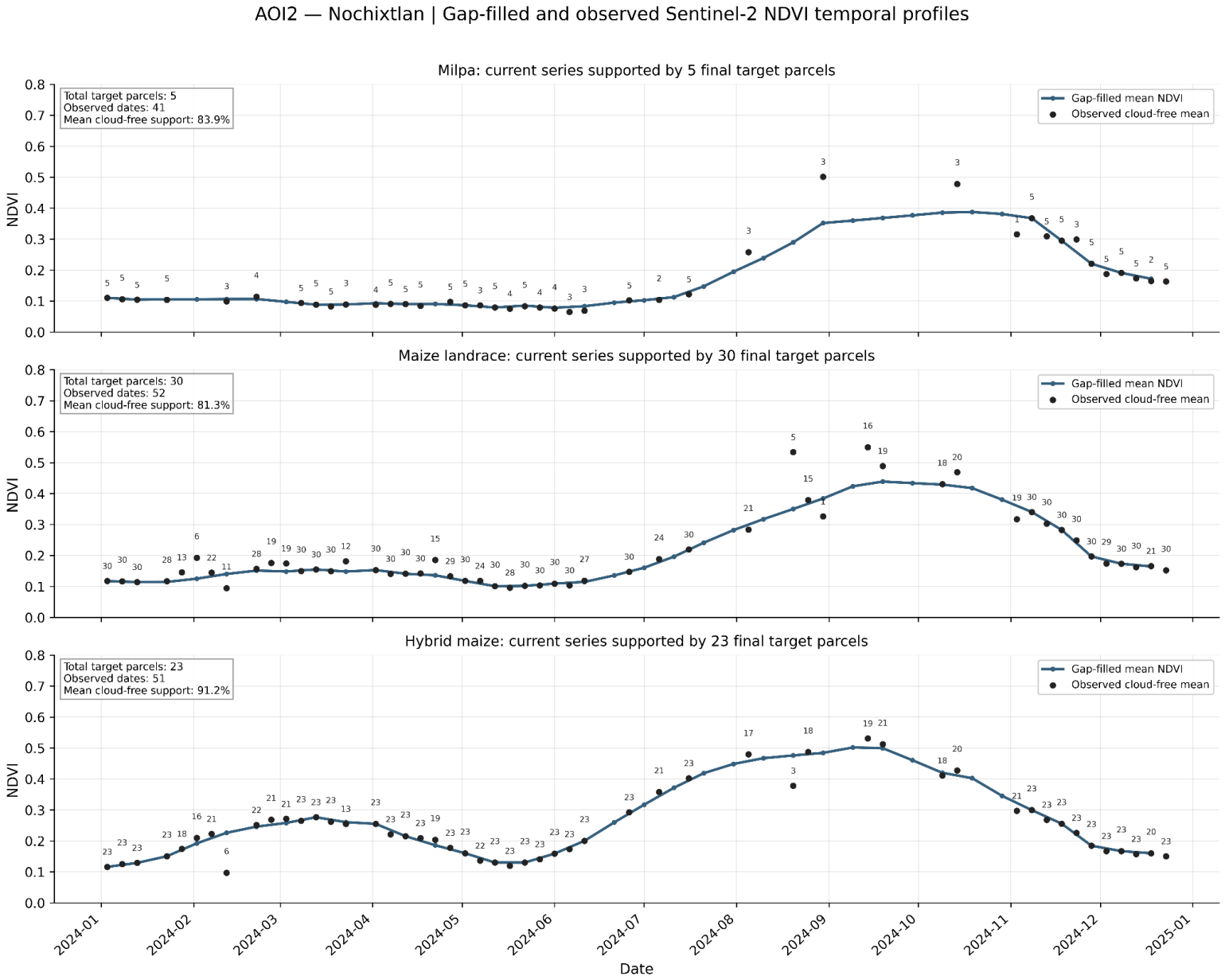
Appendix 5.2-H

AOI-specific observed and gap-filled Sentinel-2 NDVI temporal profiles

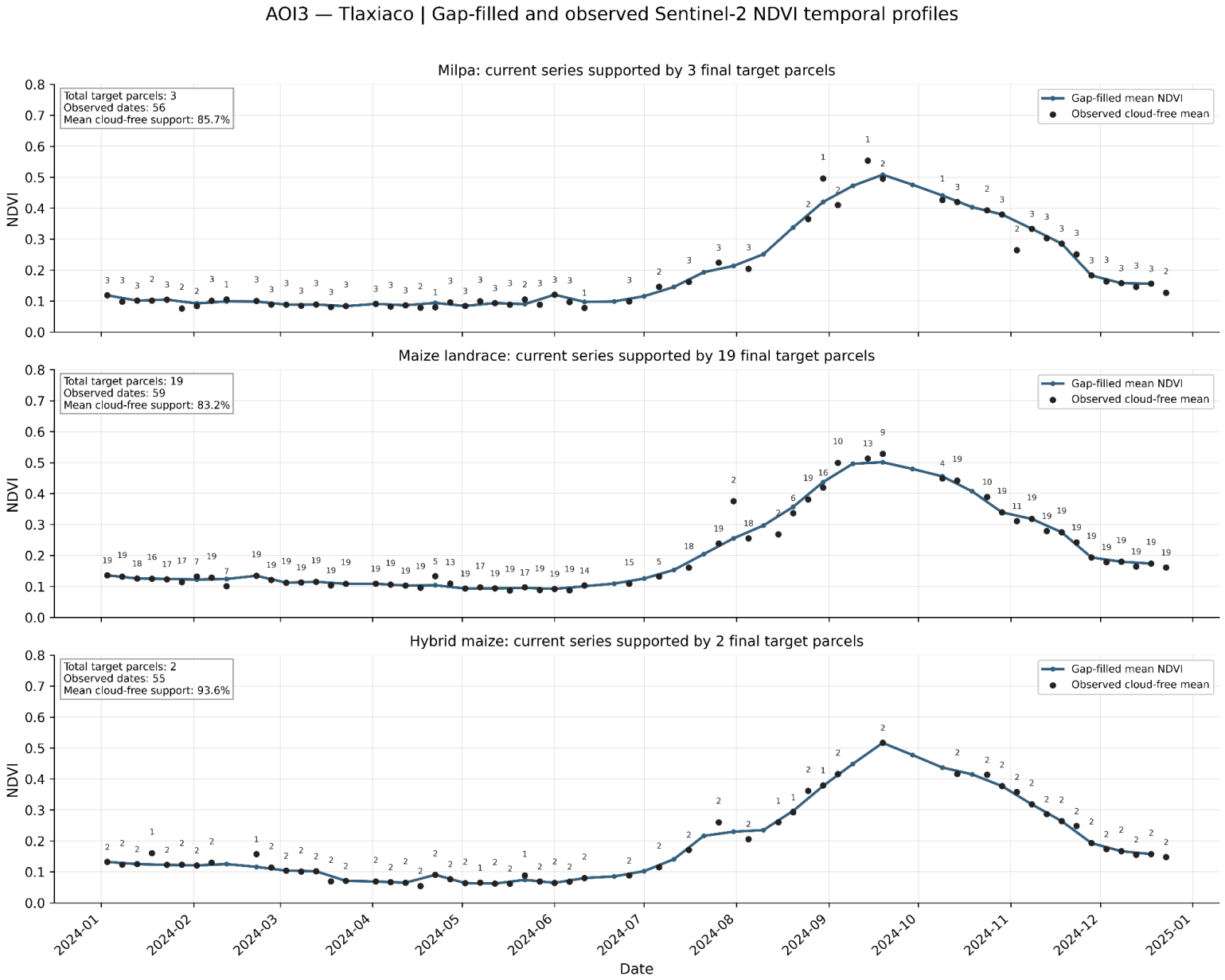
The following figures extend the descriptive temporal evidence in Figure 5.2 by showing the final 149-parcel cohort separately for AOI1-AOI4. Blue lines show class-level mean NDVI from the 36-date gap-filled Sentinel-2 series, black points show means from available cloud-free observations, and labels report the number of parcels contributing at each observed date. The figures are descriptive support and interpolation diagnostics and do not constitute independent benchmark performance.



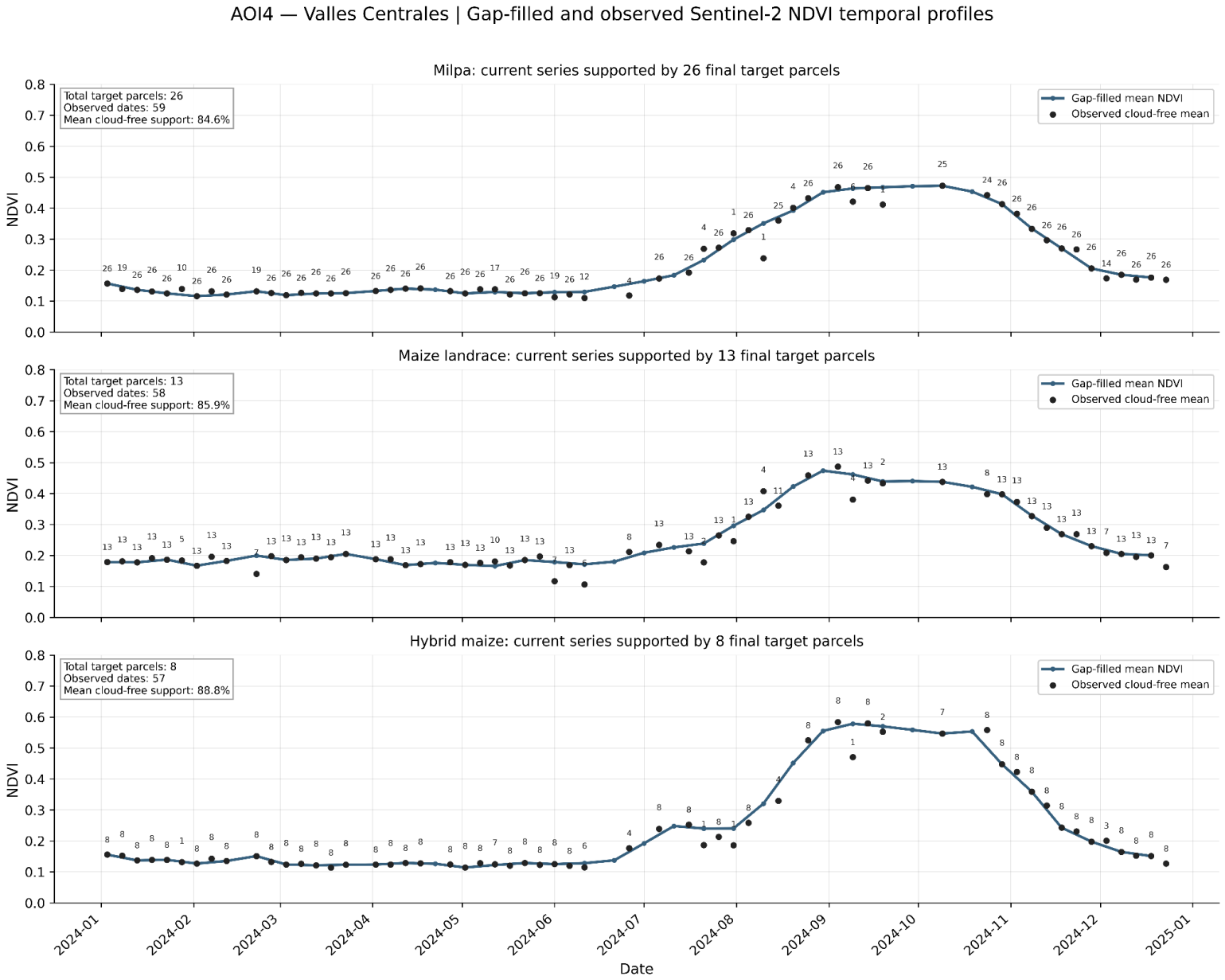
Appendix Figure A5.2-H1. AOI1 (Monte Verde) observed and gap-filled Sentinel-2 NDVI temporal profiles for the final target parcels. Blue lines show class-level mean gap-filled NDVI, black points show observed cloud-free means, and labels indicate the number of contributing parcels at each observed date.



Appendix Figure A5.2-H2. AOI2 (Nochixtlan) observed and gap-filled Sentinel-2 NDVI temporal profiles for the final target parcels. Blue lines show class-level mean gap-filled NDVI, black points show observed cloud-free means, and labels indicate the number of contributing parcels at each observed date.



Appendix Figure A5.2-H3. AOI3 (Tlaxiaco) observed and gap-filled Sentinel-2 NDVI temporal profiles for the final target parcels. Blue lines show class-level mean gap-filled NDVI, black points show observed cloud-free means, and labels indicate the number of contributing parcels at each observed date.



Appendix Figure A5.2-H4. AOI4 (Valles Centrales) observed and gap-filled Sentinel-2 NDVI temporal profiles for the final target parcels. Blue lines show class-level mean gap-filled NDVI, black points show observed cloud-free means, and labels indicate the number of contributing parcels at each observed date.

Appendix 5.3

Supporting evidence for predictor-level multi-sensor integration

This appendix provides the detailed support, class-specific, confusion-matrix and area-specific evidence supporting Section 5.3. The pooled performance summaries already reported in Tables 5.5 and 5.6 are not reproduced as standalone appendix tables. Development OOF refers to the pooled AOI3-to-AOI4 and AOI4-to-AOI3 geographic out-of-fold predictions. Benchmark refers to the frozen AOI1 and AOI2 secondary benchmark. S2 denotes Sentinel-2 and PS denotes PlanetScope.

5.3-A Sentinel-2 and SkySat predictor-level diagnostics

All 149 target parcels retained finite SkySat and Sentinel-2 predictor values. Duplicate parcel-date records were consolidated before equal-date averaging, and the final representations contained 25 Sentinel-2 predictors, 10 SkySat predictors or 35 joined predictors.

Appendix Table A5.3.1. SkySat parcel support and predictor-table verification.

Verification item	Value	Unit
Final target parcels	149	parcels
Valid parcel-scene records	316	records
Duplicate parcel-date groups consolidated	71	groups
Unique parcel-date records after consolidation	245	records
S2 predictor count	25	predictors
SkySat predictor count	10	predictors
S2 + SkySat predictor count	35	predictors
Missing or infinite values in final tables	0	values

The confusion orientation is observed class by predicted class. TN is maize monocrop correctly classified; FP is maize monocrop assigned to Milpa; FN is Milpa assigned to maize monocrop; and TP is Milpa correctly classified.

Appendix Table A5.3.2. SkySat pooled class-specific metrics and confusion counts.

Domain	Model	p	n	Milpa P	Milpa R	Milpa F1	TN	FP	FN	TP
Dev. OOF	S2	25	71	0.00	0.00	0.00	39	3	29	0
Dev. OOF	SkySat	10	71	0.07	0.03	0.05	28	14	28	1
Dev. OOF	S2 + SkySat	35	71	0.00	0.00	0.00	39	3	29	0
Benchmark	S2	25	78	0.38	0.36	0.37	56	8	9	5
Benchmark	SkySat	10	78	0.39	0.50	0.44	53	11	7	7
Benchmark	S2 + SkySat	35	78	0.36	0.29	0.32	57	7	10	4

Appendix Table A5.3.3. AOI-specific SkySat benchmark metrics.

AOI	Model	n	Accuracy	Balanced accuracy	Macro-F1	Milpa P	Milpa R	Milpa F1
AOI1	S2	20	0.65	0.61	0.56	1.00	0.22	0.36
AOI2	S2	58	0.83	0.72	0.64	0.27	0.60	0.38
AOI1	SkySat	20	0.45	0.47	0.44	0.43	0.67	0.52
AOI2	SkySat	58	0.88	0.57	0.58	0.25	0.20	0.22
AOI1	S2 + SkySat	20	0.60	0.56	0.47	1.00	0.11	0.20
AOI2	S2 + SkySat	58	0.84	0.73	0.66	0.30	0.60	0.40

Appendix Table A5.3.4. AOI-specific SkySat benchmark confusion counts.

AOI	Model	TN	FP	FN	TP
AOI1	S2	11	0	7	2
AOI2	S2	45	8	2	3
AOI1	SkySat	3	8	3	6
AOI2	SkySat	50	3	4	1
AOI1	S2 + SkySat	11	0	8	1
AOI2	S2 + SkySat	46	7	2	3

5.3-B Sentinel-2 and PlanetScope predictor-level diagnostics

Eight October 2024 PlanetScope scenes provided two planned acquisitions per AOI. All 149 parcels remained in the final predictor tables. Five AOI3 maize monocrop development parcels had one valid acquisition and used that observed acquisition without imputation, copying or zero filling.

Appendix Table A5.3.5. PlanetScope scene support and predictor-table verification.

Verification item	Value	Unit
Verified PlanetScope scenes	8	scenes
Planned parcel-scene rows	298	rows
Valid parcel-scene rows	293	rows
Zero-support parcel-scene rows excluded	5	rows
Parcels with two valid scenes	144	parcels
Parcels with one valid scene	5	parcels
Parcels retained in final tables	149	parcels
Missing or infinite values in final tables	0	values

Appendix Table A5.3.6. PlanetScope pooled class-specific metrics and confusion counts.

Domain	Model	p	n	Milpa P	Milpa R	Milpa F1	TN	FP	FN	TP
Dev. OOF	S2	25	71	0.00	0.00	0.00	39	3	29	0
Dev. OOF	PS spectral	20	71	0.11	0.07	0.08	25	17	27	2
Dev. OOF	S2 + PS	45	71	0.14	0.03	0.06	36	6	28	1
Dev. OOF	PS + spatial	23	71	0.06	0.03	0.04	26	16	28	1
Dev. OOF	S2 + PS + spatial	48	71	0.14	0.03	0.06	36	6	28	1
Benchmark	S2	25	78	0.38	0.36	0.37	56	8	9	5
Benchmark	PS spectral	20	78	0.33	0.50	0.40	50	14	7	7
Benchmark	S2 + PS	45	78	0.44	0.29	0.35	59	5	10	4
Benchmark	PS + spatial	23	78	0.27	0.43	0.33	48	16	8	6
Benchmark	S2 + PS + spatial	48	78	0.40	0.29	0.33	58	6	10	4

Appendix Table A5.3.7. AOI-specific PlanetScope benchmark metrics.

AOI	Model	n	Accuracy	Balanced accuracy	Macro-F1	Milpa P	Milpa R	Milpa F1

AOI	Model	n	Accuracy	Balanced accuracy	Macro-F1	Milpa P	Milpa R	Milpa F1
AOI1	S2	20	0.65	0.61	0.56	1.00	0.22	0.36
AOI1	PS spectral	20	0.55	0.53	0.52	0.50	0.33	0.40
AOI1	S2 + PS	20	0.60	0.56	0.47	1.00	0.11	0.20
AOI1	PS + spatial	20	0.50	0.47	0.45	0.40	0.22	0.29
AOI1	S2 + PS + spatial	20	0.60	0.56	0.47	1.00	0.11	0.20
AOI2	S2	58	0.83	0.72	0.64	0.27	0.60	0.38
AOI2	PS spectral	58	0.79	0.80	0.64	0.27	0.80	0.40
AOI2	S2 + PS	58	0.88	0.75	0.70	0.38	0.60	0.46
AOI2	PS + spatial	58	0.76	0.78	0.61	0.24	0.80	0.36
AOI2	S2 + PS + spatial	58	0.86	0.74	0.68	0.33	0.60	0.43

Appendix Table A5.3.8. AOI-specific PlanetScope benchmark confusion counts.

AOI	Model	TN	FP	FN	TP
AOI1	S2	11	0	7	2
AOI1	PS spectral	8	3	6	3
AOI1	S2 + PS	11	0	8	1
AOI1	PS + spatial	8	3	7	2
AOI1	S2 + PS + spatial	11	0	8	1
AOI2	S2	45	8	2	3
AOI2	PS spectral	42	11	1	4
AOI2	S2 + PS	48	5	2	3
AOI2	PS + spatial	40	13	1	4
AOI2	S2 + PS + spatial	47	6	2	3

The deltas below are candidate minus reference. No inferential comparison was saved; the verdicts therefore describe only the direction of the recorded metrics.

Appendix Table A5.3.9. PlanetScope representation deltas relative to the stated comparator.

Domain	Comparison	Delta accuracy	Delta BA	Delta Macro-F1	Delta Milpa R	Delta Milpa F1	Recorded outcome

Domain	Comparison	Delta accuracy	Delta BA	Delta Macro-F1	Delta Milpa R	Delta Milpa F1	Recorded outcome
Dev. OOF	PS spectral vs S2	-0.17	-0.13	-0.05	+0.07	+0.08	Mixed Metric Result
Dev. OOF	S2 + PS spectral vs S2	-0.03	-0.02	+0.01	+0.03	+0.06	Mixed Metric Result
Dev. OOF	Spatial added to PS	0.00	-0.01	-0.02	-0.03	-0.04	Non Improving
Dev. OOF	Spatial added to S2 + PS	0.00	0.00	0.00	0.00	0.00	Non Improving
Dev. OOF	Full 48-predictor vs S2	-0.03	-0.02	+0.01	+0.03	+0.06	Mixed Metric Result
Benchmark	PS spectral vs S2	-0.05	+0.02	-0.01	+0.14	+0.03	Mixed Metric Result
Benchmark	S2 + PS spectral vs S2	+0.03	-0.01	0.00	-0.07	-0.02	Non Improving
Benchmark	Spatial added to PS	-0.04	-0.05	-0.05	-0.07	-0.07	Non Improving
Benchmark	Spatial added to S2 + PS	-0.01	-0.01	-0.01	0.00	-0.01	Non Improving
Benchmark	Full 48-predictor vs S2	+0.01	-0.02	-0.01	-0.07	-0.04	Non Improving

The frozen PlanetScope benchmark verification confirmed all five expected models, 78 rows per model, 78 unique benchmark parcels, consistent reference labels and exact agreement with the saved metric summaries. It performed no retraining and produced no new predictions.

Appendix 5.4

Supporting evidence for SkySat-guided Sentinel-2 super-resolution

This appendix provides the detailed cohort, reconstruction and downstream diagnostic evidence supporting Section 5.4. Reconstruction errors are reported in normalised image space. Positive error-change percentages indicate lower CNN-SR error than bicubic interpolation; negative percentages indicate higher CNN-SR error. The downstream evidence is restricted to represented parcels and does not replace the primary 149-parcel Sentinel-2 classification baseline.

5.4-A Represented cohort and split support

Appendix Table A5.4.1. Parcel and patch representation by domain, AOI and crop class.

Domain	AOI	Class	Starting	Represented	Zero-patch	Patches
Development	AOI3	Maize monocrop	21	13	8	58
Development	AOI3	Milpa	3	2	1	6
Development	AOI4	Maize monocrop	21	19	2	94
Development	AOI4	Milpa	26	25	1	119
Benchmark	AOI1	Maize monocrop	11	10	1	36
Benchmark	AOI1	Milpa	9	8	1	27
Benchmark	AOI2	Maize monocrop	53	35	18	175
Benchmark	AOI2	Milpa	5	0	5	0

Appendix Table A5.4.2. Parcel-controlled super-resolution data split.

Split	Patches	Unique parcels
Train	222	47
Validation	55	12
Test	238	53

Of the 515 accepted paired patches, 453 used the predefined sensor pairs and 62 used the permitted ±5-day fallback; the fallback was an alternative temporal pairing route rather than imputation.

Parcel overlap was zero for train-validation, train-test and validation-test comparisons. The final branch contained 149 starting parcels, 112 represented parcels, 37 zero-patch parcels and 515 accepted paired patches.

5.4-B Detailed reconstruction assessment

Appendix Table A5.4.3. Pooled and band-specific reconstruction error on untouched test patches.

Scope	Patches	Bicubic MAE	CNN MAE	MAE change (%)	Bicubic RMSE	CNN RMSE	RMSE change (%)	Abs. bias change (%)
All bands	238	0.79	0.66	17.21	1.09	0.97	11.13	63.93
Blue	238	0.89	0.82	8.05	1.24	1.21	2.39	7.09
Green	238	1.00	0.87	13.57	1.26	1.15	9.14	47.53
Red	238	0.61	0.50	17.71	0.91	0.78	14.52	55.04
NIR	238	0.68	0.45	34.16	0.88	0.60	32.35	47.61

Appendix Table A5.4.4. AOI-specific and equal-parcel reconstruction summaries.

Weighting	Scope	n	Bicubic MAE	CNN MAE	MAE change (%)	Bicubic RMSE	CNN RMSE	RMSE change (%)
Patch-weighted	AOI1	63	0.86	0.91	-5.80	1.30	1.40	-7.73
Patch-weighted	AOI2	175	0.78	0.58	+24.97	1.01	0.79	+22.27
Equal-parcel	Overall	53	0.80	0.69	+13.90	1.05	0.94	+10.38
Equal-parcel	AOI1	18	0.85	0.90	-5.71	1.24	1.33	-6.90
Equal-parcel	AOI2	35	0.78	0.58	+24.94	0.95	0.74	+22.02
Equal-parcel	Maize monocrop	45	0.79	0.66	+17.26	1.01	0.87	+13.92
Equal-parcel	Milpa	8	0.85	0.88	-3.71	1.26	1.33	-5.63

5.4-C Downstream classification diagnostics

Appendix Table A5.4.5. AOI1 mixed-class downstream metrics.

Representation	n	Accuracy	Balanced accuracy	Macro-F1	Milpa P	Milpa R	Milpa F1	Maize FPR
Native S2 10 m	18	0.61	0.58	0.54	0.67	0.25	0.36	0.10
Bicubic S2 5 m	18	0.72	0.70	0.70	0.80	0.50	0.62	0.10
CNN-SR 5 m	18	0.61	0.56	0.48	1.00	0.13	0.22	0.00

Appendix Table A5.4.6. AOI1 mixed-class downstream confusion counts.

Representation	Maize to maize	Maize to Milpa	Milpa to maize	Milpa to Milpa
Native S2 10 m	9	1	6	2
Bicubic S2 5 m	9	1	4	4
CNN-SR 5 m	10	0	7	1

The represented AOI2 test cohort contained 35 maize monocrop parcels and no represented Milpa parcel. Balanced accuracy, macro-F1 and Milpa-specific metrics were therefore not applicable to this control.

Appendix Table A5.4.7. AOI2 maize-only false-positive control.

Representation	n	Correct maize	Maize to Milpa	Maize recall	False-positive rate
Native S2 10 m	35	18	17	0.51	0.49
Bicubic S2 5 m	35	19	16	0.54	0.46
CNN-SR 5 m	35	21	14	0.60	0.40

Appendix 5.5

Supporting evidence for classification-level integration

This appendix provides the native-model agreement, fusion correction and harm, confusion, AOI-specific and bootstrap evidence supporting Section 5.5. The 12-row pooled performance summary reported in Table 5.8 is not duplicated. Development OOF refers to the 71 AOI3 and AOI4 directional geographic out-of-fold predictions; benchmark refers to the frozen 78-parcel AOI1 and AOI2 secondary benchmark.

5.5-A Native-model agreement and complementarity

Appendix Table A5.5.1. Mutually exclusive native-model correctness patterns.

Domain	n	All agree	All agree (%)	All correct	All incorrect	Exactly one correct	Exactly two correct
Dev. OOF	71	49	69.01	21	28	12	10
Benchmark	78	45	57.69	41	4	11	22

Appendix Table A5.5.2. Native-model agreement for Milpa parcels.

Domain	Milpa n	All agree	All correct	All incorrect	Exactly one correct	Exactly two correct
Dev. OOF	29	27	0	27	1	1
Benchmark	14	6	2	4	3	5

Appendix Table A5.5.3. Parcels uniquely classified correctly by each native model.

Domain	Native model	Unique correct parcels
Dev. OOF	S2	10
Dev. OOF	SkySat	1
Dev. OOF	PS spectral	1
Benchmark	S2	5
Benchmark	SkySat	5
Benchmark	PS spectral	1

AOI3 contained disagreement among the three native models for 21 of 24 parcels. AOI4 showed agreement for 46 of 47 parcels, including unanimous misclassification of all 26 AOI4 Milpa parcels.

5.5-B Fusion correction, harm and pooled confusion counts

Appendix Table A5.5.4. Parcel-level correction and harm relative to Sentinel-2.

Fusion	Domain	Corrected	Harmed	Net correct	Milpa corrected	Milpa harmed	Fusion maize FP
S2 + SkySat fusion	Development OOF	2	3	-1	0	0	4
S2 + PS fusion	Development OOF	1	5	-4	0	0	7
Three-sensor fusion	Development OOF	1	5	-4	0	0	7
S2 + SkySat fusion	Benchmark	6	6	0	2	0	10
S2 + PS fusion	Benchmark	5	6	-1	1	1	9
Three-sensor fusion	Benchmark	7	9	-2	2	1	11

Appendix Table A5.5.5. Pooled confusion counts for the fixed classification-level fusions.

Domain	Fusion	n	TN	FP	FN	TP
Dev. OOF	S2 + SkySat fusion	71	38	4	29	0
Dev. OOF	S2 + PS fusion	71	35	7	29	0
Dev. OOF	Three-sensor fusion	71	35	7	29	0
Benchmark	S2 + SkySat fusion	78	54	10	7	7
Benchmark	S2 + PS fusion	78	55	9	9	5
Benchmark	Three-sensor fusion	78	53	11	8	6

5.5-C AOI-specific fixed-fusion metrics

Appendix Table A5.5.6. AOI-specific fixed-fusion performance.

Domain	AOI	Fusion	n	Accuracy	Balanced accuracy	Macro-F1	Milpa P	Milpa R	Milpa F1
Dev. OOF	AOI3	S2 + SkySat fusion	24	0.71	0.40	0.41	0.00	0.00	0.00
Dev. OOF	AOI3	S2 + PS fusion	24	0.58	0.33	0.37	0.00	0.00	0.00
Dev. OOF	AOI3	Three-sensor fusion	24	0.58	0.33	0.37	0.00	0.00	0.00
Dev. OOF	AOI4	S2 + SkySat fusion	47	0.45	0.50	0.31	0.00	0.00	0.00
Dev. OOF	AOI4	S2 + PS fusion	47	0.45	0.50	0.31	0.00	0.00	0.00
Dev. OOF	AOI4	Three-sensor fusion	47	0.45	0.50	0.31	0.00	0.00	0.00

Domain	AOI	Fusion	n	Accuracy	Balanced accuracy	Macro-F1	Milpa P	Milpa R	Milpa F1
Benchmark	AOI1	S2 + SkySat fusion	20	0.60	0.59	0.58	0.57	0.44	0.50
Benchmark	AOI1	S2 + PS fusion	20	0.55	0.51	0.44	0.50	0.11	0.18
Benchmark	AOI1	Three-sensor fusion	20	0.55	0.53	0.52	0.50	0.33	0.40
Benchmark	AOI2	S2 + SkySat fusion	58	0.84	0.73	0.66	0.30	0.60	0.40
Benchmark	AOI2	S2 + PS fusion	58	0.84	0.82	0.69	0.33	0.80	0.47
Benchmark	AOI2	Three-sensor fusion	58	0.83	0.72	0.64	0.27	0.60	0.37

5.5-D Three-sensor bootstrap uncertainty

The intervals are two-sided 95% percentile intervals from 10,000 paired parcel-level resamples. The same resampled parcel indices were used across compared models. The intervals are descriptive and are not presented as formal significance tests.

Appendix Table A5.5.7. Pooled three-sensor bootstrap intervals and differences from Sentinel-2.

Domain	Metric	Three-sensor estimate	95% interval	Difference vs S2	95% difference interval
Dev. OOF	Balanced accuracy	0.42	[0.37, 0.46]	-0.05	[-0.10, 0.00]
Dev. OOF	Macro-F1	0.33	[0.30, 0.35]	-0.02	[-0.05, 0.00]
Dev. OOF	Milpa recall	0.00	[0.00, 0.00]	0.00	[0.00, 0.00]
Dev. OOF	Milpa F1	0.00	[0.00, 0.00]	0.00	[0.00, 0.00]
Benchmark	Balanced accuracy	0.63	[0.50, 0.76]	+0.01	[-0.11, 0.15]
Benchmark	Macro-F1	0.62	[0.49, 0.74]	0.00	[-0.13, 0.13]
Benchmark	Milpa recall	0.43	[0.21, 0.71]	+0.07	[-0.14, 0.29]
Benchmark	Milpa F1	0.39	[0.18, 0.58]	+0.02	[-0.19, 0.23]

Evaluating Multi-Sensor Remote Sensing for Parcel-Level Mapping of Milpa Mixed-Cropping and Maize Monocrop Fields in Oaxaca, Mexico

Md Huraira Rabbi Shahriar

This thesis evaluates whether complementary information from SkySat and PlanetScope improves parcel-level discrimination of the Milpa mixed-cropping system from maize monocrop beyond a seasonal Sentinel-2 baseline in Oaxaca, Mexico. A 2024 field-reference dataset was screened to a final analytical cohort of 149 parcels across four areas of interest; 71 parcels from AOI3 and AOI4 were used for development and 78 parcels from AOI1 and AOI2 formed a secondary, non-confirmatory benchmark. Sentinel-2 time series were harmonised to 36 ten-day dates and represented using spectral, temporal, phenological, heterogeneity and observed-texture features. The primary parcel-level Random Forest achieved binary benchmark accuracy of 0.78, balanced accuracy and macro-F1 of 0.62, while Milpa recall remained 0.36. SkySat and PlanetScope provided some favourable Milpa-sensitive benchmark results, with both native representations reaching Milpa recall of 0.50, although gains were geographically inconsistent and accompanied by additional maize false positives. Predictor concatenation did not provide a consistent advantage. SkySat-guided CNN super-resolution improved image-reconstruction error relative to bicubic interpolation, but this did not translate consistently into better downstream crop-system classification. Among fixed classification-level combinations, Sentinel-2 + SkySat produced the strongest pooled benchmark fusion, while no fusion improved development Milpa recognition. Overall, finer spatial information was complementary but did not remove the difficulty of geographically transferable Milpa discrimination, highlighting the need for stronger reference data and multi-season validation.

UNIVERSITÉ CATHOLIQUE DE LOUVAIN
Faculté des bioingénieurs
Croix du Sud, 2 bte L7.05.01, 1348 Louvain-la-Neuve, Belgique | www.uclouvain.be/agro