Boundary information without parcel objects: probing the AlphaEarth annual embedding against seasonal Sentinel-1/2 features for sub-hectare coffee delineation

Antônio Campos de Abreu Filho[1,*], Marcos Antônio Timbó Elmiro[1], Marcelo Antonio Nero[1], and Rodrigo Machado Fernandes Leitão[2]

[1]Universidade Federal de Minas Gerais (UFMG), Belo Horizonte, Brazil
[2]Independent Researcher
[*]Address correspondence to: antonio@abreufilho.com.br

Abstract

Embedding products promise to replace local seasonal compositing in agricultural mapping: AlphaEarth Foundations condenses a year of multi-sensor observation into a 64-dimensional annual embedding at 10 m. Whether such a representation retains the few-pixel-wide boundary information that parcel delineation requires is untested. We test this in a demanding case: 384,496 coffee parcels in Minas Gerais, Brazil (2017 season; median 0.66 ha), using a controlled probing design. Identical lightweight probes (Random Forest plus a linear probe) are trained on four per-pixel representations: seasonal Sentinel-2, seasonal Sentinel-1 plus Sentinel-2, AlphaEarth, and their fusion. Parcel objects are extracted by a fixed watershed whose single threshold is selected on spatially held-out validation cells. On held-out test cells, the embedding carried more per-pixel boundary information than the seasonal features (boundary average precision 0.524 versus 0.406; median paired gain +0.120, $p_{\mathrm{Holm}} < 0.0001$, Holm-Bonferroni-adjusted within contrast). The signal was also linearly decodable. The advantage did not convert into better parcel objects: the embedding traded over-segmentation for under-segmentation, and seasonal features kept equal or better object metrics, especially for small parcels; the n=3 large-parcel stratum showed a possible object-quality gain but remains exploratory. Under this fixed watershed extraction, the result is a conversion gap: boundary evidence is present, but this extraction does not turn it into reliable objects. The findings caution against treating annual embeddings as drop-in inputs for sub-hectare parcel delineation.

Keywords

field boundary delineation; foundation model embeddings; AlphaEarth; Sentinel-1; Sentinel-2; representation probing; sub-hectare parcels; watershed segmentation

ORCID

ORCID: Abreu Filho 0009-0007-1615-2315; Timbó Elmiro 0000-0001-7680-3131; Nero 0000-0003-2124-5018; Leitão 0000-0002-8470-5619

1 Introduction

Field boundaries are a core enabler of agricultural monitoring: parcel-level maps support crop-type classification, yield estimation, and the operational support of food-security programmes [1, 2]. Most of the world's farms are small: more than 500 million operate on less than two hectares [3]. In small-parcel landscapes, automated delineation is both most useful and most difficult. Deep-learning boundary extractors have raised the state of the art on medium-resolution imagery [4], and dedicated architectures and datasets have pushed into small-parcel settings [5, 6], but these models are trained per task and per region, and their inputs are typically carefully engineered composites or time series of Sentinel imagery. In parallel, remote-sensing foundation models have begun to adapt self-supervised pre-training to the spectral, temporal, and geographic structure of Earth observation data: SatMAE adapts masked autoencoding to temporal and multispectral imagery [7], Prithvi scales geospatial pre-training on Harmonized Landsat–Sentinel data [8, 9], and Presto targets pixel-level remote-sensing time series [10]. Recent surveys frame these models as a rapidly growing but still task-dependent family of representations for Earth observation [11].

A newer alternative is the ready-made embedding. AlphaEarth Foundations distills a year of multi-source satellite observation, including optical, radar, and other streams, into a global 64-dimensional annual embedding at 10 m, designed so that light models trained on sparse labels can map directly from the embedding [12]. For a practitioner, the promise is concrete: skip local seasonal compositing and delineate parcels with a simple model over a product that already exists. The embedding has been evaluated chiefly on classification-style mapping tasks, where it performs strongly [12]; independent benchmarks confirm competitive accuracy on agricultural downstream tasks such as yield, tillage, and cover-crop mapping, while flagging limited spatial transferability [13]. No published evaluation targets parcel delineation. Whether the embedding preserves the fine spatial information that *delineation* needs is a different question, and an open one: parcel boundaries in small-parcel landscapes are a few pixels wide at 10 m, and annual aggregation could plausibly smooth away exactly the transient, season-dependent contrasts that separate one parcel from the next.

This study asks that question directly for a hard, well-instrumented case: coffee parcels in Minas Gerais, Brazil, where a public reference delineates 384,496 parcels with a sub-hectare median size for the 2017 season [14]. The analysis compares four per-pixel representations built from overlapping source families: seasonal optical features, seasonal optical plus radar features, the AlphaEarth annual embedding, and their fusion, under a supervised probing design. An identical lightweight probe is trained on each representation to predict boundary probability, parcel objects are extracted by a fixed watershed whose single threshold is selected on spatially held-out validation cells, and performance is measured on held-out test cells with object metrics, a threshold-free boundary signal, and per-tile paired statistics. Because the probe and every downstream step are held fixed, differences are

attributable to the representations. A secondary linear probe asks whether the embedding's boundary information is linearly decodable, following the linear-probe convention for learned representations [15].

The comparison is one of representations, not sensors: the embedding is itself trained on optical and radar streams, so the study measures the boundary information available in the ready-made representation relative to season-resolved features engineered from overlapping source families. The contributions are threefold: (i) a controlled probing design for representation comparison in parcel delineation, with spatial splits and a single validated knob; (ii) evidence of an information–object gap, where the annual embedding carries *more* per-pixel boundary information than the seasonal features and that information is linearly decodable, yet the advantage does not convert into better parcel objects under a simple extraction that trades over-segmentation for under-segmentation; and (iii) a size-stratified analysis of that gap. The stratification suggests that the gap is weakest in the large-parcel stratum, while the dominant small-parcel regime drives the object-level failure most relevant to the delineation problem.

2 Study area and data

2.1 Study area and reference parcels

The study covers coffee-growing areas of Minas Gerais (MG), Brazil, for the 2017 crop year, a region where coffee mapping from satellite imagery has a long operational history [16]. The reference is CONAB's semi-automatic mapping of Brazilian crop parcels, distributed on Source Cooperative as the `br_conab` field-boundary dataset by the Field Boundaries for Agriculture (fiboa) project [14]. Filtering the dataset to MG returned 384,496 parcels, all encoded as coffee (`MG_CAFE`) and all from the 2017 season; the parcel identifier follows the pattern `{state}_{crop}_{yy}_{n}` (e.g., `MG_CAFE_17_1`), which serves as the filtering and stratification key. No other crop or year appears for MG, so the reference fixes the scope of the study to coffee delineation in MG for the 2017 base year.

The parcels are small (Figure 1c). Because the dataset's `metrics:area` field is null, area was computed from the polygon geometry: the median parcel is 0.66 ha (interquartile range 0.13–1.98 ha), with a maximum near 472 ha. The area distribution also shows a secondary mode of sliver features below 0.01 ha: 76,203 features, or 19.8% of the count but only 0.011% of the mapped area, with a median near 4 m^2. These features are consistent with fragmentation artefacts of the semi-automatic delineation and are sub-pixel at the 10 m grid. At a 10 m ground sampling distance, the median parcel covers roughly 65 pixels; a compact parcel of that size is only about eight pixels across, so the target boundaries are just a few pixels wide. The parcels are treated as a reference with its own delineation error rather than as cadastral ground truth: the mapping is semi-automatic, and only coffee parcels are delineated. The reference is thus **sparse** because the area outside a mapped parcel is unlabelled, not confirmed non-coffee; this constrains where the method can be trained and evaluated (see Materials and Methods).

For spatial stratification and to guard against spatial leakage, each parcel was assigned, by the location of its centroid, to a 0.5° region cell, which yielded 159 cells across MG.

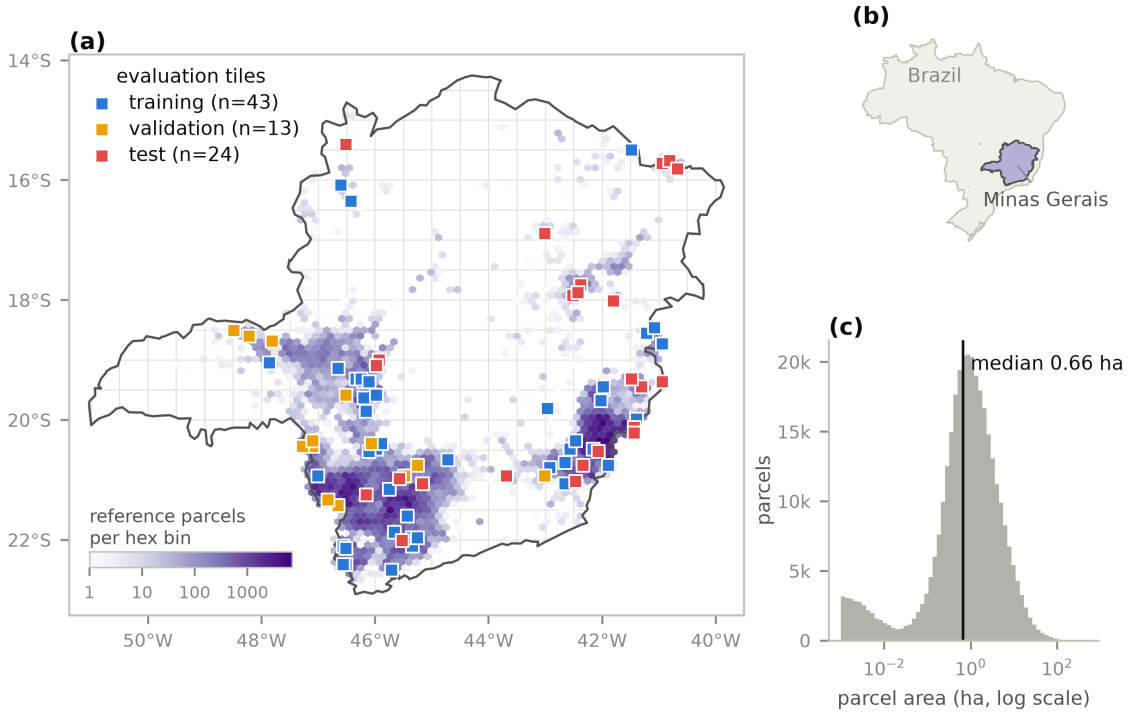

Figure 1. Study area and sampled tiles. (a) Reference coffee-parcel density in Minas Gerais and the 80 evaluation tiles by partition. The 0.5° cells define the spatial split. (b) Minas Gerais within Brazil. (c) Reference parcel-area distribution, with the median marked.

2.2 Data sources

Three feature sources cover MG for 2017, all at 10 m resolution and all accessed through Google Earth Engine [17] (Table 1). Earth Engine serves only as the data source; the analysis runs locally.

Sentinel-2 supplies the optical features [18]. The top-of-atmosphere (TOA) product (`S2_HARMONIZED`, Level-1C) was used because the surface-reflectance (Level-2A) collection in Earth Engine begins only in 2017 and covers the study area sparsely that year, leaving the Level-1C archive as the only source with full 2017 coverage. Sentinel-1 Ground Range Detected (GRD) imagery acquired in Interferometric Wide-swath (IW) mode supplies the synthetic-aperture-radar (SAR) features [19]; both Sentinel-1A and Sentinel-1B were operational throughout 2017, so the SAR record for the study year has no acquisition gap. AlphaEarth (the Satellite Embedding V1 product) supplies a 64-dimensional annual embedding [12]. The embedding is not an independent sensor: it is a learned representation trained over optical, radar, and other streams, so the comparison in this study is between representations of overlapping sources: a ready-made annual summary versus season-resolved features engineered from the same optical and radar inputs, rather than between independent sensors.

Table 1. Feature sources (Minas Gerais, 2017).

Source	Role	Resolution	Access
Sentinel-2 TOA (`S2_HARMONIZED`, L1C)	seasonal optical features	10 m	Earth Engine
Sentinel-1 GRD (IW, VV/VH)	seasonal SAR features	10 m	Earth Engine
AlphaEarth (Satellite Embedding V1)	annual 64-D embedding	10 m	Earth Engine
CONAB crop-parcel mapping (`br_conab`)	reference parcels (coffee, MG, 2017)	vector, EPSG:4674	Source Coop.

The optical and radar features were aggregated by season to follow coffee phenology in MG, whose arabica cycle spans a wet vegetative-to-fruiting phase and a drier maturation-to-rest phase [20]. A wet season (January–March plus October–December 2017) and a dry season (April–September 2017) were defined, and separate composites were computed for each. The wet-season composite thus joins two halves of adjacent phenological wet seasons within one calendar year; this is deliberate, because the AlphaEarth embedding likewise aggregates the 2017 calendar year, and aligning both representations to the same calendar window keeps the comparison fair. Aggregating within the single reference year (2017) also aligns every image with the reference and, with one year of data, replaces a temporal split with a spatial one during evaluation.

2.3 Evaluation tiles

Delineating all of MG is neither necessary nor feasible for this comparison, so the method is evaluated in tiles: square windows where the reference parcels give a local basis for evaluation. The MG bounding box was tiled into 500×500-pixel cells (about 5 km on a side), and the 3,011 cells containing at least ten reference parcels were kept as candidates. From these candidates, a stratified sample of 80 tiles was drawn by region cell, tile size class, and parcel density class so that no single condition dominates the sample (Figure 1a). Size class is assigned per tile from the median area of its reference parcels: small is <1 ha, mid is 1–3 ha, and large is ≥ 3 ha. Density class is assigned from the parcel

count in the tile: low is <50 parcels, mid is 50–200, and high is ≥200. The draw uses a fixed seed (42) for reproducibility. Each tile was downloaded separately in its local UTM projection, which keeps the pixel a true 10×10 m square (in geographic coordinates the pixel is not square at MG's latitudes). The tiles are the test sites, not the result: the study reports how the method performs within them, stratified by region and parcel size.

3 Materials and Methods

The question is whether a ready-made annual embedding preserves the parcel-boundary information that season-resolved Sentinel-1/2 features carry, or whether annual aggregation smooths it away. The analysis uses a light supervised probing framework: the reference parcels become supervision rather than only a final score, an identical probe is trained on each feature representation, and parcel objects are extracted from the probe's boundary probability. Because the probe and every downstream step are held fixed across representations, differences in the resulting objects are attributable to the features, not to the model. The pipeline runs in five stages (reference extraction, tiling, feature download, probing, and validation), each implemented as a script in `code/`, with a final script deriving the tables and figures. Figure 2 summarizes the full design, from the reference parcels through the shared probe and extraction to the paired per-tile tests.

3.1 Reference data and labels

Within each tile, the reference parcels are rasterized into three classes. Only parcels lying entirely inside the tile interior are used: the tile bounds are eroded by 10 m (one pixel) to keep parcels contained in that eroded box, so that clipping a partial parcel never creates an artificial boundary coincident with the tile edge. Each retained parcel contributes a **boundary** class, defined as its outline buffered by 10 m and clipped to the parcel so the class occupies roughly one pixel along the inner edge, and an **interior** class, the remainder of the parcel. Pixels outside any parcel are left unlabelled: the reference is sparse, so the exterior cannot be asserted as background. Labels are written per tile as a raster (0 = outside, 1 = interior, 2 = boundary), and training and boundary evaluation are confined to the labelled footprint.

3.2 Feature representations

Four feature stacks define the experiments; each is a per-pixel feature vector fed to the same probe (Table 2). E1 uses only the Sentinel-2 seasonal features; E2 adds the Sentinel-1 seasonal features; E3 uses only the AlphaEarth embedding; and E4 concatenates all three sources. E2 is the season-resolved engineered baseline, E3 is the ready-made embedding, and E4 tests whether the embedding adds anything beyond the engineered features.

Table 2. Feature stacks per experiment.

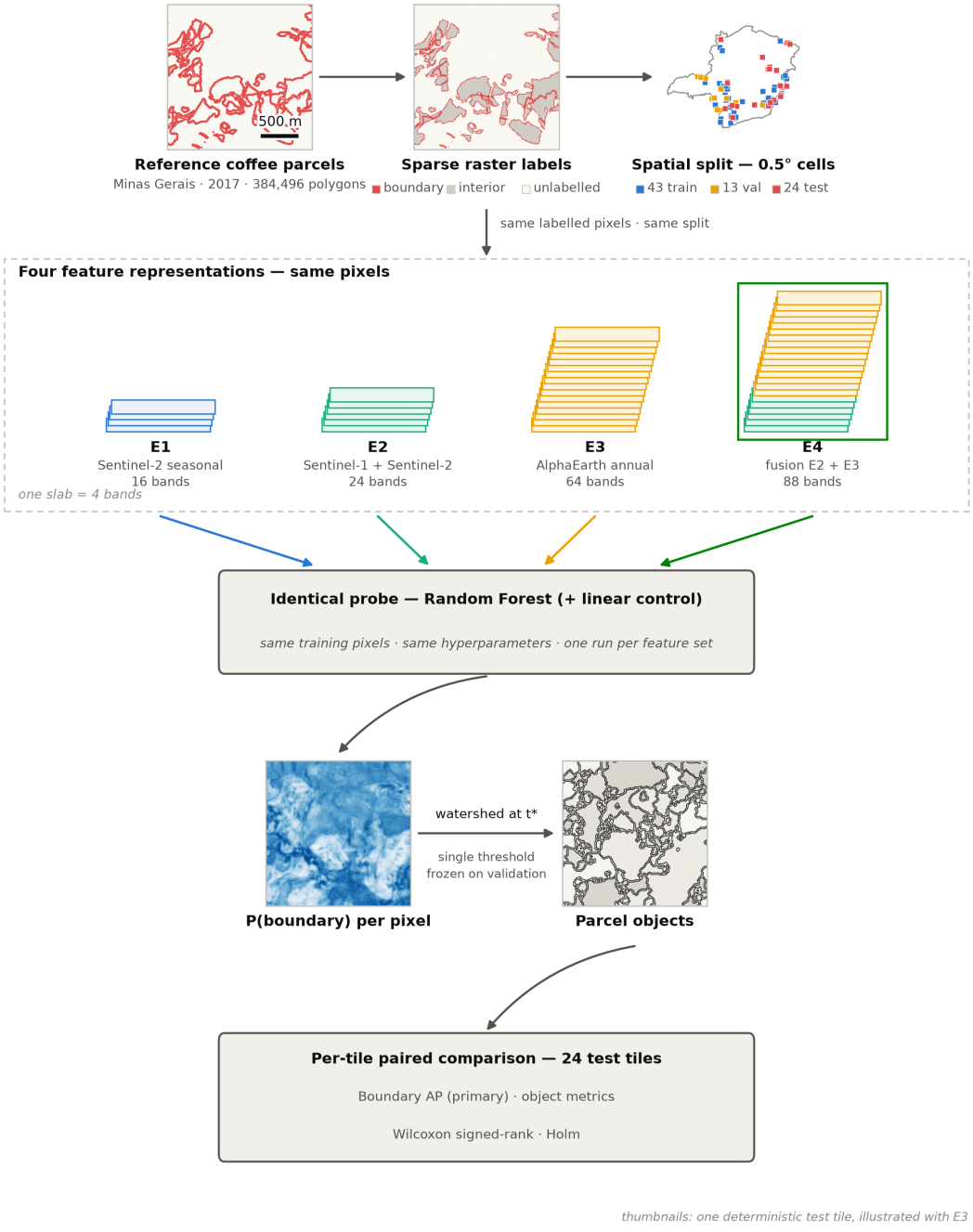

Figure 2. Experimental design. The same reference labels, probe, watershed extraction, and paired-test protocol are applied to each feature stack; only the input representation changes. Colour identifies the feature set, and grey stages are fixed across experiments.

Experiment	Sources	Bands
E1	Sentinel-2 seasonal	16
E2	Sentinel-1 + Sentinel-2 seasonal	24
E3	AlphaEarth embedding	64
E4	Sentinel-1 + Sentinel-2 + AlphaEarth	88

The Sentinel-2 stack holds eight features per season: the B2, B3, B4, B8, and B11 bands plus NDVI, NDWI [21], and EVI [22], for the wet and dry seasons, giving 16 bands. For each season, clouds were masked with the QA60 band, scenes below 60% cloud cover were kept, reflectance was scaled, and the collection was reduced to its per-pixel median before computing the indices. The Sentinel-1 stack holds four features per season: the median and standard deviation of VV and VH over IW GRD acquisitions, for both seasons, giving eight bands. Backscatter is used in decibels as distributed by Earth Engine, with no additional speckle filtering: the per-season median already suppresses speckle temporally. The AlphaEarth stack is the 64-band annual embedding mosaic for 2017. Feature values are used as provided; because the probe is invariant to per-feature scaling (see below), no cross-source normalization is applied. Pixels left without a valid observation by a seasonal composite (persistent cloud cover in a season) are assigned a feature value of zero rather than masked, so every representation is trained and evaluated on the same pixel set.

3.3 Probing

A single Random Forest (RF) classifier [23] serves as the probe, with fixed hyperparameters shared across E1–E4: 200 trees, a minimum of five samples per leaf, and balanced-subsample class weights, under a fixed random seed (42). Holding the probe and its hyperparameters identical across experiments isolates the effect of the representation. A tree ensemble is invariant to monotone per-feature rescaling, so it compares differently-scaled representations without imposing a similarity metric or a normalization choice on any of them. The probe is trained on the same set of training pixels for every experiment and predicts, per pixel, the probability that the pixel belongs to a parcel boundary.

Training pixels are drawn from the labelled footprints of the training tiles only. To keep the classes and the tiles balanced, the sample is capped at 4,000 boundary and 8,000 interior pixels per tile; in the current run this yielded about 441,000 training pixels across the 43 training tiles. Each trained probe is then applied to every validation and test tile, producing a per-pixel boundary-probability raster for each experiment. Implementation uses scikit-learn [24].

A secondary linear probe asks whether the boundary information is *linearly* decodable from each representation. A logistic regression was trained on the same training pixels as the Random Forest, with fixed hyperparameters shared across E1–E4; because a linear model is not invariant to feature scaling, each feature was standardized with the mean and standard deviation computed on the training pixels only. The linear configurations are denoted with the suffix *-lin* (E1-lin through E4-lin). The linear probe's boundary probabilities pass through the same object extraction and evaluation as the Random Forest's, including a separate threshold selection on the validation tiles.

3.4 Object extraction

Parcel objects are recovered from the boundary-probability raster by watershed segmentation [25] using scikit-image [26]. Rather than fixing an arbitrary number of segments, markers are derived from the probability itself: the connected components of the region where boundary probability falls below a threshold t seed the watershed, which then floods the probability surface to close object outlines. The number of objects emerges from the predicted boundaries. Segments smaller than five pixels are removed by sieving. Prediction and segmentation cover the whole tile, including unlabelled area: because the reference is sparse, candidate objects outside the mapped footprint cannot be suppressed without consulting the reference itself. Any resulting bias, such as a reference parcel matching a candidate that extends into unlabelled area, applies identically to all four representations, so the paired contrasts are unaffected, although absolute object metrics carry it. The threshold t is the method's single tunable knob. It is swept over the grid 0.10–0.60 in steps of 0.05 and selected on the validation tiles by maximizing mean intersection-over-union: per-parcel IoU averaged within each tile, then across tiles, with validation tiles where a threshold yields no watershed markers dropping out of that threshold's mean. Each experiment's selected t is then frozen before any test tile is touched.

3.5 Spatial split

To prevent spatial leakage, the 0.5° region cells are partitioned whole: every tile in a cell goes to the same partition. Cells are assigned to training (target 50% of tiles), validation (20%, where t is chosen), and test (30%) under a fixed seed; in the current run this yielded 43 training, 13 validation, and 24 test tiles. No pixel from a test cell is seen during training, and t is selected without touching the test tiles. Because all imagery is from the single reference year (2017), the design uses this spatial split in place of a temporal one.

3.6 Evaluation metrics and statistical testing

All metrics are computed on the test tiles only, per tile, over the parcels lying entirely within the tile interior (partial parcels are excluded to avoid an artificial boundary at the tile edge). The candidate object with the largest intersection is selected for each reference parcel, and intersection-over-union is reported together with the over- and under-segmentation measures of Clinton et al. [27]. Object recall is the fraction of reference parcels whose best candidate reaches an intersection-over-union of at least 0.5; for parcels detected at that level, the median relative area error is also reported. Boundary quality is measured by a Boundary F1 score at a 10 m (one-pixel) tolerance, computed in vector space as the overlap length between buffered boundaries, with precision restricted to the buffered reference footprint because the reference is sparse. To assess the boundary signal free of any threshold, the pixel-wise average precision (AP) of the boundary probability is computed inside the footprint, which scores the probe's output directly and does not depend on t or on the watershed. Because paired tiles share their conditions, representations are compared with a per-tile paired Wilcoxon signed-rank test [28]. The two pre-specified primary tests are the threshold-free Boundary AP contrasts E3 versus E1 and E4 versus E2, which ask whether AlphaEarth carries additional boundary signal alone and in

fusion. Object metrics and the auxiliary contrasts (E4 versus E1 and E2 versus E1) are treated as secondary diagnostics of how that signal converts into parcel objects. For transparency, the analysis reports Holm-Bonferroni-adjusted p-values across the reported metrics within each paired contrast, while interpreting secondary tests primarily by effect size and direction. All object and boundary metrics are also reported as an overall mean and stratified by region and parcel size class, since a sub-hectare median parcel makes a one-pixel boundary tolerance demanding and the difficulty varies with parcel size.

3.7 Diagnostic figure protocols

Two qualitative figures (Figures 5 and 6) illustrate the extraction mechanism on a single test tile, selected deterministically to avoid cherry-picking. The tile is drawn from the middle third of the E3–E1 Boundary AP differences, so it is representative rather than a best case. For the map comparison (Figure 5), a zoom window is chosen from the reference alone: the 640 m window containing the most reference parcels, without consulting any prediction. In-panel counts report parcels and candidates with at least 25 pixels inside the window. For the transect panel (Figure 6a), the plotted line is the highest-scoring one under a declared criterion that favours many reference crossings and at least one merge case; the two distributional panels (Figure 6b, c) instead use every pixel and object of the tile. The object-sharing counts in Figure 6c count a reference parcel as *evaluable* when its interior has at least five pixels assigned to a positive candidate, after retaining only candidates with at least half their area inside the reference footprint; because this filter and the watershed labels are experiment-specific, the evaluable denominators differ between E1 and E3 and are smaller than the full count of reference parcels.

4 Results

All results below come from the 24 test tiles, which contain 1,956 reference parcels lying entirely within tile interiors. The watershed threshold was selected per experiment on the validation tiles and frozen before any test tile was evaluated (Table 3, column t^*). Paired differences are assessed with the per-tile Wilcoxon signed-rank test described in the Materials and Methods; p-values reported for these tests are Holm-Bonferroni adjusted across the reported metrics within each paired contrast.

4.1 Overall accuracy: boundary information without object gains

Table 3 reports the test-set summaries for the four representations under the Random Forest probe and, in the lower block, the secondary linear probe. AlphaEarth produced a stronger threshold-free boundary signal: E3 reached a Boundary AP of 0.524 against 0.406 for E1, a median paired gain of $+0.120$ per test tile ($p_{\text{Holm}} < 0.0001$; Figure 3). The fusion contrast repeated this pattern, with E4 exceeding E2 by a median $+0.107$ in Boundary AP ($p_{\text{Holm}} = 0.0001$).

That additional boundary information did not become clearly better parcel objects under the same watershed extraction. E2 had the highest mean IoU among the Random Forest probes (0.282), followed by E1 (0.268), E3 (0.260), and E4 (0.258). The E3–E1 and E4–E2 mean-IoU differences were

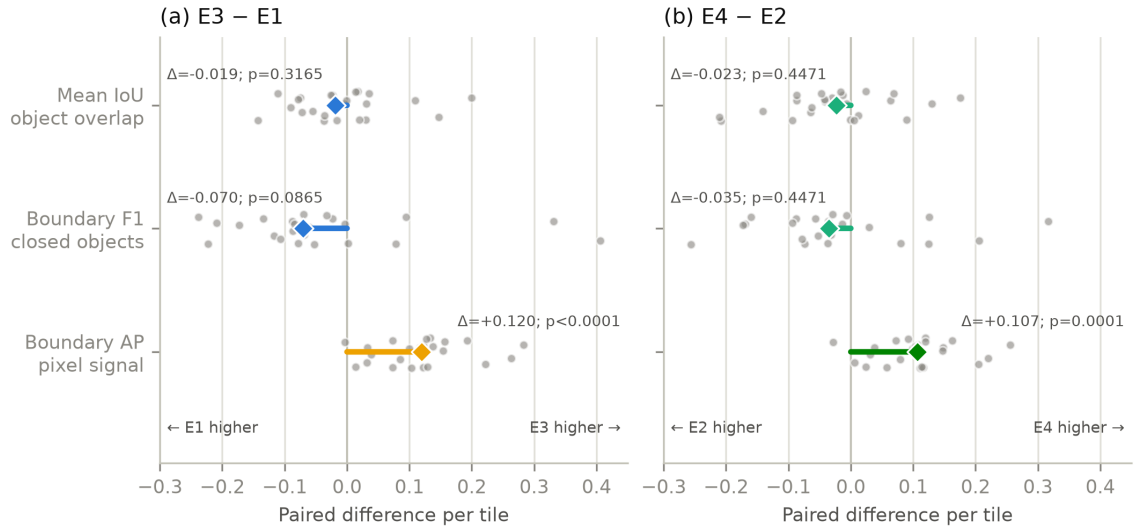

Figure 3. Paired tile-level effects for the two main contrasts: (a) E3 − E1, (b) E4 − E2. Points are tile differences, each tile at the same vertical offset in every row; diamonds and segments mark medians; annotations report median Δ and Holm-adjusted Wilcoxon p-values.

negative but not significant after within-contrast correction (median -0.019 and -0.023). Boundary F1 was lower for E3 than for E1 (0.534 versus 0.576; median -0.070; $p_{\text{Holm}} = 0.086$), consistent with weaker boundary closure, but this remains a secondary diagnostic rather than a confirmatory endpoint; the fusion contrast was weaker still (median -0.035). Among parcels detected at IoU ≥ 0.5, the median relative area error was similar for E1–E3 under the Random Forest probe (0.232–0.254) but worst for the fusion (E4, 0.311), so the added bands did not sharpen the sizing of the parcels that were found.

Table 3. Test-set performance by experiment. t^* is the validation-selected watershed threshold; metrics are test-set means except Area error, reported as the median relative area error for parcels detected at IoU ≥ 0.5. Bold marks the best value within each probe block.

Probe	Experiment	t^*	Object recall	Mean IoU	Over-seg.	Under-seg.	Area error	Boundary precision	Boundary recall	Boundary F1	Boundary AP
RF	E1 (S2 seasonal)	0.35	0.111	0.268	0.459	**0.463**	0.254	0.503	**0.712**	0.576	0.406
RF	E2 (S1+S2)	0.35	**0.165**	**0.282**	0.432	0.490	0.235	0.520	0.687	**0.578**	0.424
RF	E3 (AlphaEarth)	0.40	0.148	0.260	0.225	0.668	**0.232**	0.663	0.458	0.534	0.524
RF	E4 (fusion)	0.45	0.159	0.258	**0.215**	0.660	0.311	**0.693**	0.484	0.556	**0.529**
Linear	E1 (S2 seasonal)	0.40	0.103	0.233	0.342	0.619	0.302	0.523	0.543	0.517	0.412
Linear	E2 (S1+S2)	0.40	0.103	0.237	0.362	**0.594**	**0.246**	0.515	**0.562**	0.521	0.412
Linear	E3 (AlphaEarth)	0.35	0.154	0.254	**0.204**	0.677	0.273	**0.687**	0.433	0.523	0.530
Linear	E4 (fusion)	0.30	**0.168**	**0.268**	0.227	0.651	0.322	0.670	0.485	**0.556**	**0.540**

The component metrics show the mechanism behind this non-conversion. E3 and E4 sharply reduced over-segmentation relative to E1 and E2, but increased under-segmentation by nearly the same amount. For E3–E1, median over-segmentation decreased by -0.197 while under-segmentation increased by +0.200 (both $p_{\text{Holm}} \leq 0.0001$). Candidate counts tell the same story: the watershed produced 1,730 ± 709 E3 segments per tile and 2,397 ± 761 E4 segments, against 5,638 ± 1,298 for E1 and 4,588 ± 1,022 for E2 (mean ± SD across the 24 test tiles). The embedding fields thus support fewer, cleaner candidate regions, but they close fewer parcel boundaries between adjacent

parcels. The boundary-recall column of Table 3 reflects the same trade-off: E1's high boundary recall (0.712, against 0.458 for E3) is partly a by-product of its over-segmentation, since a dense mesh of candidate boundaries touches most reference outlines. Absolute object accuracy remained low for every representation; object recall at IoU ≥ 0.5 never exceeded 0.168.

4.2 Stratification by parcel size

The information–object gap varied with parcel size (Figure 4). The large-parcel stratum was small (three test tiles), so it is best read as exploratory; within that stratum, AlphaEarth and fusion had better boundary diagnostics, with RF E3 and E4 reaching Boundary F1 of 0.731 and 0.733, against 0.509 for E1, and Boundary AP of 0.678 and 0.689, against 0.458 for E1. The mid-parcel class was mixed: E4 matched or slightly exceeded the seasonal baselines in Boundary F1, while E3 remained lower than E1/E2. The small-parcel class drove the main failure mode. E3 and E4 kept higher Boundary AP than E1/E2, but their Boundary F1 and mean IoU fell below the seasonal baselines, consistent with adjacent small parcels merging when probability ridges are too smooth.

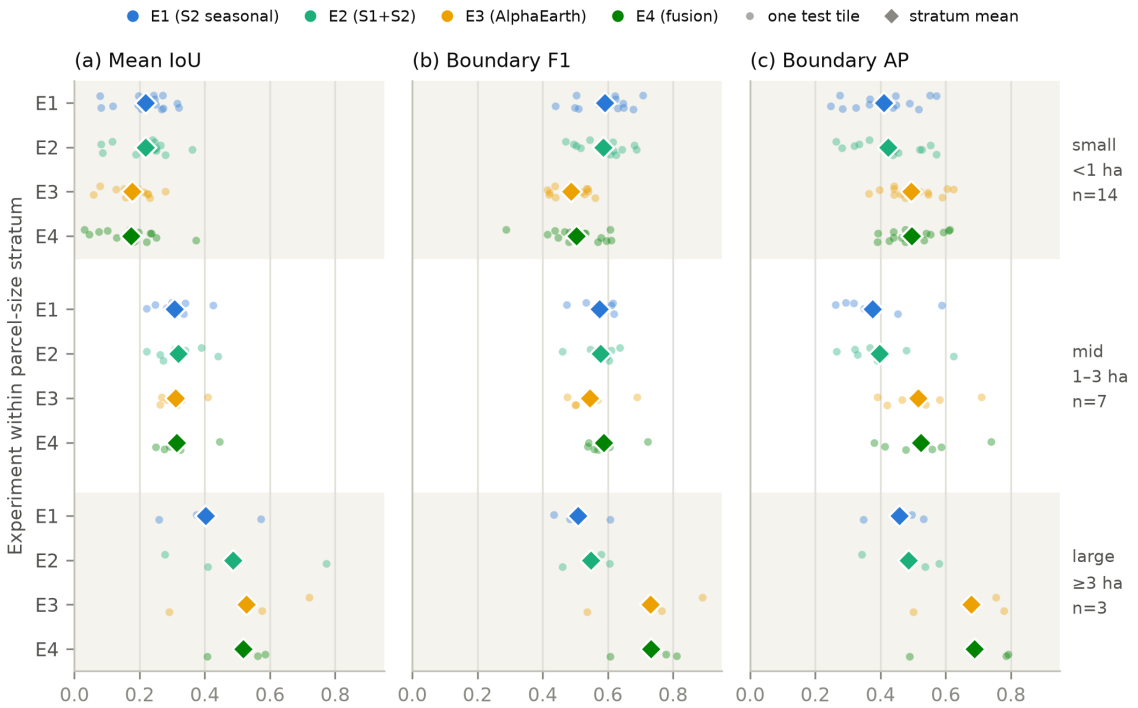

Figure 4. Size-stratified Random Forest performance on the test tiles. Rows show tile size classes (small n=14, mid n=7, large n=3), and columns show (a) mean IoU, (b) Boundary F1, and (c) Boundary AP. Small points are individual tiles; diamonds mark stratum means.

4.3 Linear decodability

The secondary linear probe confirms that the embedding's boundary signal is not dependent on a non-linear decoder. A logistic regression on E3 reached Boundary AP 0.530, essentially the same as the Random Forest on E3 (0.524), and its per-tile advantage over E1-lin was +0.121 ($p_{\text{Holm}} < 0.0001$). The best linear configuration was E4-lin, with the highest object recall (0.168) and the highest Boundary AP (0.540). Even there, however, no mean-IoU difference was detected against the linear seasonal baseline (E4-lin minus E2-lin: median +0.009, $p_{\text{Holm}} = 0.449$). The embedding makes boundary information accessible, but this simple watershed does not reliably turn that information into more accurate objects.

4.4 Qualitative comparison

Figure 5 compares the predicted boundary-probability fields and watershed candidates of E1 and E3 on a test tile drawn from the middle third of the per-tile AP differences. On this representative tile, E3 has higher Boundary AP than E1 (0.547 versus 0.447) but lower Boundary F1 and mean IoU. The seasonal-feature field is granular and produces many more candidate regions; the embedding field is smoother and produces fewer, larger candidates. The zoom row makes the consequence concrete: in the window with the densest cluster of reference parcels, twelve parcels overlap 33 E1 candidates but only 18 E3 candidates.

Figure 6 quantifies the same mechanism without maps, on the same tile. Along a transect crossing 14 reference boundaries, the E1 profile exceeds its t^* at all 14 crossings, whereas the E3 profile stays below its own t^* at 4 of them. In that condition, the watershed merges adjacent parcels. The tile-wide distributions close both ends of the dissociation: the score ECDFs separate boundary from interior pixels more cleanly for E3, consistent with its higher Boundary AP, while the area-weighted size distribution shows that half of the E3 candidate area lies in objects up to 27.8 ha against 3.1 ha for the reference parcels (E1: 2.2 ha), and 58 of 93 evaluable reference parcels share their modal E3 candidate with a neighbour (E1: 36 of 122).

5 Discussion

5.1 What the information–object gap means

The central finding is a gap between boundary information and object delineation. Judged by threshold-free Boundary AP, the annual AlphaEarth embedding carries more per-pixel boundary information than the season-resolved features engineered from Sentinel-1/2: E3 exceeds E1 by a median +0.120 per tile, and E4 exceeds E2 by +0.107 (Figure 3; $p_{\text{Holm}} < 0.0001$ and $p_{\text{Holm}} = 0.0001$, respectively, after within-contrast correction). This advantage survives under a linear decoder, which means the signal is explicit in the representation, rather than recoverable only by a non-linear Random Forest.

The object-level result is more cautious. AlphaEarth does not clearly outperform the seasonal baselines after the shared watershed extraction, and the RF mean-IoU contrasts are small and

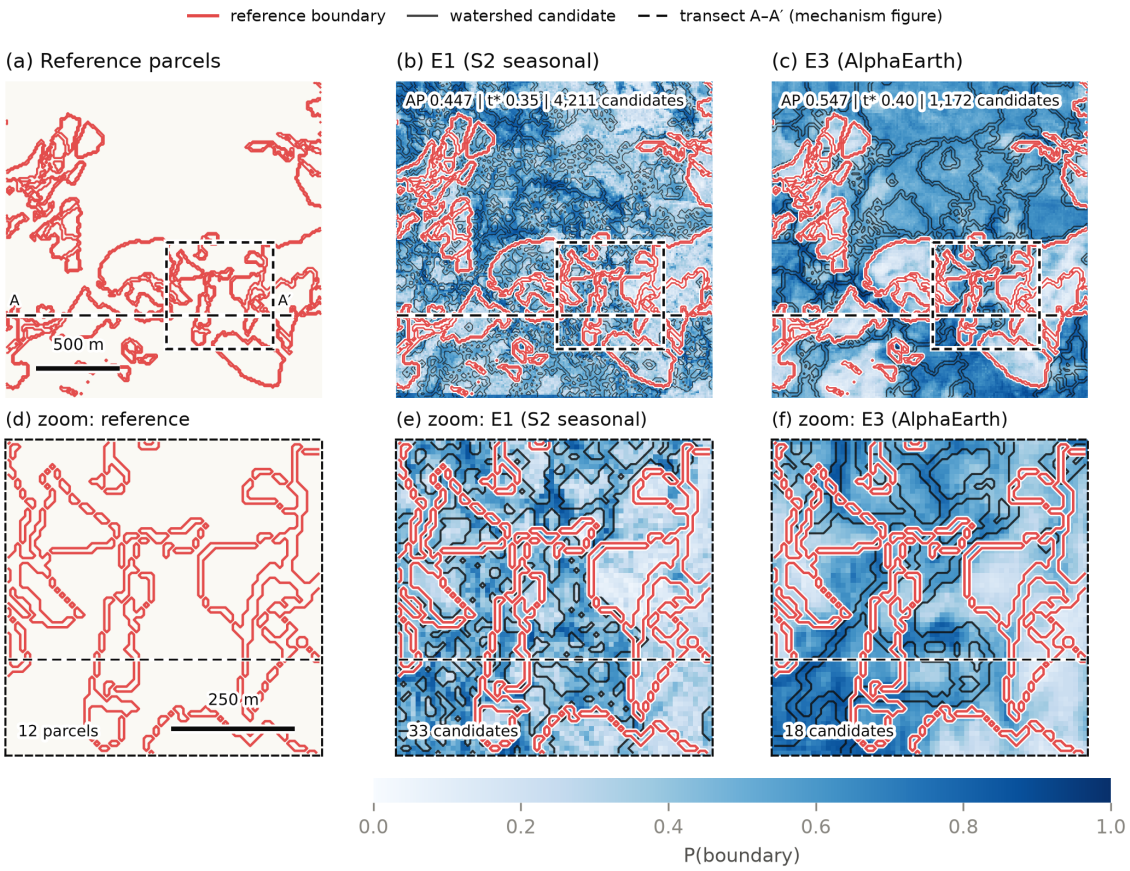

Figure 5. Boundary-probability fields and watershed candidates for the representative test tile: (a–c) communication crop, (d–f) reference-derived zoom window (dashed box). Red lines are reference parcel boundaries, black lines are candidate boundaries, and the horizontal dashed line marks transect A–A′, analysed in Figure 6.

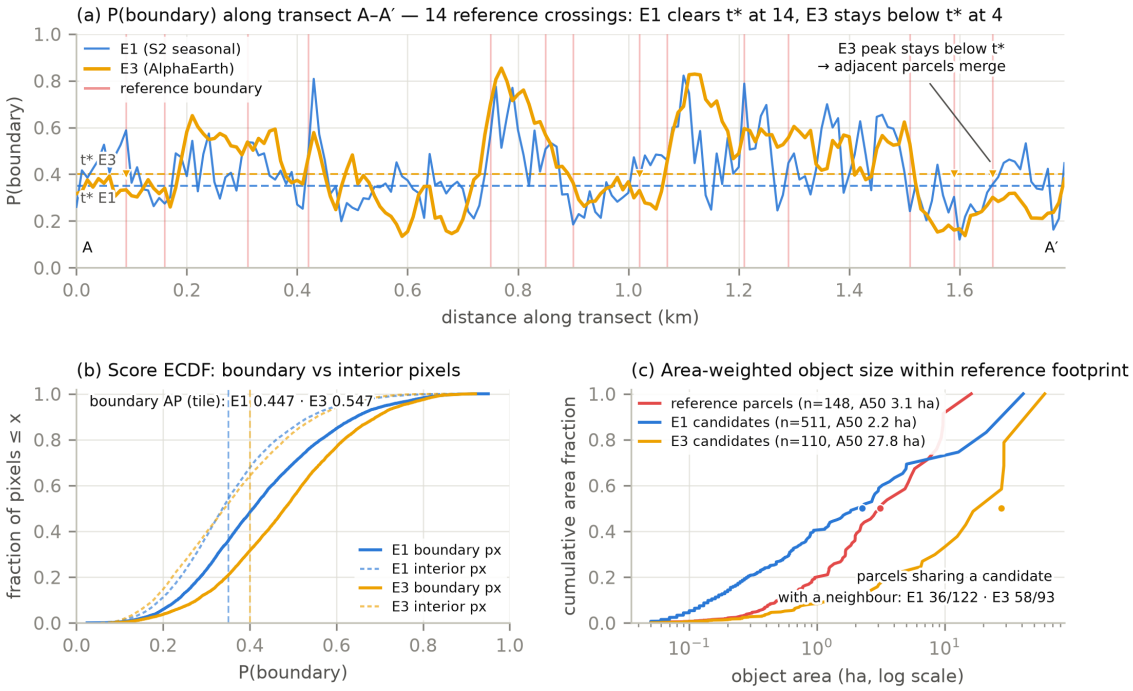

Figure 6. Non-map diagnostics for the representative test tile. Panels show (a) boundary-probability profiles along transect A–A′ (marked in Figure 5), (b) ECDFs for boundary and interior pixels, and (c) area-weighted candidate-size distributions and object-sharing counts.

negative. A non-significant contrast is not evidence of equivalence; it only means that this test set did not support a clear object-level gain. The largest object-level penalty is Boundary F1 for E3 against E1, but this remains a secondary diagnostic rather than a confirmatory endpoint. The component metrics explain why: the embedding reduces over-segmentation but increases under-segmentation. In other words, the watershed produces fewer, cleaner regions from the smoother embedding field, but adjacent coffee parcels are more likely to merge (Figures 5 and 6). The limitation exposed here is thus less about missing boundary content than about the spatial structure required to convert ranked boundary evidence into closed, correctly sized parcel objects.

The linear probe reinforces this interpretation. E3-lin reaches Boundary AP 0.530, close to the Random Forest's 0.524 on E3, and E4-lin gives the highest Boundary AP and object recall in the table. Yet neither linear nor Random Forest probing turns the embedding's AP advantage into a statistically clear mean-IoU advantage. Absolute object accuracy is also low for every configuration, with object recall at IoU ≥ 0.5 topping out at 0.168. For this sub-hectare coffee setting, the result is not an operational delineation method; it is a diagnosis of where the representation helps and where the extraction fails.

5.2 Practical implications

For the practitioner question posed in the Introduction, whether one can skip local seasonal compositing and delineate parcels from the ready-made embedding alone, the answer splits by task. For per-pixel boundary ranking within mapped coffee parcels, the embedding plus a light model is the stronger representation in this study. For parcel objects from a simple watershed, it is not a drop-in replacement for seasonal features: it improves Boundary AP, but the object metrics remain comparable or worse, especially for small parcels.

The size stratification makes this practical boundary clear (Figure 4), but its strata are descriptive rather than confirmatory. In the three large-parcel test tiles, the embedding's information advantage suggests better Boundary F1 and AP. On the small-parcel tiles that dominate the sample, E3 and E4 still rank boundary pixels better, but seasonal features keep stronger boundary closure and higher mean IoU. These are the tiles where automated parcel delineation would be most useful, so the current embedding-plus- watershed pipeline should be treated as diagnostic rather than production-ready.

The improvement opportunity follows directly. Since boundary information is present and linearly accessible, this fixed watershed exposes a conversion problem rather than a lack of ranked boundary evidence. Decoders with spatial structure, including contour completion, edge sharpening, instance segmentation, or promptable segmenters [29, 30] conditioned on the embedding, are natural candidates to test whether AlphaEarth's Boundary AP advantage can be converted into objects. The current study deliberately held the decoder simple and identical across representations so that the comparison remained attributable to the features; it therefore does not show whether a stronger decoder would recover the signal. The probing design itself carries beyond this case study: because the probe, the extraction, and the statistics are fixed, the same protocol can benchmark any new embedding product, or any other crop and region with a parcel reference, by swapping in a single feature stack.

The fusion experiment (E4) should be read in the same spirit. Concatenating the seasonal features

with the embedding is a deliberate control rather than an attempt at an optimal fusion: feeding all sources to the same fixed probe keeps every difference attributable to the inputs, at the cost of leaving the embedding's information to compete with the seasonal bands inside a model that cannot learn a joint representation. E4 accordingly tracks E3 on the boundary signal without a clean object-level gain. Learned multimodal fusion, jointly training a representation over the raw streams rather than concatenating fixed features [31], is a natural next step, but it would break the controlled attribution this study depends on and is therefore left to future work.

5.3 Limitations

Six limitations bound these conclusions.

Reference quality and sparsity. The reference is a semi-automatic mapping with its own boundary error, and it is sparse: only coffee parcels are delineated, so training, boundary precision, and Boundary AP are all confined to the mapped footprint. Object candidates, in contrast, are extracted over the whole tile, so absolute object metrics also reflect matches against unlabelled area. Because candidate density differs across representations, this sparsity can affect object-level contrasts as well as absolute scores. Metrics computed against such a reference score agreement with CONAB's delineation, not cadastral truth.

Scope. The scope is a single crop, state, and year: coffee in Minas Gerais, 2017, fixed by the reference itself; whether the same information–object gap generalizes to other crops, landscapes, or years is untested.

Minimal extraction. The object extraction is deliberately minimal: one watershed with a single validated threshold. The results quantify what a simple, fixed extraction recovers from each representation, not the ceiling of any representation under a stronger decoder.

Absolute accuracy. Absolute object accuracy is low for every configuration, reflecting how demanding sub-hectare parcels are at a 10 m grid with a one-pixel boundary tolerance.

Input asymmetries. The seasonal features are computed from TOA (Level-1C) reflectance, whereas AlphaEarth applies its own preprocessing to its source streams, an asymmetry that could penalize the seasonal baseline. The fixed Random Forest also faces representations of 16 to 88 bands, and probe capacity could interact with dimensionality; the linear probe, which reproduces the same conclusions, serves as a partial control. Finally, zero-imputation of pixels left unobserved by a composite is applied uniformly, but only the seasonal sources contain such pixels. The AlphaEarth rasters contain no missing values, so imputed zeros occur only in the seasonal representations.

Statistical scope. Holm correction is applied within each paired contrast, with the Boundary AP contrasts pre-specified as primary endpoints; this family definition is itself a choice, but the primary result would survive even a global Holm correction across all roughly 48 reported tests ($p \approx 0.0001 \times 48 \approx 0.005$). The 24 test tiles come from 17 spatial cells, so the paired tests support tile-level diagnostics under the stratified sampling design. Non-significant object contrasts should not be read as equivalence tests, and the size-stratified results are descriptive, especially the large-parcel stratum ($n = 3$).

6 Conclusion

This study asked whether a ready-made annual embedding preserves the parcel-boundary information that season-resolved Sentinel-1/2 features carry, using coffee parcels in Minas Gerais (2017) as a demanding sub-hectare test case. A supervised probing design with identical lightweight probes, a fixed watershed extraction with one spatially validated threshold, and paired per-tile statistics on held-out cells kept every difference attributable to the representations themselves.

The answer is an information–object gap. AlphaEarth carried more per-pixel boundary information than the seasonal features: under the Random Forest probe, E3 reached Boundary AP 0.524 versus 0.406 for E1, with a median paired gain of +0.120 ($p_{\text{Holm}} < 0.0001$). The same signal was linearly accessible (E3-lin Boundary AP 0.530). But this advantage did not translate into a clear object-level advantage under the shared watershed. Seasonal features retained the best RF mean IoU (E2 = 0.282) and Boundary F1 (E2 = 0.578), while AlphaEarth reduced over-segmentation at the cost of higher under-segmentation. The embedding therefore appears to preserve boundary evidence, but its smoother spatial field is not enough, by itself, to close small adjacent parcel outlines.

For practice, the ready-made embedding is promising for boundary ranking within mapped coffee parcels, but not yet for operational parcel delineation with this extraction: no configuration reached object recall above 0.168 at IoU ≥ 0.5. Future work should test whether spatially structured decoders can convert AlphaEarth's boundary-information advantage into accurate parcel objects, and whether the same pattern holds beyond coffee in Minas Gerais.

Acknowledgments

Author Contributions

- **Abreu Filho**: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Data curation, Writing – original draft, Visualization.
- **Timbó Elmiro**: Conceptualization, Methodology, Writing – review & editing, Supervision.
- **Nero**: Conceptualization, Methodology, Writing – review & editing, Supervision.
- **Leitão**: Methodology, Writing – review & editing.

Funding

This research received no external funding.

Conflicts of Interest

All authors have approved the submission and declare no competing interests.

Generative AI and Figure Provenance

OpenAI Codex (GPT-5) was used during manuscript preparation to audit the text for AI-stylistic leakage, edit wording, check consistency between the manuscript and submission files, and prepare

disclosure language. The authors reviewed and remain responsible for all scientific content, analyses, references, and conclusions. No AI-assisted tool is listed as an author or cited as a source. All figures and tables were generated from the study data and code; no AI-generated images or multimedia are included.

Data Availability

The CONAB reference parcels are openly available on Source Cooperative as the `br_conab` field-boundary dataset distributed by the Field Boundaries for Agriculture (fiboa) project (https://source.coop/fiboa/data/br_conab). Sentinel-1, Sentinel-2, and the AlphaEarth Satellite Embedding V1 are accessed through Google Earth Engine. All code to reproduce the pipeline, including reference extraction, tile sampling, feature download, probing, validation, and every table and figure, is available in the study repository at https://alphaearth-mg.abreufilho.com.br; all intermediate data are regenerable from the code with fixed seeds.

Supplementary Materials

No supplementary materials accompany this article.

References

1. d'Andrimont R, Claverie M, Kempeneers P, et al. AI4Boundaries: an open AI-ready dataset to map field boundaries with Sentinel-2 and aerial photography. Earth System Science Data 2023;15:317–29.
2. Kerner H, Chaudhari S, Ghosh A, et al. Fields of The World: A Machine Learning Benchmark Dataset for Global Agricultural Field Boundary Segmentation. arXiv preprint arXiv:2409.16252 2024.
3. Lowder SK, Skoet J, and Raney T. The Number, Size, and Distribution of Farms, Smallholder Farms, and Family Farms Worldwide. World Development 2016;87:16–29.
4. Waldner F and Diakogiannis FI. Deep learning on edge: Extracting field boundaries from satellite images with a convolutional neural network. Remote Sensing of Environment 2020;245:111741.
5. Persello C, Tolpekin VA, Bergado JR, and By RA de. Delineation of agricultural fields in smallholder farms from satellite images using fully convolutional networks and combinatorial grouping. Remote Sensing of Environment 2019;231:111253.
6. Persello C, Grift J, Fan X, et al. AI4SmallFarms: A Dataset for Crop Field Delineation in Southeast Asian Smallholder Farms. IEEE Geoscience and Remote Sensing Letters 2023;20:2505705.
7. Cong Y, Khanna S, Meng C, et al. SatMAE: Pre-training Transformers for Temporal and Multi-Spectral Satellite Imagery. In: *Advances in Neural Information Processing Systems*. Vol. 35. 2022.

8. Jakubik J, Roy S, Phillips CE, et al. Foundation Models for Generalist Geospatial Artificial Intelligence. arXiv preprint arXiv:2310.18660 2023.

9. Szwarcman D, Roy S, Fraccaro P, et al. Prithvi-EO-2.0: A Versatile Multi-Temporal Foundation Model for Earth Observation Applications. arXiv preprint arXiv:2412.02732 2024.

10. Tseng G, Cartuyvels R, Zvonkov I, Purohit M, Rolnick D, and Kerner H. Lightweight, Pre-trained Transformers for Remote Sensing Timeseries. arXiv preprint arXiv:2304.14065 2023.

11. Xiao A, Xuan W, Wang J, et al. Foundation Models for Remote Sensing and Earth Observation: A Survey. arXiv preprint arXiv:2410.16602 2024.

12. Brown CF, Kazmierski MR, Pasquarella VJ, et al. AlphaEarth Foundations: An embedding field model for accurate and efficient global mapping from sparse label data. arXiv preprint arXiv:2507.22291 2025.

13. Ma Y, Shen Y, Swatantran A, and Lobell DB. Harvesting AlphaEarth: Benchmarking the Geospatial Foundation Model for Agricultural Downstream Tasks. arXiv preprint arXiv:2601.00857 2026.

14. fiboa. br_conab: Field boundaries for Brazilian crops (CONAB). Source Cooperative dataset. `https://source.coop/fiboa/data/br_conab`. 2024.

15. Alain G and Bengio Y. Understanding intermediate layers using linear classifier probes. arXiv preprint arXiv:1610.01644 2016.

16. Moreira MA, Rudorff BFT, Barros MA, Faria VGCd, and Adami M. Geotecnologias para mapear lavouras de café nos estados de Minas Gerais e São Paulo. Engenharia Agrícola 2010;30:1123–35.

17. Gorelick N, Hancher M, Dixon M, Ilyushchenko S, Thau D, and Moore R. Google Earth Engine: Planetary-scale geospatial analysis for everyone. Remote Sensing of Environment 2017;202:18–27.

18. Drusch M, Del Bello U, Carlier S, et al. Sentinel-2: ESA's Optical High-Resolution Mission for GMES Operational Services. Remote Sensing of Environment 2012;120:25–36.

19. Torres R, Snoeij P, Geudtner D, et al. GMES Sentinel-1 mission. Remote Sensing of Environment 2012;120:9–24.

20. Camargo ÂPd and Camargo MBPd. Definição e esquematização das fases fenológicas do cafeeiro arábica nas condições tropicais do Brasil. Bragantia 2001;60:65–8.

21. McFeeters SK. The use of the Normalized Difference Water Index (NDWI) in the delineation of open water features. International Journal of Remote Sensing 1996;17:1425–32.

22. Huete A, Didan K, Miura T, Rodriguez EP, Gao X, and Ferreira LG. Overview of the radiometric and biophysical performance of the MODIS vegetation indices. Remote Sensing of Environment 2002;83:195–213.

23. Breiman L. Random Forests. Machine Learning 2001;45:5–32.

24. Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research 2011;12:2825–30.

25. Vincent L and Soille P. Watersheds in digital spaces: an efficient algorithm based on immersion simulations. IEEE Transactions on Pattern Analysis and Machine Intelligence 1991;13:583–98.

26. Walt S van der, Schönberger JL, Nunez-Iglesias J, et al. scikit-image: image processing in Python. PeerJ 2014;2:e453.

27. Clinton N, Holt A, Scarborough J, Yan L, and Gong P. Accuracy Assessment Measures for Object-based Image Segmentation Goodness. Photogrammetric Engineering & Remote Sensing 2010;76:289–99.

28. Wilcoxon F. Individual Comparisons by Ranking Methods. Biometrics Bulletin 1945;1:80–3.

29. Kirillov A, Mintun E, Ravi N, et al. Segment Anything. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 2023:3992–4003. DOI: 10.1109/ICCV51070.2023.00371.

30. Lavreniuk M, Kussul N, Shelestov A, et al. Delineate Anything: Resolution-Agnostic Field Boundary Delineation on Satellite Imagery. arXiv preprint arXiv:2504.02534 2025.

31. Hong D, Gao L, Yokoya N, et al. More Diverse Means Better: Multimodal Deep Learning Meets Remote-Sensing Imagery Classification. IEEE Transactions on Geoscience and Remote Sensing 2021;59:4340–54.