PRODUCT 1

Detailed Sampling Plan and Sampling Improvement Strategies

Technical, Methodological, and Progress Report

Improvement of Pasture Mapping in the Cerrado Biome

Project GR000153-73077-CARGILL-BRAZIL

WRI Brasil - FUNAPE Agreement, with Technical Execution by LAPIG/UFG

Scientific coordination: Prof. Dr. Laerte G. Ferreira

Goiânia, August 17, 2026

Executive Summary

Purpose of this deliverable. This report consolidates the strategy, the databases already assembled, and the quality controls planned for the collection, expansion, and curation of training and validation samples for the new 10 m pasture mapping of the Cerrado biome. The threshold analyses, the final SOM configuration, the visual inspection, and the field campaign remain under development; their results are not presented as complete.

The work is based on 701 field areas provided by Embrapa Cenargen. Of these, 608 have a defined class according to seven pasture condition categories, and 93 remain without a consolidated label. For all areas, Sentinel-2 time series and annual vectors from the AlphaEarth Foundations Satellite Embedding dataset were organized, covering 2020-2025.

The first expansion stage brings together 62,009 candidate points: 5,453 points from MapBiomas Collection 4 Sentinel-2 training samples, 6,556 points from the Landsat-based MapBiomas validation database, and 50,000 points randomly drawn from pastures persistently classified in Collection 4 between 2020 and 2025. The comparison of the 50,000 points with MapBiomas Collection 11 indicated agreement for 44,943 points and disagreement for 5,057; both scenarios - inclusive and restricted to agreement between products - will be kept in the sensitivity analysis.

The filtering is structured in two complementary stages: (i) similarity between embeddings, with tests of absolute thresholds, nearest-neighbor agreement, and margin between candidates; and (ii) organization of the Sentinel-2 time series into Self-Organizing Maps (SOM), evaluating the coherence of categories within neurons and their neighborhoods. The approved set will undergo sample inspection using pan-sharpened CBERS-4A images, with a nominal resolution close to 2 m, before feeding the Random Forest models.

The image-based inspection will serve as a general quality control for the pseudo-labels. The first version of the map will subsequently be evaluated using a specific statistical design based on imagery and field data. An initial campaign of approximately 100 localities is planned, prioritizing Goiás for logistical reasons, complemented by points in other regions of the Cerrado. This campaign will be treated as an initial, targeted field validation, not as a stand-alone estimate of accuracy for the seven categories across the entire biome.

Contractual Framing and Purpose

The technical consulting services agreement between WRI Brasil and FUNAPE, executed by LAPIG/UFG, designates this deliverable as Product 1 - Detailed sampling plan and sampling improvement strategies, with a delivery date of August 17, 2026. Annex A details the construction of the georeferenced database, the metadata, the classification criteria, the sampling uncertainty, and the corresponding technical report.

At this stage, the operational purpose is to document a traceable strategy to: expand the sample set from field references; reduce pseudo-labeling errors through independent filters in different representation spaces; select samples for visual inspection; and prepare a database suitable for Random Forest training and subsequent evaluations.

> **Scope limitation of this version.** This report does not present the similarity thresholds, the SOM parameters, the CBERS-4A inspection, the field campaign, the final training set, or the map validation as finalized. These components will be consolidated after the sensitivity analyses and the final statistical design.

Status of Activities

Table 1. Status of project activities as of August 17, 2026

Component	Status as of 08/17/2026	Evidence or Next Steps
1. Embrapa Base	Assembled	701 polygons of approximately 1 ha; 608 classified and 93 without a consolidated class.
2. Sentinel-2 Series	Extracted	Monthly composites 2020-2025 for references and candidates; final audit of valid observations and edge months in progress.
3. Annual Embeddings	Extracted	64-dimensional vectors per year; verification and correction of the mean vector norm before final similarity calculation.
4. Candidate Points	Assembled	62,009 candidates with provenance preserved; scenarios with and without the 5,057 points that disagree with Collection 11.
5. Embedding-Based Filtering	Preliminary	Top-1/Top-3, md3, and thresholds under sensitivity analysis.
6. SOM-Based Filtering	Under development	Initial 12 x 12 grid; dimensionality, normalization, neighborhood, and Bayesian criteria under evaluation.
7. CBERS-4A Inspection	Planning	Stratified sampling after the embedding-SOM selection.
8. Field Data Collection	Planning	Initial campaign of about 100 localities; spatial design and allocation still being defined.

1. Embrapa Cenargen Reference Base

Characteristics and Use

The database with the highest quality control consists of 701 sample areas surveyed in the field by Embrapa Cenargen. Each unit corresponds to a polygon of approximately 1 ha. Of the total, 608 have a defined class according to seven categories established by the Embrapa team; the 93 units without a defined class are retained with their own identification.

The 608 classified samples constitute the references for label transfer and coherence analysis. The 93 unlabeled samples may participate in unsupervised steps, such as SOM training, but cannot define neuron categories, contribute to Top-k voting, or be counted as classified references.

Harmonization of the Class Dictionary

The preliminary technical materials record different versions of the taxonomy - a five-class proposal and a later implementation with seven categories. To avoid inconsistency among labels, ecological criteria, and field names, the final database must contain a single, versioned dictionary approved by the Embrapa-LAPIG team, including: code, name, operational definition, quantitative criteria, year of observation, party responsible for the assignment, and change history.

> **Integrity rule.** No analysis of category distribution or label transfer will be presented as definitive before reconciliation between the 608 samples actually classified and the final dictionary of the seven categories.

2. Sentinel-2 Time Series

Monthly Compositing in the Reference Areas

For each Embrapa area and for each spectral index considered, a monthly series was built between January 2020 and December 2025. The value for a given month is calculated using observations from a centered moving window of approximately 60 days: the second half of the previous month, the full reference month, and the first half of the following month.

The monthly summary is the joint median of all valid pixel-date pairs within the polygon, after excluding observations contaminated by clouds and shadows. In a 1 ha polygon, an ideal 10 m grid contains approximately 100 pixels; the effective number depends on the geometry, the alignment to the grid, and the full inclusion of pixels intersected by the boundary.

$$\tilde{I}(i,k,m) = \text{median} \{ I(i,k,p,t) : p \in \text{polygon } i;\ t \in \text{window } m;\ \text{valid observation} \}$$

The joint median implicitly weights each date by the number of valid pixels available. For this reason, the metadata database must record, for each monthly composite, the total number of pixel-date pairs, the number of dates represented, the valid spatial proportion, and the presence of gaps.

Indices and Resolution

The workflow under evaluation uses indices related to vigor, moisture, senescence, and soil exposure; the configuration presented includes NDVI, EVI, NDII, NDWI, PSRI, and CAI. The final list, formulas, source bands, and resampling conventions will be consolidated in the data dictionary. Indices that use native 20 m Sentinel-2 bands require an explicit statement of resampling to 10 m.

- Source: Sentinel-2 collections accessed and processed in Google Earth Engine.
- Temporal unit: monthly composite with a moving window of approximately 60 days.
- Statistic: joint median of valid pixel-date pairs.
- Period: January 2020 to December 2025.

- Pending control: rule for January 2020 and December 2025, minimum number of observations, and masking audit.

3. AlphaEarth Foundations Satellite Embeddings

Extraction and Spatial Aggregation

For each reference area, annual AlphaEarth Foundations Satellite Embedding vectors for 2020-2025 were extracted via Google Earth Engine. The product represents each 10 m pixel with a 64-dimensional vector (A00-A63), derived from multiple Earth observation sources and referring to the trajectory of a calendar year (Brown et al., 2025).

For each polygon, the component-wise mean of the vectors of the included pixels was calculated. Thus, the result is not a new 10 m pixel, but an annual vector representative of the spatial support of the approximately 1 ha polygon.

$$\bar{e}(i,a,d) = (1 / n(i,a)) \times \text{sum of the values } e(p,a,d), \text{ for } d = 0,\ldots,63$$

Pixels partially intersected by the polygon were included in full, without weighting by the covered area fraction. This rule expands the effective spatial support and can increase land-cover mixing in narrow, irregular areas or areas near transitions.

Normalization and Similarity

The original embeddings are described as vectors with unit Euclidean norm. However, the spatial average of unit vectors does not necessarily preserve a norm equal to 1. The dot product between mean vectors is equivalent to cosine similarity only when both have unit norm.

$$\cos(x,y) = (x \cdot y) / (||x||_2 \times ||y||_2)$$

Before the final filtering, the distributions of the L2 norms of the mean vectors will be calculated. The comparison will use explicit cosine similarity or, equivalently, the dot product after L2 renormalization. This procedure avoids confusing directional similarity with the magnitude and spatial homogeneity of the mean.

The field label remains anchored to the year of observation. Vectors from other years will not automatically receive the same label without verification of temporal stability, since genuine changes in cover or condition can occur between 2020 and 2025.

4. Database Expansion: Candidate Points

Sources and Provenance

Table 2. Sources of candidate points

Source	Count	Nature and Caveats
MapBiomas Collection 4 - Sentinel-2	5,453	Pasture points visually inspected and used in generating the previous product; may carry methodological and spatial dependence.
MapBiomas validation database - Landsat	6,556	Subset of pastures interpreted within a database of approximately 85,000 samples; if used in training, it is no longer independent for validation.
Persistent pasture - Collection 4	50,000	Points sampled from pixels classified as pasture in every year from 2020-2025; these are pseudo-labeled candidates, not ground truth.
Total	62,009	Inclusive universe of candidates before the similarity, SOM, and visual inspection filters.

The 50,000 points were allocated proportionally across three vigor strata defined for 2025 based on Sentinel-2 imagery and a methodology compatible with MapBiomas: approximately 35% in low vigor, 44% in medium vigor, and 21% in high vigor. The actual integer counts drawn will be preserved in the final inventory.

Persistence over six years increases the plausibility that a point represents stable pasture, but it does not eliminate classification error and may under-represent recent conversions, rotational systems, and spectrally ambiguous areas. Each record will retain the source, the year, the vigor stratum, the persistence condition, and the identifiers needed to audit duplicates and spatial proximity.

Extraction of Candidate Points

For each candidate, the Sentinel-2 series and the embeddings were extracted using a 3 x 3 pixel window, corresponding to 900 m^2 or 0.09 ha. For the indices, each monthly composite is the joint median of the valid pixel-date pairs from the nine pixels within the approximately 60-day moving window. For the embeddings, each annual vector is the component-wise mean of the nine 64-dimensional vectors.

The 3 x 3 window may cross boundaries between pasture and other land covers, especially at edges, narrow corridors, and heterogeneous mosaics. This mixing will be recorded as a limitation and may be controlled through spatial homogeneity metrics or inspection of borderline cases.

Agreement with MapBiomas Collection 11

The 50,000 points sampled from persistent pasture in Collection 4 Sentinel-2 were compared against the Landsat-based MapBiomas Collection 11 map. Collection 11 corroborated the pasture class at 44,943 points (89.886%); at 5,057 points (10.114%) there was disagreement.

Disagreement does not prove that a point is not pasture, since neither map constitutes ground truth. It represents additional evidence of uncertainty and will be used in the sensitivity analyses.

Table 3. Sensitivity scenarios based on agreement with MapBiomas Collection 11

Scenario	Points from the 50k Source	Total Candidates	Interpretation
Inclusive	50,000	62,009	Keeps the disagreeing cases to test robustness, rejection, and out-of-domain behavior.
Agreement	44,943	56,952	Restricts the database to points corroborated as pasture by both collections.

5. Filtering Stage 1: Embedding Similarity - Comparison and Preliminary Assignment

Each candidate is compared with the Embrapa references by similarity in the 64-dimensional space. After correcting the norm of the mean vectors, the metric adopted will be cosine similarity, calculated directly or as the dot product of re-normalized vectors.

The most similar reference defines the preliminary Top-1 label. The three highest values form the Top-3, and the md3 indicator records how many of the three neighbors share the Top-1 category. An md3 of 3 represents local agreement, but not accuracy or calibrated probability.

Sensitivity Scenarios

No single threshold will be adopted a priori. Absolute thresholds, Top-3 agreement, separation margins, and agreement with Collection 11 will be compared.

Table 4. Illustrative threshold scenarios for the embedding-similarity filter

Scenario	Illustrative Criterion	Objective
A	s1 above threshold	Test the coverage obtained through absolute similarity.
B	md3 = 3	Retain neighborhoods with complete agreement among the top three.
C	s1 above threshold and md3 = 3	Combine absolute proximity and local coherence.
D	C + agreement with Collection 11	Evaluate a more conservative subset.
E	D + minimum margin s1 - s2	Reduce cases at boundaries between categories.

Retention will be reported by category, origin, vigor, region, and scenario. A high threshold increases expected purity but reduces coverage and may disproportionately eliminate rare categories. Whenever sample size permits, calibration will be category-specific and evaluated with spatial separation.

Correct interpretation. This stage selects pseudo-labels with the highest consistency in the embedding space. It does not replace independent validation, nor does it guarantee that the seven categories are distinguishable across all regions and years.

6. Filtering Stage 2: Self-Organizing Map

Objective and Initial Configuration

Candidates approved in the embedding stage will be combined with the 701 Embrapa areas and submitted to a Self-Organizing Map (SOM). The SOM organizes similar time series on a two-dimensional grid, preserving neighborhood relationships in feature space (Kohonen, 1990). The initial configuration considers a 12 x 12 neuron grid, totaling 144 units.

Each sample will be assigned to its best-matching neuron. The classes assigned by the embeddings will then be checked against the composition of categories within the neuron and its neighborhood. Since the candidates far outnumber the field references, the output will be interpreted as a spectral-temporal consistency filter, not as independent validation.

Dimensionality and Data Preparation

The statement that each neuron holds a 12-value vector is only correct for a single index over a single annual cycle. With J indices, a multivariate annual vector will have 12 x J dimensions; with six complete years, it will have 72 x J dimensions. Alternative configurations will be compared, such as one SOM per index, a multivariate annual SOM, or an inter-annual monthly climatology.

- Standardize each variable to prevent scale differences among indices from dominating the distance metric.
- Define the treatment of missing values and months with low data availability.
- Test different grid dimensions, seeds, and initializations.
- Record quantization and topographic errors, neuron occupancy, and solution stability.
- Prevent the 93 unlabeled references from contributing to neuron labeling.

Cleaning Rule

The methodological presentation considers two families of Bayesian smoothing: Normal-Normal, applied to class proportions, and Dirichlet-Multinomial, applied to category counts within the neuron and its neighborhood. The Dirichlet-Multinomial formulation is, in principle, more consistent with categorical data and will be prioritized, without excluding the sensitivity comparison.

For each neuron, both the local composition and the neighborhood-smoothed composition will be estimated. Candidates will be classified into three states: keep, forward for inspection, or reject. The thresholds, the alpha 0 pseudo-count, the neighborhood size, and the minimum number of references per neuron will be defined through sensitivity analysis.

> **Self-reference risk.** If class frequencies are calculated predominantly from the candidates' own pseudo-labels, the SOM may simply reproduce the embedding assignment. Evidence from the Embrapa references and the internal coherence of the candidates will be reported separately.

7. Visual Inspection Using CBERS-4A Imagery

Purpose

The visual inspection will be applied to a subsample of points approved in the embedding-SOM stages. Its purpose is to estimate the overall quality of the pre-selected set, identify systematic errors, and guide adjustments before the final Random Forest training. It does not correspond to map validation, which will occur after the cartographic product is generated.

Initial Sample Sizing

For an illustrative population of 40,000 candidates, a margin of error of 3%, a 95% confidence level, and a conservative proportion of p = 0.5, simple random sampling with finite population correction would yield approximately 1,040 points:

$$n = N \times z^2 \times p(1\text{-}p) \,/\, [\, e^2(N\text{-}1) + z^2 \times p(1\text{-}p) \,]$$

This sample size estimates a single global binary proportion, such as label corroborated or not corroborated. It does not guarantee a 3% margin of error for each of the seven categories individually. The final sample will be stratified by category, confidence scenario, and, when necessary, region and vigor, with minimum representation for rare classes and appropriate weights for global estimates.

Interpretation Protocol

- Use pan-sharpened CBERS-4A images with a nominal resolution close to 2 m and a date compatible with the sample year.
- Hide the pseudo-label from the interpreter whenever possible, reducing confirmation bias.
- Record the decision as corroborated, not corroborated, or inconclusive, along with the justification and image quality.
- Submit a fraction to two independent interpreters, measure agreement, and adjudicate disagreements.
- Preserve the link between image, date, interpreter, protocol version, and final decision.

CBERS-4A interpretation provides high-resolution visual evidence on structure, ground cover, presence of woody vegetation, and heterogeneity. Even so, it does not replace field observations for all categories of pasture condition and management.

8. Field Data Collection and Subsequent Map Validation

After the first version of the map is generated, image-based and field-based evaluations will be carried out following specific statistical designs. The initial estimate is approximately 100 field localities, distributed across the Cerrado, with logistical priority given to Goiás and complementary presence in other regions.

One hundred points are adequate for a pilot campaign and for identifying systematic errors, but they do not ensure high precision for seven categories across the entire biome. As a reference, an accuracy proportion close to 80% would have a margin of approximately 8 percentage points under simple random sampling, before accounting for design effects and spatial dependence.

- Maintain independence between training points and validation points.
- Seek a spatially balanced allocation across categories, confidence levels, ecoregions, and environmental gradients.
- Record inclusion probabilities and use weighting when allocation is disproportionate.
- Control clustering along routes and enforce a minimum distance between localities.
- Interpret the concentration in Goiás as a regional priority validation, not as automatic inference for the entire Cerrado.

The final sampling plan, estimators, confidence intervals, and uncertainty analysis will be defined with the guidance of Prof. Luis Rodrigo Fernandes Baumann, from IME/UFG, in coordination with the LAPIG field and remote sensing team.

Training Set Formation and Random Forest

Only candidates that meet the quality criteria defined in the embedding, SOM, and inspection analyses will be eligible for the final training set. Labels will retain levels of evidence, origin, and decision history; rejected or inconclusive samples will remain archived for auditing but will not be automatically used as positive cases.

The final set will be evaluated for balance among categories, spatial distribution, coverage of vigor strata, and redundancy. Rare categories will not be merged merely to achieve larger counts: any merging will depend on ecological coherence, spectral-temporal separability, and performance under independent validation.

The Random Forest will be trained to produce the first version of a 10 m resolution Cerrado pasture map, differentiating among the viable categories. The split between training and evaluation must use spatial blocks or another strategy that minimizes leakage from autocorrelation.

Quality Control, Traceability, and Reproducibility

Each sample must have a stable identifier and, at minimum, fields for provenance, geometry, year, source, label, confidence level, similarity metrics, SOM neuron, inspection outcome, and change history. Scripts, parameters, and collection versions will be preserved to allow the analyses to be repeated.

Table 5. Methodological risks and planned controls

Methodological Risk	Planned Control
Circularity with MapBiomas products	Preserve provenance; use multiple scenarios; do not treat pseudo-labels as ground truth; validate independently.

Methodological Risk	Planned Control
Imbalance among categories	Report retention by class; test class-specific thresholds; use weighting or controlled sampling without masking real scarcity.
Spatial mixing at edges	Record support and homogeneity; inspect borderline cases; evaluate spatial filters.
Temporal class change	Anchor the label to the field year; verify stability before propagating to other years.
Dot product of non-unit means	Calculate L2 norms and use explicit cosine similarity or renormalization before the dot product.
SOM dominated by pseudo-labels	Separate reference support from internal coherence; test grid, neighborhood, and rejection criteria.
Validation concentrated in Goiás	Limit regional inference and combine field data with a spatially comprehensive image-based sample.

Upcoming Decisions and Technical Deliverables

1. Consolidate the single dictionary of the seven categories and reconcile the 608 classified samples.
2. Audit the Sentinel-2 series: indices, masks, resampling, edge months, dates, and valid pairs.
3. Calculate the L2 norms of the mean embeddings and recompute similarity as cosine when necessary.
4. Run a sensitivity analysis of the Top-1, Top-3, md3, and margin thresholds, and agreement with Collection 11.
5. Define the SOM dimensionality and parameters based on stability, quantization/topographic errors, and reference support.
6. Size and draw the stratified sample for CBERS-4A inspection, with a blind protocol and inter-interpreter control.
7. Assemble the approved database, assess balance and separability, and define any category groupings with technical justification.
8. Train the first version of the Random Forest and implement independent validation by imagery and field data.

Conclusion

Product 1 establishes a progressive sample-expansion strategy, in which highly curated field references are used to select candidates at scale, without equating similarity or agreement between maps with ground truth. The combination of embeddings, Sentinel-2 time series, SOM, CBERS-4A inspection, and subsequent validation makes it possible to build successive quality controls and document the uncertainty associated with each pseudo-label.

As of August 17, 2026, the reference and candidate databases have been assembled and the main spectral-temporal attributes and embeddings extracted. The definition of thresholds, the final SOM configuration, the visual inspection, the definitive composition of the training set, and the design of the field campaign remain under development. This distinction between results already obtained and planned procedures is essential for the project's traceability and for the correct interpretation of future versions of the map.

References and Technical Documents

Breiman, L. (2001). Random Forests. Machine Learning, 45, 5-32. DOI

Brown, C. F. et al. (2025). AlphaEarth Foundations: an embedding field model for accurate and efficient global mapping from sparse label data. Preprint

Google Earth Engine. Satellite Embedding V1: GOOGLE/SATELLITE_EMBEDDING/V1/ANNUAL. Accessed August 2026. Catalog

Google Earth Engine. Harmonized Sentinel-2 MSI: MultiSpectral Instrument, Level-2A. Accessed August 2026. Catalog

Kohonen, T. (1990). The self-organizing map. Proceedings of the IEEE, 78(9), 1464-1480. DOI

Internal Documents Consulted

- WRI Brasil and FUNAPE. Technical Consulting Services Agreement - Project GR000153-73077-CARGILL-BRAZIL, signed in August 2026.
- LAPIG/UFG and Embrapa Cenargen. Sentinel-2 Embeddings Samples. Internal technical presentation, 2026.
- Vaz, G. R. Dot Product Analysis in Pasture Samples Using AlphaEarth Embeddings. Internal technical presentation.
- Baumann, L. R. F. Application of Self-Organizing Maps to Clean the Classification of Time Series Data. Internal technical presentation, 2026.