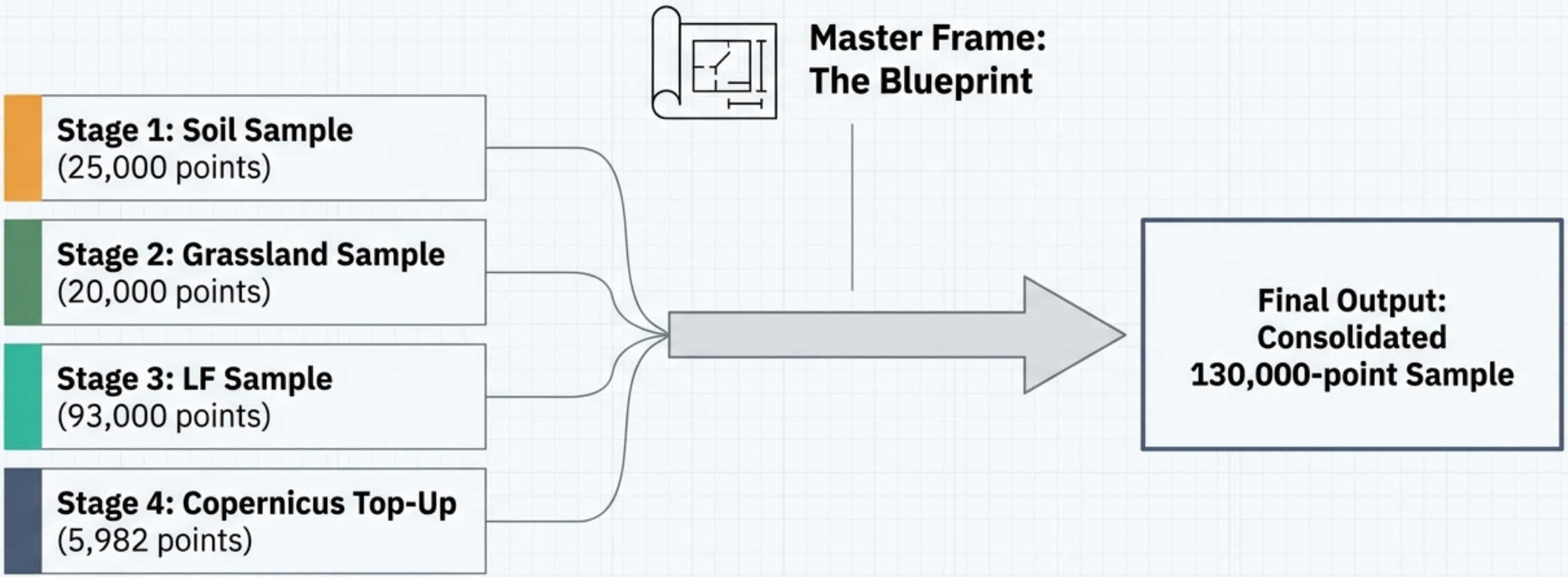
Constructing the LUCAS 2027 Sample: A Methodological Blueprint

A step-by-step guide to the assembly and validation of a high-precision, multi-module dataset.

The Assembly Process: From Raw Data to a Unified Sample

This presentation documents the meticulous, four-stage construction of the 130,000-point LUCAS 2027 sample. We will walk through the creation of each thematic module, showing how it was cleaned, weighted, and validated before being integrated into the final dataset.



Module 1: Forging the 25,000-Point Soil Sample

Objective

To produce a clean, weighted 25,000-point Soil sample, validated against the master frame for geographic and land cover representativeness.

Methodology

1

Load Raw Data: Start with the initial `lucas_soil_sample20251203.csv`.

2

```
soil_clean = soil_raw
  - points_not_in_master
  - points_in_exclusion_lists
    (XX01, AB03, etc.)
```

3

Enrich: Merge with the master frame to attach key attributes (`STR25`, `LC_pred`) and `LC_pred`) and 2022 survey weights.

4

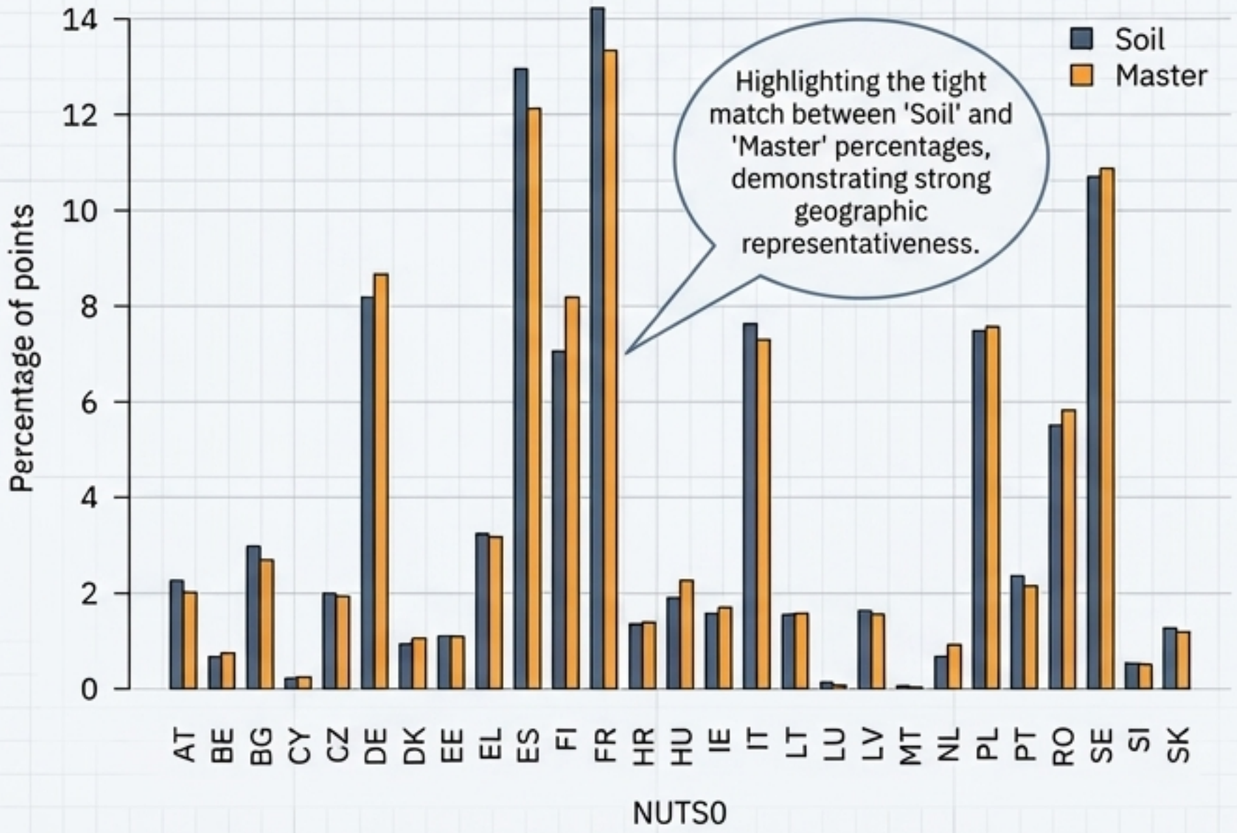
Calibrate Weights: New weights are calculated to ensure the sample's stratum distribution matches the master frame.

$$\text{Weight} = \frac{\text{(Master Count in Stratum)}}{\text{(Sample Count in Stratum)}}$$

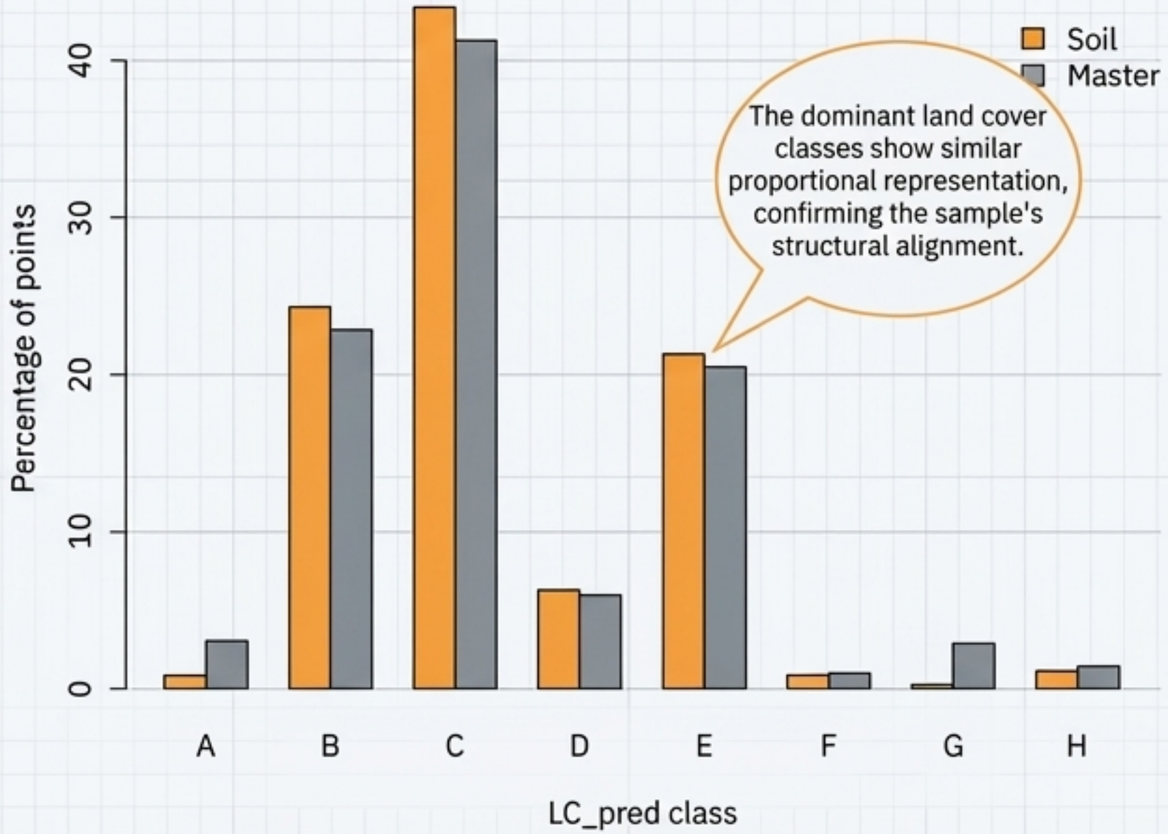
Soil Sample Validation: Alignment with the Master Blueprint

**Final Cleaned Sample: 25,000 points.

Geographic Distribution (NUTS0) Closely Mirrors the Master Frame



Land Cover Distribution (LC_pred) Shows Strong Correlation



Module 2: Assembling the Grassland Sample from Two Components

Objective: To construct a final 20,000-point Grassland sample by strategically combining points from two distinct 2022 sources.

Component 1: Non-Extended Grassland
Source: `grassland_sample_2022.csv`
Focus: Core grassland points observed in the 2022 survey.
****Yields: 11,099 points****

Component 2: Extended Grassland
Source: `grassland_extended_sample_2022.csv`
Focus: A broader set of points to supplement the core sample and reach the 20k target.
****Yields: 8,901 points****

****Final Grassland Sample: 20,000 points****

Key Methodological Note: A recurring theme in this module is the correction for "non-observation." We start with a larger potential sample but can only use points that were effectively surveyed. Weights are systematically corrected to account for this.

Grassland Part 1: Selecting the ‘Non-Extended’ Core

Methodology Flowchart

Start with 2022 Sample
Load `grassland_sample_2022.csv` (19,998 points).

↓

Isolate Observed Points
Filter for the 12,119 points that were actually observed in the field survey.

↓

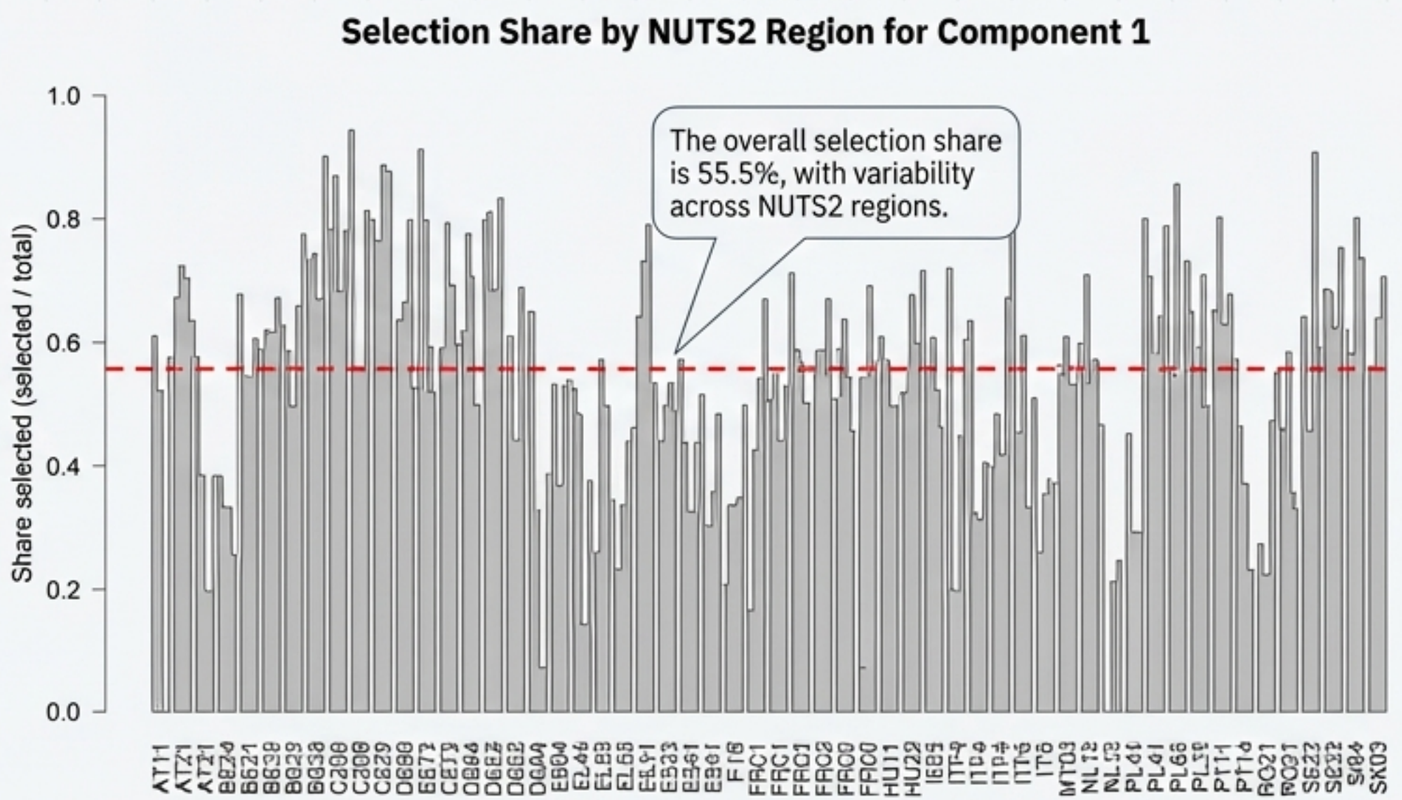
Correct Weights
Calculate a `wgt_correction` factor to re-balance the sample, compensating for the points that were not observed.

↓

Filter by Land Cover
Select only the observed points with a 2022 Land Cover (`LC1`) classification of ‘D’ (Cropland) or ‘E’ (Woodland/Shrubland).

Outcome & Validation

Component 1 Yields 11,099 Points.

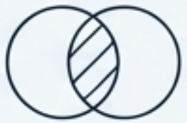


Grassland Part 2: Selecting the ‘Extended’ Remainder

Methodology Flowchart

Start with Extended Sample
Load the pool of observed ‘extended grassland’ candidates.

Crucially, remove the 6,093 points already selected for Component 1. This leaves 16,031 unique candidates.

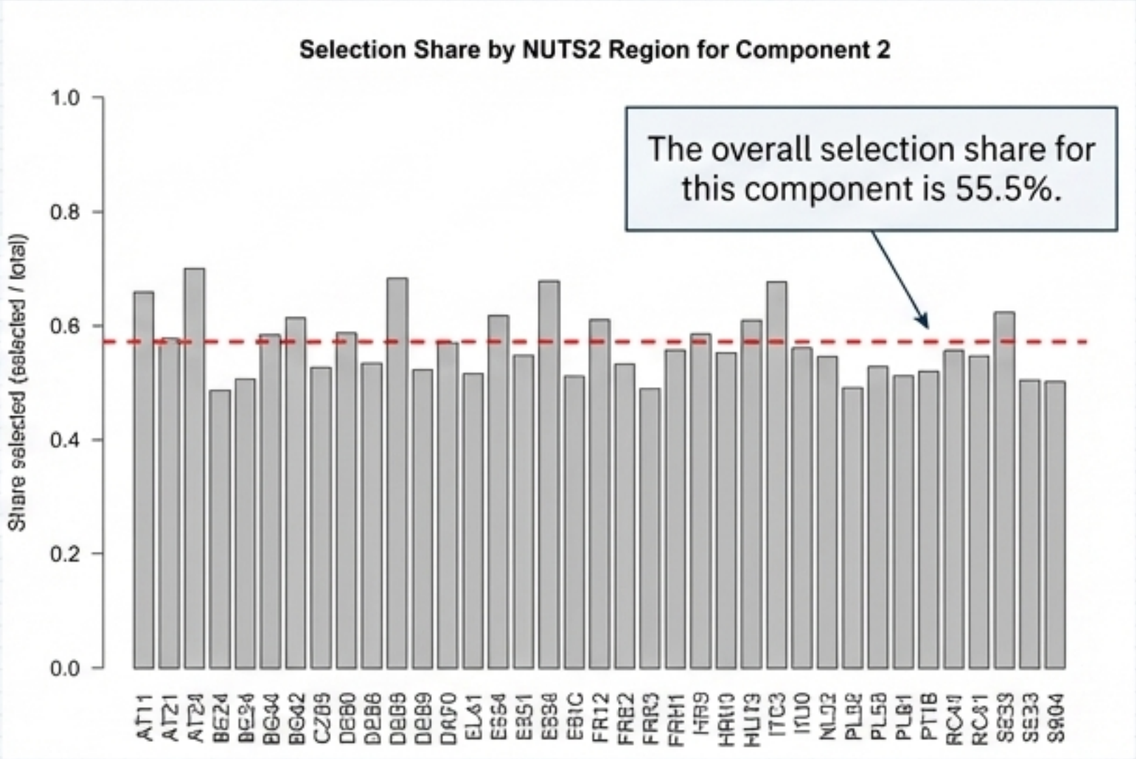


Define Target
The goal is to select the remaining points needed to reach 20,000 total: `20,000 - 11,099 = 8,901 points’.

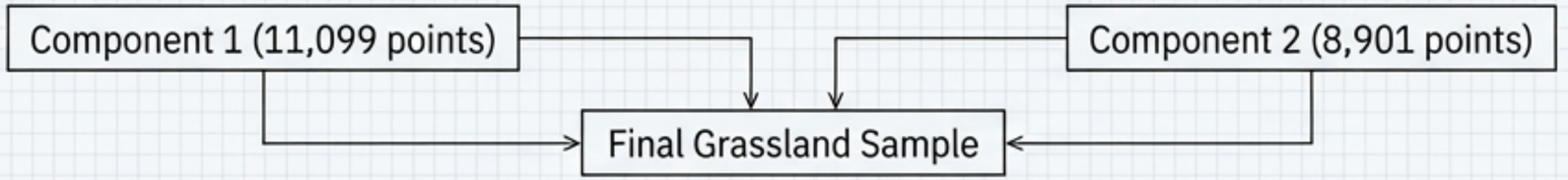
Proportional Allocation
The 8,901 points are selected from the candidates using proportional allocation by NUTS2 stratum, ensuring geographic balance. Points within each stratum are chosen based on a Permanent Random Number (PRN).

Outcome & Validation

Component 2 Adds Exactly 8,901 Points.

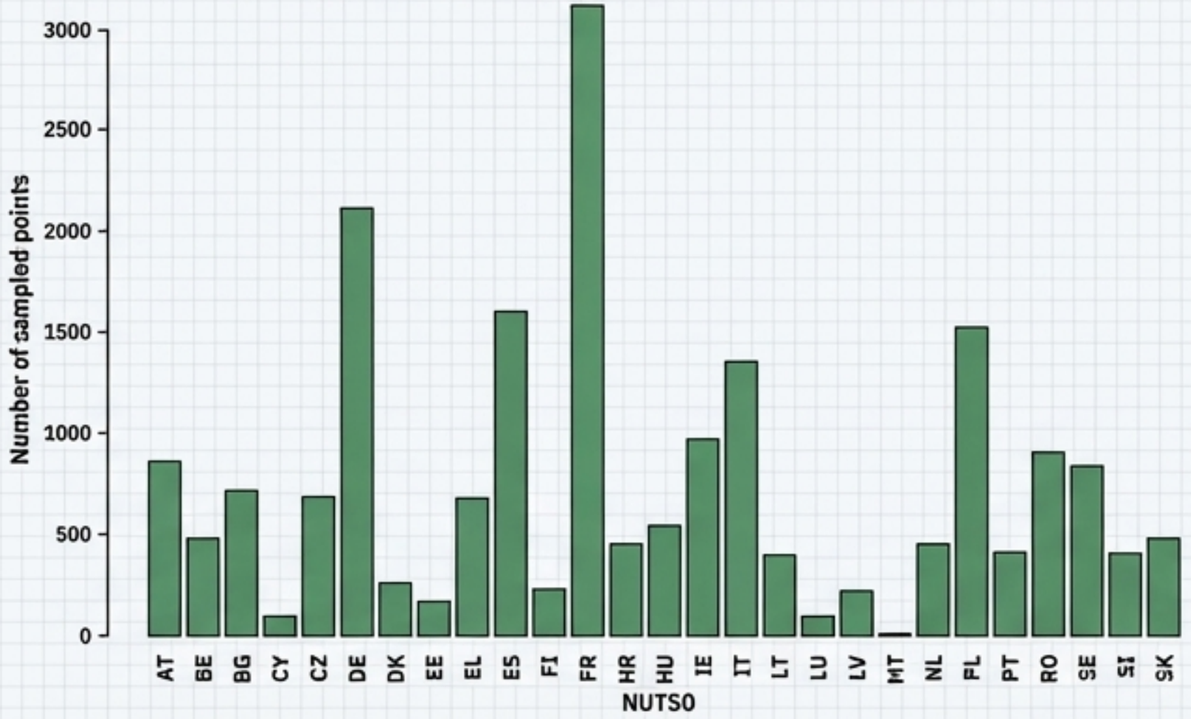


Grassland Final Assembly: A Calibrated 20,000-Point Sample

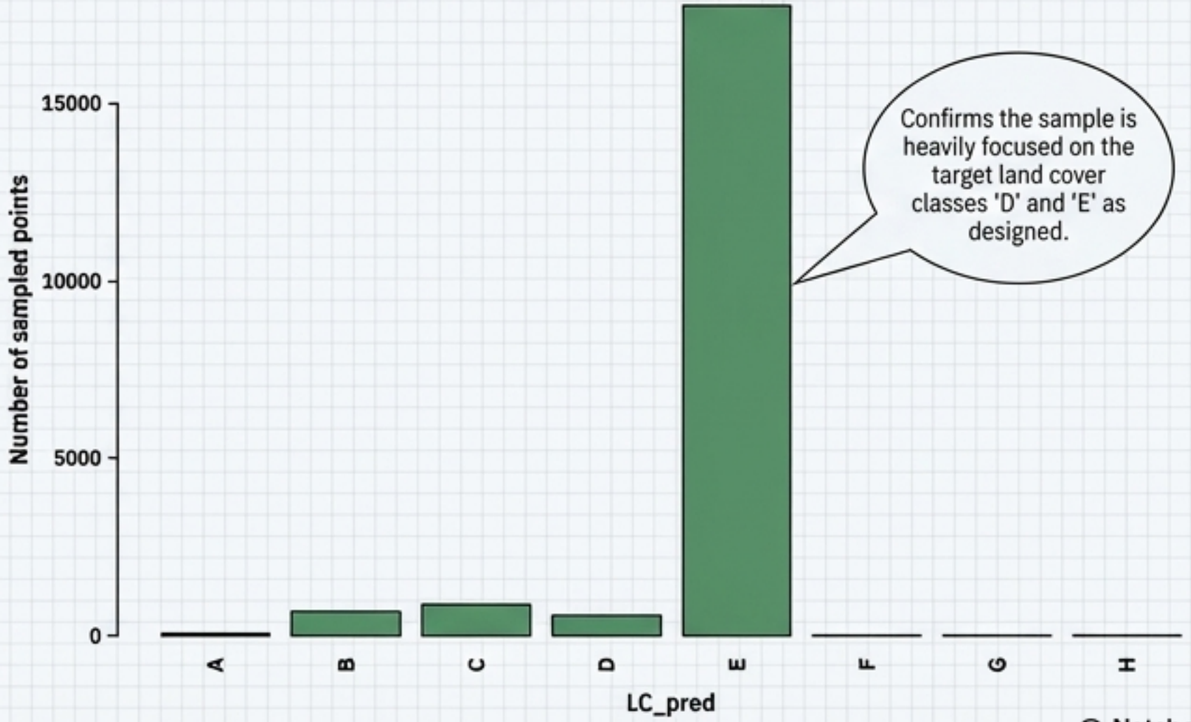


Final Step: Post-Stratification. The weights of the combined 20,000-point sample are calibrated one last time to perfectly align with the master frame's population totals for land covers 'D' and 'E'.

Final Grassland Geographic Distribution by NUTS0



Final Grassland Land Cover Distribution by LC_pred



Module 3: Engineering the 93,000-Point Linear Features (LF) Sample

Objective: To build a large-scale, 93,000-point LF sample using a sophisticated, five-stage process that optimizes for statistical precision.

1. Preparation

Create a candidate pool of 78,318 observed and eligible points from the 2022 survey.

2. Indicator Calculation

For each point, compute two key metrics: Richness ('R') and Presence ('F') of linear features.

3. Optimal Allocation

Use the indicator statistics to calculate the optimal number of points to sample from each stratum (`nh_opt`) to minimize variance.

4. Panel Selection

Select the first 46,500 points (Component 1) from the 2022 candidate pool based on the optimal allocation.

5. Non-Panel Selection

Select the final 46,500 points (Component 2) from the broader master frame, again using the optimal allocation plan.

The Core of the LF Strategy: Indicator-Driven Allocation

Instead of simple proportional sampling, the LF module uses a statistically optimized approach. By calculating the mean and standard deviation of key indicators within each stratum, we can determine precisely how many samples to draw from each to achieve our target accuracy with maximum efficiency.

1. Calculate Indicators

For 78,318 observed points, calculate:

- `R` (Richness): Average presence of 41 linear feature types.
- `F` (Presence): Binary flag (1 if R > 0).

2. Summarize by Stratum

Compute weighted Mean and Standard Deviation for R and F within each `STRATUM_LF` (NUTS2*STR25).

LF_stratum_stats.csv

3. Run Bethel Allocation

Input the stratum stats into the Bethel algorithm with a target Coefficient of Variation (CV) of 1%.

LF_strata_with_bethel_allocation.csv

This is a detailed plan specifying the optimal number of points (`nh_opt`) to select from each stratum. The total allocation is **46,500 points** for the panel.

Executing the Plan: Selecting the Panel and Non-Panel Components.

LF Panel (Component 1)

Sample Pool: 78,318 observed points from 2022.

Action: Select **46,500 points** using the optimal allocation plan (`nh_opt`), ranking candidates by PRN within each stratum.

LF2027_panel.csv

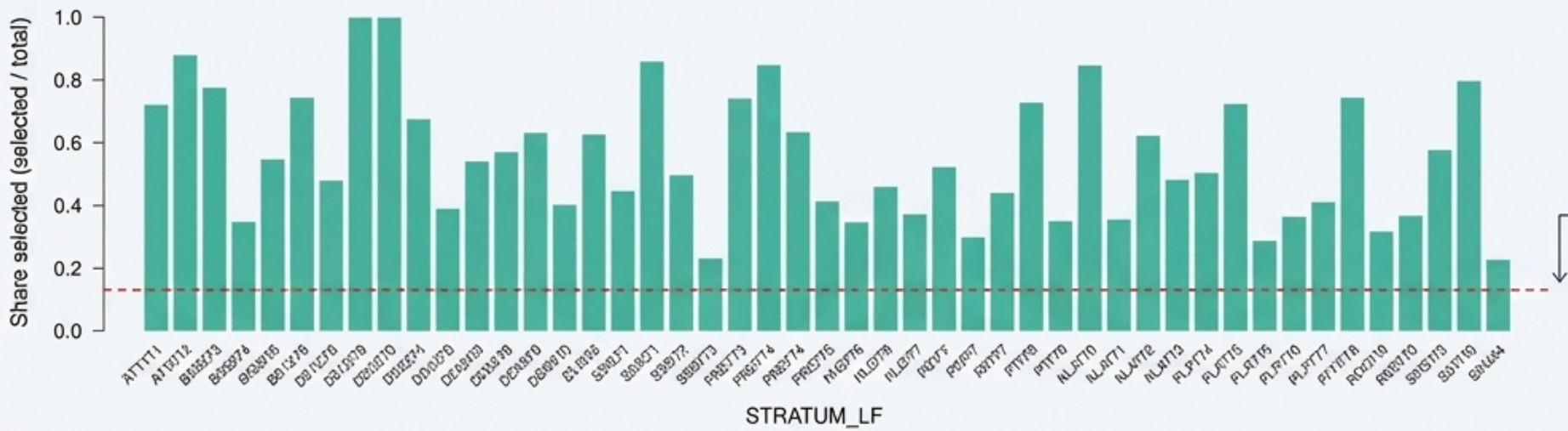
LF Non-Panel (Component 2)

Sample Pool: 360,056 eligible points from the entire master frame (after removing panel points).

Action: Select another **46,500 points** using the same optimal allocation plan, ranking candidates by PRN.

LF2027_nonpanel.csv

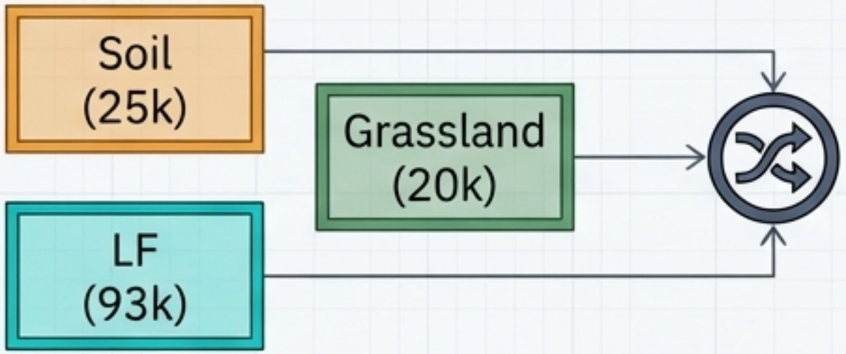
Validation: Non-Panel Selection Rate by Stratum



The overall sampling rate is ~12.9%. The variability in bar height reflects the *optimal* allocation, where some strata are sampled more intensively than others to minimize variance.

Final Assembly: Consolidating Modules and Planning the Copernicus Top-Up

Step 1: Combine Thematic Samples



After combining and removing duplicate `POINT_ID`s, the total unique sample size is **124,018 points**.

Step 2: Define the Final Target

130,000 total points.

130,000 (Target) - 124,018 (Current) = **5,982 points**

The Role of the Copernicus Sample:

1. Reach the final 130,000-point target.
2. Strategically improve the final sample's alignment with the master frame's land cover distribution.

Copernicus Allocation: A Strategic ‘Water-Filling’ Approach

The Challenge

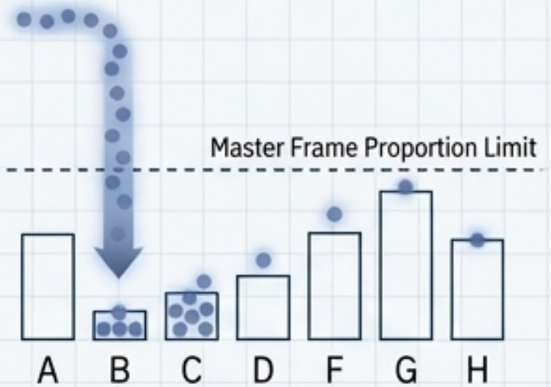
The 124,018-point sample is over-represented in some land cover classes (B, E) and under-represented in others compared to the master frame.

The Allocation Strategy

1. Fix Over-represented Classes

Land covers ‘B’ and ‘E’ are frozen.
No new points are added.

2. Equalize Under-represented Classes



Points are iteratively added to bring their absolute counts closer to each other, with one constraint: the final count cannot exceed the proportion dictated by the master frame.

3. Assign Remainder

Any remaining budget after equalization is allocated to class ‘C’.

The Final Allocation Plan

LC_pred Class	Points to Add
A	+1,686
F	+515
G	+2,545
H	+1,236
B, C, D, E	+0
Total Added	**5,982**

The Final Product: A 130,000-Point Consolidated Sample

Key Result: The four modules (Soil, Grassland, LF, and Copernicus_add) are combined and deduplicated to produce the final `LUCAS27_sample.csv` containing **130,000 unique points**.

Module Overlap Analysis

Point Composition	Count
LF only	79,748
Soil only	22,009
Grassland + LF	10,612
Grassland only	8,671
Copernicus_add (new points)	5,982
LF + Soil	2,261
Grassland + LF + Soil	379

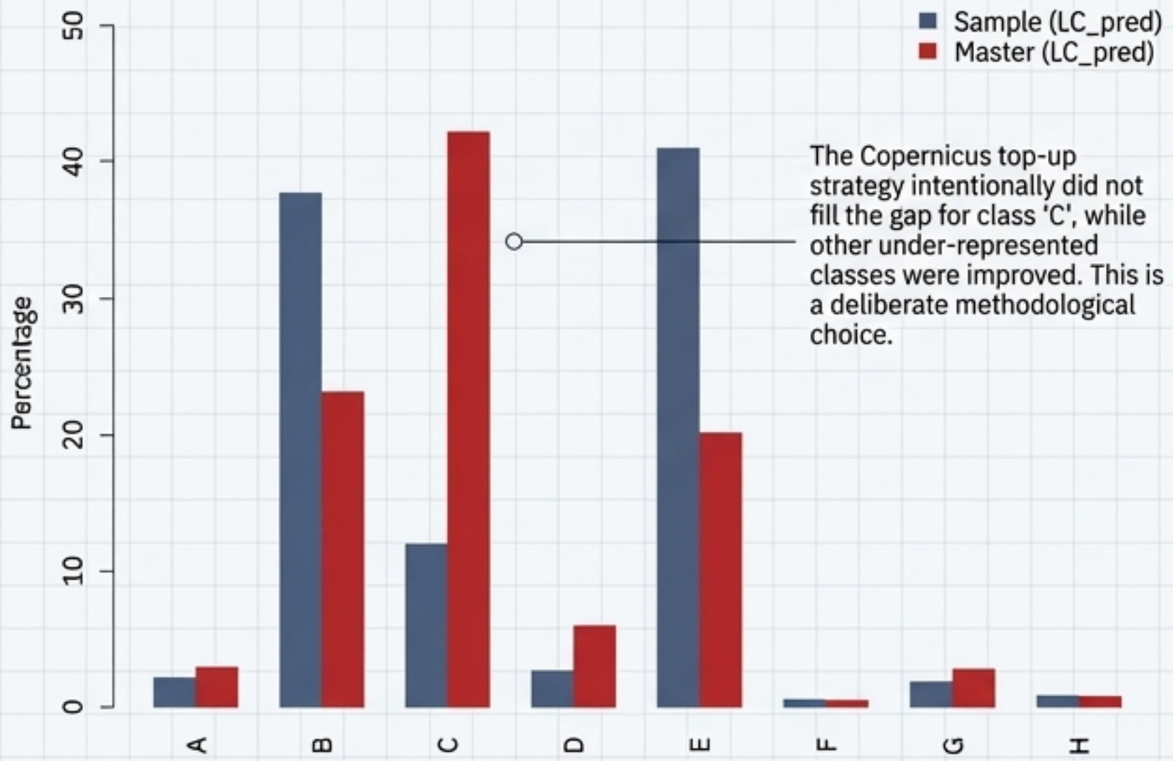
Pairwise Overlap Matrix

	SOIL	GRASSLAND	LF	COPERNICUS
SOIL	25,000	717	2,640	24,987
GRASSLAND	717	20,000	10,991	20,000
LF	2,640	10,991	93,000	93,000
COPERNICUS	24,987	20,000	93,000	130,000

Final Validation: The Assembled Sample vs. The Master Blueprint

The final validation confirms that the multi-stage construction process has produced a large, robust sample whose geographic and land cover distributions are well-aligned with the reference master frame.

Final Sample vs. Master by Predicted Land Cover (LC_pred)



Final Sample vs. Master by Geographic Region (NUTS0)