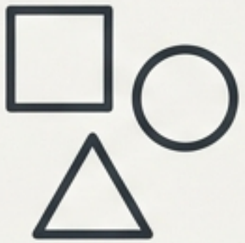
Constructing the LUCAS 2027 Sample: A Methodological Walkthrough

A step-by-step explanation of the statistical process for building the 130,000-point European survey sample.

This document details the complete data processing pipeline used to generate the LUCAS 2027 survey sample. The process is designed to be transparent, reproducible, and statistically robust.



Modularity

Building thematic samples (Soil, Grassland, LF) independently before integration.

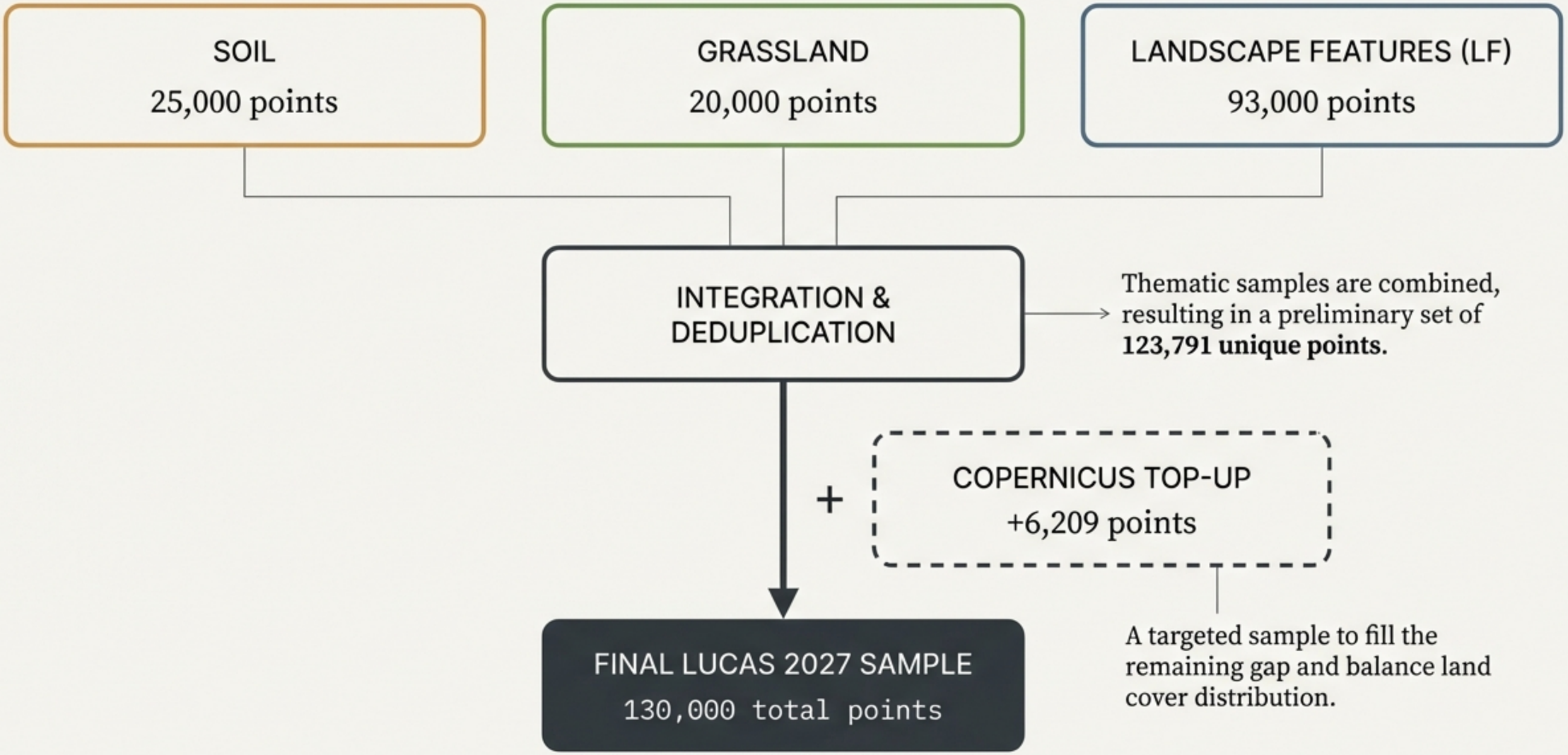
Integration

Systematically combining modules and resolving overlaps to form a cohesive whole.

Calibration

Ensuring the final sample accurately represents the master frame through rigorous weighting and a targeted top-up.

From Thematic Components to a Unified Sample



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

Input -> Process -> Output

INPUTS

`master_complete.RData`
The comprehensive master frame.

`lucas_soil_sample20251203.csv`
The raw 2027 soil sample list.

`Soil_to_be_removed.xlsx`
Exclusion lists containing points to be flagged and removed.

`Survey_2022_wgt_2nd_phase.txt`
Source of 2022 survey weights for legacy points.

PROCESS

1. **Cleaning:** Points from the exclusion lists are removed from the raw sample.
2. **Enrichment:** The cleaned sample is merged with the master frame to attach key attributes like `NUTS2`, `STR25`, and predicted Land Cover (`LC_pred`).
3. **Weight Attachment:** Weights from the 2022 survey (`WGT_LUCAS`) are attached to the corresponding points.

OUTPUT

`lucas_soil_sample_2027_clean.csv`
An intermediate file containing the clean, enriched 25,000-point sample.

The final cleaned sample contains 23,000 points with `bio2027_flag` = 0 and 2,000 with `bio2027_flag` = 1.

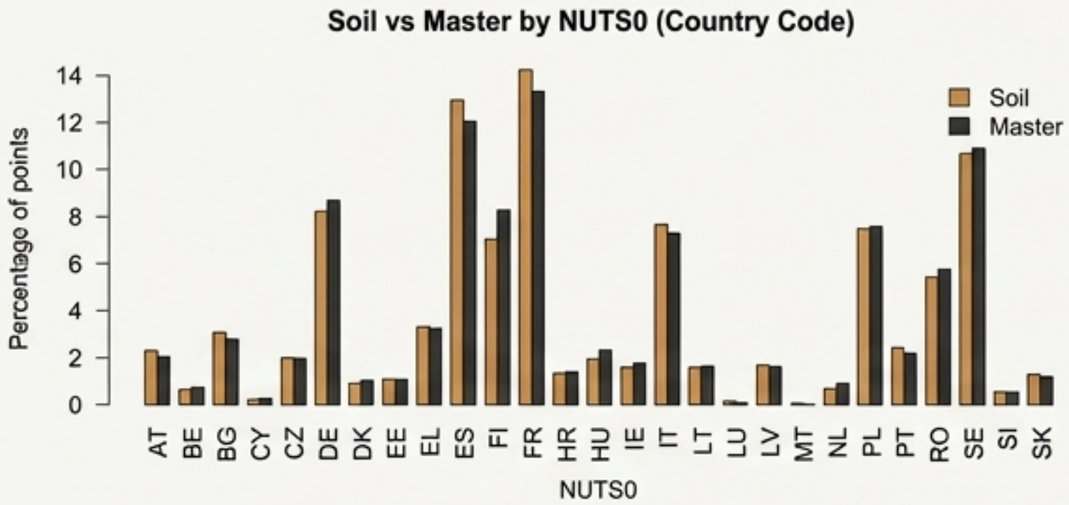
Soil Sample: Stratum Weighting and Verification

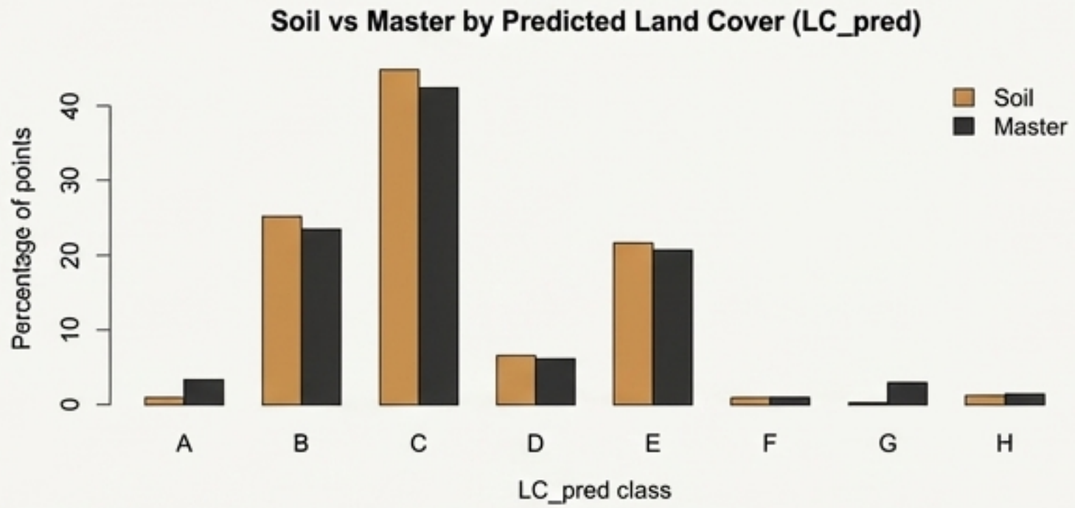
Weight Calculation

Objective: To ensure the Soil sample represents the master frame population.

Stratum Definition: `STRATUM_LC` is defined as the interaction of `NUTS2_24` and `LC_pred`.

Verification via Distribution Comparison





Key Takeaway: The bar charts confirm that the 25,000-point Soil sample is a well-distributed representation of the master frame across both geography and land cover.

Final Export: The process concludes with the export of `Soil2027_sample.csv`.

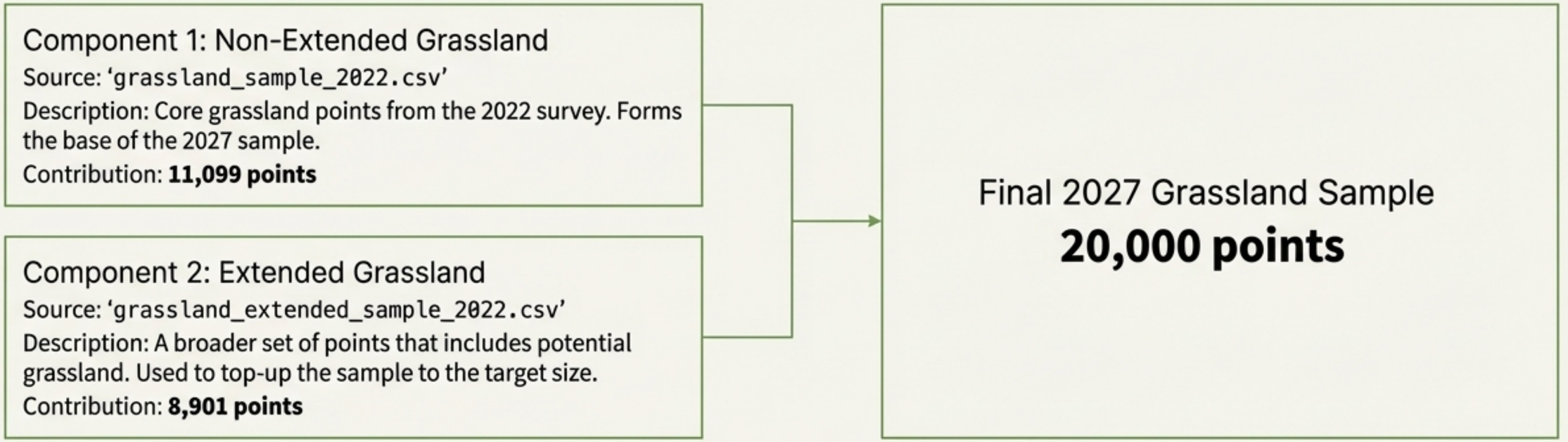
Module 2: A Two-Component Approach for the Grassland Sample

Objective & Challenge

Objective: To construct a 20,000-point sample specifically targeting land with predicted permanent grassland and pasture (`LC_pred` classes 'D' and 'E').

Total population size in master frame: 273,497 points.

The Solution: A two-component strategy.



This dual-source method allows us to maximize the reuse of previously surveyed points while ensuring the final 20,000-point sample is fully representative of the target domain.

Grassland Component 1: Building the Base from Observed 2022 Points

Process

Input: The 2022 Grassland sample (`grassland_sample_2022.csv`), containing 19,998 points.

Process Step 1: Identify Observed Points
Only points effectively surveyed in 2022 are retained. This reduces the pool from 19,998 to **12,119 observed points**.
The total weighted population represented by the observed points (111,412.5) is significantly less than the full sample's (181,848.7).

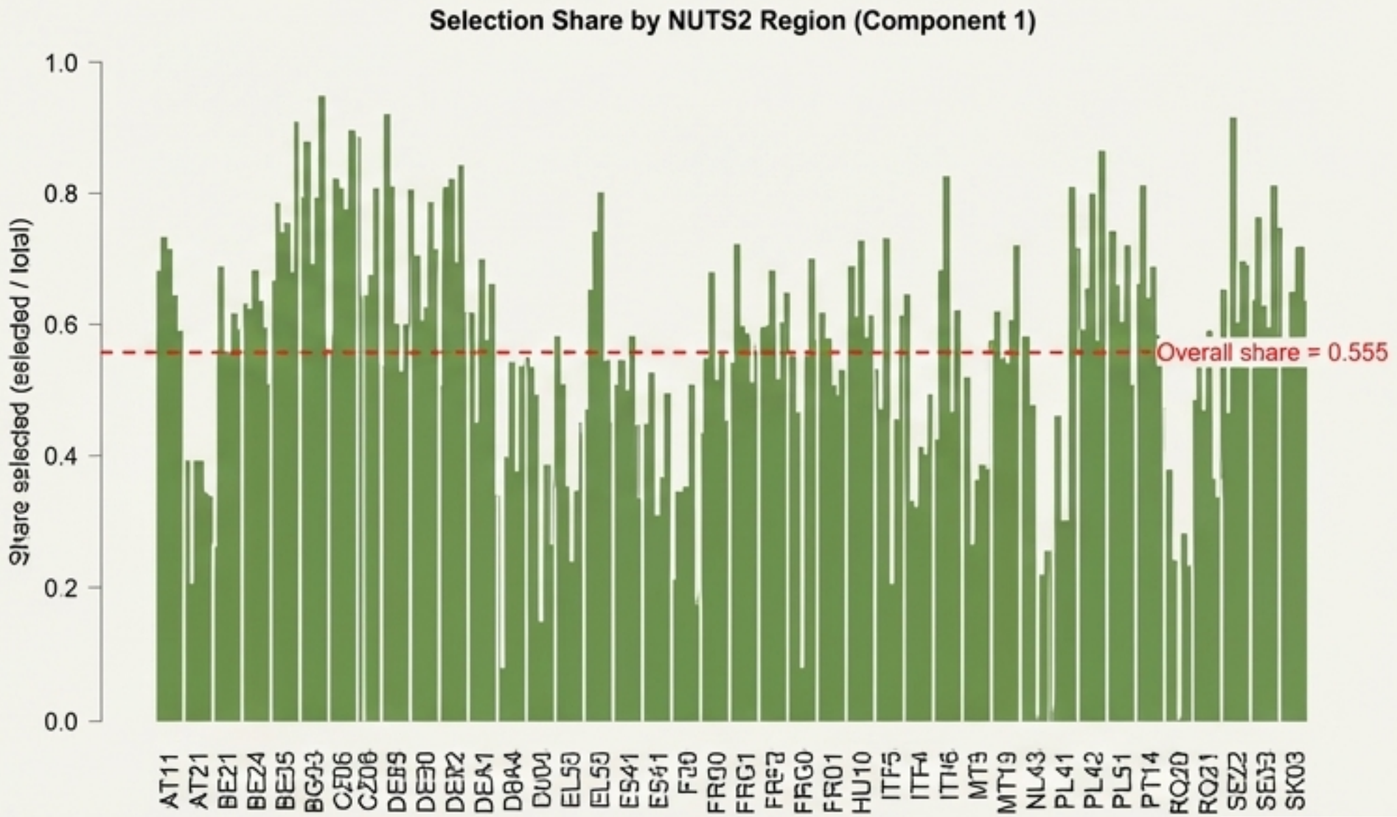
Process Step 2: Correct for Non-Response
A `wgt_correction` is calculated to adjust for the difference between the full and observed samples.
Correction is done at the stratum level (`STRATUM_GRASSLAND` and observed `LC1`), effectively up-weighting the observed points to represent the non-observed ones.

Process Step 3: Selection
From the corrected, observed points, only those with an observed land cover of 'D' or 'E' in 2022 are selected.

Output: 11,099 points form Component 1.

Visual Verification

The bar chart displays the share of selected points relative to the total available for NUTS2 regions. A dashed red horizontal line cuts across the chart, clearly labeled **'Overall share = 0.555'**. This visual demonstrates that while individual region shares vary, the selection is distributed geographically around a consistent overall rate.



Grassland Component 2: Selecting the 8,901-Point Top-Up

Process

Input: The 2022 Extended Grassland sample (`grassland_extended_sample_2022.csv`).

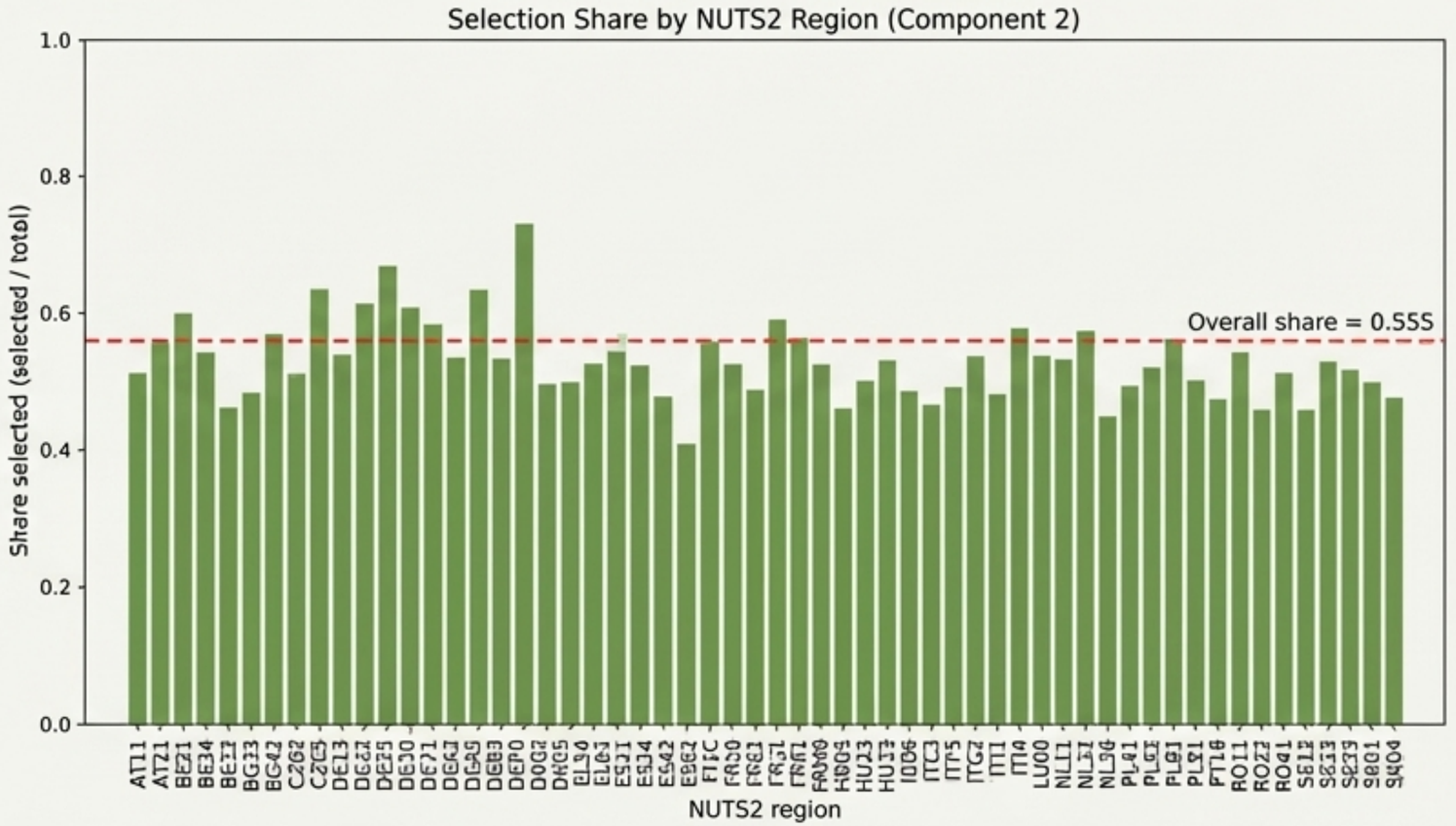
Process Step 1: Prepare the Candidate Pool
Filter for observed points with land cover 'E' or 'D'. Crucially, remove the 6,093 points already selected for Component 1, leaving **16,031 candidate points.**

Process Step 2: Proportional Allocation
Target: Select exactly **8,901 points** (20,000 total target - 11,099 from Component 1).
Method: The 8,901 points are allocated across NUTS2 regions proportionally to the number of available candidates in each region.
Rounding: Largest-remainder rounding is used to ensure integer allocations that sum perfectly to the target. A minimum of 2 points per stratum is enforced where possible.

Process Step 3: Selection
Within each NUTS2 stratum, points are selected by ranking them in descending order of their Permanent Random Number (PRN).
Output: `Grassland2027_component2.csv` containing the final 8,901 points.

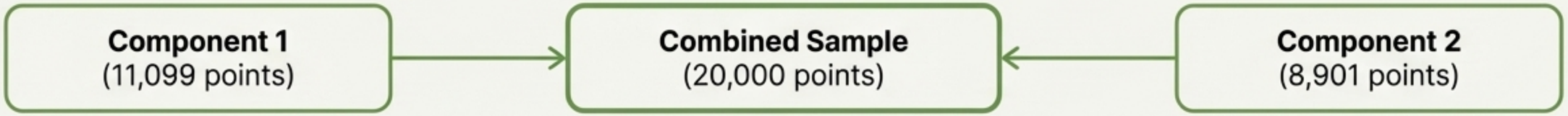
Visual Verification

The bar chart displays the share of selected points for Component 2 across NUTS2 regions. A dashed red horizontal line is again present, labeled 'Overall share = 0.555'. This demonstrates remarkable consistency in the selection pressure across both components.



Grassland Module: Integration, Final Calibration, and Output

Process

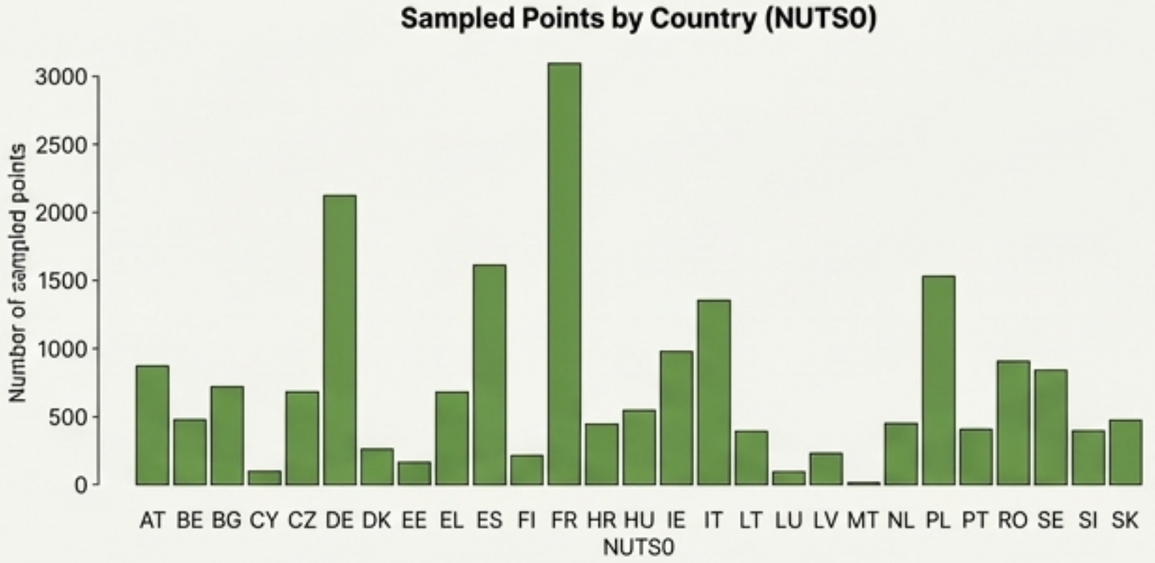


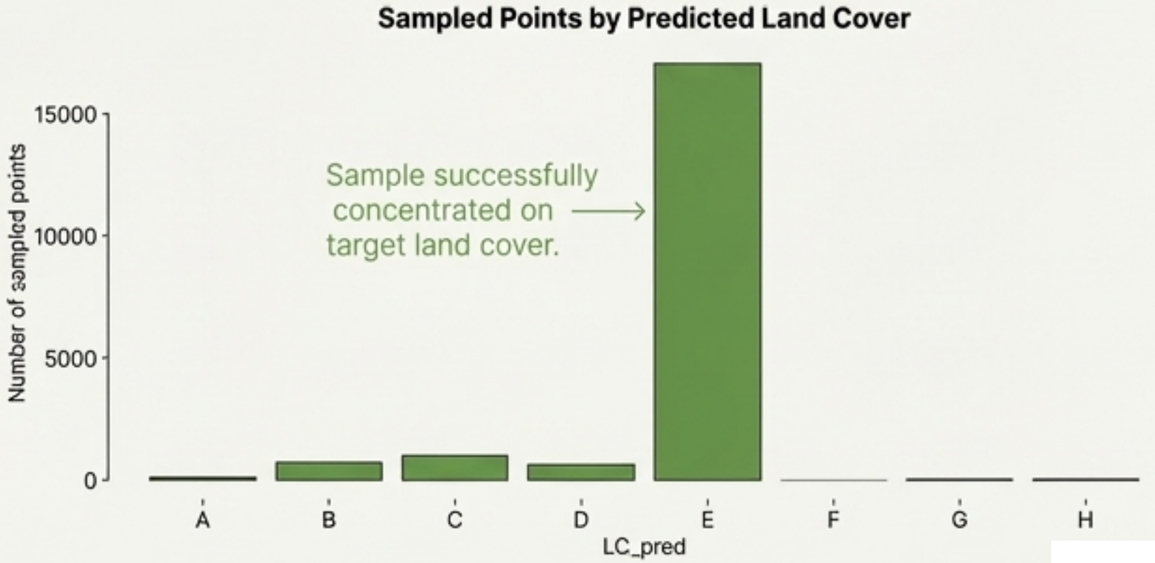
Process Step 2: Final Weight Calibration

The combined sample's weights are post-stratified to match the known population totals of the Grassland frame (all points in the master frame with `LC_pred` 'D' or 'E'). This final calibration step ensures the sample accurately represents the target population within each `NUTS2 * STR25` stratum. The sum of the final weights (`WGT_module_27`) is 259,615.

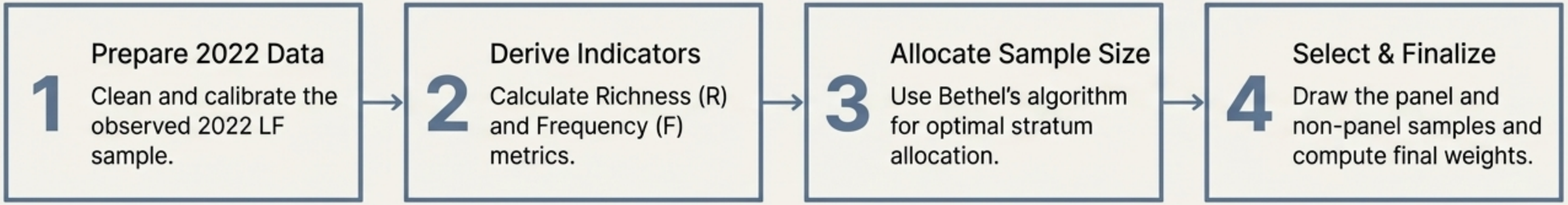
Output: The Final 20,000-point Grassland Sample

`Grassland2027_sample.csv`





Module 3: Constructing the 93,000-Point Landscape Features Sample



Deep Dive: Stage 1 - Preparing the 2022 Observed Data

- **Objective:** Create a clean, reliable dataset from the 2022 survey to serve as the basis for the 2027 sample design.
- **Inputs:** `LF_sample_2022.csv`, `effective_points_modules.xlsx` (list of observed points), `Survey_2022_wgt_2nd_phase.txt`.
- **Process:**
  - **Eligibility Filter:** The sample is restricted to points considered eligible for LF survey. A point is eligible if `STR25` is 1, 2, or 3, OR its observed land use (`LUobs`) was 'U11'.
  - **Observation Filter:** Only points that were actually observed in the 2022 LF module are retained.
  - **Weight Calibration:** A `wgt_correction` is computed to account for non-response, ensuring the 78,318 observed points represent the full 122,350 eligible points.
- **Output:** `LF_sample_2022_obs.csv` - A calibrated dataset of **78,318** observed, eligible points from 2022.

LF Stage 2: Deriving Richness (R) and Frequency (F) Indicators

Purpose & Process

Purpose: To create quantitative measures of landscape complexity for each point. These measures are essential for optimizing the sample allocation in the next stage.

How it Works:

1. **Grouping Features**: The raw 2022 LF survey data contains observations across **41 feature groups**.
2. **Flagging Presence**: For each of the 41 groups, a flag is created indicating whether any LF code (W, G, T, D, P, S, C) was present.
3. **Calculating Indicators**:
   - **Richness (R)**: The proportion of the 41 groups where at least one LF was present. `R = (Number of groups with LF) / 41`. This measures the diversity of features.
   - **Frequency (F)**: A binary indicator. `F = 1` if R > 0, and `F = 0` otherwise. This measures the simple presence or absence of any LF.

Output: Stratum-Level Statistics

The point-level R and F values are used to calculate weighted means and standard deviations for each LF stratum (NUTS2 * STR25).

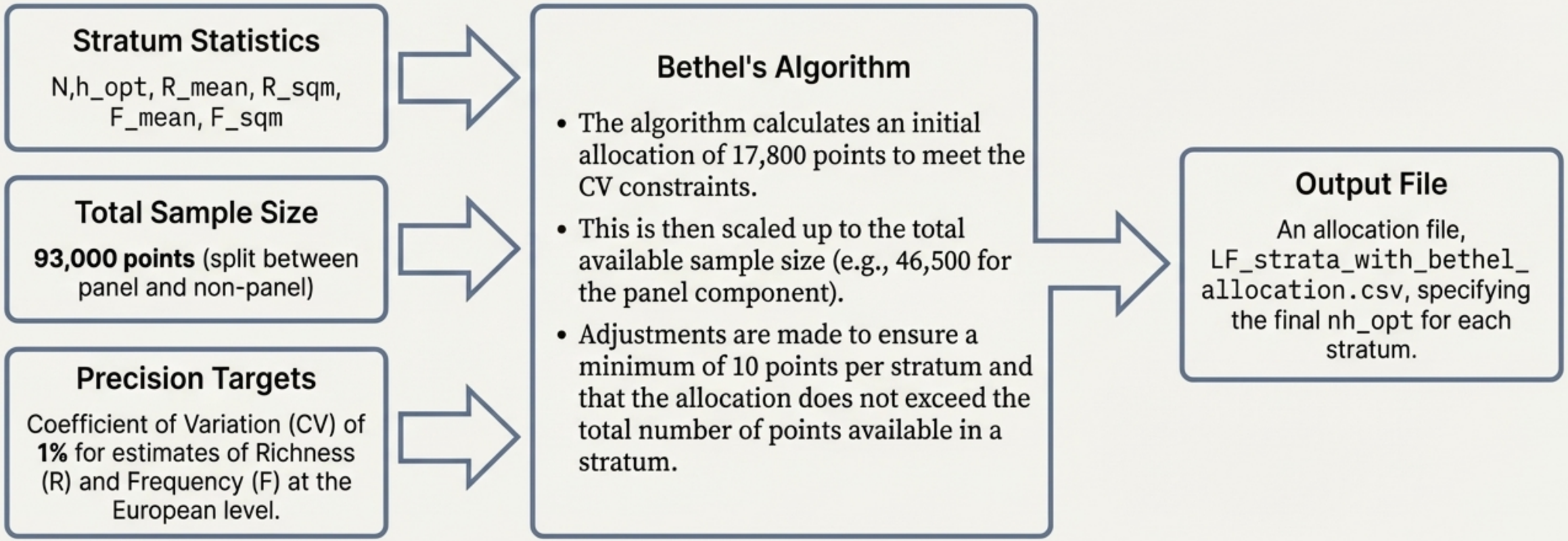
`LF_stratum_stats.csv`

STRATUM	N	R_mean	R_sqm	F_mean	F_sqm
AT11*1	139	0.040	0.066	0.410	0.492
AT11*2	8	0.052	0.079	0.500	0.500
AT11*3	19	0.053	0.066	0.526	0.499

LF Stage 3: Optimal Sample Allocation via Bethel's Algorithm

Objective: To determine the optimal number of points to sample in each stratum (`nh_opt`) to achieve the highest statistical precision for the available budget.

The Tool: The `bethel` function from the `SamplingStrata` R package is used. This is a multivariate allocation algorithm that minimizes sample size subject to precision constraints.



LF Stage 4: Panel/Non-Panel Selection and Final Weighting

Selection

Two-Component Selection: The 93,000-point LF sample is composed of two parts:

- **Panel (46,500 points):** Re-sampled points from the 2022 observed set [redacted]
- **Non-Panel (46,500 points):** New points drawn from the master frame, excluding those already in the panel [redacted]

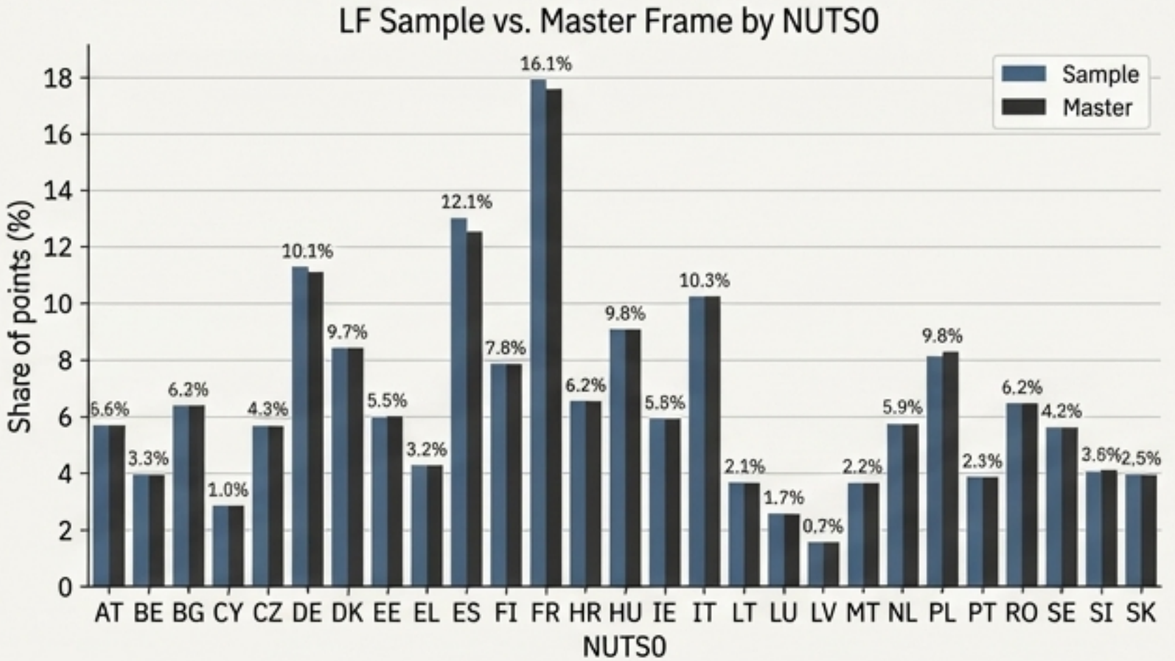
Selection Method (Non-Panel):

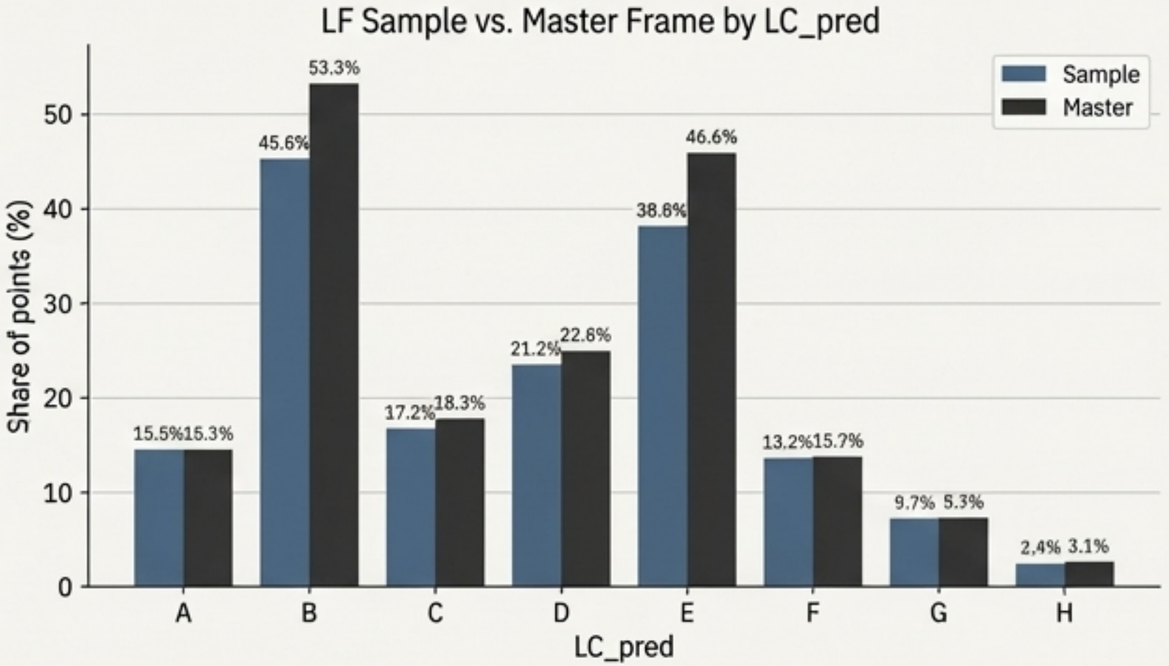
- The candidate pool consists of 360,056 eligible points from the master frame not in the panel.
- Within each stratum, points are selected by ranking them based on their Permanent Random Number (PRN). The top `nh_opt` points are chosen.

Final Weight Calibration:

- The full 93,000-point sample is combined.
- Final module weights [redacted] are calculated by calibrating the sample to the population totals of the eligible LF frame (431,917 points in the master frame).

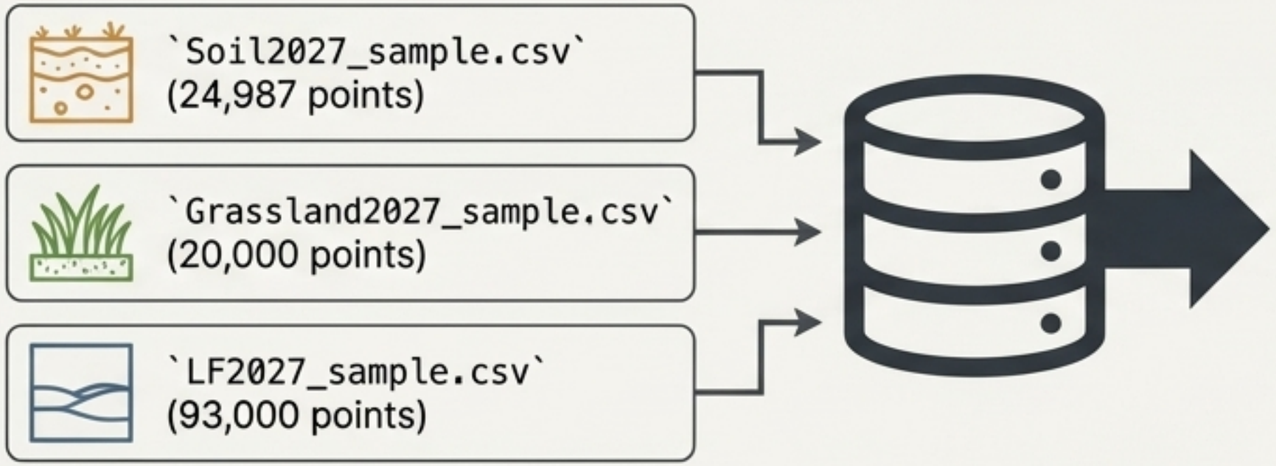
Verification





Integration: Assembling the Sample and Identifying the Final Gap

Consolidation & Deduplication



Process: Deduplication

Points can be selected for multiple modules. A deduplication step is performed on `POINT_ID` to create a list of unique points.

The combined sample consists of **123,791 unique points.**

Highlighting Point Reuse (Module Overlap)

Module Overlap: Unique Points by Combination	
Combination	**Unique Points**
LF only	79,521
SOIL only	22,031
GRASSLAND+LF	10,854
GRASSLAND only	8,429
... and other smaller combinations	...

The Final Challenge: The Gap

Overall Target: 130,000 points

Current Total: 123,791 points

Remaining Gap: **6,209 points**

The Final Top-Up: A Water-Filling Algorithm for the Copernicus Sample

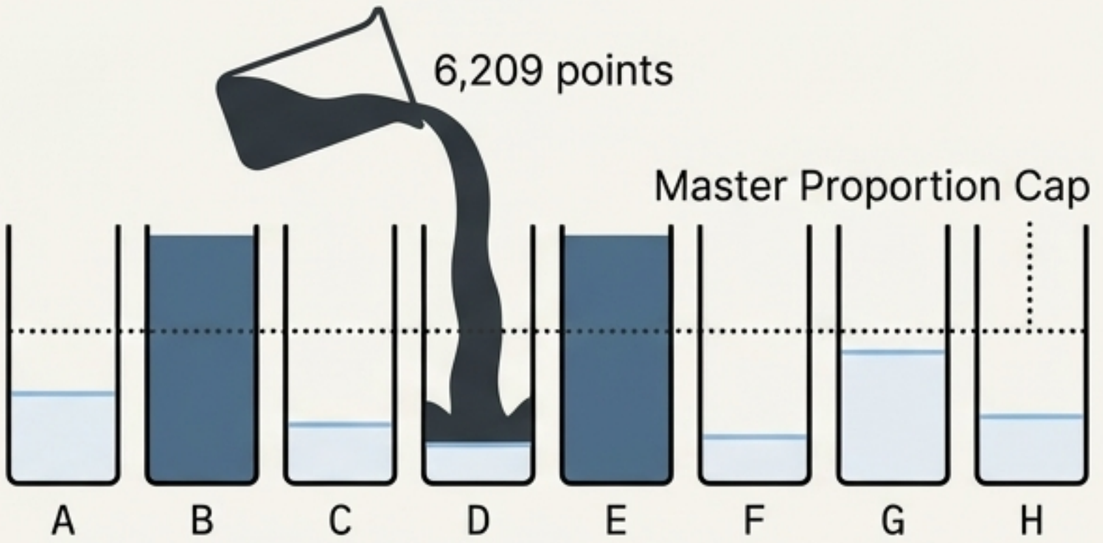
Objective and Rules

- **Objective:** To select the remaining 6,209 points in a way that aligns the final 130,000-point sample's land cover (`LC_pred`) distribution with that of the master frame.

The "Water-Filling" Allocation Strategy: A constrained, iterative algorithm is used to allocate the 6,209 points across LC_pred classes.

Constraints & Rules:

1. **Fixed Classes:** No new points are added to classes 'B' and 'E', as they are already over-represented from the thematic modules.
2. **Equalization:** The algorithm iteratively adds one point at a time to the most under-represented class among 'A', 'D', 'F', 'G', and 'H', seeking to equalize their absolute counts.
3. **Master Cap:** A class stops receiving points if its projected final count reaches the proportion defined by the master frame.
4. **Residual:** Any remaining points in the budget after the equalization process are allocated to class 'C' (capped by availability).



Output: The Allocation Plan

LC_pred Class	Points to Add (n_add)
A	1,745
F	527
G	2,705
H	1,232
B, C, D, E	0
Total Added	**6,209**

The Final 130,000-Point LUCAS 2027 Sample

Selection and Integration:

- The 6,209 "Copernicus" top-up points are selected from the available master frame points (905,870 candidates) using PRN-based ranking within strata (`LC_pred * NUTS2 * STR25` in IBM Plex Mono).
- These points are merged with the 123,791 existing points to create the final, deduplicated sample of **130,000 points**.

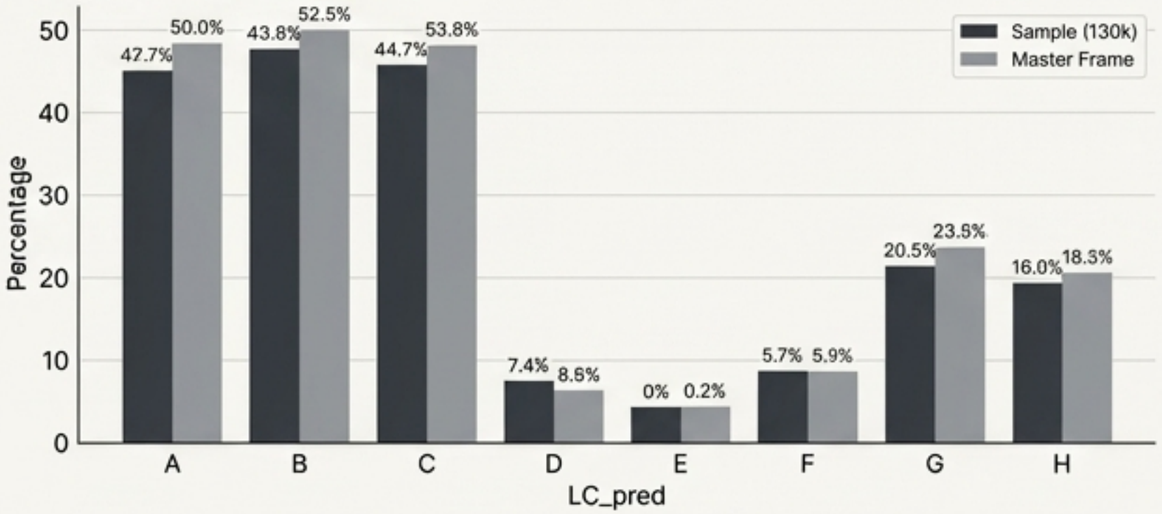
Final Output: `LUCAS27_sample.csv`

- The file contains the 130,000 points with flags indicating which module(s) each point belongs to (SOIL, GRASSLAND, LF, COPERNICUS) and their corresponding final calibrated weights (`WGT_module_27` in IBM Plex Mono).

Final Verification: Sample vs. Master Frame

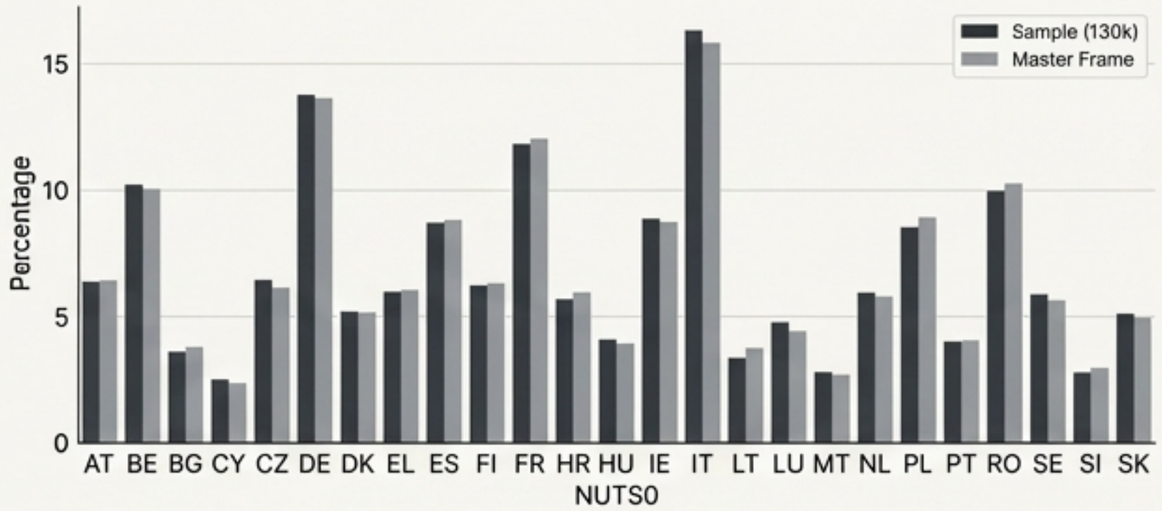
Final Land Cover Distribution (Sample vs. Master)

Sample distribution now closely mirrors the Master Frame, showing excellent alignment, particularly in classes A, G, and H.



Final Geographic Distribution (Sample vs. Master)

Geographic representativeness across NUTS0 regions remains excellent, with sample proportions matching the Master Frame.



The result is a fully integrated, multi-purpose survey sample, calibrated to provide a statistically sound basis for the LUCAS 2027 survey.