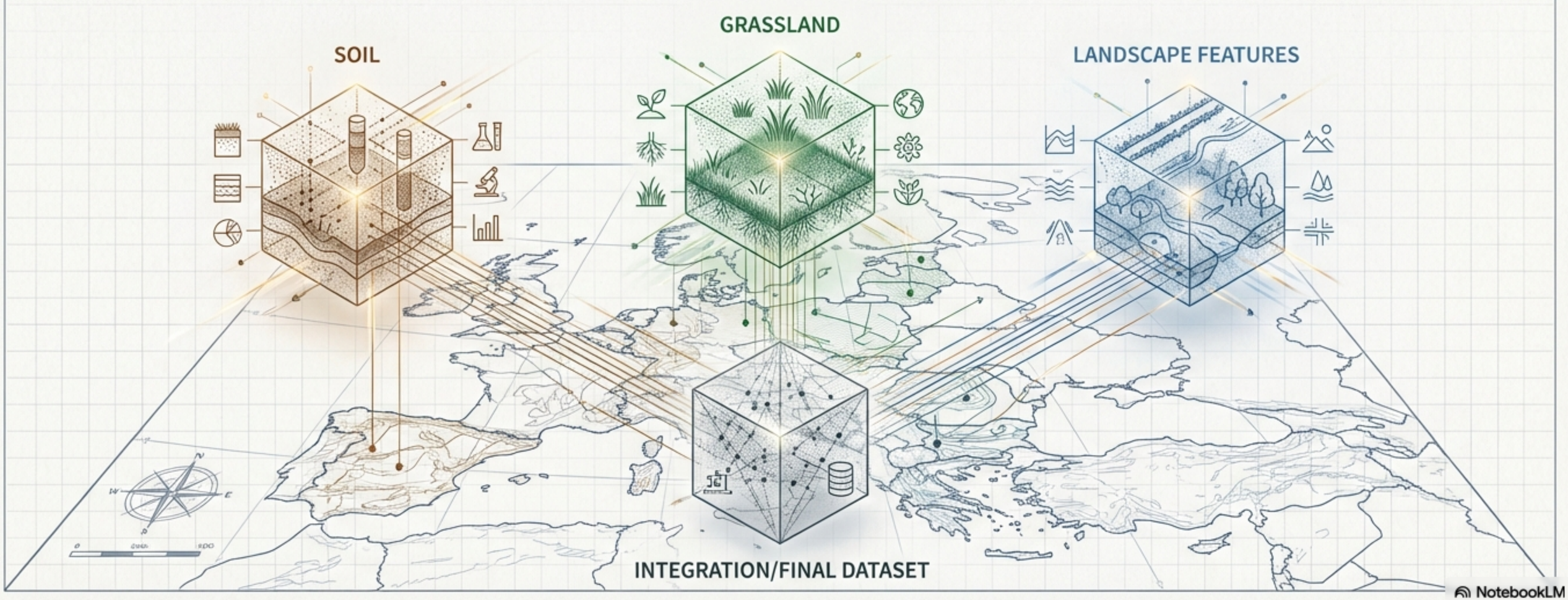
Constructing the LUCAS 2027 Survey Sample

A Methodological Journey from Thematic Modules to a Unified Dataset



A Modular Approach to Building a Robust 130,000-Point Sample

The LUCAS 2027 sample is constructed through a sequential, multi-stage process. We begin by building distinct thematic modules, each with its own sampling logic and objectives. These modules are then integrated and supplemented to create the final, comprehensive survey sample. This transparent, step-by-step methodology ensures statistical robustness and fitness-for-purpose for each thematic area.

Module 1: Preparing the Soil Sample

INPUTS

Raw Data
lucas_soil_sample20251203.csv (25,000 points)

Reference Frame
master_complete.RData (provides geographic and land cover attributes)

Ancillary Data
- Exclusion lists (Soil_to_be_removed.xlsx)
- 2022 Survey Weights (Survey_2022_wgt_2nd_phase.txt)

PROCESS

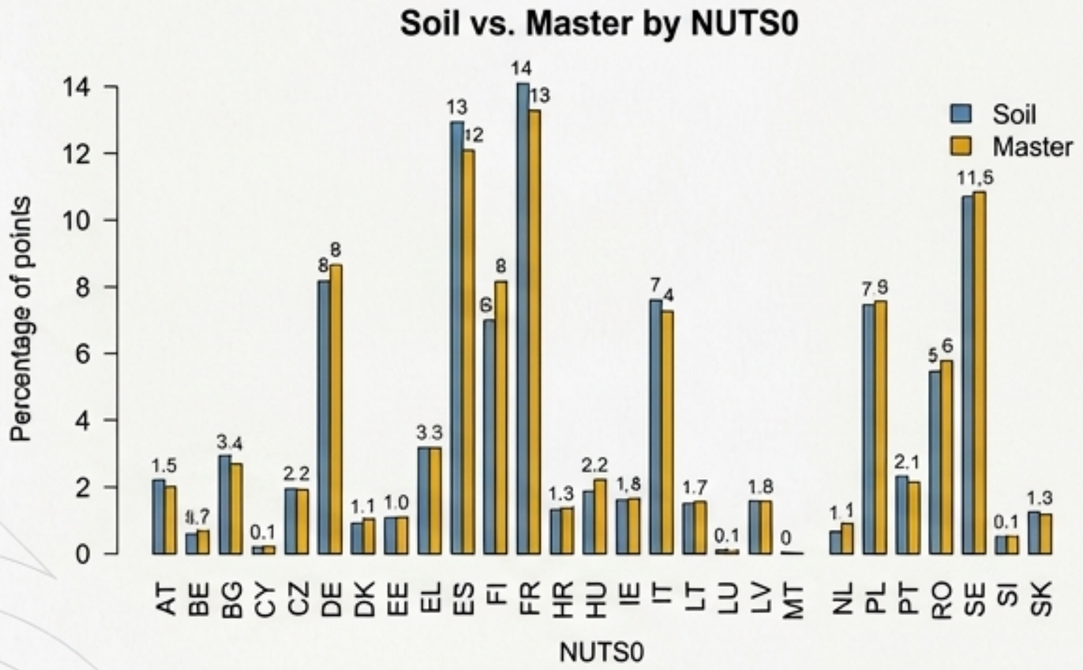
1. **Cleaning**
   Points identified in various exclusion lists are removed from the raw sample.
2. **Enrichment**
   The sample is merged with the master frame to attach key stratification variables (e.g., NUTS2_24, `LC_pred).
3. **Weight Calculation**
   New stratum-level weights (WGT_SOIL_LC and WGT_SOIL_STR25) are computed by comparing the sample counts to the master frame counts within each stratum. This ensures the sample is representative of the master population.

OUTPUTS

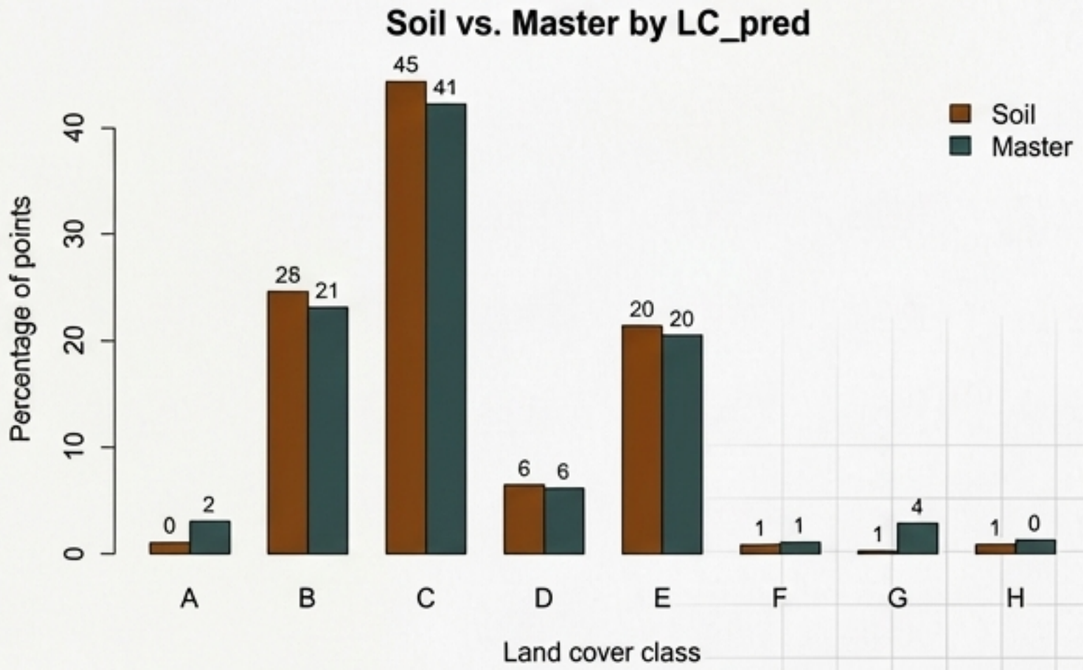
Final Module Sample
`Soil2027_sample.csv (a clean, enriched dataset of 25,000 points with calculated weights).

Quality Control: The Soil Sample Aligns with the Master Frame

A critical quality check is to compare the distribution of the prepared sample against the master frame population across key variables. The prepared 25,000-point soil sample demonstrates strong alignment with the master frame's geographic (NUTS0) and predicted land cover (LC_pred) distributions, confirming the process was successful.



The geographic distribution of the soil sample closely mirrors that of the master frame across all NUTS0 regions.



The sample's land cover distribution is highly consistent with the master frame, ensuring representativeness.

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

The Grassland module requires a more complex approach to address potential biases from non-response in the 2022 survey. To build a robust 20,000-point sample for 2027, we employ a two-component strategy that leverages the observed data from the previous survey while supplementing it with a broader sample.

Re-sampling Observed Points

Description: Starts with the 12,119 points actually observed in the 2022 Grassland survey. Uses weight correction to account for non-response. Focuses on core grassland land cover types ('D' and 'E').

Outcome: 11,099 points.

Targeted Top-Up Sample

Description: Selects additional points from the broader 2022 'Extended Grassland' sample to meet the total target size.

Outcome: 8,901 points.

Final Grassland Sample (20,000 points)

Grassland Component 1: Correcting for Non-Response from the 2022 Survey

Input

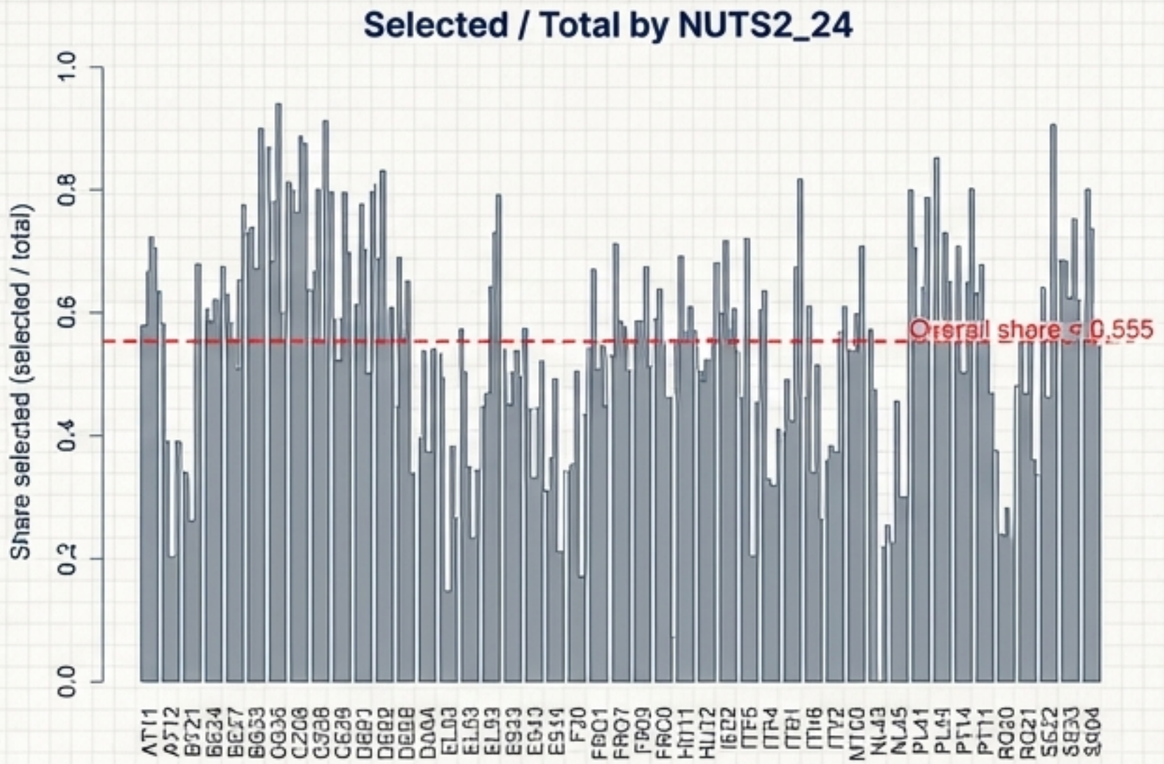
We start with the 19,998-point 2022 Grassland sample, of which only **12,119** were effectively observed in the field (`effective_points_modules.xlsx`).

Process

1. **Weight Correction**: A correction factor (`wgt_correction`) is calculated at the stratum level to inflate the weights of the 12,119 observed points. This adjustment ensures that the observed subset accurately represents the full 2022 sample, correcting for any non-response bias. The total corrected weight (`178,361.3`) closely recovers the original population total (`181,832.4`).
2. **Selection**: From the corrected set, we select all observed points where the 2022 surveyed land cover class was 'D' (Woodland) or 'E' (Shrubland).

Output

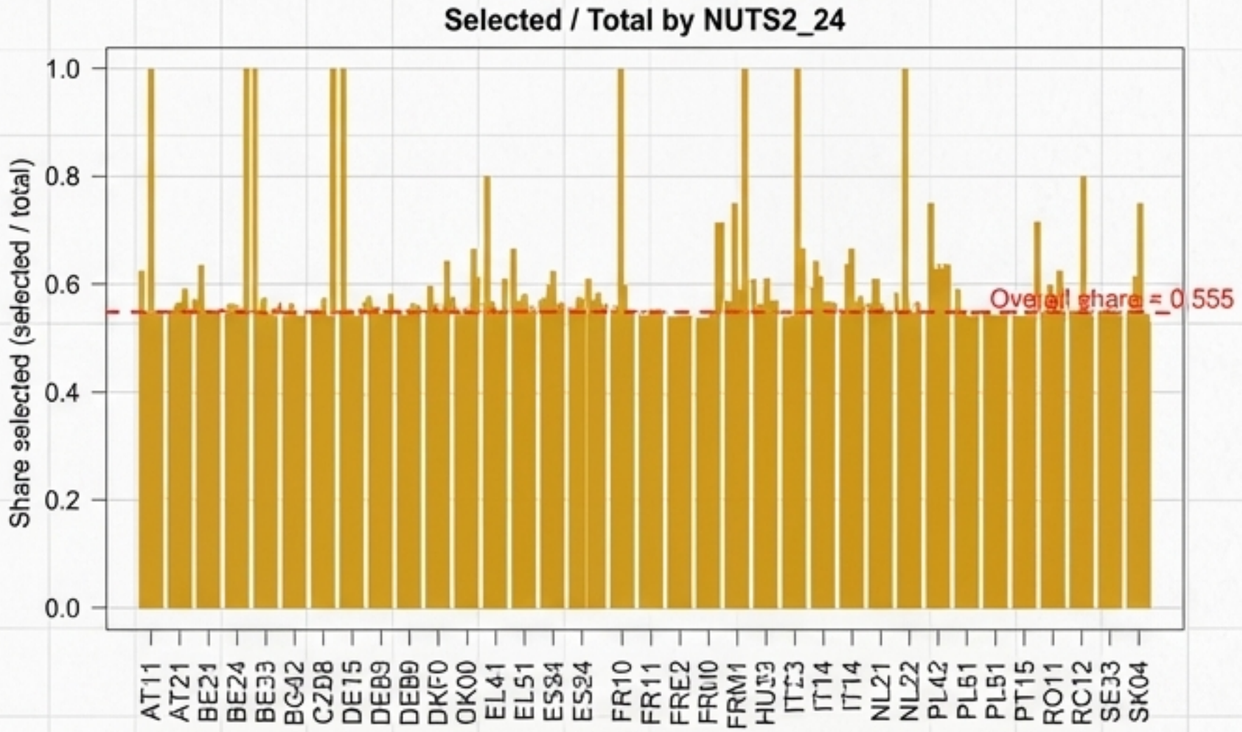
`Grassland2027_component1.csv`, a sample of **11,099** points.



The chart shows that the selection proportion varies significantly by region, with an overall share of 55.5% (indicated by the red dashed line).

Grassland Component 2: Completing the Sample with Proportional Allocation

- **Goal**: Select the remaining **8,901** points to reach the 20,000-point target.
- **Input**: The 2022 Extended Grassland sample, after removing points already selected for Component 1. This leaves a candidate pool of **16,031** points.
- **Process**: Proportional allocation is used to select the 8,901 points. The number of points taken from each `NUTS2_24` stratum is proportional to the number of available candidates in that stratum. Within each stratum, points are selected deterministically based on the highest pseudo-random number (PRN).
- **Output:** `Grassland2027_component2.csv`, containing the additional **8,901** points.



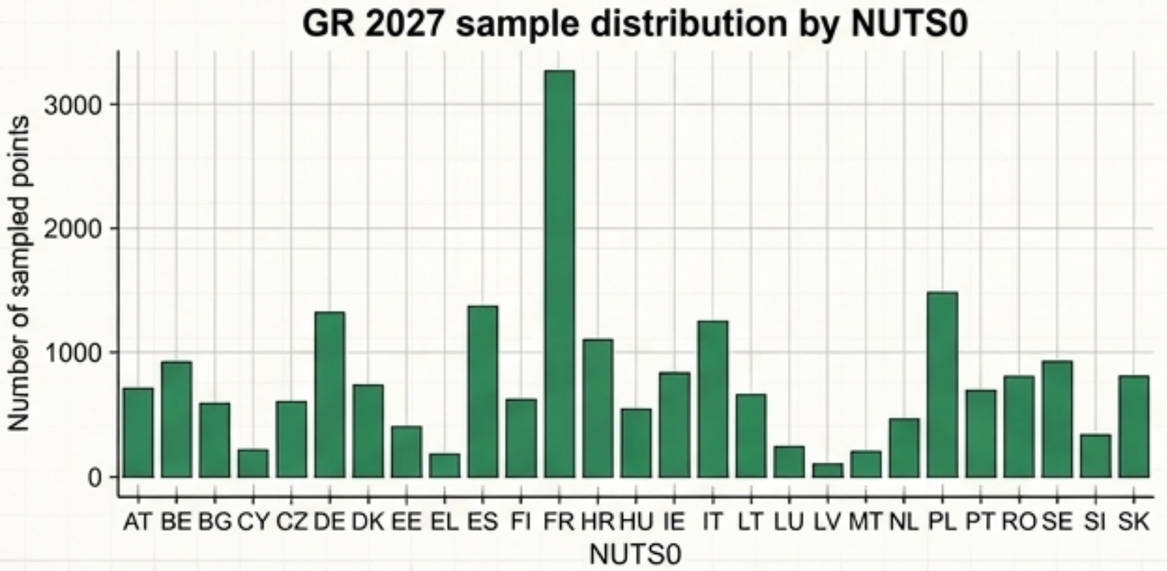
Key Takeaway: This allocation method results in a more uniform selection rate across most regions, hovering around the overall share of 55.5% (red dashed line).

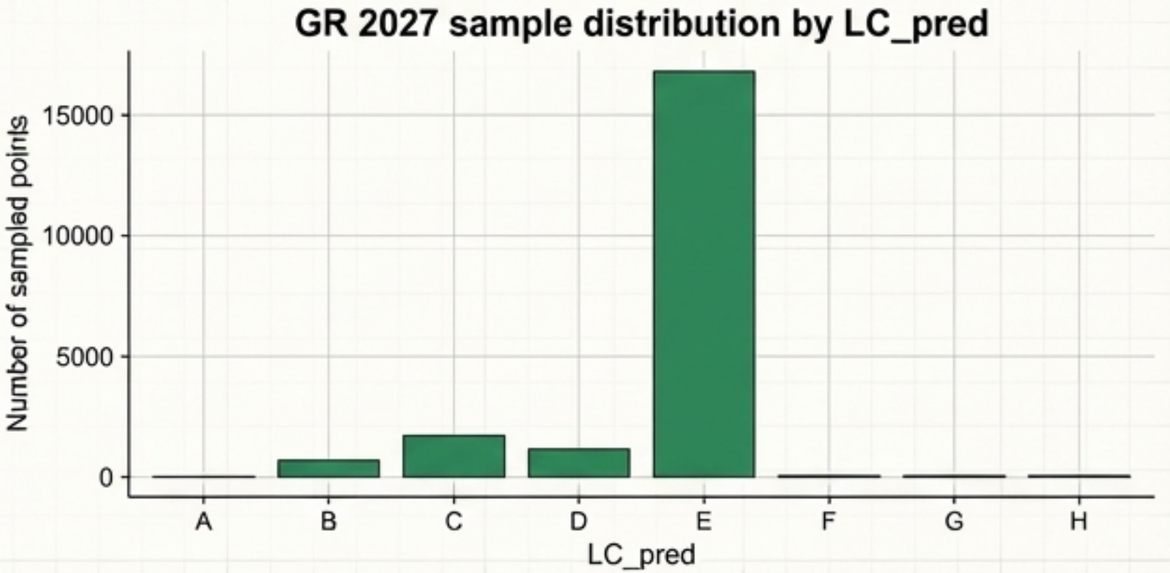
Grassland Finalization: Combining Components and Calibrating Weights

1. Combine: The ‘Component 1’ (11,099 points) and ‘Component 2’ (8,901 points) datasets are merged into a single 20,000-point sample.

2. Calibrate: A final post-stratification weight adjustment is performed. The weights of the combined sample are calibrated to match the known population totals within each stratum (NUTS2 * STR25) from the master frame. The target population consists of all ‘273,497’ master points with predicted land cover of ‘D’ or ‘E’. This ensures the final sample accurately represents the true population structure.

Output: The final ‘Grassland2027_sample.csv’, a 20,000-point sample with fully calibrated module weights (‘WGT_module_27’).

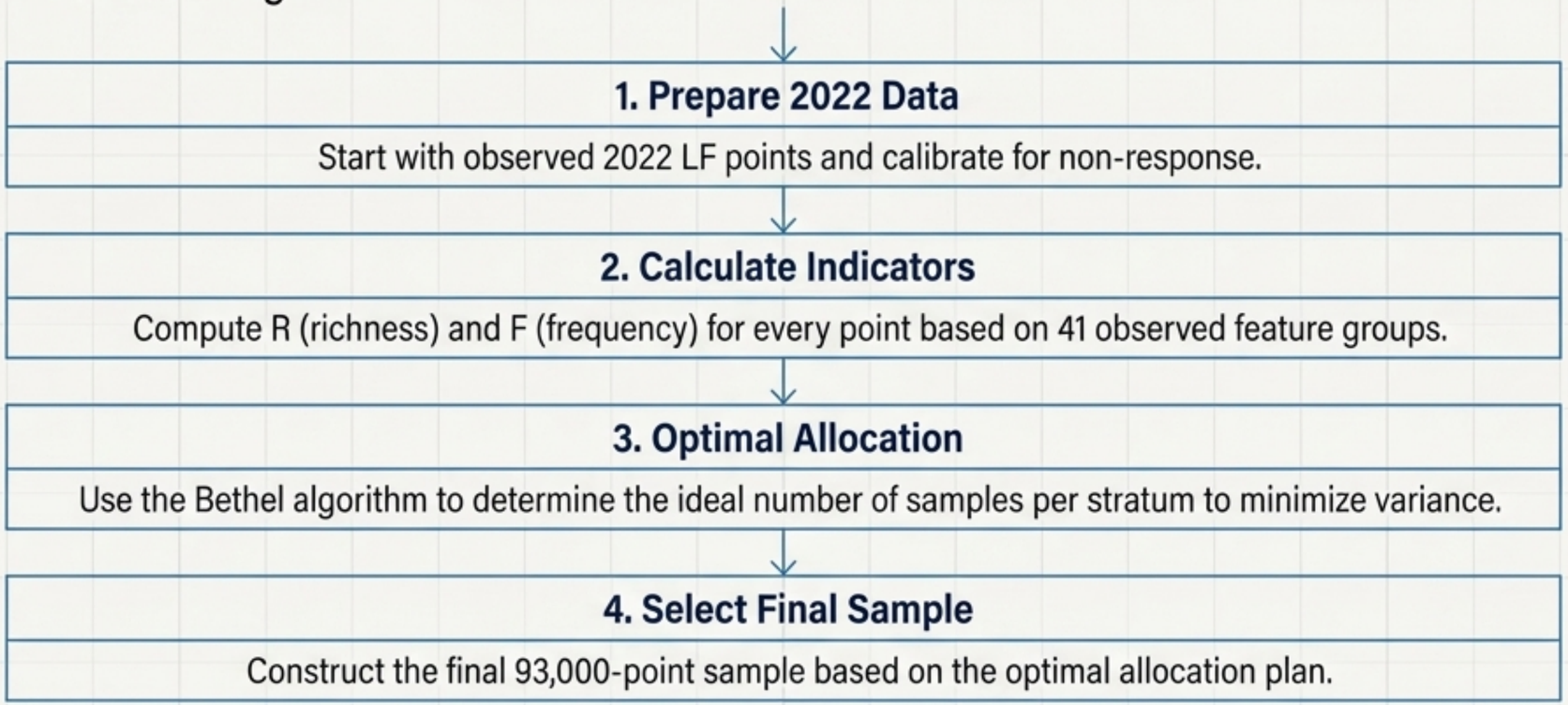




Module 3: A Statistically Optimized Sample for Landscape Features

Objective: To create a highly efficient **93,000-point** sample specifically designed to estimate the **Richness (R)** and **Frequency (F)** of landscape features with high precision.

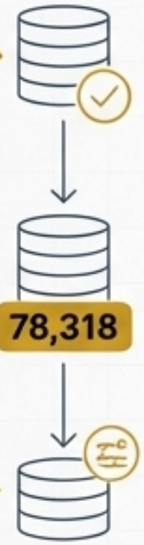
Methodological Challenge: Standard sampling is inefficient for features that may be rare or clustered. This requires an advanced, multi-step process to optimize the sample allocation, ensuring that we get the most statistical power for our budget.



LF Step 1 & 2: Preparing the Frame and Deriving Key Indicators

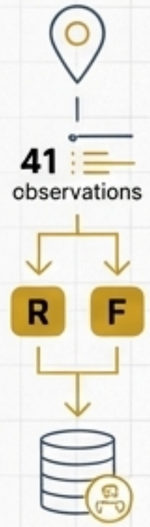
Step 1: Preparing the 2022 Observed Data (Input)

- We start with the 2022 LF sample, filtering for points that were both effectively observed and met eligibility criteria (`STR25 in c(1,2,3)` or `LUobs == 'U11'`).
- This results in a starting dataset of **78,318** observed, eligible points (LF_sample_2022_obs.csv).
- Weights are calibrated to correct for non-response within each LF stratum (NUTS2*STR25).



Step 2: Calculating Richness and Frequency Indicators (Process)

- For each of the 78,318 points, we analyze the 41 landscape feature observation fields.
- **R (Richness)** is calculated as the proportion of the 41 groups where any LF was recorded.
- **F (Frequency)** is a binary indicator, set to 1 if R > 0 (any LF present) and 0 otherwise.
- These indicators are then used to calculate the weighted mean and standard deviation of R and F for every stratum.



Output:
A stratum-level statistics file (LF_stratum_stats.csv) that serves as the direct input for the allocation algorithm.

LF Step 3: Using the Bethel Algorithm for Optimal Stratum Allocation

Objective: To determine the most efficient distribution of sample points across strata to meet a predefined level of precision for our target indicators (R and F).

Inputs

Stratum-level statistics (`**LF_stratum_stats.csv**`)

- Includes: Population size (N), Mean (M), and Standard Deviation (S) for R and F.

Constraint: Target Coefficient of Variation (CV) of **1%** for R and F estimates at the European level.

Bethel Algorithm Engine

`SamplingStrata` package in R

Output

An allocation plan (**LF_strata_with_bethel_allocation.csv**)

- Specifies the optimal sample size (`nh_opt`) for each of the 1,500+ strata.
- For a target of **46,500** points (the panel component), the algorithm distributes them optimally.

LF Step 4: Assembling and Calibrating the Final 93,000-Point Sample

A Two-Component Sample

Component 1 (Panel)

46,500 points selected from the 2022 observed LF dataset according to the optimal Bethel allocation. These points provide a link to the previous survey.

Component 2 (Non-Panel)

An additional **46,500** points are selected from the master frame (excluding panel points). This selection uses the same Bethel allocation counts to expand the sample while maintaining statistical optimality.

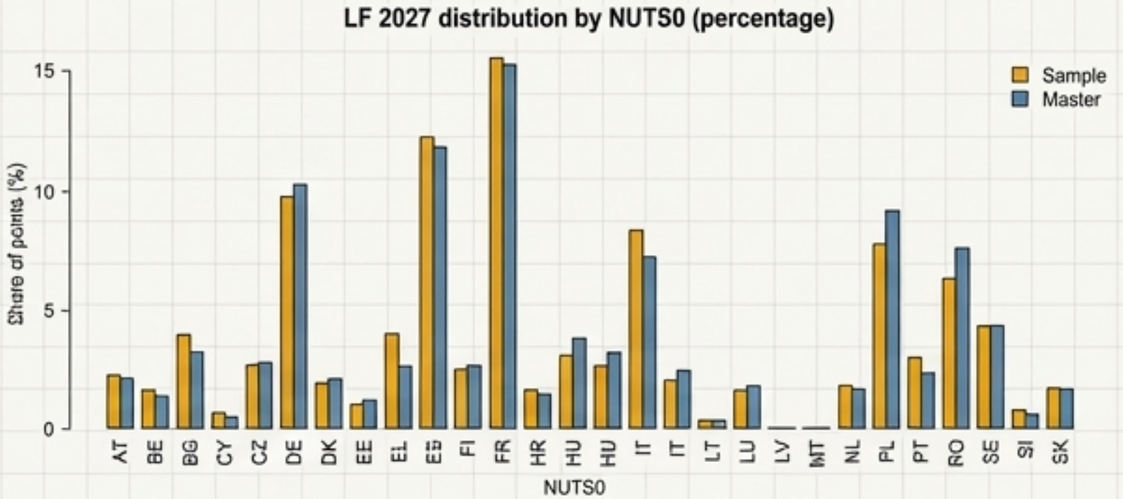
Finalization

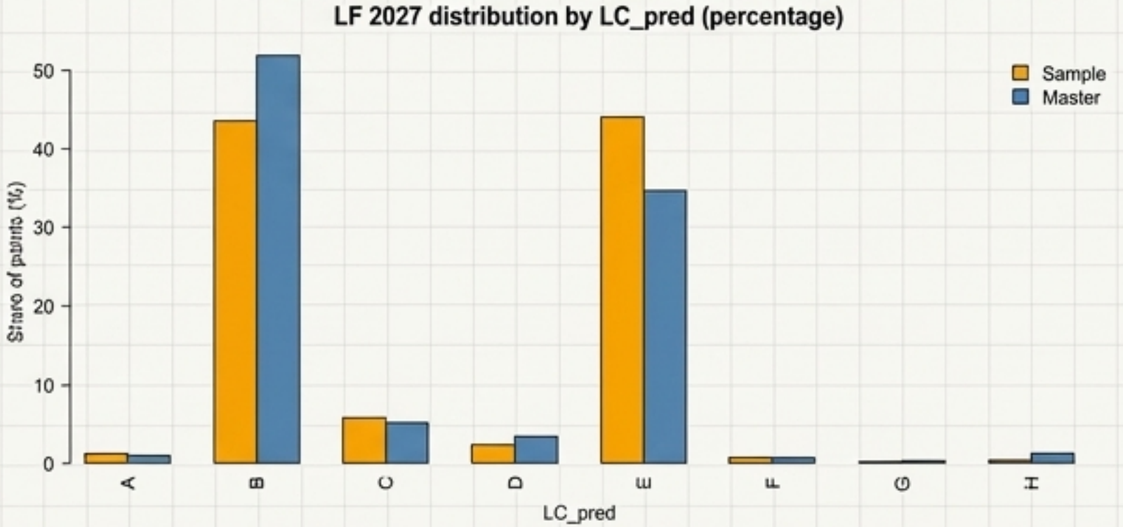
The two components are combined to form the full 93,000-point sample.

A final weight calibration is performed, adjusting the sample weights to align with the true population totals from the master frame's eligible LF population (**431,917** points).

Output

The final `LF2027_sample.csv`, and quality control charts comparing its distribution against the master frame. The charts confirm the final sample is geographically and thematically representative.





The Grand Integration: Assembling the Modules to Reach 130,000 Points

Step 1: Combine and Deduplicate Thematic Samples

- The points from the Soil (25k), Grassland (20k), and LF (93k) modules are combined.
- After removing duplicate POINT_IDs that appear in multiple modules, we arrive at a total of **123,791 unique points**.

Step 2: Calculate the Top-Up Budget

- Final Target Sample Size: 130,000 points.
- Current Unique Points: 123,791 points.
- Remaining Budget: **6,209 points**. These points will form the ‘Copernicus Top-Up” sample.

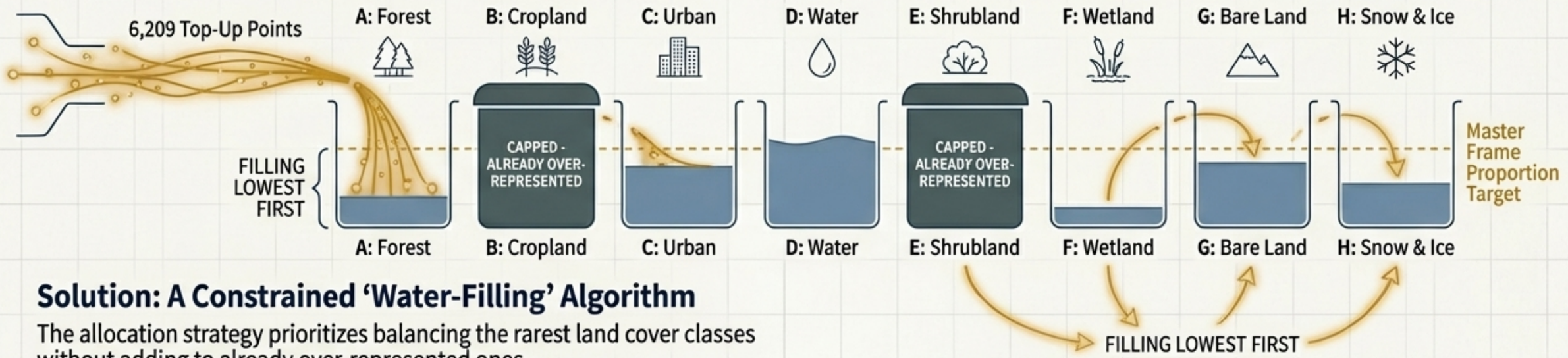
Module Overlap (Number of Common Points)

	SOIL	GRASSLAND	LF
SOIL	24,987	717	2,625
GRASSLAND	717	20,000	11,240
LF	2,625	11,240	93,000

The matrix shows how many points are shared between modules. For example, 11,240 points serve the needs of both the Grassland and LF surveys.

The Copernicus Top-Up: A ‘Water-Filling’ Algorithm to Balance the Final Sample

Challenge: How should the final 6,209 points be allocated to best improve the overall sample’s land cover representativeness compared to the master frame?



Solution: A Constrained ‘Water-Filling’ Algorithm

The allocation strategy prioritizes balancing the rarest land cover classes without adding to already over-represented ones.

- **“Fix”:** Classes ‘B’ (Cropland) and ‘E’ (Shrubland) are already over-represented in the 123,791-point sample. Their allocation is fixed at zero additional points.
- **“Equalize”:** The 6,209 points are iteratively assigned, one by one, to the most under-represented classes among ‘A’, ‘D’, ‘F’, ‘G’, and ‘H’. This process ‘fills’ the lowest levels first, bringing all classes towards a common level of representation, but is capped so as not to exceed the target proportions from the master frame.
- **“Residual”:** Any remaining budget after equalization is assigned to class ‘C’.

Final Allocation & Selection

— The algorithm determines the precise number of points to add for each class: **A (1,745)**, **F (527)**, **G (2,705)**, **H (1,232)**.

— These 6,209 points are then selected from the available master frame points using a stratified design and PRN ranking.

The Final LUCAS 2027 Sample: A Calibrated and Cohesive Dataset

The Final Product

The result of this multi-stage process is `LUCAS27_sample.csv`, a single data file containing 130,000 unique survey points.

Multi-Purpose

Each point is flagged for its inclusion in the Soil, Grassland, LF, and Copernicus modules.

Thematically Weighted

The dataset includes distinct, calibrated weights for each module ('WGT_SOIL', 'WGT_GRASSLAND', etc.), enabling valid statistical estimation for each specific theme.

Ready for Fieldwork

The sample is complete with geographic coordinates and key attributes, ready for the 2027 survey campaign.

The final sample's distribution shows excellent alignment with the master frame, demonstrating that the meticulous construction and top-up process has produced a balanced and robust dataset.