# New LUCAS 2022 sample and subsamples design: Criticalities and solutions

MARCO BALLIN, GIULIO BARCAROLI, MAURO MASSELLI

2022 edition

STATISTICAL
WORKING PAPERS

eurostat

# New LUCAS 2022 sample and subsamples design: Criticalities and solutions

MARCO BALLIN, GIULIO BARCAROLI, MAURO MASSELLI

2022 edition

Manuscript completed in June 2022



Luxembourg: Publications Office of the European Union, 2022



Theme: General and regional statistics
Collection: Statistical working paper

ISBN 978-92-76-53401-3 ISSN 2315-0807 doi: 10.2785/957524 KS-TC-22-005-EN-N

# Abstract

The Eurostat Land Use/Cover Area frame Survey (LUCAS) is mainly an in-situ survey designed to provide harmonized statistics on Land Use and Land Cover across the European Union. After the end of the pilot phase (2006), Eurostat undertook this survey every three years; the latest LUCAS survey took place in 2018 and covered all 28 EU countries by observing 340,000 out of one million points selected from a Master sample. The next round takes place in 2022, with 400,000 points observed, half of them directly in the field and the other half through photointerpretation.

In addition to the Land cover and Land Use observed at each of these points, the surveyors collect further information for specific modules added over time to assess environmental factors such as the grassland and soil quality or the quality of satellite imagery such as that of the Copernicus Programme.

The sample design for the new LUCAS in 2022 had to take into consideration not only the direct requirements of LUCAS in terms of target estimation accuracy (mainly related to Land Use and Land Cover), but also the specific requirements of the five different linked modules (Soil, Grassland, Extended Grassland, Landscape Features, Copernicus). As the Master sample served as the sampling frame for the selection of the LUCAS sample, the latter can in turn be considered as the sampling frame for each of the five module subsamples. A crucial issue was the correct determination of the different eligibility statuses for each point in order to select only points from the specific population of reference for a given module. This was possible by considering not only the observed values from previous rounds but also the predicted values, obtained by applying a machine learning approach.

This paper explains the general approach to the design of the LUCAS total sample (focusing on optimizing the stratification of the Master sample and the allocation of points to strata). For this purpose, all methods used to enrich the Master sample with the information needed to define the eligibility status for the different modules shall be reported, as well as the criteria used to select the appropriate subsamples.

Keywords: land cover and land use survey, soil, grassland, landscape features, Copernicus, optimal sampling.

JEL Classification: C83 (Survey Methods • Sampling Methods)

# Table of Contents

# List of Tables

# List of Figures

# Abbreviations

| | |
|---|---|
| AWF | Additional Woody Feature |
| CLC | CORINE Land Cover |
| COP4N2K | Copernicus for Natura 2000 |
| DG | Directorate General |
| DMT | Data Management Tool |
| EC | European Commission |
| EU | European Union |
| FI | Field observation |
| GIS | Geographical Information System |
| HRL | High Resolution Layers |
| IDs | Identifiers |
| INSPIRE | INfrastructure of SPatial InfoRmation in Europe |
| LC | Land Cover |
| LU | Land Use |
| LUCAS | Land Use/Cover Area frame Survey |
| NUTS | Nomenclature des Unités Territoriales Statistiques |
| PI | Photo Interpretation |
| SOC | Soil Organic Carbon |
| SRS | Simple Random Sample |
| SWF | Small Woody Feature |

Grassland Regions:

| | |
|---|---|
| 1 | Atlantic-North + west |
| 2 | Boreal - Scandinavia + Baltic Sea |
| 3 | Atlantic – South + East |
| 4 | Continental – North |
| 5 | Mediterranean – West + Central |
| 6 | Continental – South |
| 7 | Pannonian |
| 8 | Continental – East |
| 9 | Steppic + Black Sea region |
| 10 | Mediterranean – East |

# 1 Introduction

## 1.1. LUCAS Survey: Objectives and a short history

Within the context of quality and completeness improvement for the Land Cover and Land Use statistics, Eurostat has conducted the LUCAS survey every three years since 2006. The surveys are used to monitor the social and economic use of land as well as ecosystems and biodiversity. Sustainable development indicators and agro-environmental indicators for soil are examples of LUCAS data use, while the micro-data collected also serve to produce, verify, and validate CORINE Land Cover (CLC) and Copernicus Programme. The LUCAS surveys provide three types of information: (i) micro-data containing the statistical information collected at each sample point, (ii) point and landscape photographs, and (iii) statistical tables with aggregated results by Land Cover and Land Use at the geographical level.

The survey is based on a sample of geo-referenced points, selected from a frame of more than 1 million points belonging to the intersections of a 2 square km grid laid over the entire EU territory, the so-called Master sample or first phase sample. It is carried out by two data collection modes: the direct observations by surveyors in a small area centred around the selected point (a circle with a radius of 1, 5 meter or in some cases 20 meters) and the photointerpretation. In the first case for each point, information is collected on Land Cover (i.e. the bio-physical cover of the land, such as natural areas, forests, buildings, roads and lakes), on Land Use (i.e. the socio-economic use of the land, such as agriculture, commerce, residential use or recreation) and other variables linked to Land Cover. Surveyors also take a series of photographs of the point itself, and of what lies in all four cardinal directions (north, south, east and west). In addition to the LUCAS core data, each survey also collects some specific information called 'ad hoc modules': This was the case for the topsoil sample in 2009, 2015 and 2018, the transects in 2009, 2012 and 2015, the grassland module and the Copernicus supplementary points in 2018.

Photointerpretation (PI) is carried out both, as PI in the office for points that were not intended to be visited in the first place, according to the sampling design and depending on specific conditions known "a priori" (e.g. geographical factors), and PI in the field whenever a point is not accessible to the surveyor because of an obstacle (refusal of the owner of a private property, military area, long distance from a road, etc.). In this way, there are practically no missing units in the collected data.

The legal basis of the LUCAS survey has evolved over the years. A pilot 'Land Use and Cover Area Frame Survey (LUCAS)' was launched by DG Agriculture and Eurostat in 2000 on the basis of Decision 1445/2000/EC of 22/5/2000 of the Council and the European Parliament on the application of area frame techniques. In 2001 (postponed to 2002), the first LUCAS pilot survey was carried out in 13 of the 15 Member States of the European Union. The survey was repeated in 2003 in all 15 EU Member States as well as in Hungary, which allowed the improvement of the data collection system and the analysis of changes in Land Use and Land Cover (2001–03). The duration of the project was extended from 2004 to 2007 by Decision 2066/2003/EC from 10/11/2003.

The coverage of the EU Member States and the corresponding funding are defined in Decision 786/2004/EC from 21/4/2004. In 2006, a pilot survey was carried out in 11 Member States (Luxembourg, Belgium, Czech Republic, Germany, Spain, Poland, Italy, France, the Netherlands, Hungary and Slovakia) to test the methodology at EU level with a restricted budget by introducing the current data collection frequency: every three years. Since January 2008, LUCAS has been part of Eurostat's activities and budget as since 2012, it has been financially supported by other DGs of the Commission. The original coverage was extended to 23 EU countries in 2009 (Bulgaria, Cyprus, Malta

and Romania were not included), 27 Member States in 2012, to 28 member states in LUCAS 2015 and 2018, and to 27 countries in 2022 because of the exit of UK from the EU.

The sample size has increased accordingly. In each round of the survey, the methodology was improved to obtain further and more accurate data while maintaining the comparability between the different editions.

In LUCAS 2009, the sampling rates were fine-tuned by considering the coefficient of variation for the strata defined by NUTS2 and the variable STR05 (a classification in seven modalities of the Land Cover variable), obtained by photointerpretation for each point of the Master. In addition, certain points of the Master were excluded from sampling as non-accessible points belonging to islands not connected to the mainland or too expensive to be surveyed or located at an altitude higher than 1000 m. All these points were considered as 'non-eligible' for the field survey and the sample was selected from the complementary set of 'eligible' points. This subdivision of the Master was maintained, albeit with modifications, until the 2015 survey.

LUCAS 2012 aimed at improving the precision of the estimates by increasing the sample size by about 40 000 points and distributing them according to the diversity of the landscape resulting from the transect analysis carried out in 2009. The elevation criterion for eligibility was also raised to 1 500 meters. Furthermore, some auxiliary information, such as slope and distance to the main road, was introduced to optimize point selection. In LUCAS 2015, more sophisticated criteria for assessing the eligibility of a point was introduced by combining information derived from the CORINE Land Cover (CLC) with distance to roads and altitude. The new criteria allowed for better identification of non-eligible points and more rational use of photo-interpreted points. The bias caused by the exclusion of non-eligible points was corrected by the photointerpretation of a complementary sample of about 50 000 non-eligible points and using a calibration by classes of elevation in the estimation.

In LUCAS 2018, some important changes to the survey methodology were introduced. The distinction between eligible and non-eligible points was removed and all the points were considered available for the sample selection, requiring them to be collected either directly by the surveyors or through photointerpretation performed in the office. In this way, the focus shifted from the concept of eligibility in the Master to the mode of data collection after sample selection: photo interpretation is used when it is impossible or too costly to reach the point or it is convenient where the probability of the point to change its Land Cover characteristics is low. The rule to decide the mode for collecting data in the sampled points ('PI ex-ante' or 'in field') is based on Reachability and Propensity to Change indexes calculated for each Master point.

A second methodological revision concerns stratification. In the previous surveys, the number of strata in each sample was determined ex-ante combining the number of NUTS2 by all available modalities of STR05 in each region. In the 2018 survey, an iterative optimization algorithm that, starting from the 'atomic strata', aggregates them considering the coefficient of variations of the target variables (16 modalities of Land Cover) and the related desired sampling errors (required by Eurostat) identified the strata. The 'atomic strata' are identified by the Cartesian product of the variable STR18 (an update of the previous variable STR05 in ten modalities of the variable Land Cover obtained by a new photointerpretation of the Master points), CLC (three digits classification) and four classes of elevation. The target variables are estimated for each Master point by a logistic model based on the outcomes of the previous survey and used for calculating the CVs utilized for the strata aggregation. The optimization is performed individually for each NUTS2 domain, and the results are then aggregated at country level. The iterative algorithm optimizes the stratification, aggregating the atomic strata to minimize the overall sample size required to meet the precision constraints (the CVs of the target variables). Stratification is thus not obtained ex-ante by a fixed combination of variables but depends on the correlation between the stratification characteristics and the target variables; the combinations of stratification criteria vary according to the specificity of the NUTS2 areas, which are assumed to be, as in previous surveys, the minimum territorial study domain.

The selection of sampling points was consequently changed: In each stratum, the sampling units were selected by a simple random sample (SRS) technique, according to the optimal allocation determined jointly with the best stratification, whereas in 2015 the selection was systematic, and their number was proportional to the stratum size.

## 1.2. Main features of 2022 LUCAS sample and a short overview of the next chapters

The methodology of the 2022 sample essentially follows the innovations introduced in the 2018 survey, but some improvements have been made in different steps of the survey depending on the analysis of the previous round.

Eurostat has increased the LUCAS sample to 400 000 points for 2022, compared to around 336 000 points in 2018, while the number of countries involved has decreased by one (United Kingdom). With the same number of countries, the 'in field' sample decreases slightly (about 8 %) but the photo-interpreted points are twice as high as in 2018. This decision has the dual objective of increasing the reliability of the estimate while limiting the cost of the increased precision of the estimates. All sample sizes at country level have increased compared to 2018, on average by 25 %, and the distribution of the sample by country reflects substantially the previous one, even if some adjustments are made taking into account the sampling errors calculated using the 2018 data.

The total area classified according to the modalities Land Cover (one-digit classification) and Land Use (one -digit classification) was considered as the target of the survey, and for each of these twelve variables the desired precision was defined in close cooperation with Eurostat. In order to allocate the sample size in each stratum using Bethel's multivariate algorithm, the coefficients of variation of the twelve target variables must be calculated. To achieve this goal, each point in the Master was assigned the target variables that were observed in 2018 or 2015 (if the point was not also observed in 2018) or predicted by a Random Forest Model. The sample points were selected by a balanced spatial sampling to account for spatial correlation.

The Master (i.e. the sampling frame) was stratified by combining in each NUTS2 all available modalities of STR18 and CLC Land Cover (two-digit classification), with the binary indicator signalizing if the assigned (to the point) Land Cover is predicted or not; the resulting total number of strata is 22 173. To choose the stratification criteria, different trials have been carried out; in particular, the flag indicating if the LC variable observed or predicted in 2018 has been selected as its inclusion has the effect of reducing the variability of the same variable in the strata.

Once selected, a point must be assigned to the observation mode: directly surveyed or photo interpreted in the office. In 2022, the assignment of the observation mode is particularly important to ensure direct observation of the points belonging to the subsamples related to five 'modules' used to collect specific information. A Reachability index is calculated by the Random Forest Model for all the Master units to distinguish between direct observation and photointerpretation of a point. Moreover, the index for the sampled points is complemented by deterministic rules that take into account other factors that contribute significantly to the probability of change of a point. In addition, an application has been developed that allows the geographical visualization of LUCAS 2022 selected points to facilitate the assignment of ambiguous situations.

The selected LUCAS sample constitutes the selection list for the module subsamples (Grassland, Extended Grassland, Soil, Landscape Features and Copernicus) and it has been analysed to assess its sustainability. To ensure this function, it is necessary that all the eligibility criteria are available on the sampled points. The variables Land Cover (1 digit), the code 2 of Corinne Land Cover, STR18, the 'in field' observation type, the variable Land Use U11, Grass Percentage > 30 % and the soil organic carbon (SOC) are predicted by the Random Forest Model for all points in the Master. Hence, all LUCAS selected points have the information necessary to implement the module subsamples. The eligibility criteria also define the reference populations for the modules that are required for the selection of the probabilistic samples. This means that sampling units are selected randomly, each of them is assigned an inclusion probability and thus the calculation of sampling errors for the estimates is possible. The definition of the eligibility criteria, the design of the subsamples and their selection were carried out in constant consultation with the Eurostat and other EC DGs responsible for the analysis of the module data.

The following chapters describe all the steps involved in obtaining the LUCAS sample and the subsamples of the modules. The preparation of the sampling frame (the Master), which contains all the information for designing the samples for each of the 1 090 863 items, is presented in Chapter 2.

The previous edition of the Master was completed with the variables acquired from the Copernicus programme (CORINE Land Cover and imperviousness) that were available in the different years. In addition, the predicted value for the main target variables to correctly assess and calculate their coefficients of variation in the strata, the predicted reachability score to distinguish between in field and PI units and the variables necessary for the eligibility of sampled units for the module subsamples. Chapter 3 presents all the steps involved in designing the LUCAS sample. A first overall sample size in the 27 EU countries was defined based on the analysis of the sampling errors of the 2018 survey for the variables Land Cover and Land Use and taking into account the quantity of points of the module 'Soil'

already observed in the previous surveys. Having established the desired precision (2.5 % of the estimates) for the main variables, the choice of stratification criteria and the procedure of the best allocation of the points in the strata in every country are described. The optimal sample derived from this procedure in all the countries is adjusted by comparison with the initial sample and so the final sample is obtained. Finally, the selection of the sample points, the assignment of initial weights and observation type to the selected units, the visual check module implemented to control the position and the kind of points are outlined. Chapter 4 describes the procedures for obtaining the subsamples related to the five modules and highlights their specificities and commonalities. Chapter 5 summarizes the main topics of the previous chapters, highlighting the differences with the 2018 sample design, and provides reflections on the possible strategies for the future LUCAS survey.

# 2 Preparation of the sampling frame

## 2.1. Introduction

The enrichment and updating of the Master file was a necessary measure to make all steps of the new LUCAS process (sample design, sample selection, estimation, etc.) more efficient. Therefore, some auxiliary variables were added or updated at each point of the Master. In this context, the following groups of variables can be distinguished, in addition to those belonging to the previous editions of the survey:

- Variables acquired from the Copernicus program;
- Variables required to determine the type of observation, i.e. in-situ or photointerpretation;
- Prediction of additional variables required for modules eligibility.

## 2.2. Auxiliary variables from Copernicus program

The information provided by the Copernicus program had been collected according to a common methodology for the whole EU territory, is fully documented, freely usable, and regularly updated ([1]).

Most of these data are closely related to Land Cover (LC) and Land Use (LU) variables observed by the LUCAS program, so that they can be used for example for stratification or as predictor variables.

The following pan-European components of the Copernicus program had been attached to each point of the Master:

- CORINE Land Cover (reference years: 2000, 2006, 2012, 2018), a categorical variable with 44 classes (hierarchical 3-digit classification) describing LC and LU.
- Imperviousness (reference years: 2006, 2009, 2012, 2015, 2018), a numerical variable, indicating the degree of imperviousness in the range (0–100 %), the resolution is 20 mt for 2006–15 and 10 mt for 2018.
- Imperviousness built-up (reference years 2018), the resolution is 10 mt.
- Grassland (reference years: 2015, 2018), a binary variable (1=all types of grassland; 0=other). The resolution is 20 mt for 2015 and 10 mt for 2018.

([1]) https://land.copernicus.eu/product-portfolio/overview/

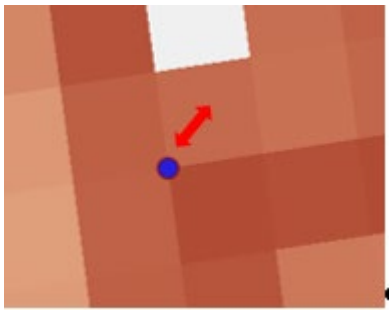
- Trees Density Cover (reference years: 2012, 2015, 2018), a numeric variable describing the level of tree cover density in a range from 0 % to 100 %. The resolution is 20 mt for 2012–15 and 100 mt for 2018.
- Forest Type (reference years: 2012, 2015, 2018), a three classes variable (0 all forest-free areas, 1 broadleaved forest, 2 coniferous forest). The resolution is 20 mt for 2012 and 2015 and 10 mt for 2018.
- Water and Wetness (reference years: 2015, 2018), a four classes variable (1 permanent water, 2 temporary water, 3 permanent wetness and 4 temporary wetness). The resolution is 20 mt for 2015 and 10 mt for 2018.
- Small Woody Feature and Additional Woody Features (reference years: 2015), a numerical variable describing the degree of density of SWF and AWF in a range from 05 to 100 %. The resolution is 100 mt.

From an operational point of view, providing the Copernicus project information to the LUCAS process means projecting each point of the LUCAS framework onto the cartographic representations of the main pan-European High-Resolution Layers (HRLs) and assigning to the points the values present in these representations. All GIS operations were performed using the following coordinate reference system:

'+proj=laea +lat_0=52 +lon_0=10 +x_0=4321000 +y_0=3210000 +ellps=GRS80 +units=m +no_defs'

In the Master, this projection is used to assign the coordinates X_LAEA and Y_LAEA to the individual points. Since the master points fall in the corners of the raster pixels, it was decided to assign the value of the upper right pixel to the point. This choice was justified by the fact that this is the convention used for naming the individual elements of the European grid.

**Figure 2.1: Example of Master point falling on the corner of a raster pixel**



## 2.3. Reachability

As reported in Ballin et al. (2018), the index of reachability was introduced to represent the difficulty an enumerator may face in reaching a given point. The index is in the range [0,1] (the higher the value, the easier the point is to reach) and was used to decide if the point should be subjected to an in-situ visit or photointerpretation as described in paragraph 1.2.

In the past, the degree of reachability was calculated using principal component analysis on the following information: the absolute difference in elevation between the altitude of the point and the one related to the nearest road (ABS_RATIO), the distance to the nearest point on a road (NEARDIST) and the angle to the nearest point on a road (NEARANGLE).

To update the index, the set of the auxiliary variables was extended and the Random Forest (James G., et al. 2013) method was used.

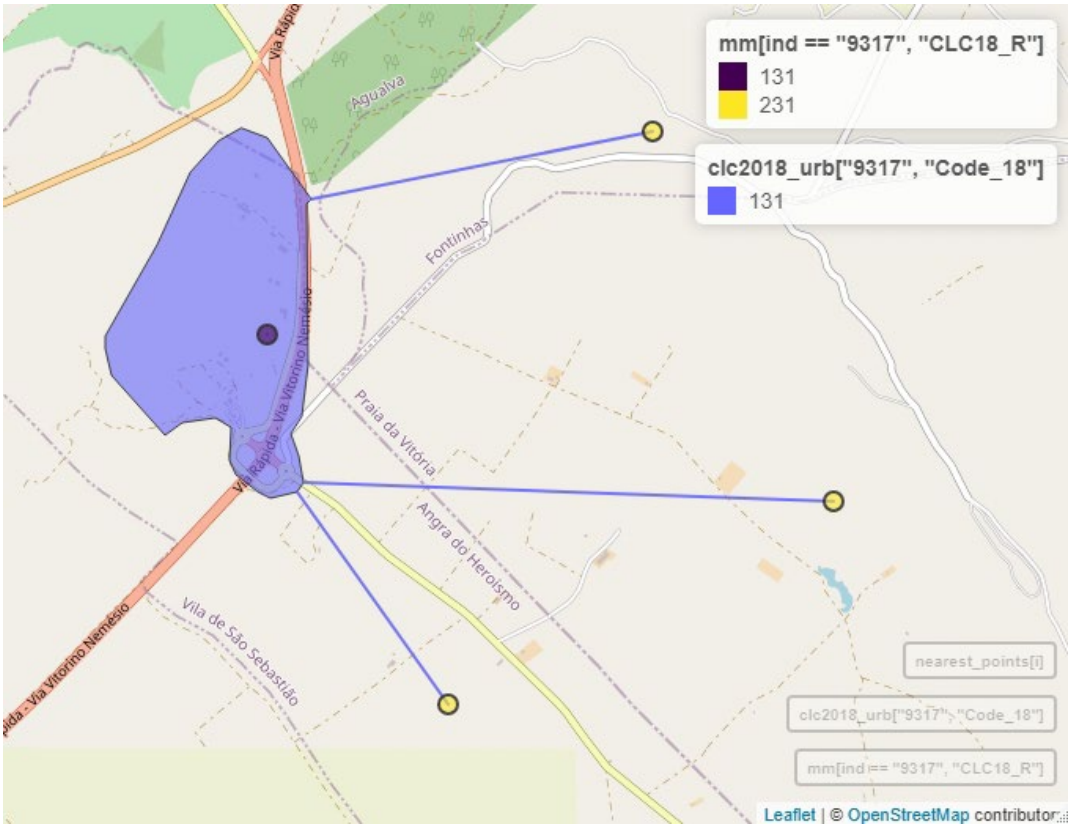
Besides the variables belonging to the HRLs program, some other structural variables of the European cartographic products such as 'SLOPE' and 'ELEVATION' ([2]) were included in the Master.

An additional variable that was elaborated concerns the distance of each point from the artificial CLC 2018. The following procedure was used:

1. The geographic file named 'CORINE Land Cover - ESRI FGDB' was downloaded from the link https://land.copernicus.eu/pan-european/corine-land-cover/clc2018?tab=download. It contains the results of the CLC 2018 in vector format (polygons);
2. From this file the artificial polygons were selected (i.e., those polygons with the first digit of the classification CLC = 1);
3. For each point of the Master, the closest point of the polygons was determined. The distance was measured in meters.

An example is shown in the following figure. An urban CLC polygon on a map downloaded from the internet (Open Street Map) was overlaid with four points of the Master. For one of these points (which lies within the polygon) the distance is zero. For the other three points, the distance is equal to the length of the straight line.

**Figure 2.2: Example of the distance between the Master points and the artificial CLC 2018 polygon**




([2]) https://land.copernicus.eu/imagery-in-situ/eu-dem/eu-dem-v1-0-and-derived-products

The following tables show some summary statistics of distance.

**Table 2.1: Summary statistics of the distances in meters between the Master points and the nearest CLC artificial polygon**

| Min. | 1stQu. | Median | Mean | 3rdQu. | Max. |
|---|---|---|---|---|---|
| 0.0 | 782.4 | 1 979.7 | 4 036.1 | 4 457.7 | 173 322.5 |

**Table 2.2: Number of Master points by distance class (km) from the nearest CLC artificial polygon**

| [0,0.5] | (>0.5,1] | (>1,2] | (>2,3] | (>3,4] | (>4,5] | (>5,10] | (>10,20] | (>20] |
|---|---|---|---|---|---|---|---|---|
| 182 353 | 126 116 | 207 131 | 135 450 | 88 609 | 61 238 | 135 575 | 59 732 | 32 826 |

*Predicting Reachability*

The complete set of auxiliary variables used to assess the reachability is as follows: elevation, slope, CLC classification, High Resolution Layers (forest, tree density cover, water and wetness, imperviousness, grassland), distance from CLC urban polygon, distance from road, elevation in meters of the nearest point on a road, the slope between the point and the nearest point on a road.

These characteristics were linked (by the Random Forest Model with 100 trees) to a dichotomous flag attached to each sampled point observed in the previous edition of the survey (2018). The flag was equal to one if the point had been reached by the enumerators and zero otherwise.

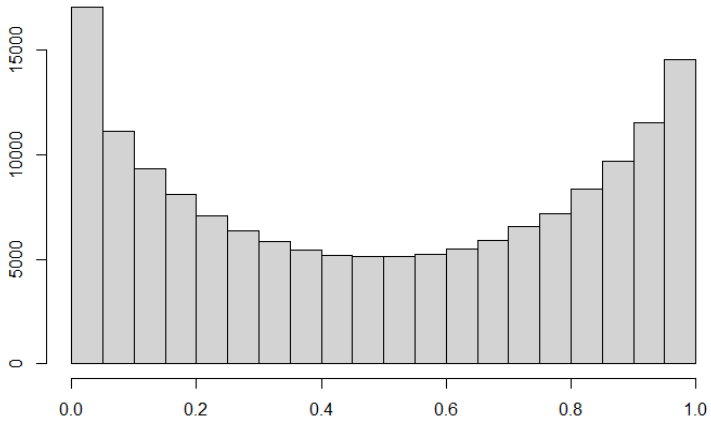
The Random Forest Model showed that reachability was mainly dependent on the following variables: NUTS0_16, NEARDIST, NEARELEV, ELEV (elevation of the master point), CLC (CORINE Land Cover class), and ABS_RATIO.

Since most of the previous variables were also used to calculate the reachability index for the previous edition of the survey, the new index can be viewed as a refinement of the previous one.

The random forest model was trained using a random sample of units (training set) with the observed reachability; the sample units included in the training set are 240 447. The complementary part (testing set) was used to measure the ability of the model to predict the reachability; it contained 80 149 units.

The following figure depicts the histogram of reachability prediction values for the test set.

**Figure 2.3: Histogram of reachability**




The reachability score was dichotomised using the median as the threshold: if the reachability score was greater than the median, the point has been classified as reachable (reachability=1); in the

opposite case (reachability score lower than the median), the point has been classified as 'not reachable' (reachability=0).

The accuracy of such classification is summarized in the following confusion matrix (calculated for the test set). As it could be seen, the achieved accuracy could be considered a good result for the following selection purposes. Only in 4 759 (12 %) out of 39 377 (4 759+34 618) cases where the point was predicted to be reachable did the enumerators encounter difficulties in making an in-situ visit during the previous editions of the survey (in the test set, the proportion of points where the enumerators encountered difficulties is 39 %).

**Table 2.3: Confusion matrix calculated with the Random Forest Model for the test set**

| | | Observed reachability | |
|---|---|---|---|
| | | 0 | 1 |
| Predicted reachability | 0 | 26 755 | 14 017 |
| | 1 | 4 759 | 34 618 |

The same methodology was used to predict the degree of reachability for each point of the master.

# 2.4. Sample - Module eligibility

To improve the efficiency of sample selection for each specific module, predicted LC and LU values were assigned to each point in the Master.

LC and LU values for all units in the Master were assigned using the following procedure:

a) For all points surveyed in 2018, allocation of observed 2018 LC and LU values (about 30 % of the total);

b) For all points collected in 2015 but not in 2018, prediction of 2018 values using also 2015 values (approximately 20 %);

c) For the remaining points (approx. 50%), prediction of 2018 values modelled using only Master variables.

For the prediction of points (b) and (c), a machine learning approach was used, setting the points observed in 2018 as training and validation sets for fitting a 'random forest' model. The explanatory variables used were the Copernicus program variables described in the previous sections and the 2018 photointerpretation (STR18). The resulting accuracies range from a minimum of 57 % for *bare land* (F) and a maximum of 93 % for *woodland* (C):

**Table 2.4: Accuracies for Land Cover predictions**

| Land Cover | Sensitivity | Specificity | Balanced accuracy |
|---|---|---|---|
| **Class: A** | 0.8319 | 0.9902 | 0.9111 |
| **Class: B** | 0.8651 | 0.9243 | 0.8947 |
| **Class: C** | 0.9349 | 0.9361 | 0.9355 |
| **Class: D** | 0.4688 | 0.9848 | 0.7268 |
| **Class: E** | 0.7072 | 0.9204 | 0.8138 |
| **Class: F** | 0.1537 | 0.9981 | 0.5759 |
| **Class: G** | 0.6762 | 0.9991 | 0.8377 |
| **Class: H** | 0.6853 | 0.9949 | 0.8401 |

For LU, the same explicative variables were used, plus the observed or predicted values of LC, leading to the following results:

**Table 2.5: Accuracies for Land Use predictions**

| Land Use | Sensitivity | Specificity | Balanced accuracy |
|---|---|---|---|
| **Class: U1** | 0.9565 | 0.6268 | 0.7917 |
| **Class: U2** | 0.0066 | 0.9999 | 0.5032 |
| **Class: U3** | 0.6757 | 0.9785 | 0.8271 |
| **Class: U4** | 0.5111 | 0.9683 | 0.7397 |

For the grassland and landscape features modules, the first digit prediction was not sufficient. It was necessary to predict some land cover and land use modalities with a more disaggregated classification level. In particular, predictions were made for:

- LC equal to B7x (fruit trees): Prediction accuracy was 66 %
- LC equal to C1x (broadleaved woodland): Prediction accuracy was 95 %
- at least 30 % of grass (yes/no): Prediction accuracy was 91 %
- LU equal to U11 (agriculture): predicted with 95 % accuracy

Using the variables predicted or observed in the previous survey, it was possible to translate the eligibility criteria for each module into five eligibility flags assigned to every point in the Master. As these flags were available for all points in the Master, once selected, then LUCAS 2022 sample could be automatically used as the Master sample for selecting the subsamples of the five modules.

# 3 LUCAS sample design and selection

The following steps had been undertaken in the design and selection of the LUCAS sample:

1. Determination of the allowable sample size per country;
2. Optimization of stratification and allocation of points in the strata;
3. Selection of the sample;
4. Attribution of weights.

After the execution of the above activities, the 'observation type' was assigned to each point in the selected sample, i.e., it was decided whether the selected point should be observed in the field, or it has to be photo-interpreted. This is a very important task, not only for the LUCAS sample, but also because 'observation type = field' is one of the conditions always part of the eligibility criteria for the selection of the five subsamples.

## 3.1 Determination of the allowable sample size per country

The determination of the allowable sample size for each country was based on:

- the analysis of the LUCAS 2018 results in terms of precision of the estimates of the main target variables (Land Cover and Land Use),
- the total number of points available (400 000),
- the absence of the UK in the group of Member States,

It was decided to redefine the total sample size in the different EU 27 Member States. Taking into account the fact that due to the requirements of the Soil Module a certain number of points has to be selected in each case, the final situation regarding the sample sizes is presented in Table 3.1.

**Table 3.1:** LUCAS planned sample size

| Country | Planned size | Points fixed for Soil module | Points free for sample selection |
|---|---|---|---|
| **Belgium** | 7 402 | 127 | 7 275 |
| **Bulgaria** | 12 908 | 463 | 12 445 |
| **Czechia** | 9 108 | 435 | 8 673 |
| **Denmark** | 6 985 | 164 | 6 821 |
| **Germany** | 46 179 | 698 | 45 481 |
| **Estonia** | 2 453 | 182 | 2 271 |
| **Ireland** | 8 600 | 101 | 8 499 |
| **Greece** | 18 881 | 498 | 18 383 |
| **Spain** | 31 025 | 3 079 | 27 946 |
| **France** | 55 937 | 2 589 | 53 348 |
| **Croatia** | 5 821 | 73 | 5 748 |
| **Italy** | 34 833 | 1 183 | 33 650 |
| **Cyprus** | 1 347 | 51 | 1 296 |
| **Latvia** | 6 770 | 295 | 6 475 |
| **Lithuania** | 5 427 | 342 | 5 085 |
| **Luxembourg** | 644 | 10 | 634 |
| **Hungary** | 7 713 | 296 | 7 417 |
| **Malta** | 80 | 3 | 77 |
| **Netherlands** | 7 870 | 84 | 7 786 |
| **Austria** | 9 294 | 611 | 8 683 |
| **Poland** | 35 083 | 1 149 | 33 934 |
| **Portugal** | 6 746 | 346 | 6 400 |
| **Romania** | 23 269 | 240 | 23 029 |
| **Slovenia** | 3 965 | 123 | 3 842 |
| **Slovakia** | 5 640 | 188 | 5 452 |
| **Finland** | 11 536 | 994 | 10 542 |
| **Sweden** | 34 485 | 1 626 | 32 859 |
| **Total** | 400 001 | 15 950 | 384 051 |

## 3.2 Optimization of stratification and allocation of points in the strata

As in 2018, the optimization of the sample design was carried out using the R package SamplingStrata (Ballin and Barcaroli, 2013) (Barcaroli, 2014).

## SamplingStrata

This R package SamplingStrata (Barcaroli et al., 2020) provides an approach for determining the best stratification of a sampling frame, i.e., the one that ensures the minimum sample cost under the condition that precision requirements are met in a multivariate and multidomain case. This approach is based on the use of the *genetic algorithm*: each solution (i.e., a particular partition in strata of the sampling frame) is considered as an individual in a population; the fitness of all individuals is evaluated applying the Bethel algorithm to calculate the sample size that satisfies the precision constraints on the target estimates.

A complete documentation can be found at **https://barcaroli.github.io/SamplingStrata/**

Unlike 2018, when the package was used in its entirety (and left to determine both the optimal stratification and the allocation of sampling units in the final strata), for the LUCAS survey it was decided to strictly control how many strata should be considered in the allocation of points in the Master.

This decision was made to avoid the problem related to the high number of final strata obtained in 2018 and the resulting excessive variability in inclusion probabilities observed in the 2018 LUCAS sample.

The optimization task was based on the following steps:

1. Determination of the precision constraints;
2. Choice of stratification and determination of the best allocation;
3. Sample size adjustment and final allocation.

It should be noted that the full application of the genetic algorithm implemented in the optimization step of SamplingStrata was also used for the optimization of the Soil Module (see Section 4.4).

### 3.2.1. Determination of the Precision Constraints

The target estimates were set as:

- The total area classified by the 8 values of the Land Cover variable (one digit):
  - A: artificial land
  - B: cropland
  - C: woodland
  - D: shrubland
  - E: grassland
  - F: bare land
  - G: water areas
  - H: wetlands
- The total area classified by the 4 values of the Land Use variable (one digit):
  - U1: Primary sector (agriculture, forestry, aquaculture and fishing, mining and quarrying, other)

- U2: Secondary sector (energy, industry and manufacturing)
- U3: Tertiary sector (transport, utilities and residential)
- U4: Unused and abandoned areas

Therefore, we have 12 target estimates for each of the 27 Member States, for 324 total precision constraints.

Each constraint corresponds to the maximum expected value for the coefficient of variation of the target estimate, i.e., the ratio between its standard deviation and the mean. Each of them was set to 0.025, which means that we expect a maximum value of the coefficient of variation of 2.5 % in each country and for each target variable.

## 3.2.2. Choice of stratification and determination of the best allocation

As mentioned above, for this new survey round, it was decided to define the stratification of the Master frame a priori and determine the best allocation in the strata on this basis, instead of letting the optimisation algorithm determine the best stratification starting from an initial 'atomic' stratification, as was the case in 2018.

For any given region (NUTS2 level), stratification is determined by cross classifying the following variables:

1. CORINE Land Cover 2 digits (15 values);
2. STR18 (9 values);
3. Flag indicating whether LC value matches the predicted value (flag=1) or not (flag=0).

These variables determine a total number of 22 173 strata for the entire Master.

Each stratum is characterized by a unique combination of the values of the three variables above in a particular region. For instance, consider the Italian region 'Piemonte' (NUTS2 = 'ITC1'). In the Master, 6 345 points belong to this region. We can classify these points by CLC, STR18 and the flag, obtaining 130 different strata.

For example, the one consisting of the combination of flag = 0 (LC assigned values differ from predicted values), CLC=21 (Arable Land) and STR18 = 3 (Grassland) contains 17 points (Nh). For this stratum (as for all others), the means and standard deviations of the 12 target variables are calculated, and these indicators form the basis for deciding on the allocation of the $n_h$ sampling units to the stratum, according to the Bethel methodology.

The choice of these variables was the result of various trials. In particular, the last variable (flag indicating whether the LC variable was observed or predicted in 2018) was chosen as its inclusion greatly reduces the variability of Land Cover variable in the strata.

To determine the best allocation with a given stratification, a new function had been developed for the SamplingStrata package, namely the 'procBethel' function, which contains the following input

- the points in the Master which are 'free',
- the other points that must be included in any sample (the 15 950 points for the Soil module)
- the precision constraints,

and determines the best allocation by applying the Bethel algorithm separately for each region (NUTS2 level) of the 27 Member States.

## Bethel algorithm

The Bethel algorithm (Bethel, 1989) is a generalization of the multivariate case of the Neyman approach for the optimal determination of the total sample size and allocation of sampling units in a stratified design. It allows determining both the total sample size and the allocation of units in strata, in order to minimise costs under the constraints of certain precision levels of the estimates. The input to this algorithm utilises the information on the distributional characteristics of the target variables in the population strata.

### 3.2.3. Adjustment of sample size and final allocation

For each Member State, the resulting sample size is compared with the planned sample size for that state (see table 3.1), and consequently adjusted.

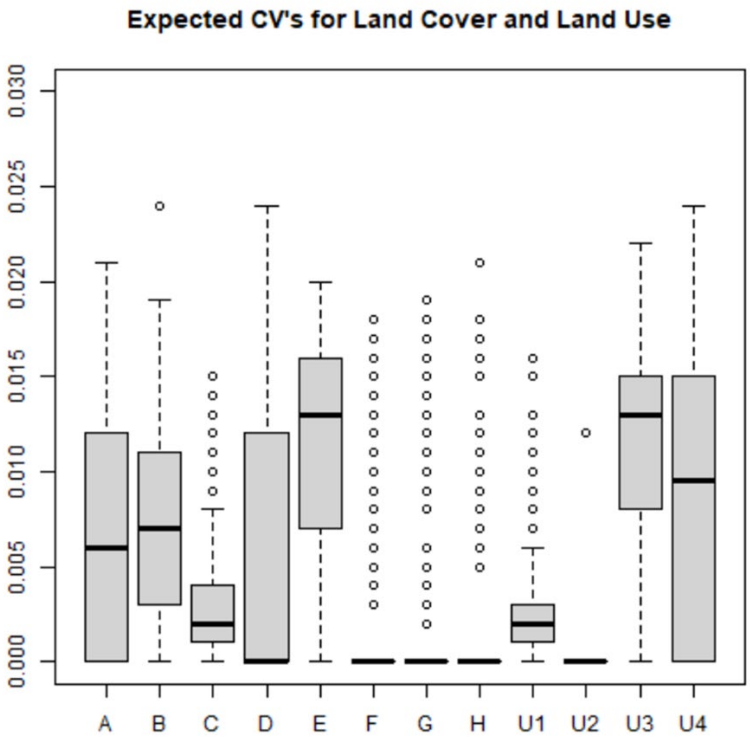
The adjustment was made in each stratum by decreasing or increasing the initial allocation in that stratum, proportionally to the difference between the total allocation and the planned allocation at the state level. The adjustments are shown in Table 3.2.

**Table 3.2:** Optimal size and adjustments

| Countries | Optimal size | Adjusted size | Difference | Fixed points for Soil | Total sample |
|---|---|---|---|---|---|
| **Belgium** | 4 737 | 7 175 | 2 438 | 127 | 7 302 |
| **Bulgaria** | 8 755 | 12 436 | 3 681 | 465 | 12 901 |
| **Czechia** | 5 964 | 8 670 | 2 706 | 435 | 9 105 |
| **Denmark** | 4 889 | 6 818 | 1 929 | 164 | 6 982 |
| **Germany** | 30 409 | 45 455 | 15 046 | 698 | 46 153 |
| **Estonia** | 1 741 | 2 269 | 528 | 182 | 2 451 |
| **Ireland** | 6 145 | 8 496 | 2 351 | 101 | 8 597 |
| **Greece** | 12 646 | 18 362 | 5 716 | 498 | 18 860 |
| **Spain** | 22 272 | 27 929 | 5 657 | 3 079 | 31 008 |
| **France** | 39 096 | 53 331 | 14 235 | 2 588 | 55 919 |
| **Croatia** | 4 409 | 5 743 | 1 334 | 73 | 5 816 |
| **Italy** | 23 719 | 33 623 | 9 904 | 1 184 | 34 807 |
| **Cyprus** | 604 | 1 291 | 687 | 51 | 1 342 |
| **Latvia** | 6 217 | 6 474 | 257 | 295 | 6 769 |
| **Lithuania** | 3 586 | 5 082 | 1 496 | 342 | 5 424 |
| **Luxembourg** | 358 | 633 | 275 | 10 | 643 |
| **Hungary** | 5 038 | 7 404 | 2 366 | 296 | 7 700 |
| **Malta** | 69 | 72 | 3 | 3 | 75 |
| **Netherlands** | 4 920 | 7 750 | 2 830 | 84 | 7 834 |
| **Austria** | 6 070 | 8 672 | 2 602 | 613 | 9 285 |
| **Poland** | 23 068 | 33 919 | 10 851 | 1 149 | 35 068 |
| **Portugal** | 4 613 | 6 389 | 1 776 | 346 | 6 735 |
| **Romania** | 17 153 | 23 026 | 5 873 | 240 | 23 266 |
| **Slovenia** | 2 661 | 3 841 | 1 180 | 123 | 3 964 |
| **Slovakia** | 3 724 | 5 450 | 1 726 | 188 | 5 638 |
| **Finland** | 8 891 | 10 534 | 1 643 | 994 | 11 528 |
| **Sweden** | 26 138 | 32 854 | 6 716 | 1 626 | 34 480 |
| **Total** | 277 892 | 383 698 | 105 806 | 15 954 | 399 652 |

As observed, the optimal size required to be compliant with the precision constraints of a 2.5 % CV for all target estimates, is always lower than the affordable sample size. This factor ensures that the expected CVs for all countries are always below the 2.5 % threshold (see Figure 3.1).

**Figure 3.1:** Expected CVs after the adjustment



# 3.3 Selection of the sample and attribution of weights

## 3.3.1 Selection of the Sample

The sample was selected using the function 'selectSampleSpatial' of the R package SamplingStrata, which is a wrapper of the function 'lpm2_kdtree' in the package SamplingBigData, in order to ensure a spatially balanced sample. This function implements the local pivotal method.

### Local Pivotal Method

The local pivotal method (Grafström et al., 2012) (Lisic and Grafström, 2020) provides a way to perform balanced sampling. This implementation replaces the linear search in lpm2, with k-d trees. K-d trees are binary trees used for effective search in high dimensional spaces, and they reduce the average computational complexity of lpm2 from O(N^2) to O(N log(N)).

The local pivotal method has been implemented in the R package SamplingBigData (Lisic and Grafström, 2018).

## 3.3.2 Attribution of the weights

The sample selection function automatically assigns initial weights ('WEIGHTS') that are equal to the inverse of the inclusion probabilities, i.e., the ratio between the number of points selected and the number of points in each stratum.

The 15 950 points specified as 'fixed' points for the Soil module, were assigned an inclusion probability equal to 1, since these points are required by the Soil module to be included anyway. Their weight is therefore equal to 1. However, since these points will most likely be included in the panel component again in the next rounds of LUCAS, it was decided to assign them a weight closely related to the sampling design. For this reason, the weights were re-computed as the inverse of the rate of sample points in the total number of points in the sampling strata.

Here is an example of how this re-computation of weights was done.

Let us consider a region with NUTS2 = 'AT8': in the LUCAS selected sample there are 1,369 units belonging to this region, distributed in 89 different strata, each characterized by a different combination of values of the variables 'flag', 'CLC' and 'STR18'.

Let us take the stratum where flag = 1, CLC = 31 and STR = 4. It contains 413 points, of which

- 23 have been taken from the pool of SOIL module, so these points have weights = 1;
- 390 have been selected with simple random sampling within the stratum, with a weight = 2.223077.

Note that the sum of weights in this stratum is 890, equal to the number of points in the Master belonging to this stratum.

The new weights (LUCAS_WGT) were now calculated in a simple way as the ratio between the number of units in the Master (890) and the number of units in the sample (413):

w'=N_h/n_h

which is a value of 2.15496 for all units in the stratum.

## 3.4 Attribution of the type of observation to the selected units

Observation of a LUCAS sample point can be done in two ways: by direct observation carried out by an enumerator, in-situ visit (or field observation, FI), or by photointerpretation (PI) carried out by a photo interpreter on the most recent orthophoto available.

In general, the first method can be considered more accurate, but also more expensive and, sometimes, impossible to implement. The second one is less accurate, not all information can be collected, and sometimes the available orthophoto is not very recent, but it is less expensive.

For the main LUCAS, in-situ visits should be preferred, but given the limited resources, some of the sample points have to be used for photointerpretation. For the five planned modules, FI observation is mandatory.

Half of the approximately 400 000 points selected in the LUCAS sample will be observed in the field, the remaining half will be photo interpreted in the office.

The way in which the observation type (field / photo interpreted) is attributed consists of utmost importance for the following reasons:

- Efficiency of the LUCAS sample:
  a. the number of field points for which it will not be possible to observe the variables of interest at a reasonable distance should be minimized;
  b. since photointerpretation will be based on time-lagged images, the probability of their change should, on average, be lower than that of the field points;
- Efficiency of the subsamples: the reachability of selected points is the fundamental condition for the five different modules (in particular for the Soil module, where soil samples have to be collected, but also for all other modules).

A reachability score (a value included in the range of [0,1]) has already been calculated for all points in the Master (see par.2.3), by using a Random Forest model.

Independently from reachability, or other considerations, 17 156 points defined as fixed points (due to the Soil module requirements) were assigned a ‘field’ observation mode.

To reduce the difficulty for enumerators or to limit the number of points to be observed in the field even if their status is very unlikely to change, some of the reachability scores were reduced to 0 or 1 on the basis of some characteristics of the points.

The score was reduced to 0 for:

- Points that are far away from the nearest road or fall into CLC classes with a very low probability of change;
- A random proportion of points that fall into a particular class of a CLC urban polygon.

A specific deterministic rule was applied to ensure that a point more distant than 750 meters from the nearest road was not assigned the value 'field'. This rule assigns the value '0' to obs_type_score:

```
obs_type_score <- ifelse(NRDIST17 > 750, 0, obs_type_score)
```

Then, the same 0 assignment for the points for which the probability of a change is very low was done, i.e.

```
obs_type_score <- ifelse (CLC18_R == 124 # airports
                          | CLC18_R == 335 # glaciers and perpetual snow
                          | CLC18_R == 512 # water bodies
                          | CLC18_R == 521 # coastal lagoons
                          | CLC18_R == 522 # estuaries
                          | CLC18_R == 523 # sea and oceans
                           ), 0, obs_type_score)
```

For these other cases, we assign zero only in 90% of the cases:

```
obs_type_score <- ifelse (CLC18_R == 111 # continuous urban fabric
                          | CLC18_R == 112 # discontinuous urban fabric
                          | CLC18_R == 121 # industrial or commercial unit
                          | CLC18_R == 122 # road and rail networks
                          | CLC18_R == 511 # water courses
                          & samptot$rnd < 0.9)   # 90% of cases
                           ), 0, obs_type_score)
```

These rules have the effect of imposing the value 'photo interpreted' on the observation type of the point. The opposite, i.e. forcing the value 'field', is achieved by applying the following rule:

```
obs_type_score <- ifelse(obs_type_score > 0.5
                                & NRDIST17 < 750
                                & DIST_URBAN == 2  # peri-urban
                                & rnd < 0.5,       # 50% of cases
                  1, samptot$obs_type_score)
```

Therefore, if the current score is already greater than 0.5, and the nearest road is no more than 750 meters away, and we are in a peri-urban area (points located in a range greater than zero and up to 500 meters from CORINE artificial polygon), then in 50 % of the time, we assign the value 1 to obs_type_score, to be sure that the point is assigned the value 'field' as the observation type.

After applying the above rules, the LUCAS selected points that have a value of obs_type_score greater than the median value will be assigned a 'field', otherwise 'PI'.

The following table depicts the "Field or PI" assignment in relation to the distance from urban areas, based on the CORINE artificial polygon.

**Table 3.3:** Distribution of LUCAS selected points according to observation type

| Country | FI | PI | Country | FI | PI |
|---|---|---|---|---|---|
| **Belgium** | 4 879 | 2 423 | **Latvia** | 3 385 | 3 384 |
| **Bulgaria** | 5 047 | 7 854 | **Luxembourg** | 519 | 124 |
| **Czechia** | 4 553 | 4 552 | **Hungary** | 3 683 | 4 017 |
| **Denmark** | 5 148 | 1 834 | **Malta** | 51 | 24 |
| **Germany** | 23 077 | 23 076 | **The Netherlands** | 4 723 | 3 111 |
| **Estonia** | 1 229 | 1 222 | **Austria** | 5 482 | 3 803 |
| **Ireland** | 4 299 | 4 298 | **Poland** | 17 534 | 17 534 |
| **Greece** | 9 430 | 9 430 | **Portugal** | 4 174 | 2 561 |
| **Spain** | 17 727 | 13 281 | **Romania** | 7 867 | 15 399 |
| **France** | 27 959 | 27 960 | **Slovenia** | 2 784 | 1 180 |
| **Croatia** | 2 908 | 2 908 | **Slovakia** | 3 701 | 1 937 |
| **Italy** | 17 403 | 17 404 | **Finland** | 5 405 | 6 123 |
| **Cyprus** | 955 | 387 | **Sweden** | 12 012 | 22 468 |
| **Lithuania** | 3 893 | 1 531 | **European Union** | 199 827 | 199 825 |

## 3.5 Geographical visualization of the selected points

An application has been developed that allows the geographical visualization of LUCAS 2022 selected points. It could be accessed, along with instructions for its installation and use, via the link:

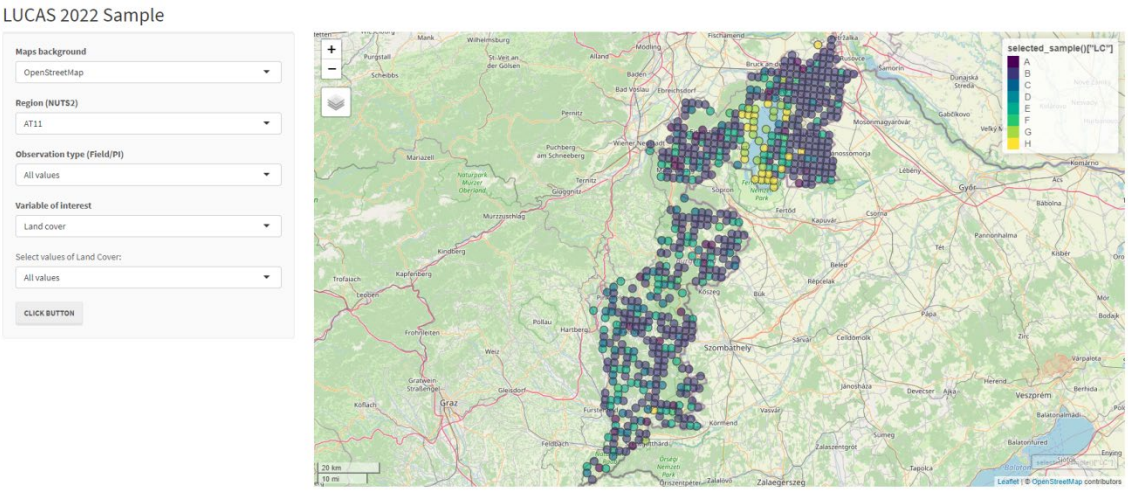
https://github.com/barcaroli/LUCAS2022ShinyApp

The user must specify the following parameters:

1. Map to be used as background,
2. Region to be visualized (NUTS2),
3. Observation type (all/field/photo interpreted),
4. Variable to be visualized (Land Cover/Land Use).

When choosing Land Cover, the user also has the option to select one of the eight values of this variable (see Figure 3.2).

**Figure 3.2: Geographical visualization of selected points with OpenStreetMap**




The background map can be of three different types:

1. OpenStreetMap
2. Esri.WorldImagery
3. OpenTopoMap

While the first one is the default map, the Esri.WorldImagery map is very useful when the objective is to check the coherence of the Land Cover type in relation to the physical characteristics of the points by zooming in on the image (see Figure 3.3).

The use of the OpenTopoMap, on the other hand, is practical when the aim is to check the reachability of selected points, as this background highlights the road network (see Figure 3.4).

**Figure 3.3: Geographical visualization with Esri.WorldImagery**

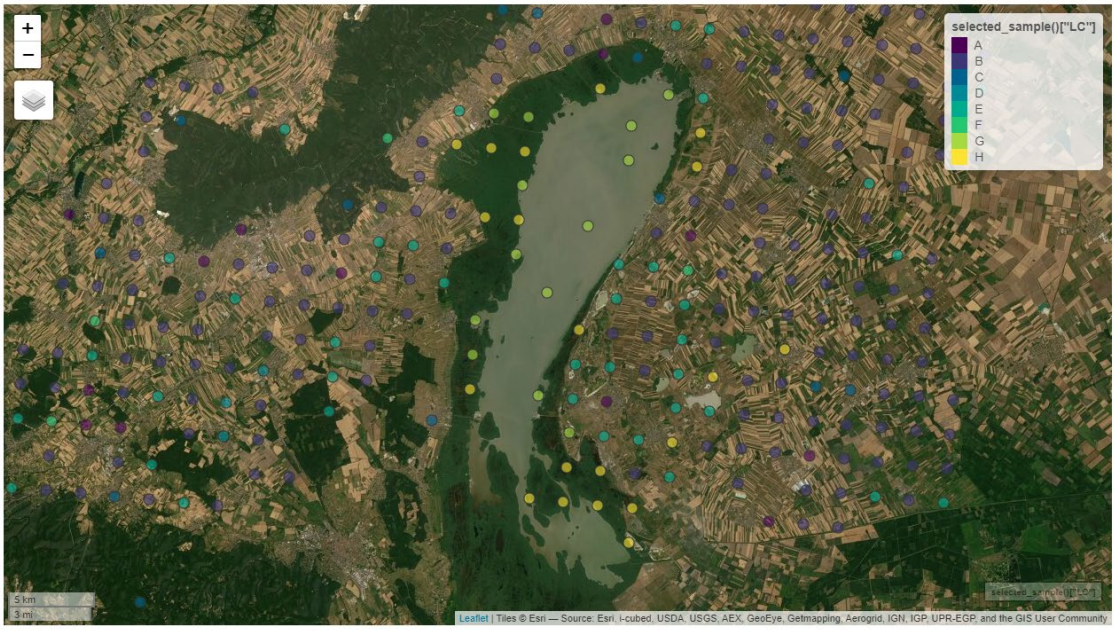
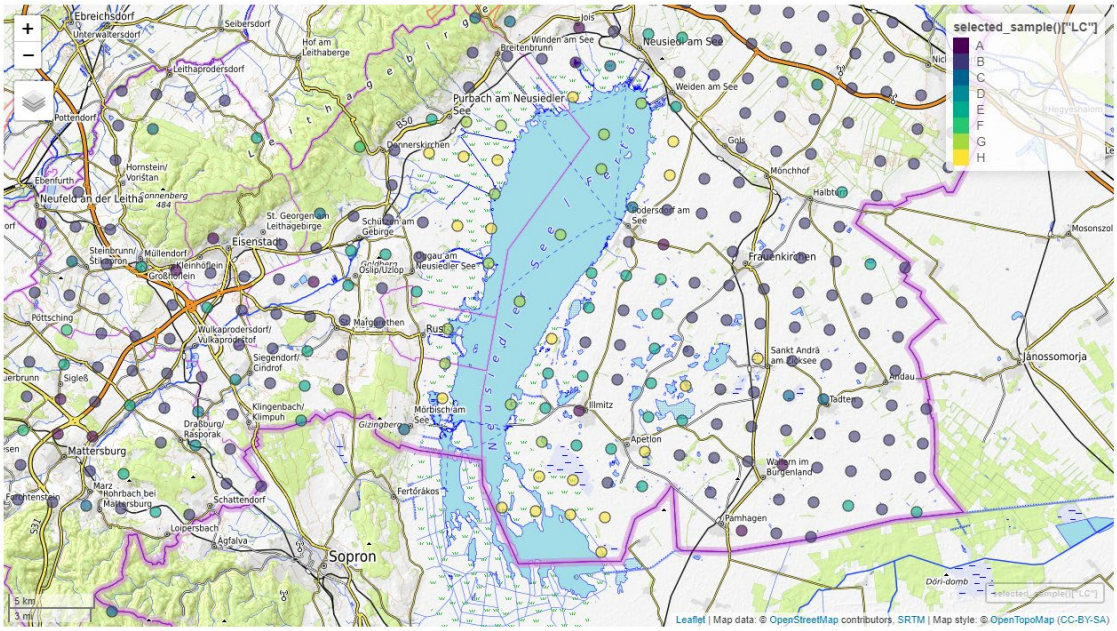


**Figure 3.4: Geographical visualization with OpenTopoMap**



# 4 Design and selection of subsamples

The LUCAS sample plays the role of a Master sample for the design and selection of five different subsamples for the modules of Grassland, Extended Grassland, Landscape Features, Soil and Copernicus.

During the subsamples design and selection, close cooperation was established between Eurostat, other relevant DGs, and Member States to take into account different requirements, always keeping the solutions within a methodological framework.

The complete sample design of each module could be viewed as a two-phase sample design. The LUCAS sample represents the first phase and each module could be considered as the second phase (Cochran, W.G. (1977), chapter 12), (Hansen et al., (1993), chapter 11). The resulting sample designs are compliant with the requirements of the stakeholders responsible for the modules and expressed by the following eligibility criteria applicable to each selected LUCAS sample unit:

| Subsample | Eligibility criteria |
|---|---|
| Grassland | ((Land Cover = D OR Land Cover = E) |
| | AND Grass Percentage > 30%) |
| | AND Observation Type = Field |
| Extended Grassland | Land Cover = E |
| | AND Observation Type = Field |
| Landscape Features | (Land Use = U11 OR |
| | STR18 = 1,2,3 OR |
| | CLC18_R = 2) |
| | AND Observation Type = Field |
| Soil | (Land Cover = B |
| | OR Land Cover = C |
| | OR Land Cover = E |
| | OR Land Cover = H) |
| | AND Observation Type = Field |
| Copernicus | Observation Type = Field |

The final inclusion probability for each selected unit in a given module subsample is the result of the product:

- The inclusion probability of each selected LUCAS point (i.e., the inverse of its first phase weight $w_i^{LUCAS}$
- The inclusion probability associated to each selected point in the subsample.

In general, further stratification of the eligible points by significative (for the module) variables had been implemented. Hence, the inclusion probability for each *s* module can be expressed by the following equation:

$$p_i^s = 1/w_i^{LUCAS} . \frac{m_h^s}{n_h^s}$$

Where $n_h^s$ is the number of LUCAS selected units in stratum *h*, eligible for module *s* according to the eligibility criteria defined above, and $m_h^s$ is the number of module *s* units allocated in stratum *h* (strata *h* are defined specifically for each module).

An exception to the approach followed, which ensures the completely probabilistic nature of each selected subsample, is the Soil module, where a set of 17 156 points observed in previous rounds of the survey or belonging to land cover H remained in the 2022 sample; these units were assigned an inclusion probability of 1 in the second phase.

It could be assumed that the data collected for a given module will be primarily used for modelling and analysis. However, if estimates are to be produced for the entire population of interest specific to the module, it is suggested that the calibration should be made using known totals calculated from the Master summarizing the area (or number) of points identified by the eligibility criteria for the given module, possibly distributed according to the same strata used in the design of the module. The points in the Master identified by the eligibility criteria related to a specific module constitutes the reference population for this module. Moreover, the calibration allows to deal with the total non-response, and to take into account that the selection of points was carried out only on LUCAS points with the observation type 'in field'.

The following paragraphs report for each of the modules:

- The initial requirements;
- The approach to the design and selection of the subsample;
- The characteristics of the selected subsample.

# 4.1 Grassland

Eligible points are those with a grass coverage of at least 30% if they fall into land cover classes D (shrubland) and E (grassland), and of course with a 'field' observation type. The final LUCAS sample contains 60351 eligible observations. The desired sample size of the Grassland module is 20 000, divided into 2 sets: 15 000 (main) to be distributed to provide results at Member State level and 5 000 (specific) for assessing specific questions. The aim of a grassland analysis over time is to detect with sufficient precision the absolute variations of more than 10% of the main grassland related estimates. The specific sample was intended to ensure that information covered the needs of Natura2000, *dehesas* and the project "Copernicus for Natura 2000" (COP4N2K). The selection was done in two phases, i.e.:

1. Firstly, the main sample was selected and the distribution of selected points was analysed;
2. Based on the previous step, the needs for the specific sample were defined and the selection was carried out.

In contrast to other modules, the selection of points in grassland was not carried out using the balanced spatial sampling, as it was not possible to ensure the given allocations using this method. The 'systematic random selection' method was therefore used, because the selection had to be compliant

with the given allocations. Besides, the spatial selection already applied in overall LUCAS sample allowed indirectly their spatial distribution.

The main sample was analysed for its distribution across countries, grassland regions and altitude classes, as well as for its belonging to specific groups of interest (Natura2000, *dehesas* and COP4N2K). The analysis of the main grassland sample revealed the following:

- The coverage of countries and grassland regions was ensured;
- To allow an assessment of the main sample by country/grass regions for IT, DE, FR, ES and altitude classes, additional 2 000–3 050 points were needed (to ensure 200/220/250 points per reference area);
- For Natura2000, additional 2 225–2 938 points were needed (to provide 200/220/250 points per country/grass region). 2052 points for Natura2000 are already included in the main grassland sample;
- To cover the *dehesas,* about 300–400 points were needed.

It was decided that:

- 2 000 points should be allocated to allow an analysis of altitude classes, per country and, where appropriate, grassland region (DE, IT, FR, ES);
- 2 500 points should be allocated for an assessment of Natura2000, aiming at 220 points in Natura2000 per country and, where appropriate, grassland region (DE, IT, FR, ES);
- 300–400 points should be allocated to *dehesas* (200 in Spain, 200 in Portugal);
- The remaining points should be allocated to COP4N2K (about 200–400 points).

Therefore, the above elements comprised the additional specific grassland sample.

**Table 4.1:** Allocation of Grassland module subsample points

| Country | Allocation for main sample | Selection with specific sample | Addition of COP4N2K points | Addition of dehesas points | Addition of soil points | Total allocated |
|---|---|---|---|---|---|---|
| **Belgium** | 330 | 509 | 0 | 0 | 0 | 509 |
| **Bulgaria** | 700 | 813 | 65 | 0 | 7 | 885 |
| **Czechia** | 320 | 535 | 0 | 0 | 2 | 537 |
| **Denmark** | 200 | 230 | 0 | 0 | 0 | 230 |
| **Germany (DE03)** | 260 | 285 | 0 | 0 | 4 | 289 |
| **Germany (DE04)** | 530 | 699 | 37 | 0 | 13 | 749 |
| **Estonia** | 260 | 265 | 0 | 0 | 0 | 265 |
| **Ireland** | 550 | 669 | 0 | 0 | 2 | 671 |
| **Greece** | 930 | 1 059 | 16 | 0 | 2 | 1 077 |
| **Spain (ES03)** | 260 | 704 | 0 | 12 | 5 | 721 |
| **Spain (ES05)** | 1 080 | 1 208 | 129 | 138 | 18 | 1 493 |
| **France (FR03)** | 400 | 618 | 10 | 0 | 10 | 638 |
| **France (FR04)** | 680 | 1 028 | 48 | 0 | 11 | 1 087 |
| **France (FR05)** | 400 | 508 | 0 | 0 | 0 | 508 |
| **Croatia** | 460 | 611 | 5 | 0 | 0 | 616 |
| **Italy (IT05)** | 970 | 1 076 | 52 | 0 | 4 | 1 132 |
| **Italy (IT06)** | 660 | 985 | 0 | 0 | 5 | 990 |
| **Cyprus** | 210 | 163 | 0 | 0 | 0 | 163 |
| **Latvia** | 260 | 278 | 0 | 0 | 2 | 280 |
| **Lithuania** | 260 | 323 | 0 | 0 | 2 | 325 |
| **Luxembourg** | 240 | 144 | 0 | 0 | 0 | 144 |
| **Hungary** | 530 | 670 | 34 | 0 | 1 | 705 |
| **Malta** | 130 | 15 | 0 | 0 | 0 | 15 |
| **Netherlands** | 260 | 277 | 0 | 0 | 3 | 280 |
| **Austria** | 530 | 837 | 0 | 0 | 3 | 840 |
| **Poland** | 460 | 697 | 107 | 0 | 20 | 824 |
| **Portugal** | 660 | 768 | 13 | 92 | 3 | 876 |
| **Romania** | 950 | 1 077 | 28 | 0 | 13 | 1 118 |
| **Slovenia** | 400 | 490 | 0 | 0 | 0 | 490 |
| **Slovakia** | 400 | 609 | 0 | 0 | 1 | 610 |
| **Finland** | 260 | 305 | 0 | 0 | 3 | 308 |
| **Sweden** | 460 | 617 | 2 | 0 | 6 | 625 |
| **Total EU** | 15 000 | 19 072 | 546 | 242 | 140 | 20 000 |

Table 4.1 could be interpreted as follows. As mentioned above, requirements for the specific sample were established based on the analysis of the main sample selected with the allocation defined in column (1). These requirements were mainly to ensure the reliability of the analyses in relation to the altitude classes and Natura2000, plus *dehesas* regions. In order to meet these requirements, 4 072 points were allocated and selected, resulting in a total number of 19 072 points (column (2)). The

remaining points were used to increase the number of COP4N2K points, *dehesas* and soil points (the latter to be observed together with the Soil module).

Tables 4.2, 4.3 and 4.4 illustrate respectively the distribution of selected points in Grassland regions, by elevation class and by interest group.

**Table 4.2: Selection of Grassland module subsample points across Grassland regions**

| | Grassland regions | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **Country** | **Total** | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** | **9** | **10** |
| **Belgium** | 509 | 0 | 0 | 211 | 298 | 0 | 0 | 0 | 0 | 0 | 0 |
| **Bulgaria** | 889 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 843 | 46 | 0 |
| **Czechia** | 537 | 0 | 0 | 0 | 537 | 0 | 0 | 0 | 0 | 0 | 0 |
| **Denmark** | 230 | 0 | 230 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **Germany** | 1 039 | 0 | 0 | 289 | 750 | 0 | 0 | 0 | 0 | 0 | 0 |
| **Estonia** | 265 | 0 | 265 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **Ireland** | 671 | 671 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **Greece** | 1 078 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 078 |
| **Spain** | 2 201 | 0 | 0 | 721 | 0 | 1 480 | 0 | 0 | 0 | 0 | 0 |
| **France** | 2 237 | 0 | 0 | 637 | 1 092 | 508 | 0 | 0 | 0 | 0 | 0 |
| **Croatia** | 617 | 0 | 0 | 0 | 0 | 118 | 499 | 0 | 0 | 0 | 0 |
| **Italy** | 2 135 | 0 | 0 | 0 | 0 | 1 145 | 990 | 0 | 0 | 0 | 0 |
| **Cyprus** | 163 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 163 |
| **Latvia** | 280 | 0 | 280 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **Lithuania** | 325 | 0 | 325 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **Luxembourg** | 144 | 0 | 0 | 0 | 144 | 0 | 0 | 0 | 0 | 0 | 0 |
| **Hungary** | 702 | 0 | 0 | 0 | 0 | 0 | 0 | 702 | 0 | 0 | 0 |
| **Malta** | 15 | 0 | 0 | 0 | 0 | 15 | 0 | 0 | 0 | 0 | 0 |
| **Netherlands** | 280 | 0 | 0 | 280 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **Austria** | 840 | 0 | 0 | 0 | 558 | 0 | 282 | 0 | 0 | 0 | 0 |
| **Poland** | 819 | 0 | 0 | 0 | 819 | 0 | 0 | 0 | 0 | 0 | 0 |
| **Portugal** | 876 | 0 | 0 | 0 | 0 | 876 | 0 | 0 | 0 | 0 | 0 |
| **Romania** | 1 116 | 0 | 0 | 0 | 0 | 0 | 0 | 63 | 937 | 116 | 0 |
| **Slovenia** | 490 | 0 | 0 | 0 | 0 | 0 | 490 | 0 | 0 | 0 | 0 |
| **Slovakia** | 610 | 0 | 0 | 0 | 397 | 0 | 0 | 213 | 0 | 0 | 0 |
| **Finland** | 308 | 0 | 308 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **Sweden** | 624 | 0 | 624 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **Total EU** | 20 000 | 671 | 2 032 | 2 138 | 4 595 | 4 142 | 2 261 | 978 | 1 780 | 162 | 1 241 |

**Table 4.3:** Selection of Grassland module subsample points by elevation class

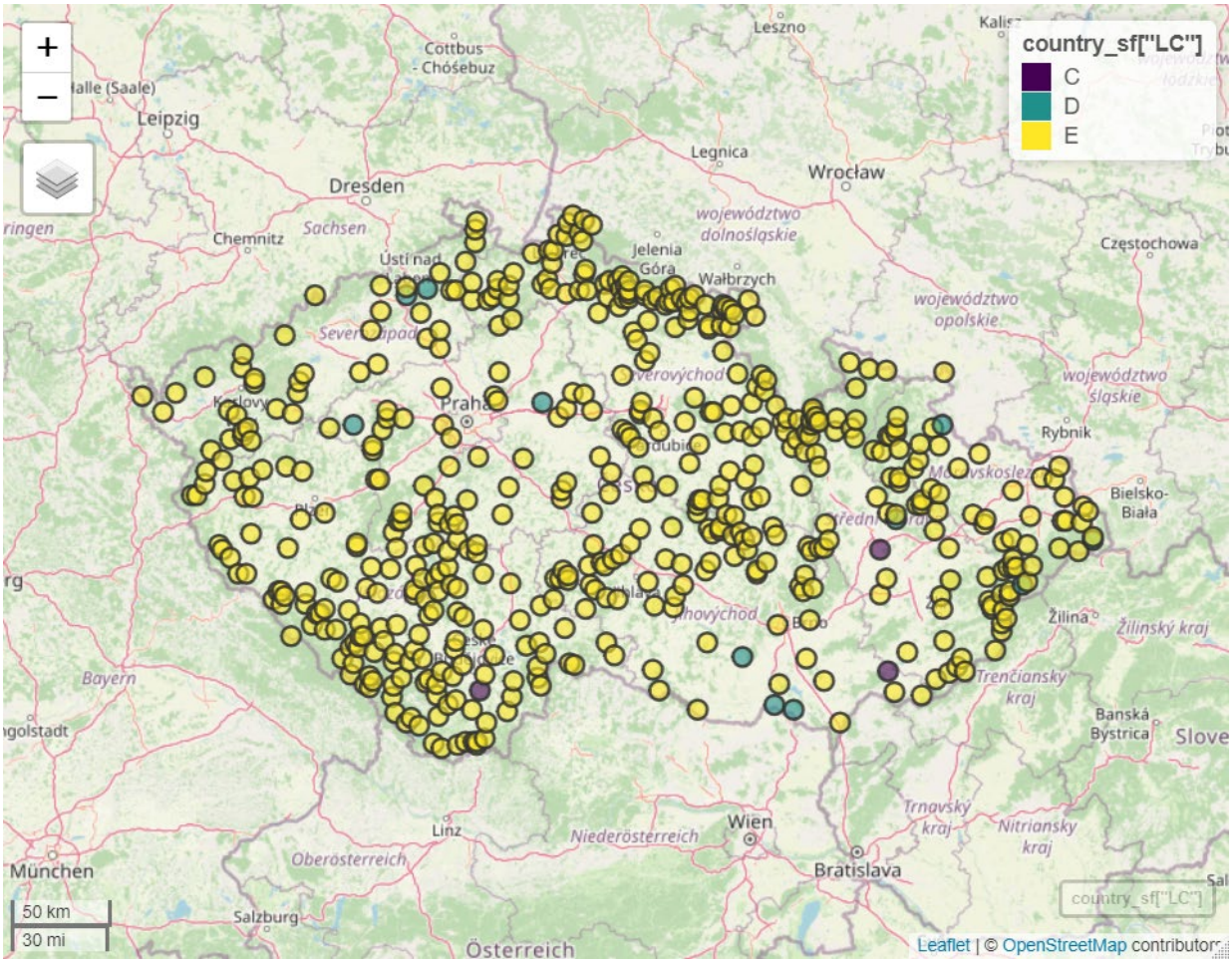
| Countries | Elevation classes | | | | | |
|---|---|---|---|---|---|---|
| | Total | <200 | 200-499 | 500-999 | 1000-1499 | 1500+ |
| Belgium | 509 | 242 | 214 | 53 | 0 | 0 |
| Bulgaria | 889 | 262 | 356 | 219 | 50 | 2 |
| Czechia | 537 | 22 | 250 | 263 | 2 | 0 |
| Denmark | 230 | 230 | 0 | 0 | 0 | 0 |
| Germany (DE03) | 289 | 285 | 4 | 0 | 0 | 0 |
| Germany (DE04) | 750 | 216 | 306 | 224 | 4 | 0 |
| Estonia | 265 | 263 | 2 | 0 | 0 | 0 |
| Ireland | 671 | 533 | 138 | 0 | 0 | 0 |
| Greece | 1 078 | 432 | 245 | 264 | 104 | 33 |
| Spain (ES03) | 721 | 230 | 204 | 230 | 54 | 3 |
| Spain (ES05) | 1 480 | 215 | 392 | 594 | 253 | 26 |
| France (FR03) | 637 | 386 | 206 | 32 | 10 | 3 |
| France (FR04) | 1 092 | 207 | 418 | 259 | 193 | 15 |
| France (FR05) | 508 | 206 | 139 | 129 | 25 | 9 |
| Croatia | 617 | 290 | 210 | 113 | 4 | 0 |
| Italy (IT05) | 1 145 | 435 | 359 | 250 | 97 | 4 |
| Italy (IT06) | 990 | 253 | 195 | 235 | 202 | 105 |
| Cyprus | 163 | 66 | 69 | 28 | 0 | 0 |
| Latvia | 280 | 269 | 11 | 0 | 0 | 0 |
| Lithuania | 325 | 277 | 48 | 0 | 0 | 0 |
| Luxembourg | 144 | 6 | 135 | 3 | 0 | 0 |
| Hungary | 702 | 532 | 170 | 0 | 0 | 0 |
| Malta | 15 | 14 | 1 | 0 | 0 | 0 |
| Netherlands | 280 | 280 | 0 | 0 | 0 | 0 |
| Austria | 840 | 57 | 233 | 316 | 196 | 38 |
| Poland | 819 | 450 | 237 | 131 | 1 | 0 |
| Portugal | 876 | 372 | 277 | 212 | 15 | 0 |
| Romania | 1 116 | 431 | 457 | 204 | 24 | 0 |
| Slovenia | 490 | 51 | 245 | 187 | 7 | 0 |
| Slovakia | 610 | 177 | 205 | 227 | 1 | 0 |
| Finland | 308 | 253 | 55 | 0 | 0 | 0 |
| Sweden | 624 | 421 | 196 | 7 | 0 | 0 |
| Total EU | 20 000 | 8 363 | 5 977 | 4 180 | 1 242 | 238 |

**Table 4.4:** Selection of Grassland module subsample points by groups of interest

| Countries | Natura2000 | COP4N2K | dehesas |
|---|---|---|---|
| **Belgium** | 147 | 45 | 0 |
| **Bulgaria** | 292 | 231 | 0 |
| **Czechia** | 163 | 82 | 0 |
| **Denmark** | 121 | 19 | 0 |
| **Germany (DE03)** | 87 | 10 | 0 |
| **Germany (DE04)** | 198 | 75 | 0 |
| **Estonia** | 19 | 12 | 0 |
| **Ireland** | 114 | 6 | 0 |
| **Greece** | 243 | 59 | 0 |
| **Spain (ES03)** | 137 | 69 | 24 |
| **Spain (ES05)** | 368 | 328 | 204 |
| **France (FR03)** | 194 | 44 | 0 |
| **France (FR04)** | 259 | 147 | 0 |
| **France (FR05)** | 82 | 6 | 0 |
| **Croatia** | 212 | 46 | 0 |
| **Italy (IT05)** | 286 | 222 | 0 |
| **Italy (IT06)** | 188 | 108 | 0 |
| **Cyprus** | 28 | 0 | 0 |
| **Latvia** | 49 | 7 | 0 |
| **Lithuania** | 88 | 4 | 0 |
| **Luxembourg** | 19 | 5 | 0 |
| **Hungary** | 247 | 118 | 0 |
| **Malta** | 4 | 3 | 0 |
| **Netherlands** | 69 | 5 | 0 |
| **Austria** | 160 | 29 | 0 |
| **Poland** | 307 | 213 | 0 |
| **Portugal** | 216 | 74 | 194 |
| **Romania** | 233 | 72 | 0 |
| **Slovenia** | 155 | 49 | 0 |
| **Slovakia** | 179 | 11 | 0 |
| **Finland** | 21 | 1 | 0 |
| **Sweden** | 48 | 28 | 0 |
| **Total EU** | 4 933 | 2 128 | 422 |

Figure 4.1 illustrates an example of geographic distribution of Grassland module selected points by Land Cover.

**Figure 4.1: Grassland selected points in Czech Republic**



## 4.2 Extended Grassland

For this module, 40 000 points had to be selected from the LUCAS sample where the eligibility criteria were Land Cover class equal to E and 'field' observation type. In LUCAS sample, 57 474 points were found that fulfilled these criteria. The distribution of the selected points is reported in Table 4.5.

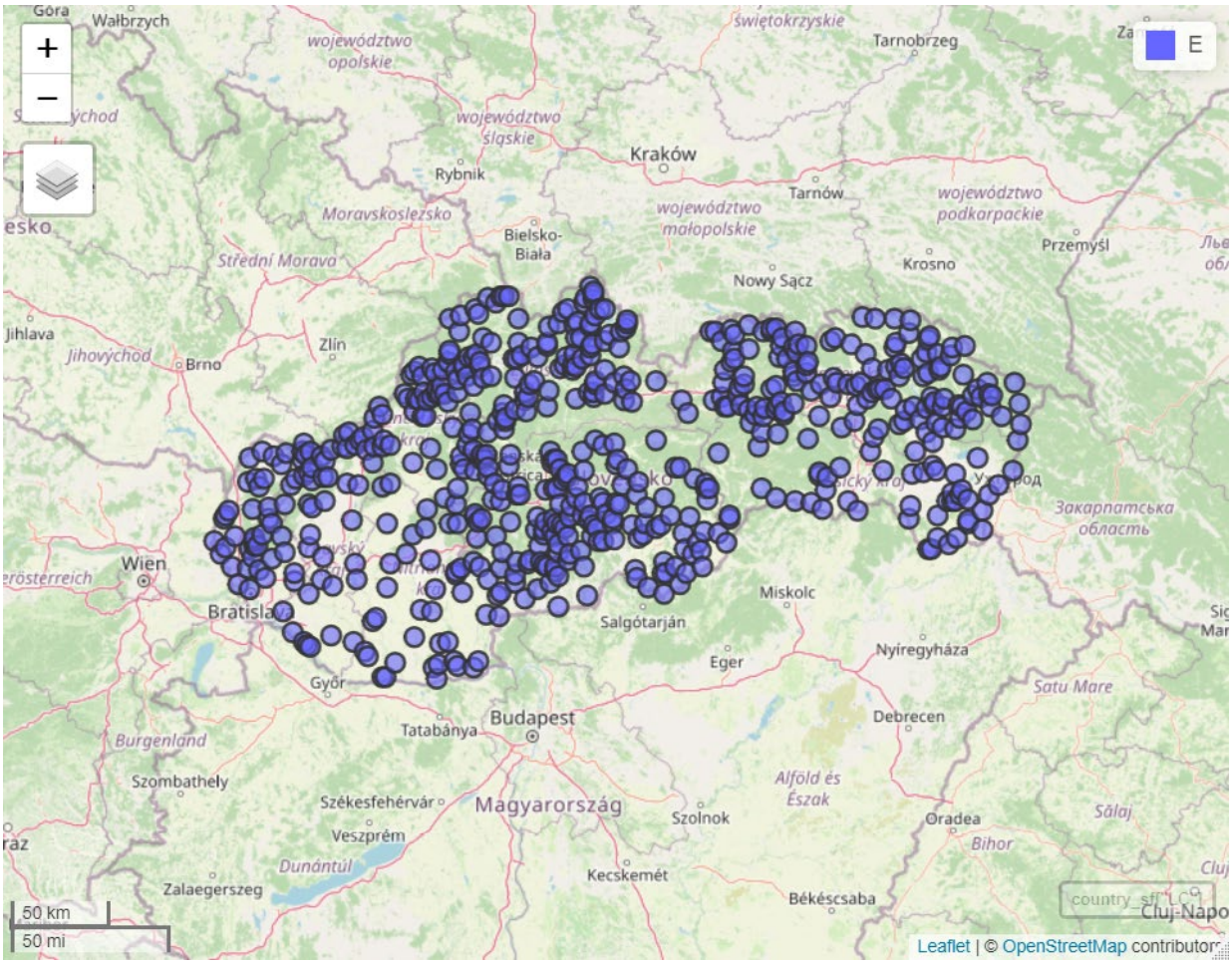
The selection method was of the 'spatial' type, i.e., taking into consideration the coordinates of individual points in order to obtain a spatially balanced sample. In order to ensure a spatially balanced sample, the subsample was selected using the function 'selectSampleSpatial' of the R package SamplingStrata,

**Table 4.5: Distribution of selected points in Extended Grassland module by countries**

| Countries | Eligible points | Selected points | Sampling rate |
|---|---|---|---|
| **Belgium** | 1 369 | 945 | 0.7 |
| **Bulgaria** | 1 469 | 1 023 | 0.7 |
| **Czechia** | 1 446 | 1 007 | 0.7 |
| **Denmark** | 970 | 676 | 0.7 |
| **Germany** | 5 955 | 4 140 | 0.7 |
| **Estonia** | 392 | 275 | 0.7 |
| **Ireland** | 2 631 | 1 831 | 0.7 |
| **Greece** | 2 215 | 1 541 | 0.7 |
| **Spain** | 4 439 | 3 090 | 0.7 |
| **France** | 9 684 | 6 743 | 0.7 |
| **Croatia** | 952 | 663 | 0.7 |
| **Italy** | 4 289 | 2 982 | 0.7 |
| **Cyprus** | 133 | 93 | 0.7 |
| **Latvia** | 645 | 446 | 0.7 |
| **Lithuania** | 1 474 | 1 027 | 0.7 |
| **Luxembourg** | 138 | 97 | 0.7 |
| **Hungary** | 1 155 | 801 | 0.7 |
| **Malta** | 12 | 8 | 0.7 |
| **Netherlands** | 1 935 | 1 345 | 0.7 |
| **Austria** | 1 959 | 1 372 | 0.7 |
| **Poland** | 5 194 | 3 623 | 0.7 |
| **Portugal** | 1 230 | 858 | 0.7 |
| **Romania** | 2 782 | 1 937 | 0.7 |
| **Slovenia** | 530 | 367 | 0.7 |
| **Slovakia** | 884 | 611 | 0.7 |
| **Finland** | 1 092 | 759 | 0.7 |
| **Sweden** | 2 500 | 1 740 | 0.7 |
| **Total EU** | 57 474 | 40 000 | 0.7 |

Figure 4.2 illustrates an example of geographic distribution of Extended Grassland module selected points.

**Figure 4.2: Expected CVs after the adjustment**




# 4.3 Landscape Features

For this module, approximately 93 000 points had to be selected from the LUCAS sample. The proposed eligibility rule was as follows: if a unit satisfied the following conditions (in OR) then the point is eligible for Landscape Features module:

- Field LU = U11 (currently used agricultural land, temporarily unused agricultural land, kitchen gardens)
- STR18 = (1, 2 or 3): arable, permanent crops, grassland
- CLC18 = 2: arable, permanent crops, pastures, heterogeneous agricultural

In the LUCAS sample, 137 768 points were found to meet these criteria. The allocation was determined proportionally to the eligible points in the Master

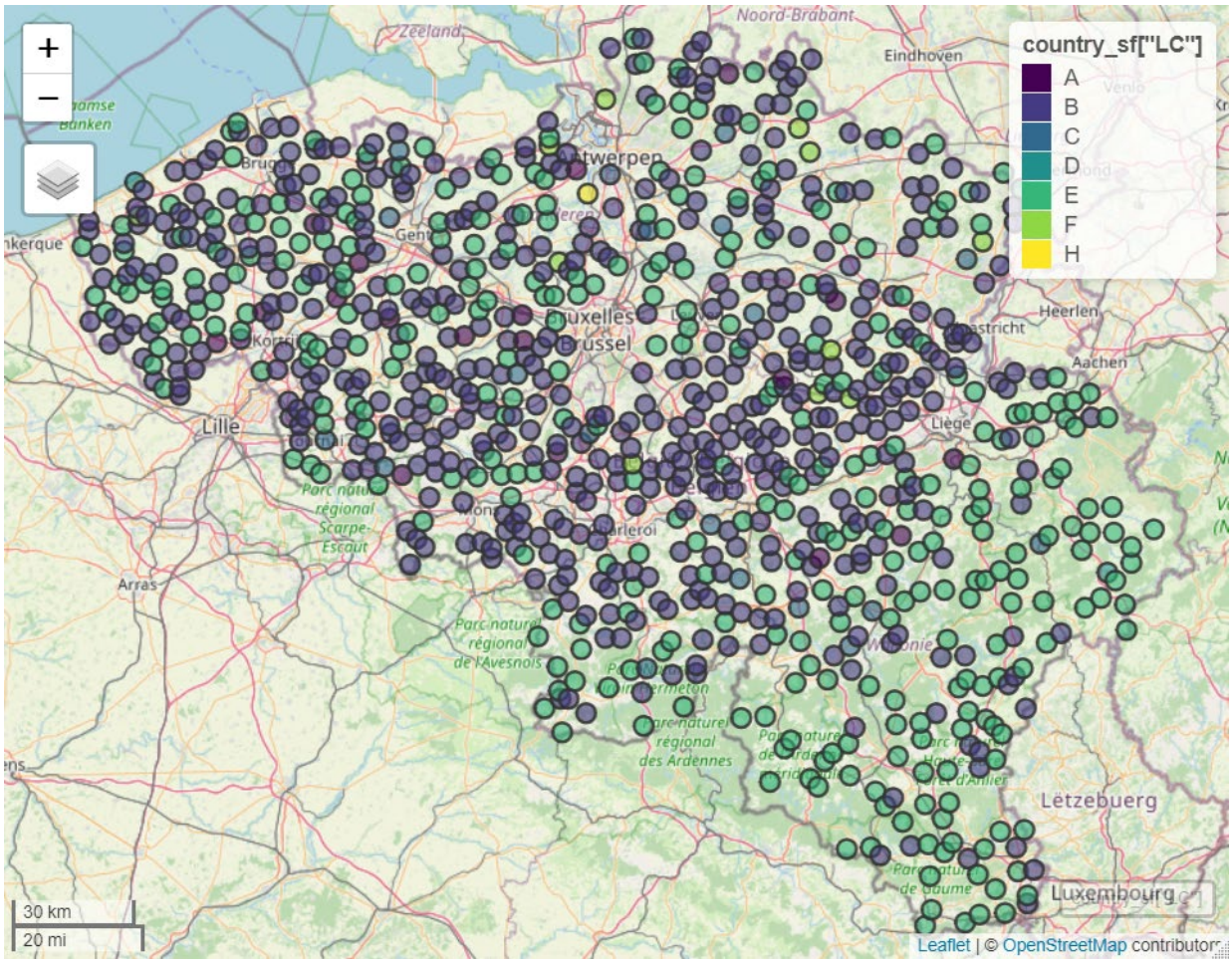
Table 4.6 depicts the distribution of the selected points by country. This subsample was selected using the same function, as mentioned above, in order to ensure a spatially balanced sample.

Figure 4.3 illustrates an example of geographical distribution of the selected points of the module 'Landscape Features' by Land Cover in Belgium.

**Table 4.6: Distribution of Landscape Features selected points**

| Countries | Eligible points | Allocated points | Selected points | Sampling rate |
|---|---|---|---|---|
| **Belgium** | 3 652 | 886 | 885 | 0.2 |
| **Bulgaria** | 3 509 | 2 723 | 2 721 | 0.8 |
| **Czechia** | 3 768 | 2 075 | 2 076 | 0.6 |
| **Denmark** | 4 300 | 1 483 | 1 483 | 0.3 |
| **Germany** | 16 776 | 9 589 | 9 588 | 0.6 |
| **Estonia** | 769 | 737 | 737 | 1 |
| **Ireland** | 3 567 | 2 320 | 2 320 | 0.7 |
| **Greece** | 5 327 | 2 840 | 2 843 | 0.5 |
| **Spain** | 12 615 | 12 294 | 12 291 | 1 |
| **France** | 21 157 | 15 467 | 15 466 | 0.7 |
| **Croatia** | 1 616 | 1 087 | 1 086 | 0.7 |
| **Italy** | 9 749 | 7 764 | 7 767 | 0.8 |
| **Cyprus** | 486 | 228 | 228 | 0.5 |
| **Latvia** | 1 247 | 1 242 | 1 243 | 1 |
| **Lithuania** | 2 949 | 1 744 | 1 743 | 0.6 |
| **Luxembourg** | 299 | 120 | 120 | 0.4 |
| **Hungary** | 2 413 | 2 890 | 2 413 | 1 |
| **Malta** | 35 | 53 | 35 | 1 |
| **Netherlands** | 3 737 | 1 114 | 1 111 | 0.3 |
| **Austria** | 4 304 | 1 647 | 1 653 | 0.4 |
| **Poland** | 14 860 | 8 559 | 8 567 | 0.6 |
| **Portugal** | 2 642 | 2 081 | 2 083 | 0.8 |
| **Romania** | 6 862 | 6 569 | 6 569 | 1 |
| **Slovenia** | 920 | 348 | 343 | 0.4 |
| **Slovakia** | 2 701 | 1 094 | 1 088 | 0.4 |
| **Finland** | 2 364 | 2 284 | 2 281 | 1 |
| **Sweden** | 5 144 | 3 890 | 3 893 | 0.8 |
| **Total EU** | 137 768 | 93 128 | 92 633 | 0.7 |

**Figure 4.3:** Landscape Features selected points in Belgium




# 4.4 Soil

The main objective of this module is to estimate soil organic carbon (SOC) content:

- For cropland at NUTS 2 level;
- For grassland and woodland at NUTS 0 level.

Initially, the JRC used the Raosoft calculator([3]) to determine the recommended size of the soil set to estimate SOC with the following precision: 90% confidence and 10% margin of error for cropland at NUTS 2 level, 90% confidence and 5% margin of error for cropland, grassland and woodland at NUT 0 level.

A total of 41,000 points distributed by LC class as follows:

- Cropland: 26,000 points
- Grassland: 6,000 points

([3]) http://www.raosoft.com/samplesize.html

- Woodland: 8,000 points
- Wetland: 1,000 points

This approach did not take into account the information on the distribution of the target variable (Soil Organic Carbon) in the strata / domains (defined by NUTS levels and land cover values 'cropland', 'woodland', 'grassland' and 'wetland'). The allocation should be proportional to the variability in the strata and to the desired precision levels in domains. In this approach, only the precision levels in the domains were considered to determine the allocation to the strata.

The proposed approach is a refinement of JRC initial allocation, based on the idea of using the available information related to SOC distribution by the land covers involved. This implies that a prediction of the SOC value is made for all points in the Master Dataset.

This was also done in a previous survey (D. de Brognieza et al., 2015), but the correlation between the observed and predicted values was not high (R2 = 0.27).

This exercise was repeated using all the information available in the Master Dataset, obtaining better results.

Based on the above considerations, the whole procedure consisted of the following steps:

1. Prediction of Soil Organic Carbon (SOC) for all the points in the Master;
2. Evaluation of model variance;
3. Optimization of both stratification and allocation;
4. Selection of the Soil module sample.

## 4.4.1. Prediction of Soil Organic Carbon (SOC)

The availability of the set of 12,516 Soil Organic Carbon observations in the Soil module subsample in 2018 suggested the modelling the SOC variable with a number of variables as explicative ones, available in the Master sample, i.e.:

- NUTS0_16: country
- ELEV: elevation
- CLC18_R: CORINE Land Cover 2018
- TDC15: Forest 2015
- FTY15: Forest Type 2015
- WAW2015: Water 2015
- GRA2015: Grass 2015
- imp15_cl: Imperviousness 2015 (class)
- STR18: Photo Interpreted Land Cover in 2018
- Slope: inclination
- LC: assigned values of Land Cover accordingly to par. 2.1.1.

The fitting of a Random Forest model was done on the train set (50 % of the 2018 Soil observations), and the significance of the different explicative variables was as follows:

| | %IncMSE |
|---|---|
| NUTS0_16 | 35.483785 |
| ELEV | 28.893416 |
| TDC15 | 25.272457 |
| CLC18_R | 24.795137 |
| LC | 17.404893 |
| SLOPE | 14.356055 |
| STR18 | 13.378475 |
| FTY15 | 9.514700 |
| GRA2015 | 8.710819 |
| WAW2015 | 4.865614 |
| imp15_cl | 2.476520 |

In the test set (remaining 50%), the correlation coefficient between observed and predicted values was 53%. We can compare this value with the other value previously obtained in a similar study (D. de Brognieza et al., 2015), which was 27 %.

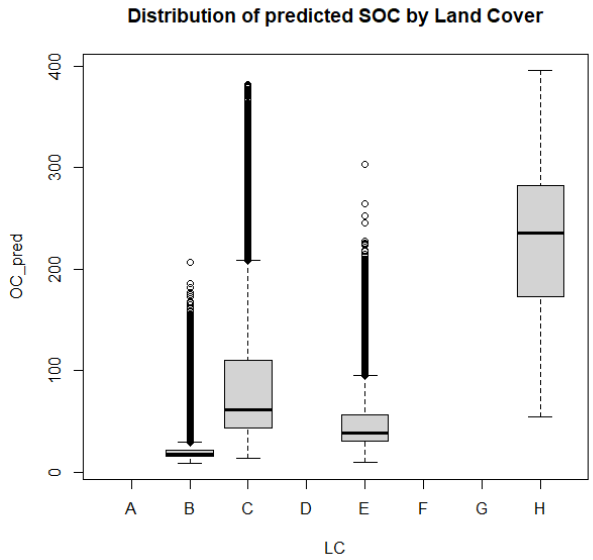
Considering that the actual target variables of the Soil module are the SOC in the different LC areas pertaining to classes B, C, E and H, four different models were fitted, again with the same Random Forest model and the same explicative variables, one for each of these classes. This is very convenient given the very different distribution of SOC values in the LC classes (see Figure 4.4).

The correlation coefficients between observed and predicted values in each class were:

- Land Cover B (cropland): 64.7 %
- Land Cover C (woodland): 59.8 %
- Land Cover E (grassland): 62.5 %
- Land Cover H (wetland): 87.7 %

Whereas the above values could be considered very satisfactory.

**Figure 4.4: Distribution of predicted SOC by Land Cover in Master sample**



## 4.4.2. Evaluation of model variance and heteroscedasticity

In order to account for the so-called 'anticipated variance' in the strata formed and optimized in the sampling frame, an assessment of the model variance must be performed. For each of the target variables, a linear model was fitted between observed and predicted values. Making use of the 'computeGamma' function, the total variance and a heteroscedasticity index were calculated for each model.

The information for the four models is structured in this way:

```
      beta      sig2   type     gamma
1 1.174005  1.961326 linear 0.8317806
2 1.142319 10.128160 linear 0.7180787
3 1.172472  8.260275 linear 0.7081315
4 1.626680  1.995340 linear 0.7420506
```

Where the meanings of the column derive from the following model with heteroscedasticity:

$$Z = \beta Y + \epsilon$$

Where $\epsilon \sim N(0, \sigma^2 Y^{2\gamma})$ , $Z$ is the target variable and $Y$ is the predicted one.

Consequently, beta=β is the regression coefficient, $sig2 = \sigma^2$ is the model variance and gamma=$\gamma$ is the etheroschedastic parameter.

These values, along with other parameters, were passed to the 'optimStrata' function of the SamplingStrata package to correctly evaluate the strata variance of the four target variables.

## 4.4.3. Optimization of stratification and allocation

The optimization step (in R package SamplingStrata) operates in this way:

1. Each target variable (the four of them: SOC in cropland, SOC in woodland, SOC in grassland and SOC in wetland) is also considered as a stratification variable;
2. The strata of the sampling frame are defined by the cross-product of each stratification variable class (obtained by randomly cutting their definition interval);
3. In each stratum so obtained, the variance of the target variables is calculated by using the predicted values and inflating the result (to account for model variance);
4. For each stratification so obtained, the sample size required to be compliant with the precision constraints is calculated;
5. Steps 2-4 are repeated for a given number of iterations following the logic of a genetic algorithm.

At the end, the stratification that yields the minimum sample size is retained as the optimal stratification.

As input for the optimization step, which is performed independently in each country, there are:

- The eligible points in LUCAS sample (158 020 points)
- The 'fixed' points, i.e., the 15 950 indicated by the commitment, plus all eligible points in class 'H' (wetland, in order to ensure an adequate number of points for this class), i.e., 17 156 points in total.

The parameter set is the desired number of strata for each country, determined in a previous step (by applying a kmeans-based algorithm that can set an appropriate number of final strata), with a maximum

number of strata of 10 in each case. That was implemented in order to obtain the final number of strata.

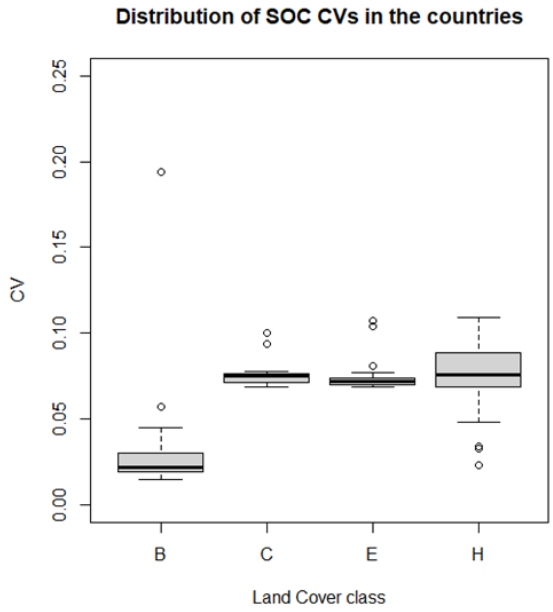
The precision constraints, expressed as maximum expected coefficients of variation with regard to the target estimates, i.e., the SOC content in each one of the Land Cover classes of interest, were set as follows:

1. SOC in cropland: 1 %
2. SOC in woodland: 5 %
3. SOC in grassland: 5 %
4. SOC in wetland: 5 %

This favours the accuracy of the first target variable, as requested by the commitment. The optimization identified a sample size of about 61 000, while the affordable size is 41 000.

Thus, an adjustment step was needed to retain the desired sample size. This adjustment obviously implied an increase in the expected CVs, beyond the values of the precision constraints. Their distribution is visualized graphically in Figure 4.5.

**Figure 4.5: Distribution of expected coefficients of variation**




## 4.4.4. Selection of the subsample

The selection method was of the 'spatial' type, i.e., the coordinates of each point were taken into account to obtain a spatially balanced sample.

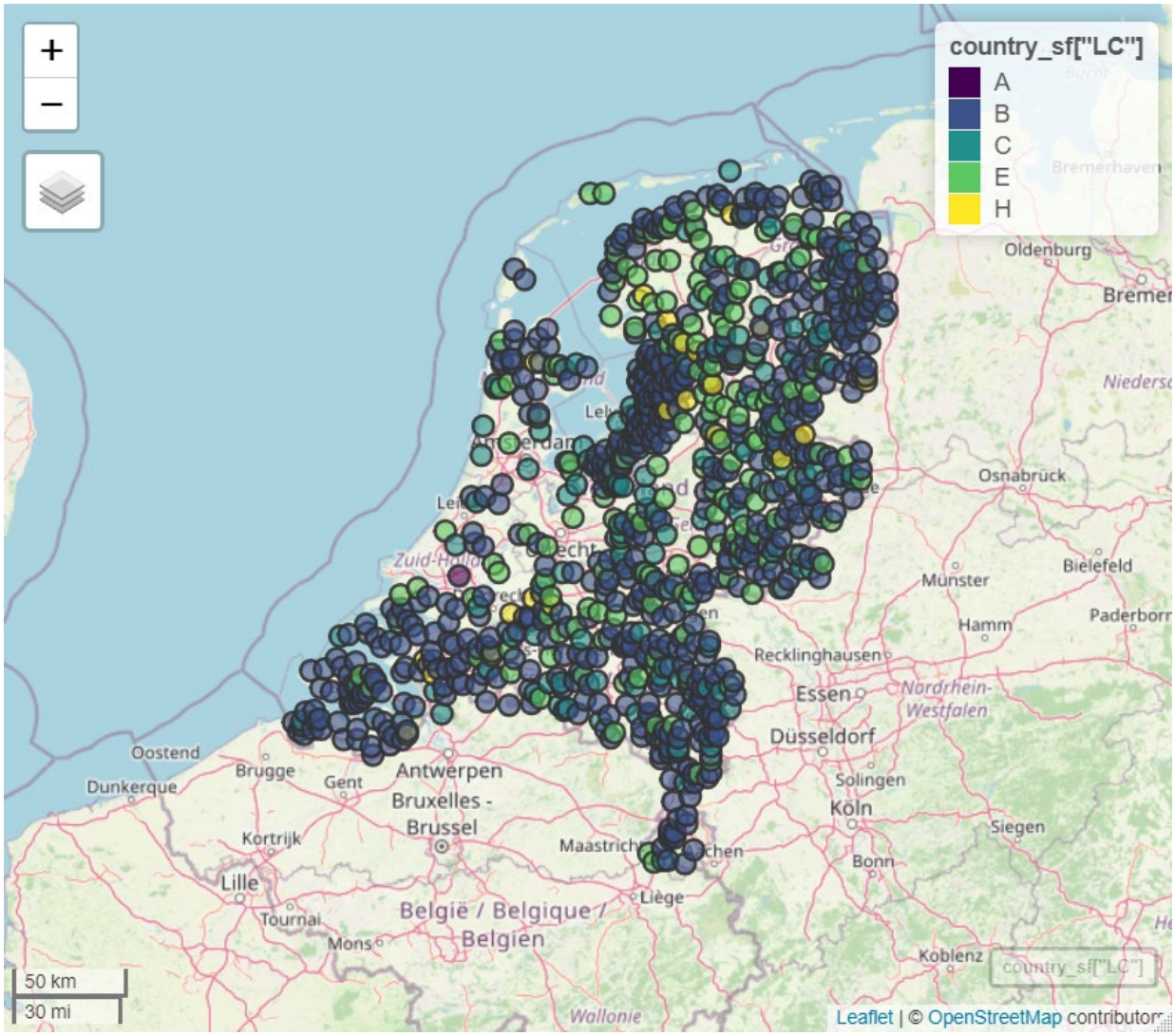
The distribution of selected points according to different characteristics is shown in Table 4.7.

**Table 4.7:** Distribution of Soil module selected points by country and Land Cover class

| Country | A | B | C | D | E | F | H | Total |
|---|---|---|---|---|---|---|---|---|
| **Belgium** | 1 | 808 | 162 | | 177 | | 10 | 1 158 |
| **Bulgaria** | 1 | 847 | 259 | 1 | 241 | | 7 | 1 356 |
| **Czechia** | 1 | 972 | 189 | | 239 | | 13 | 1 414 |
| **Denmark** | 1 | 1 082 | 132 | | 109 | | 24 | 1 348 |
| **Germany** | | 2 077 | 378 | | 349 | 1 | 40 | 2 845 |
| **Estonia** | 1 | 208 | 136 | | 112 | | 4 | 461 |
| **Ireland** | | 315 | 75 | | 198 | | 152 | 740 |
| **Greece** | 1 | 957 | 309 | 3 | 274 | 2 | 59 | 1 605 |
| **Spain** | 1 | 2 775 | 797 | 16 | 732 | 16 | 25 | 4 362 |
| **France** | | 3 023 | 789 | 9 | 902 | 15 | 38 | 4 776 |
| **Croatia** | | 298 | 156 | 3 | 144 | 1 | 5 | 607 |
| **Italy** | 1 | 1 672 | 435 | 7 | 439 | 1 | 24 | 2 579 |
| **Cyprus** | 1 | 155 | 68 | 3 | 60 | 1 | 2 | 290 |
| **Latvia** | 1 | 302 | 242 | | 159 | | 13 | 717 |
| **Lithuania** | | 665 | 179 | | 235 | | 31 | 1 110 |
| **Luxembourg** | 1 | 74 | 71 | | 55 | | | 201 |
| **Hungary** | 1 | 504 | 195 | | 186 | | 25 | 911 |
| **Malta** | | 11 | 2 | | 7 | | | 20 |
| **Netherlands** | 1 | 583 | 119 | | 169 | | 23 | 895 |
| **Austria** | 2 | 855 | 266 | 7 | 358 | 1 | 23 | 1 512 |
| **Poland** | | 2 281 | 444 | | 469 | 1 | 35 | 3 230 |
| **Portugal** | | 505 | 259 | 2 | 221 | 1 | 10 | 998 |
| **Romania** | | 1 154 | 146 | | 275 | 1 | 38 | 1 614 |
| **Slovenia** | 1 | 177 | 191 | 2 | 140 | | 1 | 512 |
| **Slovakia** | 1 | 718 | 174 | | 183 | | 4 | 1 080 |
| **Finland** | 2 | 507 | 910 | | 194 | 1 | 204 | 1 818 |
| **Sweden** | | 759 | 1 463 | 3 | 199 | 1 | 420 | 2 845 |
| **Total EU** | 19 | 24 284 | 8 546 | 56 | 6 826 | 43 | 1 230 | 41 004 |

As an example of the geographical distribution of selected points, an example for the Netherlands is provided in Figure 4.6.

**Figure 4.6:** Geographical distribution of Soil module selected points in the Netherlands




# 4.5 Copernicus

The only explicit requirement of this module was the total number of points to be selected i.e. 150 000. In order to make the module subsample representative of the whole Master Dataset, the allocation of units had been proportionally implemented to the predicted Land Cover.

This resulted in the following inclusion probabilities and allocation:

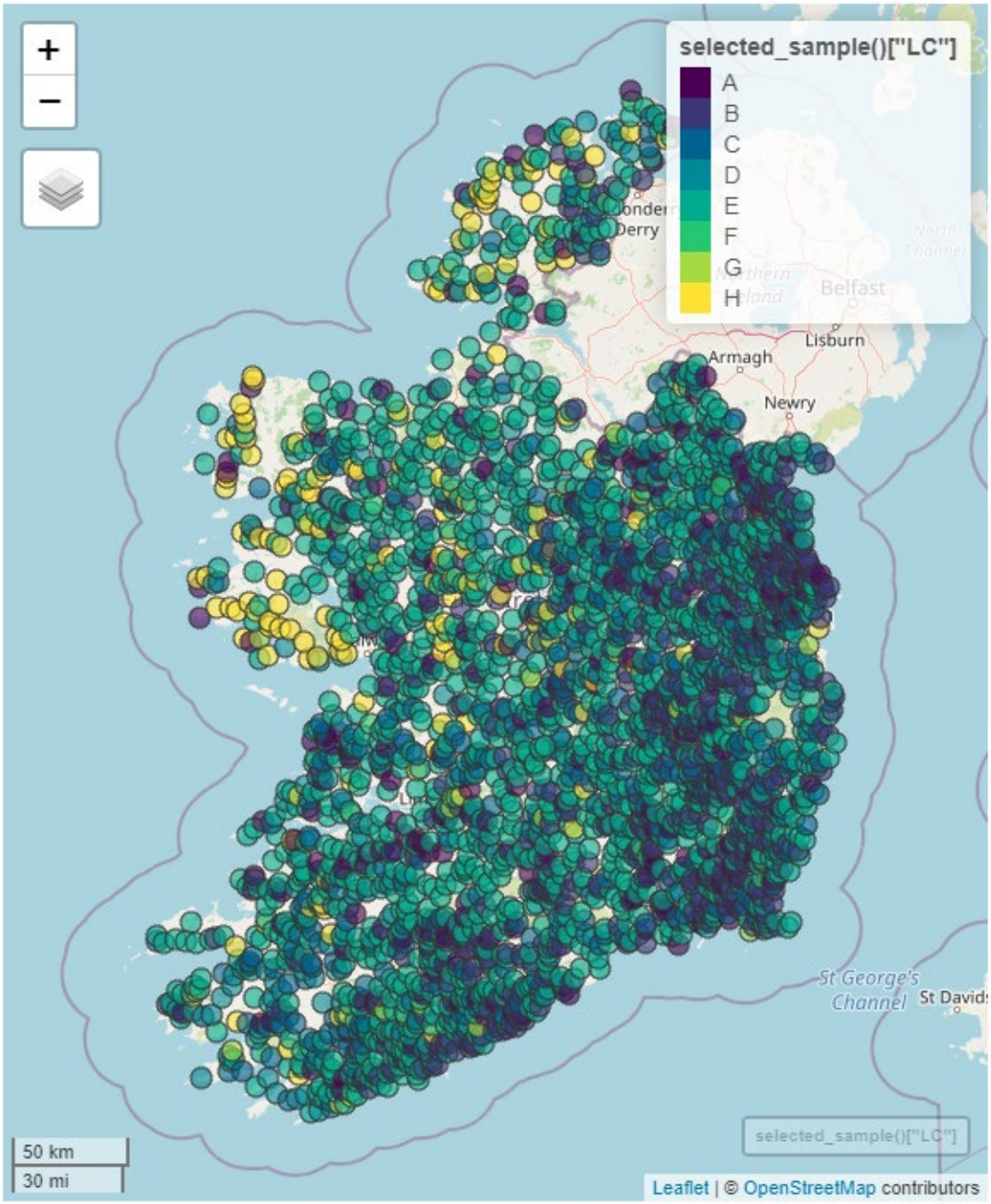
| LC | probs | allocation |
|---|---|---|
| A | 0.9259214 | 8106 |
| B | 0.6982757 | 69617 |
| C | 1.0000000 | 48155 |
| D | 1.0000000 | 7850 |
| E | 0.5361779 | 59231 |
| F | 0.7957581 | 3666 |
| G | 1.0000000 | 1966 |
| H | 1.0000000 | 1236 |

The selection method was of the 'spatial' type, i.e., taking into consideration the coordinates of each point to obtain a spatially balanced sample. The distribution of selected points according to different characteristics is depicted in Table 4.8.

**Table 4.8: Distribution of Copernicus module selected points and sampling rate by country**

| Countries | Available points | Selected points | Sampling rate |
|---|---|---|---|
| **Belgium** | 4 879 | 3 571 | 0.7 |
| **Bulgaria** | 5 047 | 3 766 | 0.8 |
| **Czechia** | 4 553 | 3 177 | 0.7 |
| **Denmark** | 5 148 | 3 710 | 0.7 |
| **Germany** | 23 077 | 17 092 | 0.7 |
| **Estonia** | 1 229 | 925 | 0.8 |
| **Ireland** | 4 299 | 2 846 | 0.7 |
| **Greece** | 9 430 | 7 608 | 0.8 |
| **Spain** | 17 727 | 13 457 | 0.8 |
| **France** | 27 959 | 20 075 | 0.7 |
| **Croatia** | 2 908 | 2 246 | 0.8 |
| **Italy** | 17 403 | 13 849 | 0.8 |
| **Cyprus** | 955 | 791 | 0.8 |
| **Latvia** | 3 385 | 2 913 | 0.9 |
| **Lithuania** | 3 893 | 2 764 | 0.7 |
| **Luxembourg** | 519 | 404 | 0.8 |
| **Hungary** | 3 683 | 2 799 | 0.8 |
| **Malta** | 51 | 38 | 0.8 |
| **Netherlands** | 4 723 | 3 327 | 0.7 |
| **Austria** | 5 482 | 3 868 | 0.7 |
| **Poland** | 17 534 | 12 189 | 0.7 |
| **Portugal** | 4 174 | 3 234 | 0.8 |
| **Romania** | 7 867 | 5 399 | 0.7 |
| **Slovenia** | 2 784 | 2 426 | 0.9 |
| **Slovakia** | 3 701 | 2 739 | 0.7 |
| **Finland** | 5 405 | 4 576 | 0.9 |
| **Sweden** | 12 012 | 10 211 | 0.9 |
| **Total EU** | 199 827 | 150 000 | 0.8 |

**Figure 4.7:** Geographical distribution of Copernicus module selected points in Eire



# 4.6 Overlap of the subsamples

The five selected subsamples have common points. The overlap situation is shown in Table 4.9.

**Table 4.9: Overlap of subsamples**

| Overlap of subsamples | | |
|---|---|---|
| **Number of points interested by at least one subsample** | 192 268 | 96.1% |
| **Number of points interested by only one subsample** | 85 516 | 42.8% |
| **Number of points interested by two subsamples** | 69 756 | 34.9% |
| **Number of points interested by three subsamples** | 30 369 | 15.2% |
| **Number of points interested by four subsamples** | 6 173 | 3.1% |
| **Number of points interested by all subsamples** | 634 | 0.3% |

The composition of the overlaps is presented in Table 4.10.

**Table 4.10: Composition of overlap**

| Module | Grassland | Ext.grassland | LF | Soil | Copernicus |
|---|---|---|---|---|---|
| **Grassland** | 20 000 | 12 853 | 12 544 | 2 941 | 11 477 |
| % | 100 | 64 | 63 | 15 | 57 |
| **Ext.grassland** | 12 853 | 40 000 | 25 279 | 4 685 | 21 372 |
| % | 32 | 100 | 63 | 12 | 53 |
| **LF** | 12 544 | 25 279 | 92 633 | 21 669 | 60 852 |
| % | 14 | 27 | 100 | 23 | 66 |
| **Soil** | 2 941 | 4 685 | 21 669 | 41 004 | 30 389 |
| % | 8 | 14 | 41 | 20 | 100 |
| **Copernicus** | 11 477 | 21 372 | 60 852 | 30 389 | 150 000 |
| % | 8 | 14 | 41 | 20 | 100 |

For instance, when conducting the Grassland module, of the 20 000 points observed, 12 853 are also observed for Extended Grassland, 12 544 for the Landscape Features module, 2 945 for Soil and 11 477 for Copernicus.

# 5 Conclusions

Some methodological changes have been introduce in the design of LUCAS 2022 sample, based on the empirical findings of the previous round or on new needs, essentially linked to the development of five subsamples, related to the corresponding modules, integrated into the main LUCAS sample.

First, Eurostat has increased the LUCAS sample size for 2022 to 400 000 points, compared to almost 336 000 in 2018. All sample sizes at country level have been averagely increased by 25 % compared to 2018. The percentage distribution of the sample by country substantially reflects the 2018 campaign, although some adjustments have been made to take into account sampling errors calculated using 2018 data. The proportion of photo-interpreted points increased from around 30 % in 2018 to 50 % in 2022. In addition, the number of involved countries decreased by one (United Kingdom). With the same number of countries, the 'in field' sample rate decreases slightly (about 8 %), but the photo-interpreted points are almost doubled in comparison with the 2018. This choice has the dual objective of increasing the reliability of the estimates, while limiting the cost of the increased precision of the estimates.

Unlike previous surveys, the design of the sample included Land Use modalities (4) along with Land Cover modalities (8) in the set of target estimates, as these variables are becoming increasingly important in the dissemination process.

The stratification procedure has also changed. In the 2018 survey, stratification started with clustering all Master points into 'atomic' strata obtained from all combinations of the stratification criteria, i.e. STR18, CORINE Land Cover, and elevation variables. An optimization algorithm aggregated them iteratively to optimize the stratification by minimizing the total sample size required to meet the precision constraints. Stratification was thus not produced ex-ante by a fixed combination of variables, but depended on the correlation between the stratification characteristics and the target variables. The combinations of the stratification criteria varied according to the specificities of the NUTS2 territory. In the 2022 sample, strata are formed ex-ante by combining in each NUTS2 all available modalities of STR18, CLC Land Cover, and a binary indicator signalizing if the target variable values in the Master were imputed or observed. The fixed stratification ex-ante represents, conceptually, a return to the 2015 survey logic, although in this case the stratification is much more detailed than the 2015 stratification. The fixed stratification is more appropriate than the 2018 stratification, in case some kind of panel structure (e.g. rotating panel) will be introduced in the next LUCAS surveys.

In 2018, sampling units in each stratum were selected using a Simple Random Sample (SRS) procedure; in 2022, the drawing of the sampling points was carried out utilizing balanced spatial sampling in the strata to account for spatial correlation.

As in the previous round, once a point is selected, it must be assigned to the observation mode: directly surveyed or photo-interpreted in the office. The distinction of sample points in the two sets is always required before the start of the survey. In 2022, the assignment of observation mode is particularly important, as direct observation is required for the corresponding points in all modules. While in the previous LUCAS rounds the assignment was based on two indices (Reachability and Propensity to Change), the distinction between observation modes in the 2022 survey is based on a revised Reachability index and a set of deterministic rules supported by a geographical visualisation application for the selected points.

The selected LUCAS sample (or practically all the points directly observable), constitutes the selection list for the module subsamples and, to ensure this function, it is necessary that the eligibility criteria are satisfied by an appropriate amount of the LUCAS sampled points. The eligibility criteria also define the reference populations in the Master for the modules, needed to consider the subsamples as probabilistic, a characteristic specifically requested in the 2022 survey unlike in the previous LUCAS rounds.

The integration of the five modules into the 2022 LUCAS survey has some implications for data collection and data processing. In the previous rounds, the points that could not be directly observed during data collection were replaced by photointerpretation, which ensured the maintenance of the planned sample size. In the 2022 survey, the treatment of missing units will remain the same for the LUCAS sample, but no replacement or photointerpretation will be possible for the modules; non-response will reduce the planned subsample sizes and some kind of treatment should be implemented. This predictable increase in the number of missing units depends on two reasons. Firstly, the increase in statistical burden due to the joint data collection of LUCAS and one or more modules (96.1 % of the LUCAS sample is joined with at least one module and the 18.6 % with three or more modules) for all situations where direct observation is conducted in a private area as the 'agricultural' points (in the 2018 survey they were about the 55 % of the direct observations in the sample). Secondly, the uncertainty around the actual eligibility of points for a given module, which in many cases was defined according to the predicted values rather than observed values.

The changes in 2022 represent a turning point for LUCAS. The integration of modules and their planned continuity into the future will require a major change in survey design, especially if the requirement, common to all the modules, to have a panel of points in the next rounds is to be met.

In this context, almost half of the sample and exactly the part of the directly observed points (since the module subsamples represent 48.1 % of the LUCAS sample) will have a panel structure, while the remaining part of the sample could have a cross-sectional design.

In the LUCAS survey, part of the sample units belong to the previous sample, which should reduce the standard error of the estimates of variations at different points in time by the covariance between the common points.

However, since longitudinal estimates are not currently produced according to a sampling scheme, it is generally only possible to conduct longitudinal analysis in a descriptive manner. It is therefore desirable to change the current cross-sectional design by introducing a panel structure that focuses more on estimating 'changes' over time.

The panel component selected on the bases of a sample design allows the calculation of variation estimates by aggregating differences at the unit level; the related standard errors are lower than those of the variation of independent estimates are in a cross-sectional scheme of the same size. The longitudinal component implies a profound revision of data collection procedures: fieldwork, checks on collected data, periodicity of massive photointerpretation, data treatment, and estimation procedure.

All these changes present LUCAS with new challenges and new opportunities. So far, LUCAS has been considered a direct survey more focused on providing Land Cover statistics harmonized at European level. Nevertheless, new reliable and user-friendly tools to produce LC statistics are continuously being released by public or private companies (for example the World Cover viewer released by ESA[4] or the S2-GLC[5]), most of which are based on remote earth observation data, such as those from Sentinel 1 and 2. Their results could be updated more frequently and this offers LUCAS the opportunity to play a new role that is not limited to the production of Land Cover statistics, but also gives greater importance to other issues that cannot be covered remotely (Land Use, the information collected by the modules, the landscape diversity, fragmentation, etc.).

LUCAS could also play a complementary role by producing quality indicators for statistics from Earth Observation and, in particular, by ensuring the production of confusion matrices and related indicators,

[4] https://viewer.esa-worldcover.org/worldcover

[5] https://s2glc.cbk.waw.pl/

given the accuracy of data collection in the field and the finer classification of the different Land Cover and Land Use modalities. Under this hypothesis, the part of the sample collected through direct observation needs to be renewed in all rounds, and this necessity should be harmonized with the panel structure required for the modules, for example through a rotating panel. Aligning, even if only partially, the definitions of LUCAS and CORINE Land Cover, would allow important synergies; for example, the totals at some territorial levels from CORINE could be used to calibrate (in statistical terms) the LUCAS data substituting the corresponding LUCAS estimates, thus achieving a more accurate estimate and avoiding duplication of information on the same topic.

Further improvements can be envisaged for future LUCAS. Leaving room for countries to add the observation of phenomena of national interest, may, on the one hand, improve the quality of the survey thanks to specific local knowledge and, on the other hand, lead to a greater participation in the data estimation and analysis phase.

As far as model-based analyses are concerned, LUCAS could make a fundamental contribution to the tuning of models for the spatialization of phenomena or new classifications of the territory. The models commonly used for this purpose are *supervised*; they require good quality field observations that are also consistent with the type of models to be interpolated and the format of the available explanatory variables. These requirements increase costs and demand resources that are too burdensome even for some European institutions; the synergies with LUCAS could make it possible to mitigate these problems and open up new areas of research.

# 6 References


d'Andrimont, R., Verhegghen, A., Meroni, M., Lemoine, G., Strobl, P., Eiselt, B., Yordanov, M., Martinez-Sanchez, L., van der Veld, M. (2021), 'Lucas Copernicus 2018: Earth observation relevant *in-situ* data on land cover throughout the European Union', *Earth Syst. Sci. Data*, 13, 1119–1133.

d'Andrimont, R., Yordanov, M., Martinez-Sanchez, L., Eiselt, B., Palmieri, A., Dominici, P., Gallego, J., Reuter, H.I., Joebges, C., Lemoine, G., van der Velde, M. (2020), 'Harmonised LUCAS in-situ land cover and use database for field surveys from 2006 to 2018', *European Union. Sci Data.*, 2020 Oct,16;7(1):352. doi: 10.1038/s41597-020-00675-z. PMID: 33067440; PMCID: PMC7567823.

Ballin, M., Barcaroli, G., Masselli, M. Scarnò, M. (2018), 'Redesign sample for Land Use/Cover Area frame Survey' (LUCAS) 2018. *Statistical Working Paper*, Eurostat

Ballin, M, Barcaroli, G. (2013), 'Joint determination of optimal stratification and sample allocation using genetic algorithm', *Survey Methodology*, Vol. 39, n.2/2013

Barcaroli, G. (2014), 'SamplingStrata: An R Package for the Optimization of Stratified Sampling', *Journal of Statistical Software*, 61(4), 1–24.

Barcaroli, G., Ballin, M., Odendaal, H., Pagliuca, D., Willighagen, E., Zardetto, D. (2020). SamplingStrata: Optimal Stratification of Sampling Frames for Multipurpose Sampling Surveys. R package version 1.5-2 URL https://cran.r-project.org/web/packages/SamplingStrata/index.html.

Bethel, J.W. (1989), 'Sample Allocation in Multivariate Surveys', *Survey Methodology*, Vol. 15, pp. 47–57.

de Brogniez, D., Ballabio, C., Steven, A., Jones, R. J. A., Montanarella, L., van Wesemael, B. (2015), 'A map of the topsoil organic carbon content of Europe generated by a generalized additive model', *European Journal of Soil Science*, January 2015, 66, 121–134.

Cochran, W.G. (1977), *Sampling Technique*, 3nd Edition, John Wiley and Sons Inc., New York.

Eurostat, *Technical reference document c-1: Instructions for surveyors*. https://ec.europa.eu/eurostat/documents/205002/8072634/LUCAS2018-C1-Instructions.pdf (2018).

Eurostat, *LUCAS Quality Report 2015*, https://ec.europa.eu/eurostat/documents/205002/769457/LUCAS+Quality+Report+2015

Eurostat, *LUCAS Quality Report 2018*, https://ec.europa.eu/eurostat/documents/205002/769457/LUCAS-2018-Quality-Report

Grafström, A., Lundström, N.L.P., Schelin, L. (2012), 'Spatially balanced sampling through the pivotal method', *Biometrics,* 68(2):514–520.

Hansen, M.H., Hurwitz, W.N., *Madow, W.G. (1993), Sample Survey Methods and Theory*, John Wiley & Sons Inc, New York

Lisic, J,, Grafström, A .(2018). SamplingBigData: Sampling Methods for Big Data. R package version 1, https://CRAN.R-project.org/package=SamplingBigData.

James, G., Witten, D., Hastie, T., Tibshirani, R. (2013), 'An Introduction to Statistical Learning with application in R', *Springer,* New York.

Lisic, J., Grafström, A. (2020), *SamplingBigData: Sampling Methods for Big Data,* https://CRAN.R-project.org/package=SamplingBigData.

Palmieri, A. (2016), 'Integrating statistical and geographical information: LUCAS survey, a case study for land monitoring in European Union', CONFERENCE OF EUROPEAN STATISTICIANS Workshop on Statistical Data Collection 'Visions on Future Surveying' 3-5 October 2016, The Hague, Netherlands.

Weigand, M., Staab, J., Wurm, M., Taubenböck H. (2020), 'Spatial and semantic effects of lucas samples on fully automated land use/landcover classification in high-resolution sentinel-2 data', *International Journal of Applied Earth Observation and Geoinformation* 88,102065 (2020).

**GETTING IN TOUCH WITH THE EU**

**In person**
All over the European Union there are hundreds of Europe Direct information centres. You can find the address of the centre nearest you at: https://europa.eu/european-union/contact_en

**On the phone or by email**
Europe Direct is a service that answers your questions about the European Union. You can contact this service:
– by freephone: 00 800 6 7 8 9 10 11 (certain operators may charge for these calls),
– at the following standard number: +32 22999696 or
– by email via: https://europa.eu/european-union/contact_en

**FINDING INFORMATION ABOUT THE EU**

**Online**
Information about the European Union in all the official languages of the EU is available on the Europa website at: https://europa.eu/european-union/index_en

**EU publications**
You can download or order free and priced EU publications at: https://op.europa.eu/en/publications. Multiple copies of free publications may be obtained by contacting Europe Direct or your local information centre (see https://europa.eu/european-union/contact_en).

**EU law and related documents**
For access to legal information from the EU, including all EU law since 1952 in all the official language versions, go to EUR-Lex at: http://eur-lex.europa.eu

**Open data from the EU**
The EU Open Data Portal (http://data.europa.eu/euodp/en) provides access to datasets from the EU. Data can be downloaded and reused for free, for both commercial and non-commercial purposes.

KS-TC-22-005-EN-N

# New LUCAS 2022 sample and subsamples design: Criticalities and solutions

**MARCO BALLIN, GIULIO BARCAROLI, MAURO MASSELLI**

The Eurostat Land Use/Cover Area frame Survey (LUCAS) is mainly an in-situ survey designed to provide harmonized statistics on Land Use and Land Cover across the European Union. LUCAS 2022 survey covers all EU Member States by observing 400,000 selected points. In addition to the Land cover and Land Use observed at each of these points, further information is collected in order to assess environmental aspects such as the grassland and soil quality.

This paper delineates the design of the LUCAS 2022 sample and for the specific sub-samples.

**For more information**
**https://ec.europa.eu/eurostat/**

Publications Office of the European Union

ISBN 978-92-76-53401-3