Redesign sample for Land Use/Cover Area frame Survey (LUCAS) 2018

MARCO BALLIN, GIULIO BARCAROLI, MAURO MASSELLI, MARCO SCARNÓ

2018 edition

STATISTICAL WORKING PAPERS

eurostat

Redesign sample for Land Use/Cover Area frame Survey (LUCAS) 2018

MARCO BALLIN, GIULIO BARCAROLI,
MAURO MASSELLI, MARCO SCARNÓ

2018 edition

Manuscript completed in September 2018

Neither the European Commission nor any person acting on behalf of the Commission is responsible for the use that might be made of the following information.

Luxembourg: Publications Office of the European Union, 2018

The information and views set out in this publication are those of the authors and do not necessarily reflect the official opinion of the European Union. Neither the European Union institutions and bodies nor any person acting on their behalf may be held responsible for the use which may be made of the information contained therein.

PDF ISBN 978-92-79-93921-1 ISSN 2315-0807 doi: 10.2785/132365 KS-TC-18-006-EN-N

Abstract

Eurostat’s Land Use/Cover Area frame Survey (LUCAS) provides harmonised statistics on land use and land cover across the European Union. Land use shows the socio-economic use of a given land: agriculture, commerce, industry, residence, etc. while Land cover refers to the bio-physical coverage of the land: crops, forest, buildings, roads, etc. LUCAS is characterised by the unique, in-situ information that it provides.

Since the end of the pilot phase (2006), Eurostat has carried out this survey every 3 years (2009, 2012, 2015) with the next LUCAS survey, covering all the 28 European Union (EU) countries, taking place in 2018. Each round has been an opportunity to adapt, adjust and constantly improve the LUCAS methodology in order to obtain the most accurate and harmonised data possible.

In this context, the LUCAS 2018 survey focused on (i) a different specification and use of the non-eligibility concept, (ii) a review of the rules for assigning photo-interpreted and field points in the sample, (iii) a finer stratification.

This paper shows the characteristics of the new sampling design for 2018, and the innovative methodology which is applied to it.

After a first chapter on the features of the LUCAS survey, chapter 2 details the data preparation (statistical models for land cover variation prediction, probability to change and reachability, use of satellite information) implied by the new methodology. The latter is presented in the next chapter, outlining the 2018 sample design and sampling procedure, while chapter 4 discusses the main differences between the sample designs of the last round in 2015 and the next one for 2018. Finally, in chapter 5, the conclusions are presented.

Keywords: optimal sampling, land cover and land use survey

Authors:

Marco Ballin, Giulio Barcaroli, Mauro Masselli, Marco Scarnó.

Acknowledgement:

This work has been carried out by under the supervision of the Eurostat project manager Alessandra Palmieri. Special thanks to Yohann Langonne for taking care of the layout and the backstopping support.

The information and views set out in this publication are those of the author(s) and do not necessarily reflect the official opinion of the European Union. Neither the European Union institutions and bodies nor any person acting on their behalf may be held responsible for the use which may be made of the information contained therein.

Contents

Figures

Tables

Abbreviations

CLC	Corine Land Cover
CRAN	Comprehensive R Archive Network
CVs	Coefficients of Variation
DG	Directorate General
EC	European Commission
ELEV	Elevation
EU	European Union
FN	False Negative
FP	False Positive
GOPA	Gesellschaft für Organisation, Planung und Ausbildung mbH
GPS	Global Positioning Satellite
INSPIRE	INfrastructure for SPatial InfoRmation in Europe
LC	Land cover
LUCAS	Land Use/Land Cover Area frame Survey
NUTS	Nomenclature des Unités Territoriales Statistiques
PI	Photo-interpretation
SRS	Simple Random Sample
STR	Strata
TN	True Negative
TP	True Positive
TW	Transitional Water
UN	United Nations

1 LUCAS survey

1.1 Introduction

Land is the basis for most biological and human activities on the earth. Agriculture, forestry, industry, transport, housing and other services all use land as a natural and/or economic resource. Land is also an integral part of ecosystems and indispensable for biodiversity and the carbon cycle.

Changes in land cover, biophysical attributes of the earth's surface, and land use, human purpose or intent applied to these attributes, significantly affect key aspects of Earth System functioning. The capability of monitoring those changes is linked to the availability of information on the coverage and the use of the land.

The European Commission (EC) uses land cover/use data and statistics for purposes such as nature protection, forest and water management, urban and transport planning, agricultural policy, natural hazards prevention and mitigation, soil protection and mapping, monitoring climate change, biodiversity, etc. Land cover and use data also form the base for spatial and territorial analysis which have become increasingly crucial for policy planning in many respects.

In order to improve the quality and the completeness of the land cover/use statistics, every 3 years since 2006, Eurostat has implemented the LUCAS survey, which is an area-frame statistical survey on land use and cover.

LUCAS stands for Land Use/Cover Area Frame Survey; its data contributes to some of the major EU policy areas as its findings offer a comprehensive and comparable overview on the state and the dynamics of land use and cover in the European Union.

LUCAS is carried out by direct observations of surveyors in a small area centred on the selected point. The survey is a multipurpose in-situ platform. The core survey consists of measuring the nature of land cover (arable, grassland, artificial, water, etc.), its use (industry, agriculture, transport, housing, etc.), as well as environmental parameters associated to the single surveyed points and a set of pictures taken on the point and in the cardinal directions.

In addition to the core LUCAS, some specific information, so-called "ad hoc modules" are also collected in each survey (it was the case of the top soil sample in 2009 and 2015, and the transects in 2009, 2012 and 2015).

The LUCAS surveys generate three types of information: (i) micro-data containing the statistical information collected in every sample point (ii) point and landscape photos and (iii) statistical tables with aggregated results by land cover and land use at geographical level.

The LUCAS surveys are used to monitor social and economic use of land as well as to monitor ecosystems and biodiversity. Sustainable Development Indicators and Agro Environmental indicators on soil are examples of LUCAS data use, while the micro-data collected in it also serve to produce, verify and validate Corine Land Cover (CLC) and Copernicus.

The survey consists of a two phases area sample; in the first phase a frame of more than 1 million geo-referenced points (the so-called Master sample or first phase sample) is systematically selected from a 2 square km grid built all over the EU territory; the frame is stratified according to land cover classes. From the Master sample, a second phase sample is selected; on these points statistical information is collected by surveyors in the field or by photo interpretation in the office.

1.2 The historical perspective

LUCAS survey was initially developed to provide early crop estimates for the European Commission. The survey started as a pilot survey across a limited number of EU Member States; the first survey was held in 2001.

Over time, the survey has become a key tool for policymakers and statisticians alike, with increasing amounts of data on different forms of land use and land cover in the EU. In 2006, the sampling methodology changed and its focus shifted from an agricultural land survey to a broader land cover, land use and landscape survey. In the same year, a three-yearly interval was introduced as the frequency for carrying out the survey.

The legal base 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, based on Decision 1445/2000/EC of 22/5/2000 of the Council and the European Parliament, dealing with 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 carried out again in 2003 in all EU-15 Member States plus Hungary, allowing improvement of the data collection system and analyses of land use and land cover changes (2001-2003). The project was extended in duration from 2004 to 2007 by Decision 2066/2003/EC of 10/11/2003. The coverage of the EU Member States and the related financing is laid down by Decision 786/2004/EC of 21/4/2004. In 2006, a new pilot survey was carried out on 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 starting the current data collection frequency: every three years. From January 2008 onwards, LUCAS has been part of Eurostat's activities and budget. As from 2012, it has been supported financially by other DGs of the Commission.

The initial coverage was extended to 23 EU countries in 2009 (Bulgaria, Cyprus, Malta and Romania were not included), 27 member states in 2012 and finally 28 member states in LUCAS 2015 and 2018. The number of points collected has thus increased accordingly.

1.3 Methodological challenge - accessibility

LUCAS process relies on two main aspects: it is a sample survey where the unit of observation is the point (more precisely the circle around the point); and the physical location that the surveyor has to reach is identified by geographical coordinates that can be everywhere in a region, therefore the accessibility of the point is a critical element for the data collection.

Each round of the survey brings the challenge to improve the methodological aspects of the survey in order to obtain further and more accurate data while maintaining the comparability between the different editions.

Till 2015 survey, points of the Master were divided into "eligible" and "not eligible" for field data collection.

In LUCAS 2009, the sampling rates were fine-tuned by considering the coefficient of variation for each major land cover initial classification and NUTS2 region. In addition, some conditions for the exclusion of specific points were introduced, such as considering as non-accessible the points belonging to islands not connected to the mainland, too expensive to be surveyed, and the points above 1 000 metres of elevation. Those points were considered "not eligible" for field survey, permitting the second phase sample to be selected from the "eligible" points. Despite these measures, almost 25% of the selected points in 2009 resulted to be not suitable for in situ data collection (because of geographical factors not known at the moment of designing the sample, access denied, etc.), and hence it was necessary to consider their photo-interpretation as a secondary method for gathering information. The photo-interpretation was carried out both in the office for points for which, "a priori", it was not worth reaching them, and in field operations (like in the successive surveys) when the point was not accessible to surveyor because of an obstacle (private property, military area, long distance from road, etc.).

LUCAS 2012 aimed to improve the precision of estimates by increasing the sample size with about 40 000 points and allocating them according to the diversity of the landscape resulting from the transect analysis carried out in 2009. The maximum level of elevation above which the points were automatically excluded from a field visit was also raised to 1 500 meters. In addition, some auxiliary information such as slope and distance to the main road was introduced for the first time to optimise the point selection.

The survey design of LUCAS in 2015 introduced more sophisticated criteria to evaluate the eligibility of a point. It combined information derived from the Corine Land Cover (CLC) with the distance to roads and altitude. Hence, the accessibility of each point was classified as “potentially easy” or “difficult”. In a further step combined with the altitude and/or the distance to roads (less or greater than 600 metres), it permitted to better identify a set of non-eligible points for a field visit and to rationalise the use of photo-interpreted points in the 2015 survey. In 2015, an additional methodological improvement consisted of a new estimation method that took into account the secondary information collected for land cover/use. The bias resulting from the exclusion of the non-eligible points was corrected by the photo interpretation of a complementary sample of the non-eligible points.

1.4 LUCAS 2018 main features

The survey design for 2018 has been fine-tuned in several aspects summarised in the following paragraphs.

1.4.1 Master sample update

A first important improvement is the update of the information related to the Master sample: each point of the 2 by 2 Km grid was assigned with an updated stratification and all the related geographical and administrative information available. In comparison to 2005, date of the previous grid stratification, the variable "STR05" collected in the year 2005 has been replaced and updated by a new variable "STR18" including an enlarged classification. STR18 is indeed classified in 10 modalities: the modalities of STR05 “wooded area and shrubland” is split into two (“wooded area” and “shrubland”), while two new modalities “transitional water” (estuaries, intertidal areas, coastal lagoons, etc.) and “impossible to photo-interpret” have been inserted (see Chapter 2). As a result, a fair proportion of the points (about 26.5%) changed classification.

1.4.2 Coverage: LUCAS refers to the NUTS area

6 975 points were excluded from the second phase sample: points with a stratification code equal to “transitional water” and points outside the reference NUTS area. Hence, as with the 2015 data, it will be possible to produce two different estimates of total area for each country, one excluding the “transitional water” (TW) areas and the other including them (concerning TW points, the estimates are provided by the photo-interpretation operation - see Chapter 2).

1.4.3 Survey design 2018: the concept

In the 2018 survey, data collection will be carried out by a twofold mode: either directly by surveyors in the field (in situ) or by photo-interpretation (PI) in the office. The choice to assign one modality to a selected point is however now done after the sample selection and not, as in the previous surveys, by dividing the Master into eligible and non-eligible points and then going to the selection step from these two subpopulations.

Photo-interpretation is needed if it is impossible or too costly to reach the point. Besides, it can be convenient when the probability of the point to change its land cover characteristics is low. Therefore, a point is considered in situ or photo-interpreted on the basis of two indices calculated for all the points of the Master: the reachability index and the propensity to change index (see Chapter 2). The choice depends also on the constraints of the PI quotas in each country, fixed by the technical specifications of each contract. So the thresholds of the above-mentioned indices are different from one country to another (see Chapter 3). It is to be noted that the probability, or propensity, to change has been calculated for land cover variables because the sample design has been required to take into account, as target variables, some modalities of land cover and related desired errors. However, the estimates are also required for other collected information, e.g. land use. Consequently, in some cases, points with a low probability to change in land cover could nevertheless have a high propensity to change in land use.

The LUCAS 2018 survey also faces some challenges and opens debates. For example, the photo-interpretation could produce underestimates when assessing the changes of land cover characteristics, because the available photos were taken in a previous year. To avoid or to reduce the risk of biases, the use of photo-interpretation should be limited to unchangeable points or those with a very low probability to change during the time between surveys.

1.4.4 The Model

The most probable Land Cover is assigned to each point of the Master, forecasted by a linear logistic regression model, estimated on the basis of the real data from the 2015 LUCAS survey, and considering about 16 covariates. This information is used to calculate the coefficients of variation (CVs) for the 16 target variables (see Chapter 2).

1.4.5 Dynamic stratification of the second phase sample

The second phase sample design is a stratified one but the stratification is not fixed like in the previous surveys (given by the combinations of NUTS level 2 regions (NUTS2) by STR05) but rather obtained in a dynamic way. Starting from the “atomic strata” (given by the Cartesian product of STR18, CLC and ELEV (elevation) classifications) the final strata and sample size are identified by aggregating the atomic strata with an iterative algorithm that optimises the CVs of the target variables at NUTS2 level and taking into account the related, desired sampling errors fixed ex ante. Therefore, the final stratification depends on the most correlated combinations of modalities of the stratification characteristics with the target variables; the stratification “criteria” vary according to the specificity of the country and NUTS2 territories. Finally, as the sample size corresponding to the optimised solution does not equalise the predetermined contractual amount of units to be selected in each country, in a second step this sample size is adjusted accordingly by proportionally decreasing or increasing the allocation in each stratum according to the sign of the difference between optimised and acceptable sample sizes (see Chapter 3).

1.4.6 LUCAS survey 2018: the figures

In the 2018 LUCAS, a total of 337 854 points were selected for the second phase sample including 240 174 for the in situ data collection and 97 680 to be photo-interpreted in the office. The data collection is carried out in situ in the period between March and September 2018, while the photo-interpretation in the office will be concluded by March 2019. The survey not only collects data on land cover and land use but also includes:

1. an extended soil module where a topsoil sample is collected on a maximum of 26 014 points. Out of these points, some 9 000 points will be evaluated for bulk density (this evaluation is done by the surveyor). On 1 000 locations out of these 9 000 points, a sample for assessing soil biodiversity is also to be taken. Additionally, on 1 470 points, the depth of the organic horizon is to be measured by the surveyor (up to 40 cm),
2. a test module for grassland on a maximum of 3 734 points,
3. an additional observation on 94 013 points for the Copernicus programme.

The coordinates of the points needed to gather information for the grassland and soil modules are considered in the second phase sample.

The aim of this paper is to describe the new methodology of 2018 sample design and sampling procedure.

2 Data preparation and new information added

2.1 Introduction

The LUCAS Master data set is obtained by using a 4 km² grid (2x2 km) which includes around 1 100 000 points covering the EU-28 territory. Each of these points was classified (during a specific activity conducted in 2017) into 10 land cover categories (the strata), on the basis of photointerpretation (PI) of aerial photos or satellite images. Beyond the geographical characteristics of the point (i.e. its GPS coordinates, the values of the corresponding NUTS3, NUTS2, NUTS1 and of the country), some specific information was added to each point; in particular the elevation, the distance to the nearest road, the population density in the most internal 1 km², etc.

To develop the sampling strategy described in this work, it was necessary to estimate the most probable land cover (LC) that could be observed in each point. Such information is important because it permits to estimate the distribution of the target variables in the different strata, according to the algorithm that was previously described.

Moreover, due to the necessity that each point of the sample has to be associated to an in-situ visit or to be photo-interpreted, two additional information were added to the records of the Master data set; these related to an indicator of reachability and to the probability that the predicted land cover can change in the next three years.

It has to be noted that the points associated to a TW were deleted from the Master data set. These areas correspond to what is defined in the water framework directive (Directive 2000/60/EC) and refer to bodies of surface water near river mouths which are partly saline in character because of their proximity to coastal waters but which are substantially influenced by freshwater flows. They also include water surfaces in estuaries (the wide portion of rivers at their mouths subject to the influence of the sea into which the water course flows) and lagoons (water areas cut off from the sea by coastal banks or other forms of relief with, however, certain possible openings). These areas are not part of the NUTS definition and therefore excluded from the LUCAS reference area.

2.2 The estimated Land Cover

The most probable land cover for each point of the LUCAS Master data set was obtained by generalizing a predictive model based on the results of the LUCAS 2015 sample survey.

In particular, it was assumed that it is possible to estimate the land cover by referring to a proper classification model in which the value in 2015 could be derived considering some covariates, like the strata to which the point was classified in 2005, in 2017, the land cover as from Corinne 2012, etc. Once the parameters of the model were estimated, these are applied to all the information in the Master data set, to obtain the predicted probability to observe a given land cover for all its records.

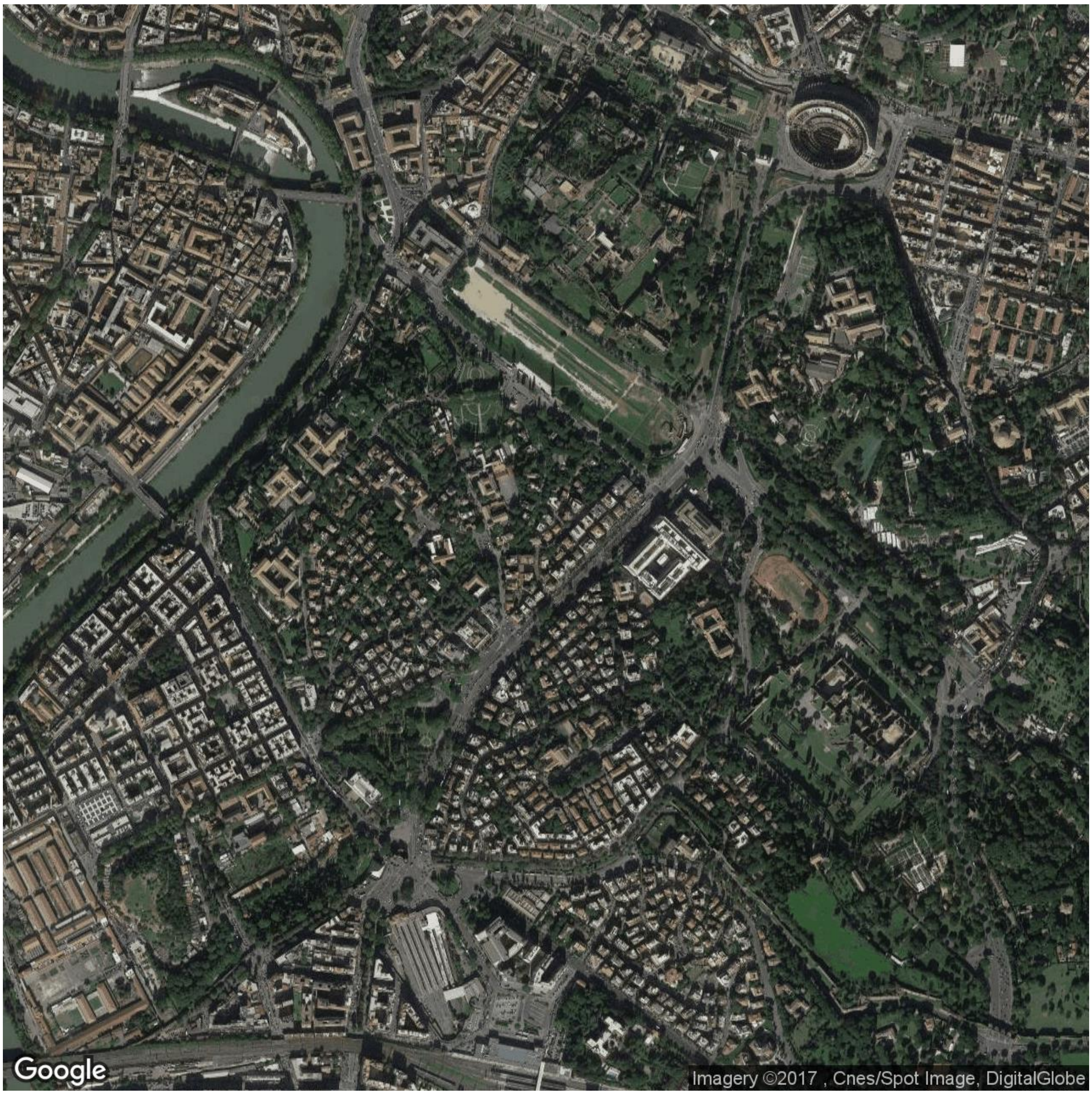
It has to be noted that the land cover can assume different values, considering all the possible bio-physical coverage of land (for example: natural areas, forests, buildings, roads or lakes, etc.). In our case we considered the classification referred to the upper-bound of the expected errors for the next LUCAS survey. This leads to have 16 classes, as in table 1.

Table 1: **Rules used to classify the LUCAS land covers in the typologies referred to the upper-bounds expected errors**

Name of the Recoded LC	Land cover	Original classification of land cover accordingly to the LUCAS standards (two digits)
A	Roofed built-up areas	A1
B	Artificial non-built up areas	A2
C	Cereals	B1
D	Root, non permanent industrial crops, dry pulses, etc.	B2, B3, B4 and B5
E	Permanent crop	B7, B8
F	Broadleaved woodland	C1
G	Coniferous woodland	C2
H	Mixed woodland	C3
I	Shrubland with sparse tree cover	D1
L	Shurbland without tree cover	D2
M	Grassland with sparse tree/shrub cover	E1
N	Grassland without sparse tree/shrub cover	E2
O	Spontaneously re-vegetated surfaces	E3
P	Bare land and lichens/moss	F
Q	Water areas	G
R	Wetlands	H

Concerning the active variables used in the model, we added new information derived by an automatic synthesis of the satellite image centred in each point of the Master data set.

Figure 1: **Image derived from Google Satellite (1)**



2.2.1 Additional information derived from Google Satellite

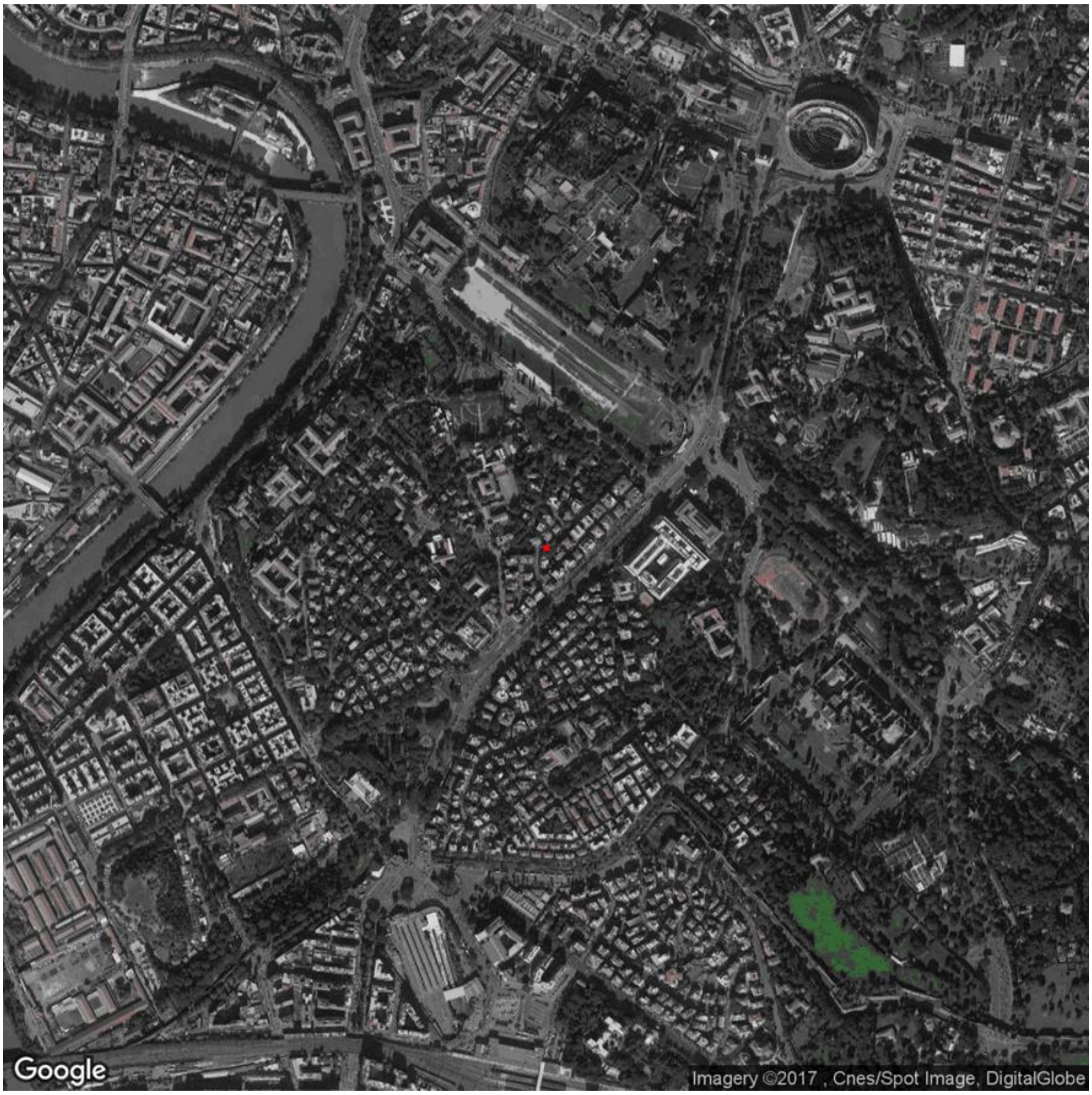
Object of this step is the addition of the characteristics of the main pixels centred in the coordinates of each LUCAS point. For instance, it is possible to consider the point:

- X_WGS84: 12.485160653,
- Y_WGS84: 41.882493321,

The above coordinates are the centre of the image as in figure 1, obtained from Google Maps (satellite view).

A proper elaboration of the image permits to extrapolate the RGB colors of the 7*7 pixels in its center (the red dot in figure 2).

Figure 2: **Image derived from Google Satellite (2)**



From the 49 (7x7) pixels it is possible to obtain the following statistics:

- mean and standard deviations of the values referred to each colour channel (Red, Green and Blue); these referred to the 7*7 square but, also, to the 5*5 and 3*3 sub squares,
- luminance for the 7*7, 5*5 and 3*3 squares, where this value represent the synthesis of all the colour channels and is evaluated by considering: 0.299 * MEAN_RED + 0.587 * MEAN:GREEN + 0.114 * MEAN_BLUE.

Such statistics were evaluated for all the points of the LUCAS master data set by a proper procedure able to download automatically the image (in JPEG format) centred at their GPS coordinates.

2.2.2 The other covariates considered in the model

The following variables were considered as covariates in the model used to estimate the land cover:

- value of the stratification as for 2005 (STR05),
- value of the stratification as for 2005 (STR18),
- CLC classification of each point,
- indicators if the point is on wet area,
- indicators if the point is associated to an artificial land cover or used for industrial/commercial/transport/residential,
- distance buffer (0-2.5 km) from the Corinne Land Cover value,
- distance buffer (2.5-5 km) from the Corinne Land Cover value,
- secondary value of the stratification variable,
- elevation of the nearest road,
- angle referred to the nearest road,
- distance to the nearest road,
- elevation of the point,
- population density in the related 1 km^2 grid cell of the Population Grid 2011 (this variable was considered 0 if null or not valid)-

All the statistics referred to the image as described in the previous paragraph.

2.2.3 Characteristics of the model used to estimate the land cover

The estimates of the land cover(s) are obtained by generalizing to all the records the parameters of a linear logistic regression model in which the real LC (as in the LUCAS 2015) was derived by considering some covariates.

It has to be noted that:

- the statistical approach used to estimate the LC refers to a multi-modal supervised algorithm: this implies to have estimated different models, one for each LC,
- in the estimation of the land cover, the model gives a “score” (between 0-1) referred to each of the possible LCs. Having introduced a threshold for such score, the point could be then associated to one (or more) land cover(s),
- the information derived from the analysis of the Google satellite images entered as covariates in the model. These refer to the characteristics of the main pixels centred in the coordinates of each LUCAS point. In particular:
    - mean and standard deviations of the values referred to each colour channel (Red, Green and Blue); these referred to the 7*7 square but, also, to the 5*5 and 3*3 sub squares,
    - luminance for the 7*7, 5*5 and 3*3 squares, where these values represent the synthesis of the three colour channels; such synthesis is evaluated by considering: 0.299 * MEAN_RED + 0.587 * MEAN:GREE + 0.114 * MEAN_BLUE

2.2.4 Indexes to evaluate the results of the model

The classification capacity of each model was tested by considering in a first step all the records belonging to the LUCAS 2015 survey (for each selected country). For these records, as the land cover is known; it permitted to split the data set in two parts (of almost equal size). The first, called *train*, was used to estimate the parameters of the model, the second, *test*, to verify its classification performance.

In a second step, after having evaluated the capacity of the model, all the records of the selected Country were considered, thus permitting to estimate the parameters of the final linear logistic regression model (for all the records that belong to the LUCAS 2015 survey). Then, the model was applied to all the remaining records, having the score of presenting a given land cover.

It is important to observe that the score of the linear logistic regression was transformed in a specific value of land cover by means of a threshold, able to reproduce the original ratio of points having the considered land cover in the train set.

Moreover, specific indexes were considered to evaluate the capacity of the models; these are based on the confusion matrixes obtained at the end of the estimation of each of the above described steps (*train and test* and *all the records belonging to the 2015 survey*). It has to be noted that each column of the confusion matrix represents the instances in the predicted class while each row represents those of the observed one. In particular:

Table 2: General confusion matrix

		Predicted class		
		0	1	Total
Observed class	0	True Negative (TN)	False Positive (FP)	TN+FP
	1	False Negative (FN)	True Positive (TP)	FN+TP
	Total	TN+FN	FP+TP	N

The indexes that were considered are those usually adopted to evaluate the results of a classification model:

- accuracy: (TN+TP)/N,
- error rate: (FP+FN)/N,
- sensitivity: TP/(TP+FN),
- specificity: TN/(TN+FP),
- ratio of original positive: (FN+TP)/N,
- ratio of predicted positive: (FP+TP)/N.

Except for the *Error rate*, higher values of the indexes show a good discriminant classification.

2.2.5 Results of the model

In table 3 are the confusion matrixes obtained for each land cover; it has to be noted that these results refer to all the European countries and to all the points in the Master data set having an observed value of LC in the 2015 survey.

Table 3: Confusion matrix for each land cover

Land cover	Observed:0, predicted: 0	Observed:0, predicted: 1	Observed:1, predicted: 0	Observed:1, predicted: 1
Roofed built-up areas	332 180	1 879	1 896	2 420
Artificial non-built up areas	326 687	1 964	1 968	7 756
Cereals	274 153	18 003	17 829	28 390
Root, non permanent industrial crops, dry pulses, etc.	304 267	12 692	15 009	6 407
Permanent crop	325 439	466	468	12 002
Broadleaved woodland	272 216	13 241	15 529	37 389
Coniferous woodland	278 177	17 934	17 946	24 318
Mixed woodland	292 469	16 827	16 825	12 254
Shrubland with sparse tree cover	324 939	3 193	5 692	4 551
Shurbland without tree cover	317 183	4 148	4 149	12 895
Grassland with sparse tree/shrub cover	321 608	6 182	6 193	4 392
Grassland without sparse tree/shrub cover	277 751	13 220	14 259	33 145
Spontaneously re-vegetated surfaces	324 319	5 541	5 789	2 726
Bare land and lichens/moss	324 361	3 708	3 813	6 493
Water areas	330 473	4	2	7 896
Wetlands	333 684	368	34	4 289

Concerning the indexes from which it is possible to better evaluate the classification performance of the models, it is possible to consider the table 4.

Table 4: Classification performance

Land cover	Accuracy	Error rate	Sensitivity	Specificity	Original percentage (%)	Percentage from model
Roofed built-up areas	0.989	0.011	0.561	0.994	0.013	0.013
Artificial non-built up areas	0.988	0.012	0.798	0.994	0.029	0.029
Cereals	0.894	0.106	0.614	0.938	0.137	0.137
Root, non permanent industrial crops, dry pulses, etc.	0.918	0.082	0.299	0.960	0.063	0.056
Permanent crop	0.997	0.003	0.62	0.999	0.037	0.037
Broadleaved woodland	0.915	0.085	0.707	0.954	0.156	0.150
Coniferous woodland	0.894	0.106	0.575	0.939	0.125	0.125
Mixed woodland	0.901	0.099	0.421	0.946	0.086	0.086
Shrubland with sparse tree cover	0.974	0.026	0.444	0.990	0.030	0.023
Shurbland without tree cover	0.975	0.025	0.757	0.987	0.050	0.050

Land cover	Accuracy	Error rate	Sensitivity	Specificity	Original percentage (%)	Percentage from model
Grassland with sparse tree/shrub cover	0.963	0.037	0.415	0.981	0.031	0.031
Grassland without sparse tree/shrub cover	0.919	0.081	0.699	0.955	0.140	0.137
Spontaneously re-vege surfaces	0.967	0.033	0.320	0.983	0.025	0.024
Bare land and lichens/moss	0.978	0.022	0.630	0.989	0.030	0.030
Water areas	1.000	0.000	1.000	1.000	0.023	0.023
Wetlands	0.999	0.001	0.992	0.999	0.013	0.014

2.3 The index of reachability

The index of reachability was introduced to represent the difficultness that an enumerator can have in reaching a given point. More precisely, it synthesizes the possibility that the point is far from a road, or on a cliff, etc.; such information should be considered discriminant in selecting the point for an in-situ visit or for a photo-interpretation of its content.

According to the variables in the Master data set, the following ones were considered useful in determining such index:

- the absolute difference in elevation between the altitude of the point and the one referred to the nearest road (ABS_RATIO),
- the distance to the nearest point on a road (NEARDIST),
- the angle to the nearest point in a road (NEARANGLE).

The index, in particular, is obtained by combining these variables with proper coefficients, that were estimated by means of a Principal Component Analysis. Such statistical technique permits to obtain combinations of the active variables that took into account their correlation structure; these combinations, considered as net variables, will be orthogonal (not correlated) between them.

In our case we selected the first component, that resulted able to well represented the index of reachability.

Concerning the relation of this component with the original variables, the Pearson’s correlation index is:

- 0.85 for ABS_RATIO and NEARDIST,
- -0.01 for NEARANGLE.

It is possible to observe that the first combination is positively correlated with all the variables except for the angle to the nearest point in a road. Moreover, this new variable explains almost 90% of the variability of all the variables.

The values of the projections can be further reduced in the interval [0:1] by taking into account the normalization considering their minimum (-0.79) and maximum (17.1) values.

With this procedure we were able to add the index of reachability that assumes higher values for those points resulting more difficult for an in-situ visit.

Moreover, we added special values to this new variable in order to take into account some notes that are able to a-priori identify those points that were previously observed as of difficult access. In particular, the following conditions were considered (and for these the value of the index was imposed as 1):

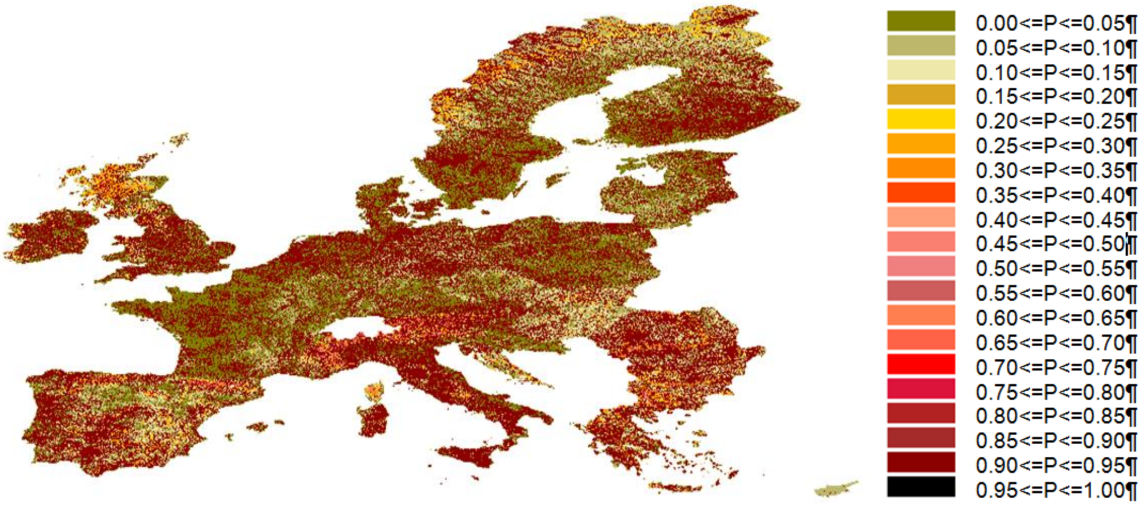
Previously considered as a point to be photo-interpreted (EXANTEPI=TRUE);

Value of the stratification variable specifying that the points should be photo-interpreted;

Points with difficult access comment, or points that landowner refused access or points that landowner refused to collect SOIL data.

The next image represents the distribution of such index for all the points in the master data set.

Figure 3: **Distribution of index**
(Probability)



2.4 The probability of change

Another additional information added to each record of the LUCAS master data set refers to the propensity of change in the estimated land cover. Such propensity depends, also, on the type of land cover that was associated to the point. For instance, it could be considered that the propensity to change for a point associated to an “artificial land” should be less than the one associated to a “Crop” or “Grassland”.

To estimate such variable, a linear logistic regression model was introduced; the dependent variable was obtained by considering the results observed in the LUCAS surveys related to the years 2009, 2012 and 2015. In particular, it was supposed to have the same land cover if:

- the land cover in LUCAS 2015 was the same as the one observed in LUCAS 2012,
- the land cover in LUCAS 2015 was the same as observed in LUCAS 2009 (and the point was not observed in 2012),
- the land cover in LUCAS 2012 was the same as observed in LUCAS 2009 (and the point was not observed in 2015).

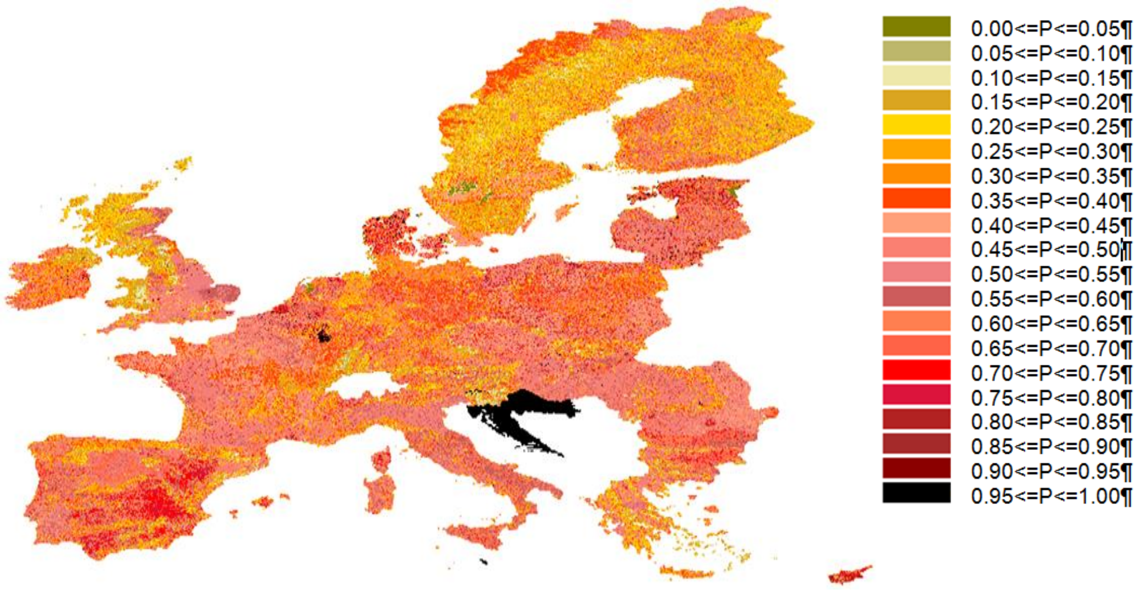
Instead, all the records observed in at least two LUCAS surveys were associated to a change in the land cover if any of the above conditions was not met.

The covariates of the linear logistic regression model were the same of those used when estimating the land cover, except that we did not consider the characteristics of the satellite images, while the estimated land covers entered as independent variables.

It has to be noted that the estimated score is related to the “not change” in land cover; its complement to 1 could be, instead, intended as the probability of change. Moreover, we did not transform such value in a class, leaving it able to represent a score applied to each LUCAS point.

The results of this model are analysed by considering the graphical representation of this probability as distributed in the European countries (next figure).

Figure 4: Probability not to change in the European countries
(Probability)



3 The sampling procedure

3.1 Optimisation of stratified sampling

A sample can be defined as optimal both in terms of its costs (i.e. the number of units to be interviewed) and its accuracy (related to the sampling variance of target estimates).

In order to optimise a stratified sampling design of a given population of interest, its members must be assigned to groups, called *strata*, that should be homogeneous with respect to the target variables, whose estimation is the aim of the survey. Simple random sampling can then be applied within each stratum, having defined the overall *allocation*, i.e. the number of units to be selected in each stratum (Cochran, 1977). The allocation is in general proportional to the variability of target variables in strata (Neyman, 1934).

If the variables chosen to form strata are such to explain the variability of the target estimates of the sampling survey (i.e. variability of the target variables *within* strata is minimised, while their variability *between* strata is maximised), then the representativeness of the sample is increased, and the sampling error of estimates reduced.

Many studies dealing with the problem of stratified sample design optimization have been conducted; a general review of the proposed methods is contained in Gonzales (2010}.

Basically, the optimisation can be conceived as based on:

- an *objective functio*n: it can be defined in terms of minimisation of costs, or maximisation of the estimates precision,
- *constraints*: defined on minimum precision required on target estimates, or on maximum affordable cost given the available budget,
- *parameters*: related to the distribution of target variables in the population,
- *decision variables*: quantities that have to be determined in order to optimise the objective function, in our case, how many population units have to be selected in each stratum.

In general, optimisation of stratified sampling has been considered as a two-step process: first, a stratification is chosen by exploiting all the auxiliary information available on sampling units, or only a subset, selected on the basis of known correlations between target and stratification variables. Then, given the chosen stratification, the problem of allocation is solved (Dalenius and Hodges, 1959).

Well-known solutions in the multivariate case (more than one target variable) are the ones given by Bethel (1985 and 1989) and Chromy (1987). Together with many others, these solutions assume that stratification of population is given.

This assumption can greatly penalise the success of the optimisation process: the way the population is stratified is of the greatest importance as the relationships between the survey target variables and the stratification variables are at the basis of the stratified sampling, and in order to take maximum advantage of these relationships, choices regarding the way we define population strata should enter into the optimisation process together with the allocation choices.

The Lavallée and Hidiroglou method for the stratification of a skewed population (Lavallée,1988) allows the determination of both strata boundaries and best allocation, but only in the univariate case, and having assumed a pre-determined number of strata.

The approach followed in the optimisation process of LUCAS sampling design is based on the joint determination of the optimal stratification of a sampling frame, together with the optimal sample size determination and allocation. This approach is the most general one, as it can operate in the full multivariate case (i.e. with regards to both stratification and target variables), without being obliged to choose the number of strata. Its implementation is based on the use of the genetic algorithm. The general procedure has been implemented in an R package named SamplingStrata, which is available on the CRAN (Barcaroli et al, 2018).

3.2 The approach based on genetic algorithm

We assume that in the population frame (F) a set of M auxiliary variables X_m (m=1,…, M) are available. This set may contain different typologies of variables (nominal, ordinal, or continuous). We also assume that continuous auxiliary variables are split into classes by applying suitable transformation algorithms.

All these variables can potentially be used to stratify the units in the frame.

Under these assumptions, we can assign to each auxiliary variable a vector $d_m = \{x_1,...,x_{k_m}\}$ of contiguous integer values, each of them representing an original value in the domain set.

Then, the most detailed stratification of F can be considered as the result of the Cartesian product $CP = X_1 \times X_2 \times ... \times X_M$.

The maximum number of strata will be $K = \prod_{m=1}^{M} k_m - I^*$, where I^* is the number of impossible or missing combinations of values in the frame. So, the most detailed stratification of the frame is such that it contains K strata, corresponding to all possible combinations of values in the M auxiliary variables. We call atomic strata the strata belonging to this particular stratification. Each atomic stratum is characterised by a unique combination of values of the M auxiliary variables. We can assign a label l_k (k=1,…,K) to each atomic stratum.

If we consider the labelled set of atomic strata $L = \{l_1, l_2,..., l_K\}$, we can define the set of all its possible partitions $P_1, P_2,..., P_B$, where B can be calculated by using the Bell formula:

$$B_K = \sum_{i=0}^{K-1} \binom{K-1}{i} \cdot B_i \qquad (B_0 = 1)$$

We define the set $\{P_1, P_2,..., P_B\}$of partitions of L as the space of stratifications.

The efficiency of each stratification belonging to this space can be evaluated in terms of the size of the sample needed to satisfy a given set of precision constraints on target variables.

Given a partition P_i of L, characterized by H strata, let N_h and $S^2_{h,g}$, h=1,..,H, g=1,...,G be respectively the number of units and variances in stratum h of the G different survey target variables $Y_1,...,Y_G$. Assuming a simple random sampling of n_h units without replacement in each stratum, the variance of the Horvitz-Thompson estimator of the total of the g-th target variable ($\hat{T}_g$) is

$$Var\left(\hat{T}_g\right)=\sum_{h=1}^{H}N_h^2\left(1-\frac{n_h}{N_h}\right)\frac{S_{h,g}^2}{n_h}, \quad \text{g=1,..,G} \qquad (1)$$

Consider the following cost function

$$C(n_1,..,n_H)=C_0+\sum_{h=1}^{H}C_h n_h , \qquad (2)$$

where C_0 indicates a fixed cost (not dependent on the sample size) and C_h represents the average cost of observing a unit in stratum h.

Given V_g (g=1,..,G), the upper bounds for the expected sampling variance for $\hat{T}_1,...,\hat{T}_G$, the classical optimal multivariate allocation problem (Bethel, 1985) can be defined as the search for the solution of the minimum (with respect to n_h) of the linear function C under the convex constraints $Var\left(\hat{T}_g\right)\le V_g$ g=1,..,G:

$$\begin{cases}\min C(n_1,..,n_H)=C_0+\sum_{h=1}^{H}C_h n_h \\ Var(\hat{T}_g)=\sum_{h=1}^{H}N_h^2\left(1-\frac{n_h}{N_h}\right)\frac{S_{h,g}^2}{n_h}\le V_g \qquad g=1,...,G\end{cases} \qquad (3)$$

An algorithm that is proved to converge to the solution (if it exists) was provided by Bethel by applying the Lagrangian multipliers method to this problem.

The optimisation approach here illustrated a continuous solution, which must be rounded to provide integer stratum sample sizes. The implementation we made of the Bethel algorithm provides the n_h values as the values $1/x_h$ rounded up to the upper integer.

It should be noted that the same approach can be used to deal with the multi-domains problem. Let us consider the usual transformation for the domain estimation problem:

$$Y_i^d=\begin{cases}Y_i & \text{if the unit } i \text{ belongs to domain } d \\ 0 & \text{otherwise}\end{cases}$$

If the quantities previously defined to describe the Bethel approach are computed using the variables Y^d (d=1,…,D), then the multivariate allocation solution is also the solution for the multi-domain case.

In order to choose the best stratification of a given frame, i.e. the one that ensures the minimum cost $C(n_1,..,n_H)$ associated with a sample whose total size and allocation are compliant to precision constraints, it is possible to proceed as follows:

1. generate the most detailed stratification associated with F, which is the set L of atomic strata;
2. enumerate all partitions P_i of L;
3. for each partition P_i , solve the corresponding allocation problem, which is equivalent to determining the vector $(n_1,..,n_H)$, and calculating the value $C_i(n_1,..,n_H)$ associated to P_i ;
4. choose the partition P_i for which $C_i(n_1,..,n_H)$ is minimised.

By doing so, the optimisation of the solution is obtained by considering the whole universe of stratifications.

Unfortunately, this procedure is applicable only in situations where the dimension K of L is low: in fact, the number of partitions (given by the Bell formula) grows very rapidly (for example, B_4 = 15, B_{10} = 115,975 and $B_{100} \approx 4.76 \times 10^{115}$). Therefore, in most cases, the complete enumeration of the space of the solutions is not feasible. The approach based on the genetic algorithm, allows us to explore the universe of stratifications and to identify the one that is expected not to be far from the optimal.

Genetic algorithms belong to the class of evolutionary algorithms that make use of techniques based on concepts derived from biology, such as *inheritance*, *mutation*, *crossover*, *fitness* and *selection*. (Dejong, 2006) (Vose, 1999).

The application of genetic algorithm to the specific sample size and allocation problem has already been attempted (Keskinturk and Er, 2007) (Day, 2006 and 2010), but assuming the stratification as given.

In order to apply the genetic algorithm to the problem of jointly finding the best stratification and the best allocation, the following setting has been adopted:

1. a given stratification is considered as an *individual* in a population (or generation of individuals),
2. an individual is characterised by a *genome* that is optimised in the course of the evolution,
3. the genome is represented by a vector whose dimension is given by the number of atomic strata (K),
4. an atomic stratum is assigned to each position in this vector,
5. an integer value lying in the interval (1, K) is assigned randomly to each element in the vector: atomic strata that share the same integer value collapse in an aggregate stratum,
6. the *fitness* of each individual is evaluated by solving the system reported in Equation (3) (using the Bethel algorithm),
7. in the passage from one generation to the next, the fittest individuals are privileged;
8. a percentage of those with the highest fitness are directly moved to the next generation, the others are randomly selected with probability proportional to their fitness, in order to let them procreate children,
9. each child is procreated by applying crossover to their parents (a swap of the genes contained in the two genomes), and applying mutation to the resulting genome.

At the end of the evolution (the chain of generations), the individual with the absolute best fitness will be chosen: the genome of this individual represents the optimal stratification in which all or some of the atomic strata have been aggregated.

It is worth while noting that if we set C_0 = 0, and C_h = 1 for all the atomic strata, then the value of the cost function simply coincides with the sample size required to satisfy precision constraints.

Under this approach, the optimisation of the sampling design starts by considering the available population frame, defining the target estimates of the survey and establishing precision constraints on them. It is then possible to determine the best stratification and the optimal allocation. Finally, the sample can be drawn from the frame stratified to the optimal stratification accordingly.

A more detailed description of the above approach can be found in (Ballin and Barcaroli, 2013) and (Ballin and Barcaroli, 2016).

The R package “SamplingStrata” implements the approach previously described (Barcaroli et al, 2018). Comparable evaluations of the performance of “SamplingStrata” with respect to the R package “stratification” (Baillargeon and Rivest 2012 and 2014) that implements variants of the Lavallée-Hidiroglou approach (Kozak and Wang, 2010) are reported in (Barcaroli, 2014) and (Ballin et al 2016).

This package was used to develop the sampling procedure that has been applied in each EU country.

3.3 LUCAS sampling procedure

The optimisation of the sampling design starts by making the sampling frame available, defining the target estimates of the survey and establishing the precision constraints on them. It is then possible to determine the best stratification and the optimal allocation. Finally, the selection of the sample can be carried out. When formalising the description above, these are the required steps:

1. *analysis of the frame data*: identification of available auxiliary information,
2. *manipulation of auxiliary information*: in case auxiliary variables are of the continuous type, they must be transformed into a categorical form,
3. *construction of atomic strata*: on the basis of the categorical auxiliary variables available in the sampling frame, a set of strata can be constructed by calculating the Cartesian product of the values of all the auxiliary variables and assigning the information on the distributions of the target variables (means and standard deviations) to each stratum;
4. choice of the *precision constraints* for each target estimate, possibly differentiated by domain,
5. *optimisation of stratification* and *determination of required sample size and allocation* in order to satisfy precision constraints on target estimates,
6. *adjustment of the final sampling size*,
7. *selection of units* from the sampling frame with a stratified random sample selection scheme,
8. *evaluation of the found optimal solution* in terms of expected precision.

In the following sections, we will illustrate each step considering one of the countries participating in the LUCAS project, namely Ireland.

3.3.1 Construction of the frame dataset

As a first step, a “frame” dataset is defined and populated by the records belonging to a given country selected in the overall Master dataset, and containing the following information:

1. a unique identifier of the unit,
2. the values of m auxiliary variables (named from X1 to Xm),
3. the values of p target variables (named from Y1 to Yp),
4. the value of the domain of interest for which we want to produce estimates (named 'domainvalue').

The unique identifier is the variable POINT_ID.

The variables STR18, CLC12 and ELEV have been chosen as auxiliary variables because of their strong correlation with the target variables. The first two are already categorical, while the third one is continuous and needed to be transformed into categories by applying the k-means algorithm (Hartigan and Wong, 1979).

The target variables have been chosen among those added in the Master. In particular Y1--Y16 correspond to the added variables "pred_a",…..,"pred_r". It is important to recall that the values of the target variables are not the observed ones (otherwise we would not need a survey), but they have been predicted in previous steps.

NUTS2_13 is the variable indicating the “domain”: optimisation of the frame stratification, which will be carried out domain by domain.

The overall frame dataset will be divided into two different datasets:

- the first one (“framesamp”) is the real sampling frame, where the units will be selected when drawing the sample,
- a second one (“framecens”) will contain the units to be contained in any case in the final sample, without any selection step.

The partition is made on the basis of the values of the variable CENSUS (0: can be sampled; 1: must be selected).

3.3.2 Construction of atomic strata

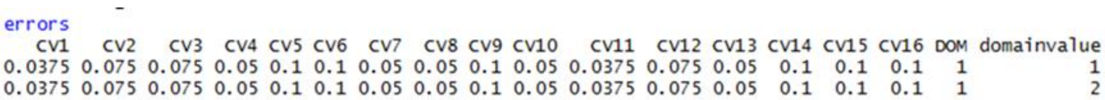
The “strata” dataset reports information regarding each stratum in the population. There is one row for each stratum. The total number of strata is given by the number of different combinations of Xs values in the frame.

For each stratum, the following information is required:

1. the identifier of the stratum (named 'stratum'), concatenation of the values of the X variables,
2. the values of the m auxiliary variables (named from X1 to Xm) corresponding to those in the frame,
3. the total number of units in the population (named 'N'),
4. a flag (named 'cens') indicating if the stratum is to be censused (=1) or sampled (=0),
5. a variable indicating the cost of interviewing per unit in the stratum (named 'cost'),
6. for each target variable y, its mean and standard deviation,
7. the value of the domain of interest to which the stratum belongs ('DOM1').

The “strata” dataset is automatically generated by a specific function in the “SamplingStrata” package (namely, function “buildStrataDF”).

The information contained in this dataset is fundamental for the optimisation step: the higher the variability in a given strata, the higher the sampling rate in that strata is required to contribute to the compliance of the precision constraints. During the optimisation step, the strata with the highest variability will be decomposed or aggregated to other strata in order to try to decrease the overall internal variability of strata.



```
errors
   cv1   cv2   cv3  cv4 cv5 cv6  cv7  cv8 cv9 cv10   cv11  cv12 cv13 cv14 cv15 cv16 DOM domainvalue
0.0375 0.075 0.075 0.05 0.1 0.1 0.05 0.05 0.1 0.05 0.0375 0.075 0.05  0.1  0.1  0.1   1           1
0.0375 0.075 0.075 0.05 0.1 0.1 0.05 0.05 0.1 0.05 0.0375 0.075 0.05  0.1  0.1  0.1   1           2
```

3.3.3 Choice of the precision constraints

A precision constraint on each target variable by each domain can be expressed in terms of maximum expected coefficient of variation. These constraints are organised in a dataset where each row is related to accuracy constraints in a particular subdomain of interest, identified by the DOM1 value. Here is an example where it is supposed that the country has two values of NUTS2 (domain value=1,2).

For instance, the values related to CV1 means that for the first variable (“Roofed built-up areas”), we can accept a maximum coefficient of variation equal to 3.75% in both domains.

It is worth noting that we halved the values of the CVs initially chosen in order to take into account the uncertainty derived from the fact that the values of the Ys are not observed, but predicted.

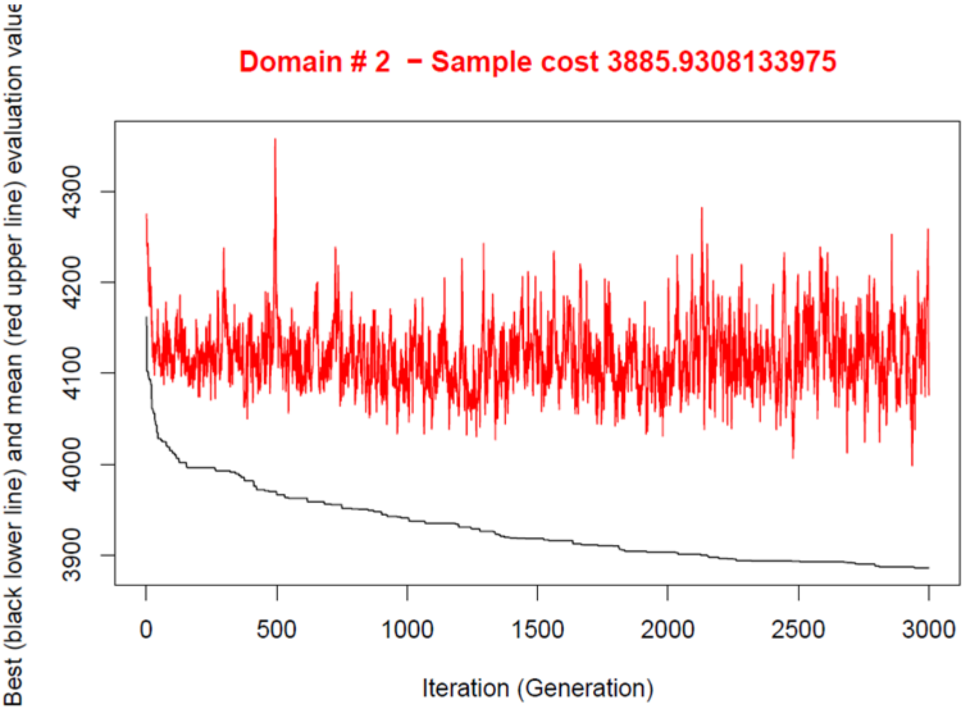
3.3.4 Optimisation of stratification and determination of required sample size

Once the strata and the constraints datasets have been prepared, it is possible to look for the optimal stratification and allocation. In this case, the search has been implemented using an R function called “optimizeStrata”. This function operates on all domains, identifying the best solution for each of them.

The fundamental parameters to be passed to “optimizeStrata” are:

- *errors*: the (mandatory) dataset containing the precision levels expressed in terms of maximum allowable coefficients of variation on the estimates related to the target variables of the survey,
- *strata*: the (mandatory) dataset containing the information related to 'atomic' strata, i.e. the strata obtained by the Cartesian product of all auxiliary variables Xs. Information concerns the identifiability of strata (values of Xs) and variability of Ys (for each Y, mean and standard deviation in strata),
- *cens*: the (optional) dataframe containing the information related to the 'take-all' strata, those strata whose units must be selected in whatever sample. It has same structure as the “strata” dataset,
- *minnumstr*: indicates the minimum number of units that must be allocated in each stratum (two in our case),
- *iter*: indicates the number of iterations (= generations) to be performed by the genetic algorithm. It has been set to 3 000,
- *pops*: the dimension of each generation (=iteration) in terms of individuals (stratifications to be generated and evaluated). It has been set to 20,
- *mut_chance* (mutation chance): for each new individual (possible stratification), the probability to change each single ‘chromosome’, i.e. one bit of the solution vector. High values of this parameter allow a deeper exploration of the solution space, but at the price of a slower convergence, while low values permit a faster convergence, but the final solution can be distant from the optimal one. It has been set to 0.01,
- *elitism_rate*: this parameter indicates the rate of better solutions that must be preserved from one generation to another. It has been set to 0.02.

Figure 5: **Optimisation carried out for the second region of Ireland**



The decision on the number of iterations is very important. A value that is too high implies a long (sometimes very long) time of executions (even days for a single country). A value that is too low may yield a solution that is too far from the optimal. This is the reason why it is important to inspect the graphs produced during the execution.

For instance, figure 5 reports the optimisation carried out for the second region of Ireland.

The black line indicates the best solution found until the i-th iteration. Considering that after the 2 000th iteration, the gain in the saving of units to be sampled in order to guarantee the precision constraint is negligible, we can conclude that 3 000 is an adequate number of iterations.

3.3.5 Adjustment of the final sampling size

After the optimisation step, the final sample size is the result of the allocation of units in optimised strata. This allocation is such that the precision constraints are expected to be satisfied.

Actually, three possible situations may occur:

1. the resulting sample size is acceptable,
2. the resulting sample size is too high, it is not compatible with the available budget,
3. the resulting sample size is too low, the available budget permits an increase in the number of units.

In the first case, no action is required.

In the second case, it is necessary to reduce the number of units, by equally applying the same reduction rate in each stratum.

In the third case, we could either set more tight precision constraints, or proceed to increase the sample size by applying the same increase rate in each stratum.

This increase/reduction process is iterative, as by applying the same rate we could find that in some strata there are not enough units to increase or to reduce. The function “adjustSize” permits us to obtain the desired final sample size.

3.3.6 Frame update and sample selection

Once the optimal stratification and allocation have been obtained, we need to accomplish the following three steps:

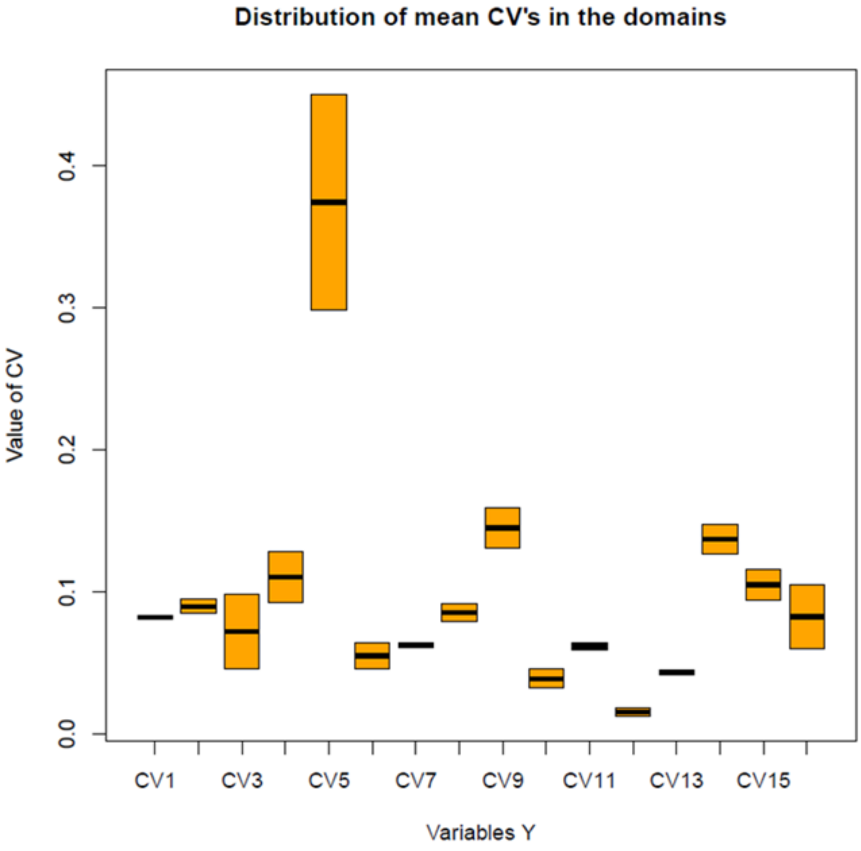
1. to update the frame units with new stratum labels (combination of the new values of the auxiliary variables Xs),
2. to select the sample from the frame,
3. to evaluate the solution.

The third step has been carried out by simulation, selecting 100 samples from the frame to which the stratification identified as the best has been applied.

For each drawn sample, the estimates related to the Y's are calculated. Their mean and standard deviation are also computed, in order to produce the CV related to each variable in every domain. These CVs are stored in an external csv (“expected_cv.csv”), and also plotted (see figure 6).

Analysing this plot, there is a clear problem on expected precision related to the 5th variable, which is “Permanent crop”.

Figure 6: Distribution of expected CVs for each target variable in the different domains



A general analysis of the compliance of expected CVs in the different NUTS2 of each country has been carried out.

We recall that, as the sample size corresponding to the optimised solution does not equalize the predetermined contractual amount of units to be selected in each country, this sample size is adjusted in a second step by proportionally varying the allocation in each stratum (by decreasing or increasing it accordingly to the sign of the difference between adjusted and optimal sample sizes). These differences are reported in the following table (table 5).

The cells highlighted in green are those in which there is an acceptable situation: the optimised sample size is lower or not too much higher than the adjusted sample size.

Table 5: Contractual, optimal and adjusted sample size by country

Country	Points in Master	Contractual Sample size	Optimal sample size	Adjusted sample size	(Adjusted-Optimal) / Optimal (%)
Belgium	7 673	3 659	5 522	3 659	-33.7
Bulgaria	27 731	7 680	13 512	7 680	-43.2
Czech Republic	19 716	5 713	10 069	5 713	-43.3
Denmark	10 771	3 703	6 422	3 703	-42.3
Germany	89 399	26 777	50 196	26 777	-46.7
Estonia	11 322	2 665	2 874	2 665	-7.3
Ireland	17 399	4 975	7 206	4 975	-31.0
Greece	32 817	12 622	13 388	12 622	-5.7
Spain	124 543	45 314	40 016	45 314	13.2
France	137 047	48 215	61 786	48 215	-22.0
Croatia	14 141	4 239	6 835	4 239	-38.0
Italy	75 034	28 294	36 338	28 294	-22.1
Latvia	16 135	5 376	2 695	5 376	99.5
Lithuania	16 234	4 584	4 685.7	4 584	-2.2
Luxembourg	644	340	463	340	-26.6
Hungary	23 267	5 513	11 824	5 513	-53.4
Netherlands	8 882	5 011	5 837	5 011	-14.2
Austria	20 982	8 840	8 509	8 840	3.9
Poland	77 964	23 086	32 265	23 086	-28.4
Portugal	22 144	7 168	9 377	7 168	-23.6
Romania	59 558	16 723	16 828	16 723	-0.6
Slovenia	5 064	1 923	2 252	1 923	-14.6
Slovakia	12 265	2 898	5 711	2 898	-49.3
Finland	84 316	16 182	9 279	16 182	74.4
Sweden	112 385	26 709	20 197	26 709	32.2
United Kingdom	61 038	17 253	36 260	17 253	-52.4

Note: Contractual size: the size fixed before the running of the procedure to assign the batches;
Optimal size: the size calculated by the procedure on the basis of CVs of estimated target variables;
Adjusted size: the calculated size normalised with the contractual ones;

In some situations, the negative differences have caused an increase of the CVs in some countries, as reported in table 6.

The highlighted cells are those where expected values exceed the precision constraints of the amount indicated in the cell. Note that these deviations depend only on the fact that the contractual sample sizes are below what was necessary to ensure the compliance with precision constraints at NUTS2 level. In general, constraints at country level should be respected.

Table 6: Expected coefficient of variations (CVs) for the estimated target land cover by country

	Roofed built-up areas	Artificial non-built up areas	Cereals	Root, non permanent industrial crops, dry pulses, etc.	Permanent crop	Broadleaved woodland	Coniferous woodland	Mixed woodland	Shrubland with sparse tree cover	Shrubland without tree cover	Grassland with sparse tree/shrub cover	Grassland without sparse tree/shrub cover	Spontaneously re-vegetated surfaces	Bare land and lichens/moss	Water areas	Wetlands
Belgium	0.085	0.057	0.009	0.047	0.094	-0.022	0.252	0.102	0.236	0.197	0.083	0.037	0.074	0.058	0.147	0.163
Bulgaria	0.143	0.042	-0.015	0.065	0.029	-0.078	0.156	0.064	-0.015	0.062	0.019	0.018	0.040	0.067	0.041	0.120
Czech Republic	0.093	0.030	-0.020	0.039	0.098	-0.009	0.013	0.048	0.074	0.198	0.079	0.021	0.045	0.136	0.029	0.283
Denmark	0.092	0.041	-0.028	0.037	0.096	-0.010	0.044	0.118	0.119	0.314	0.089	0.011	0.059	0.131	0.042	0.059
Germany	0.084	0.047	0.019	0.408	0.121	-0.007	0.082	0.084	0.191	0.288	0.099	0.010	0.074	0.143	0.056	0.236
Estonia	0.013	0.004	-0.029	0.006	0.054	-0.051	0.000	-0.018	-0.028	0.010	0.004	-0.012	0.001	0.016	-0.025	-0.015
Ireland	0.045	0.015	-0.003	0.060	0.275	-0.045	0.013	0.036	0.046	-0.011	0.025	-0.060	-0.007	0.038	0.005	-0.018
Greece	0.055	0.005	0.030	0.044	-0.021	-0.058	0.006	0.089	-0.008	0.004	0.004	0.003	0.008	0.004	0.012	0.046
Spain	-0.004	-0.005	-0.040	-0.009	-0.024	-0.052	-0.002	-0.002	-0.004	-0.009	-0.004	-0.019	-0.003	-0.015	-0.018	-0.026
France	0.037	0.012	-0.032	0.021	0.020	-0.063	0.063	0.126	0.048	0.099	0.017	-0.021	-0.005	0.022	0.029	0.061
Croatia	0.037	0.030	0.020	0.052	0.024	-0.087	0.028	0.015	-0.041	0.045	0.019	-0.006	0.033	0.075	0.055	0.055
Italy	0.065	0.024	0.059	0.056	0.020	-0.053	0.062	0.043	0.043	0.026	0.024	-0.002	0.043	0.038	0.041	0.208
Latvia	-0.025	-0.032	-0.054	-0.030	-0.084	-0.079	-0.016	-0.025	-0.033	-0.017	-0.015	-0.030	-0.037	-0.088	-0.034	-0.033
Lithuania	0.011	-0.006	-0.055	-0.019	0.022	-0.059	-0.002	-0.009	-0.010	0.010	0.008	-0.041	-0.029	0.001	-0.014	-0.015
Luxembourg	0.060	0.033	-0.003	0.059	0.154	-0.050	0.052	0.048	0.046	0.073	0.060	-0.008	0.040	0.078	0.262	-0.100
Hungary	0.110	0.062	-0.030	0.049	0.082	-0.051	0.160	0.188	0.059	0.126	0.053	0.025	0.029	0.133	0.062	0.110
Netherlands	0.064	0.002	-0.009	0.063	0.180	-0.008	0.115	0.073	0.096	0.045	0.204	0.005	0.076	0.107	-0.006	-0.033
Austria	0.010	-0.009	-0.013	0.018	-0.009	-0.022	-0.014	0.003	-0.002	0.036	-0.002	-0.017	-0.002	-0.004	-0.026	-0.002

	Roofed built-up areas	Artificial non-built up areas	Cereals	Root, non permanent industrial crops, dry pulses, etc.	Permanent crop	Broadleaved woodland	Coniferous woodland	Mixed woodland	Shrubland with sparse tree cover	Shrubland without tree cover	Grassland with sparse tree/shrub cover	Grassland without sparse tree/shrub cover	Spontaneously re-vegetated surfaces	Bare land and lichens/moss	Water areas	Wetlands
Poland	0.060	0.026	-0.044	0.023	0.049	-0.036	0.000	0.016	-0.005	0.046	0.026	-0.014	0.006	0.045	0.025	0.046
Portugal	0.038	0.025	0.058	0.049	0.004	-0.054	0.028	0.069	0.020	0.007	0.014	-0.009	-0.050	0.039	0.030	0.064
Romania	0.010	-0.001	-0.040	0.053	-0.003	-0.072	0.008	0.046	-0.021	0.017	0.016	-0.027	-0.005	0.037	-0.015	0.016
Slovenia	0.060	0.018	0.019	0.114	0.024	-0.042	0.013	0.006	0.041	0.073	0.012	0.003	0.060	0.028	-0.010	0.074
Slovakia	0.156	0.068	-0.004	0.043	0.088	-0.061	0.128	0.097	0.093	0.140	0.101	0.053	0.043	0.271	0.080	0.229
Finland	-0.030	-0.037	-0.054	-0.031	-0.100	-0.057	-0.021	-0.023	-0.052	-0.035	-0.024	-0.038	-0.031	-0.045	-0.062	-0.050
Sweden	-0.026	-0.015	-0.042	-0.027	-0.100	-0.019	-0.023	-0.012	-0.020	-0.024	-0.010	-0.039	-0.020	-0.020	-0.048	-0.028
United Kingdom	0.127	0.085	0.112	0.168	0.302	0.022	0.271	0.211	0.225	0.167	0.118	0.016	0.175	0.194	0.227	0.263

3.3.7 Attribution of the enumeration mode

Once the sample for a given country has been drawn, it is necessary to assign to each selected point the enumeration mode, that is, the indication that it has to be observed directly in the field, or by photo-interpretation (PI).

The balance between the total number of field and PI observations is determined a priori for each country. For instance, in Ireland, a total number of points to be enumerated has been fixed at 4 975. Of these, 3 427 have to be observed in the field and 1 548 by photo-interpretation.

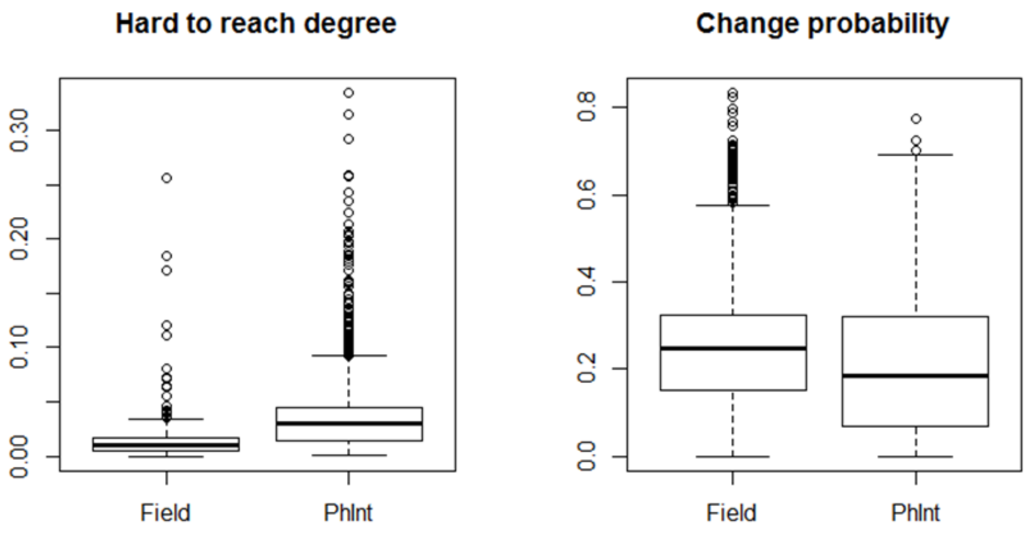
The idea is to assign the field mode to those points that are relatively easy to reach, and with a higher probability of change. Conversely, photo-interpreted points will be those that are difficult to reach and with a lower probability of change.

To obtain this, the following steps are carried out:

1. sampled units are ordered by increasing reachability, and 70% of the total quota of PI mode is assigned to the first units in the ordered list,
2. remaining sampling units are ordered by increasing probability of change, and the remaining 30% of the total quota of PI mode is assigned to the first units in this other ordered list.

As a result, the distributions of reachability and probability of change in the two subsets of sampled units are differentiated (more for reachability, less for probability of change), as illustrated in figure 7.

Figure 7: Distributions of reachability and probability of change in the two subsets of units in the Ireland sample (values equal to 1 have been excluded from the graphs)



4 Analysis of resulting sample and comparison to previous sample designs

4 000 000 points obtained using the grid resulting from the INSPIRE (INfrastructure for SPatial InfoRmation in Europe) recommendations. From the frame is selected the LUCAS first phase sample, the Master, of more than 1 000 000 points sized, by a systematic drawing. Each point of the master sample has been photo-interpreted and hence classified by a variable used to stratify the second phase sample "the stratification variable".

This scheme in 2018 is essentially the same as the previous survey but some changes have been implemented in the Master, mainly in the second phase sample. In this chapter are described the main differences between 2015 and 2018 sample designs of LUCAS survey regarding:

- the new stratification variable in Master 2018,
- eligibility and photo interpretation,
- the use of photo interpretation,
- the stratification of the second phase sample,
- the calculation of the sample size and the allocation of the sampling units.

4.1 The new stratification variable in Master 2018

The variable for stratification has been updated: the 7 modalities in 2015 survey points were reclassified in 10 modalities in 2018 Master as in the following Table 7; the modality "wooded area" and "shrub land" were divided in two ones while the modalities "transitional water" and "impossible to PI" were completely new ones.

Table 7: Classification of "stratum" variable in 2018 and in 2015 surveys

2018	2015
1-Arable land	1-Arable land
2-Permanent crops	2-Permanent crops
3-Grassland	3-Grassland
4-Wooded	4-Wooded areas and shrub land
5-Shrub land	
6-Bare land	5-Bare land, low or rare vegetation
7- Artificial	6-Artificial land
8 – Inland water	7-Water
9 – Transitional water	
10 – Impossible to PI	

The stratum variable adopted in 2015 was collected in the year 2005 (STR05). In 2018 survey, the classification was modified and updated (STR18); so part of the points were classified differently because they actually changed their characteristics. Moreover, in every LUCAS survey the points are codified by a second code for STR05 or Str18 in case of uncertainty in classifying them with the first one.

In the following Table 8 are reported some indicators on the reclassifying process; they are calculated by considering the “cleaned” Master, used as final sampling frame, and, obviously, taking back STR18 to STR05. Moreover, this table, similarly to other tables and graphs, is calculated without considering Cyprus and Malta data because for both countries all the points will be surveyed and, hence, are outside of sample design considerations.

In 2018 survey, with respect to 2015 one, the points have been reclassified (see the column ratio) increasing “arable land” (9.2%), “permanent crops” (6.2%), “wooded areas” (6.0%) and “water” (4.2%); conversely the remaining modalities decreased, “grassland” (-26.4%), “wooded area and shrub land” (-45.5%), “artificial” (-6.3%). The percentage of changes is about 26% in total but it greatly varies among the modalities, from 55.6% for “bare land” to 19.1% for “wooded area and shrub land”. To check the uncertainty of the classification in the two years, the second information, used for codifying the points, has been considered. In 2015 the percentage of the points classified by a second code was 6.26% while in 2018 the percentage is reduced to 0.3%. This difference shows that the uncertainty of the classification is substantially reduced in the 2018 classification operations. All the second codes in STR18 are combined with the first one “arable land” and in most cases they are classified as “wooded” and “shrub land” (0.24%).

Table 8: **Distribution of stratum variable in 2015 and 2018 surveys - cleaned Master**

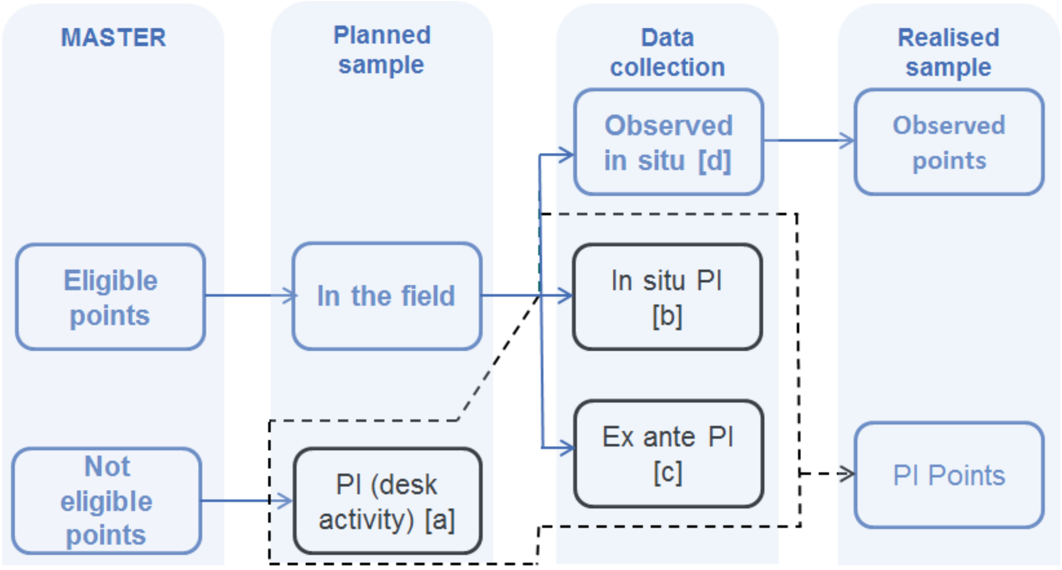
Codes		First classification code					Second classification code			
		Number of points		Ratio (%)	Number of changes	% changes (%)	Number of points		Percentage over total points (%)	
2015	2018	N2015	N2018	(N2018 –N2015) / N2015 *100	STR18≠STR05	Changes/ N2018	2015	2018	2015	2018
1	1	294 230	321 193	9.2	86 054	26.8	10 972	0	1.01	0
2	2	29 486	31 300	6.2	11 440	36.5	3 314	125	0.30	..
3	3	177 219	130 400	-26.4	67 770	52.0	23 043	1	2.11	..
4	4+5	493 260	514 953	6.0	86 851	19.1	22 087	2 664	2.02	0.24
5	6	21 693	11 815	-45.5	6 564	55.6	3 662	0	0.34	0
6	7	44 544	41 722	-6.3	16 763	40.2	3 549	446	0.33	0.04
7	8+9	30 431	31 706	4.2	6 200	19.6	1 608	0	0.15	0
Total		1 090 863	1 083 089	--	281 642	25.8	68 235	3 236	6.26	0.30

4.2 Eligibility and photo interpretation

LUCAS survey can be considered as a multimode survey: information on territory's characteristics is collected both directly by surveyors and by photo interpretation, where directly means "to see and to codify" and the photo interpretation can be carried out ex ante in the office or by the surveyors in the field.

In 2015 survey, photo interpretation is strictly related with the concept of eligibility as the following figure 8 shows.

Figure 8: Eligibility and PI in survey 2015



[a] Points that are part of the sample of points to be photo-interpreted in the office, covering areas excluded from field survey (not eligible points): the points difficult to be reached or those that was not surveyed because in the past surveys the access was denied by land owners or because of emergency issues (such as land mines in Croatia, military areas, UN buffer area in Cyprus, etc..).

[b] Points that had to be assessed in the field and therefore were approached by a surveyor but that were not visible in the field (e.g. hidden by a high wall delimiting the property) and therefore had to be photo-interpreted in the field.

[c] Points that had to be assessed in the field but were identified as impossible to be reached (e.g. military area) during the planning of the survey and therefore in agreement with ESTAT were not approached by a surveyor and were directly photo-interpreted in the office ex-ante, without trying to approach it in the field.

[d] Points directly surveyed by surveyors.

In previous LUCAS surveys eligibility is used with the meaning of "points on which it is possible to directly collect information by surveyors"; so the opposite to eligibility, not-eligible points, are unattainable units or units that are too costly to be reached. The not eligibility concept, translated into opportune indicators, was used to divide in two parts the sampling frame before proceeding to the point selection. The eligible points are selected by a systematic sample design and their characteristics are collected directly or where impossible, by photo interpretation by the surveyors; part of the not eligible points are collected by photo interpretation ex ante in the office, part are excluded from the selection and are considered missing responses (e.g. the points with an elevation greater than a threshold varying in the different rounds). In the estimation step, because the estimates are given at level of the whole NUTS area, all the collected points are put together and a post-stratification by class of elevation, NUTS2 and stratum variable is adopted in order to treat the missing units and to limit the risk of biases.

In 2018 survey all the points are considered suitable to be selected and surveyed by a twofold modes: directly by surveyors and by photo interpretation. The perspective has moved from the concept of eligibility in Master to the mode of data collection after the sample selection; photo interpretation is needed if 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. Therefore no partition of the Master is done but for each point of the master a reachability and a propensity to change indexes are calculated; these indexes are used, after the sample selection, to assign the points as PI ex ante or "in field", given the constraint of PI quotas in each country fixed by the contract.

4.3 The use of photo interpretation

In both surveys (2015 and 2018) photo interpretation is used as a reasonable method (i) to deal with the missing units found in field work and (ii) to take into account unattainable units or units that are too costly to be reached.

In this second case the criteria used in 2015 survey are, substantially, the same considered in building up the reachability index in 2018 survey but a difference can be found in their application. In 2015 survey the criteria are applied in the same way in all the country; in 2018 survey the same threshold cannot be applied because the number of PI points depends on the country batches fixed ex ante. So the threshold is different in different countries and depends on the distribution of index.

The PI could produce underestimate when estimating the changes of land cover characteristics, because the available photos are kept at a preceding year. To avoid or to reduce the risk of biases, the use of PI should be limited to points unchangeable or with a very low probability to change during the time.

In 2018 survey a "propensity to change index" has been calculated and it is used in selecting the set of PI points whose number has been fixed by contract in every country. Given this constraint, 70% of these points are obtained by considering those with more difficult access, the remaining those with the lower probability of change.

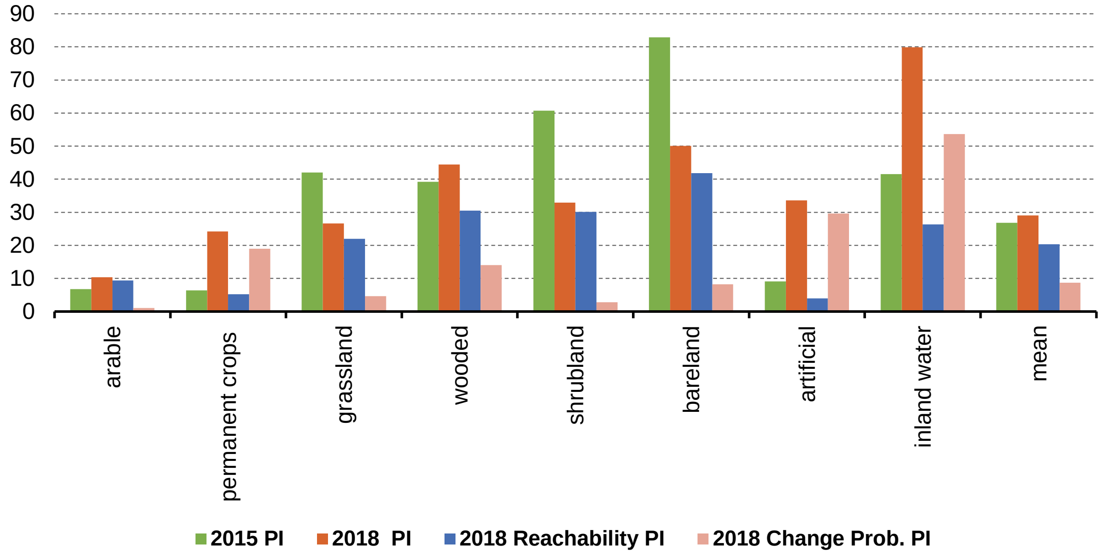
It is to be noted that the probability, or propensity, to change has been calculated for land cover variable because the sample design has been required to take into account, as target variables, some modalities of land cover and related desired errors. But the estimates are required also for other collected information, e.g. land use; so in some cases, points with a low probability to change in land cover could have an high propensity to change in land use.

The two different approaches in using photo interpretation followed in 2015 and 2018 surveys, lead to different distributions of PI points in the two samples. In the following graphs the percentages' ratios of PI points over the totals of two structural characteristics, stratum variable STR18 and elevation, are showed; because in 2018 survey the PI points are assigned for two different reasons: the probability of change (PI_prob) and the reachability (PI_reach). These features are reported as well.

Figure 9 reports the ratios by STR18; in average the percentage of PI in the two samples are similar (about 29% and 27% respectively) but the modalities of STR18 show different figures.

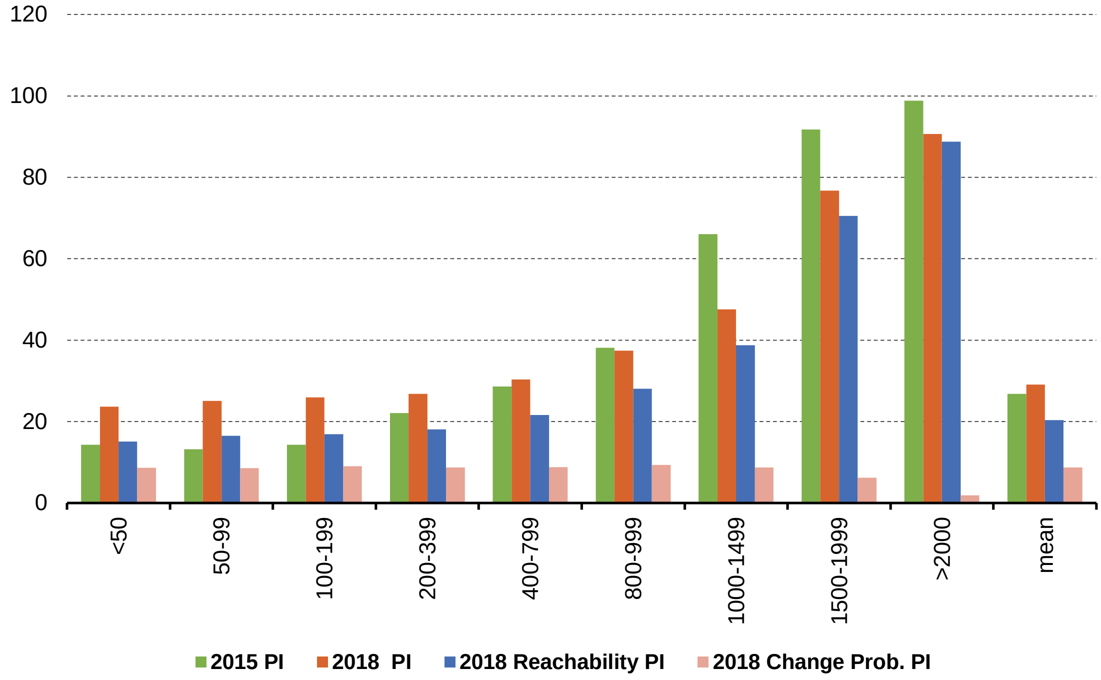
The STR18 modalities can be divided in two groups, according to the greater differences showed by the graph; in the first group - *permanent crops, artificial and inland water* - the percentage of PI in sample 2018 are higher than in the 2015 one, while in the second group - *grassland, shrubland and bareland* - the vice versa holds. The figures in the two groups are correlated to the different composition of PI_prob and PI_reach; in the first group the PI_Prob are much more than the PI_Reach while the contrary occurs in the second group.

Figure 9: **Percentage ratios of PI points over totals by STR18 - 2015 and 2018 samples**
(Percentage)



In figure 10 the same analysis is carried out for the variable class elevation. The ratios in 2018 sample are higher than the ones in 2015 survey up to 800 meters of elevation then the trend inverts; below this threshold the ratios are substantially steady in both the samples while above it they strongly increase. This pattern for 2018 depends on the PI assigned because the reachability index while the ratios for the PI assigned for the probability to change are slightly decreasing.

Figure 10: **Percentage ratios of PI points over totals by class of elevation - 2015 and 2018 samples**
(Percentage)



4.4 The stratification of the sample

In 2015 survey the sample was stratified considering, in each country, the variables NUTS2 and the stratum variable STR05; so the number of strata is fixed ex ante and it was given by the Cartesian product (combinations) of the number of NUTS2 by all the available modalities of STR05 in each region. In this schema, furthermore, the regions (NUTS2) are the minimum territorial study domain.

In 2018 survey the strata are identified by an iterative optimization algorithm that, starting from the "atomic strata", aggregates them considering the coefficient of variations of the target variables and the related desired sampling errors (given by Eurostat). The optimization is carried out distinctly for each value of NUTS2 domain, and then aggregating the results at country level. For each NUTS2 value, the atomic strata are given by the Cartesian product of STR18, CLC and ELEV classifications. As ELEV is a continuous variable, a preliminary step of categorization has been performed, utilizing the K-means algorithm to produce four distinct classes for this variable. The coefficients of variation are related to the estimates of the 16 target variables, whose values have been previously predicted for each point of the master by a logistic model. The iterative algorithm optimizes the stratification, aggregating the atomic strata with the aim of minimizing the overall sample size required to fulfil the precision constraints (the CVs of the target variables). So the stratification is not produced by a fixed combination of variables but it depends on the most correlated combinations of modalities of the stratification characteristics with the target variables; the stratification "criteria" vary according to the specificity of the country and of the NUTS2 territories, that are assumed to be, as in 2015 survey, the minimum territorial study domain.

In the following Table 9 a comparison, by country, among the actual number of strata and the one obtained only considering the combinations of NUTS2 and STR18 (instead of STR05 as in 2015, that is using the same 2015 criteria) is given.

Table 9: Number of strata according to the 2018 actual stratification and the hypothetical one obtained using the 2015 criteria

Country	Strata number in		ratio : (a)/(b)
	Actual stratification (a)	Hypothetical stratification (b)	
Belgium	400	82	4.9
Bulgaria	586	48	12.2
Czech Republic	526	63	8.3
Denmark	406	40	10.2
Germany	1 930	294	6.6
Estonia	147	8	18.4
Ireland	251	16	15.7
Greece	1 374	104	13.2
Spain	2 172	128	17.0
France	1 920	176	10.9
Croatia	236	16	14.8
Italy	1 967	167	11.8
Latvia	171	8	21.4
Lithuania	189	8	23.6
Luxembourg	55	8	6.9
Hungary	568	55	10.3
Netherlands	418	89	4.7
Austria	558	70	8.0
Poland	1 279	128	10.0
Portugal	725	40	18.1

Romania	841	64	13.1
Slovenia	152	15	10.1
Slovakia	284	32	8.9
Finland	372	39	9.5
Sweden	657	62	10.6
United Kingdom	1 776	284	6.3
Total	19 962	2 060	9.7

Note: Cyprus and Malta are not reported because they are entirely collected

As the table shows, the number of strata in the actual stratification is higher (about 10 times) than the number of the hypothetical one (implemented by the same criteria of 2015 survey).

Given the same sample size (and related percentages of PI and direct data collection) and the way the PI are chosen, the strata are in average smaller and the sample units are much more spread and mixed (PI and direct data collection) over the countries. In the following figures 2 and 3 is reported an example (Italy) of the distribution of samples according to the data collection modes.

Figure 11: Distribution of sample units in 2015 survey – Italy

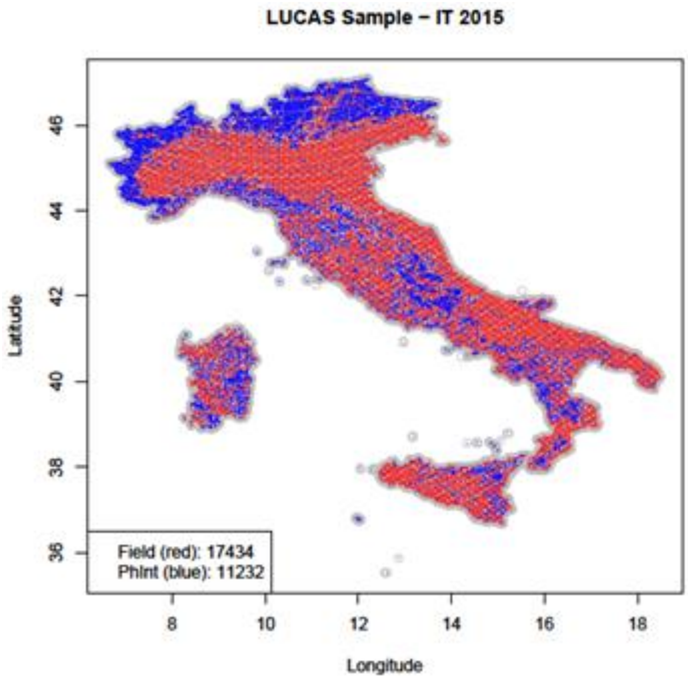
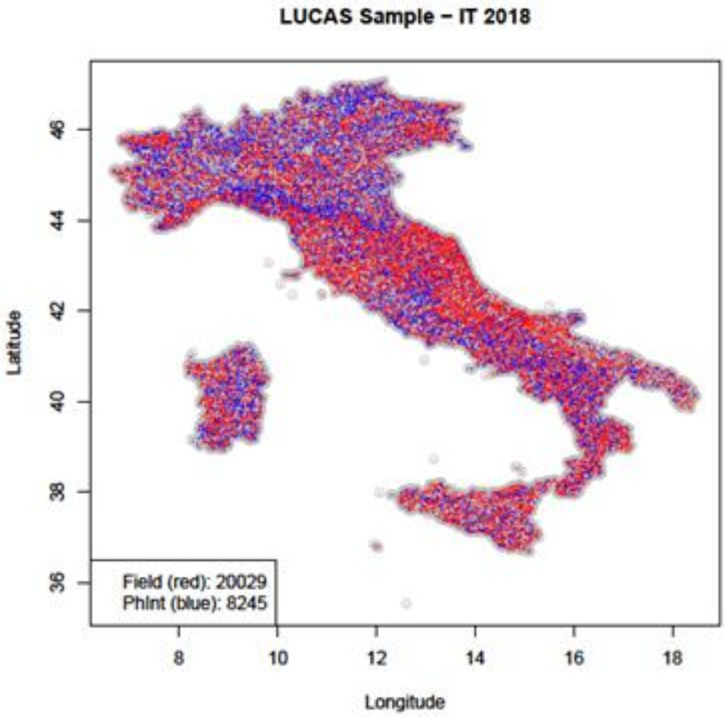


Figure 12: Distribution of sample units in 2018 survey – Italy



4.5 The calculation of the sample size and the allocation of the sampling units

In 2015 survey the sample size was calculated according to the requested precision at level NUTS1 for the more important modalities of land cover (table 10) on the basis of the previous survey results. Once the sample size was fixed, in every country the allocation of the sample units in the strata (identified as the combinations of regions by the STR05 variable available in the Master) was more or less proportional to the strata population because the points were taken by a systematic selection (excluding strata related to smaller subpopulations).

In 2018 survey, given the budget and the timing of the contractual steps to assign the batches, a calculation was made in order to confirm the 2015 sizes in terms of direct and PI data collection; these data states the constraints to be respected in the final allocation of sampling units by country.

Table 10: 2015 Survey-requested expected accuracy (relative error) by different land cover modalities at territorial level NUTS1

Land cover class	Relative error	Land cover class	Relative error
A	0.15	**C**	0.15
B	0.15	**C1**	0.2
B1	0.15	**C2**	0.2
B2	0.25	**C3**	0.2
B3	0.25	**D**	0.02
B5	0.25	**E**	0.075
B7	0.25	**F**	0.2
		G	0.2

New sampling precisions has been successively given by Eurostat (table 11) as ”desired” target precision at region level; they are used, by the same procedure for optimizing the stratification, to calculate the desired sample size, according to the desired accuracy, and to allocate the number of units in the strata of each country. According to the sample size in every strata, the sampling units are selected from the corresponding population in the strata by a simple random selection procedure.

Table 11: 2018 survey-desired accuracy (relative error) by different land cover modalities at territorial level NUTS1/NUTS2

Land cover code	Land cover	Relative error (%)
A10	ROOFED BUILT-UP AREAS	15
A20	ARTIFICIAL NON-BUILT UP AREAS	15
B10	CEREALS	15
B2-B5	ROOT, NON-PERMANENT INDUSTRIAL CROPS, DRY PULSES, VEGETABLES AND FLOWERS, FODDER CROPS	20
B7-B8	PERMANENT CROP	20
C10	BROADLEAVED WOODLAND	20
C20	CONIFEROUS WOODLAND	20
C30	MIXED WOODLAND	20
D10	SHRUB LAND WITH SPARSE TREE COVER	20
D20	SHRUB LAND WITHOUT TREE COVER	20
E10	GRASSLAND WITH SPARSE TREE/SHRUB COVER	15
E20	GRASSLAND WITHOUT TREE/SHRUB COVER	15
E30	SPONTANEOUSLY RE-VEGETATED SURFACES	20
F00	BARE LAND AND LICHENS/MOSS	20
G00	WATER AREAS	20
H00	WETLANDS	20

The land cover modalities (the same reported in table 11, on which the procedure calculates the coefficients of variations in the optimization procedure) were estimated for each point in the Master from the previous LUCAS surveys by a statistical model. Therefore their values are not the observed ones but the predicted values, the sampling errors and the generated sample size are hypothetical ones given under the condition that the statistical model is adequate.

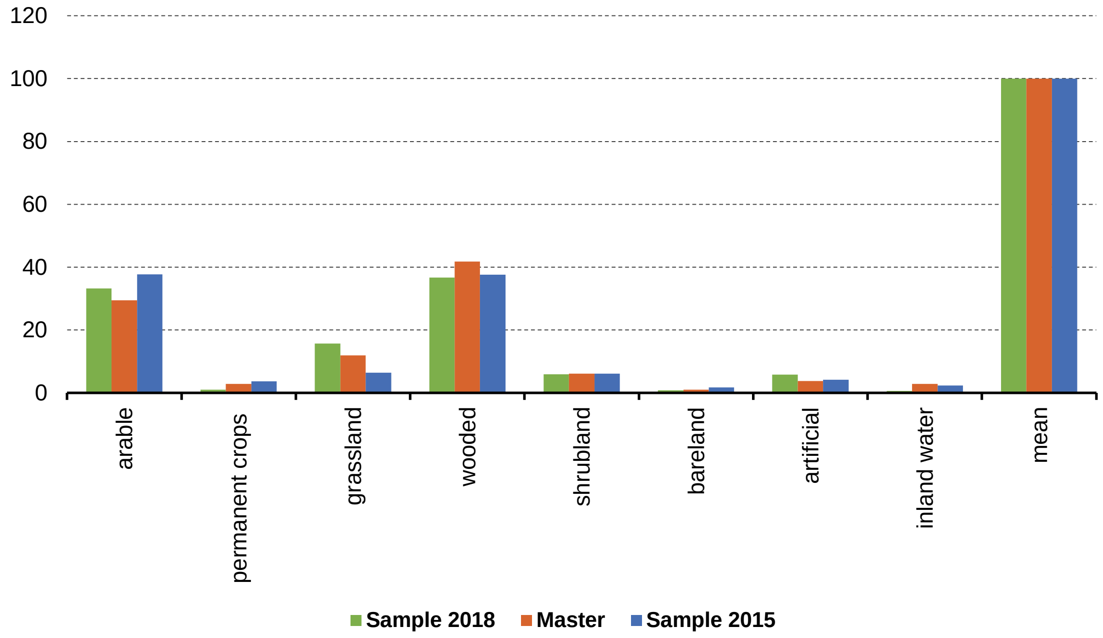
Summarising, while the sample size at level of country has been fixed with the same procedure in 2015 and 2018 surveys, the allocation of the units in strata is quite different in the two LUCAS occasions; in 2015 survey the units have been allocated more or less proportionally while in 2018 the allocation is done by an optimisation algorithm. Moreover, the selection procedure in 2015 is systematic while in 2018 survey the units are selected by a simple random sample (SRS) procedure.

Nevertheless in both surveys the level of CVs, calculated for observed variables in 2015 survey and for the predicted target variables in 2018 sample design, and hence the related sample size, are not sufficient to guarantee the desired precision at territorial level NUTS2; as available the actual data for 2018 survey, should be advisable.

The effects of the different techniques of sample allocation are described by the following graphs that compare the distributions of the 2015 and 2018 sample units with the point of Master with regard to the main characteristics. Both the collected points by field operations and by photo-interpretation in the office ex ante are considered here. The differences depend on various factors: stratification criteria, the allocation procedure used, the estimated variability of the target variables, the use of PI points etc.

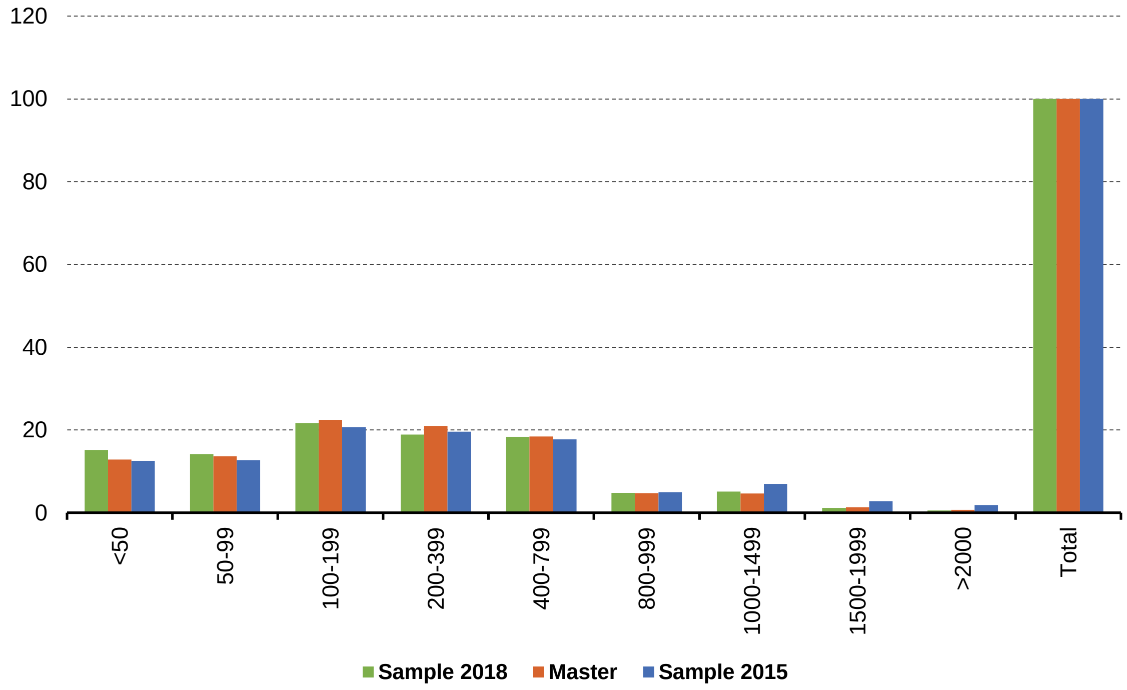
In figure 13 are reported the percentage distributions of points in the 2015 and 2018 samples and in the Master by the stratum variable STR18. For “arable land” and “artificial” the two samples are over represented with respect to the Master while for “wooded area” and “inland water” the percentage in the Master is higher than in the two samples. For “permanent crops” and “bareland” the points in 2015 sample are more than in Master and in Sample 2018 while the vice versa holds for “grassland”.

Figure 13: Percentage distributions of points in Master and in 2015/2018 samples by STR18
(Percentage)



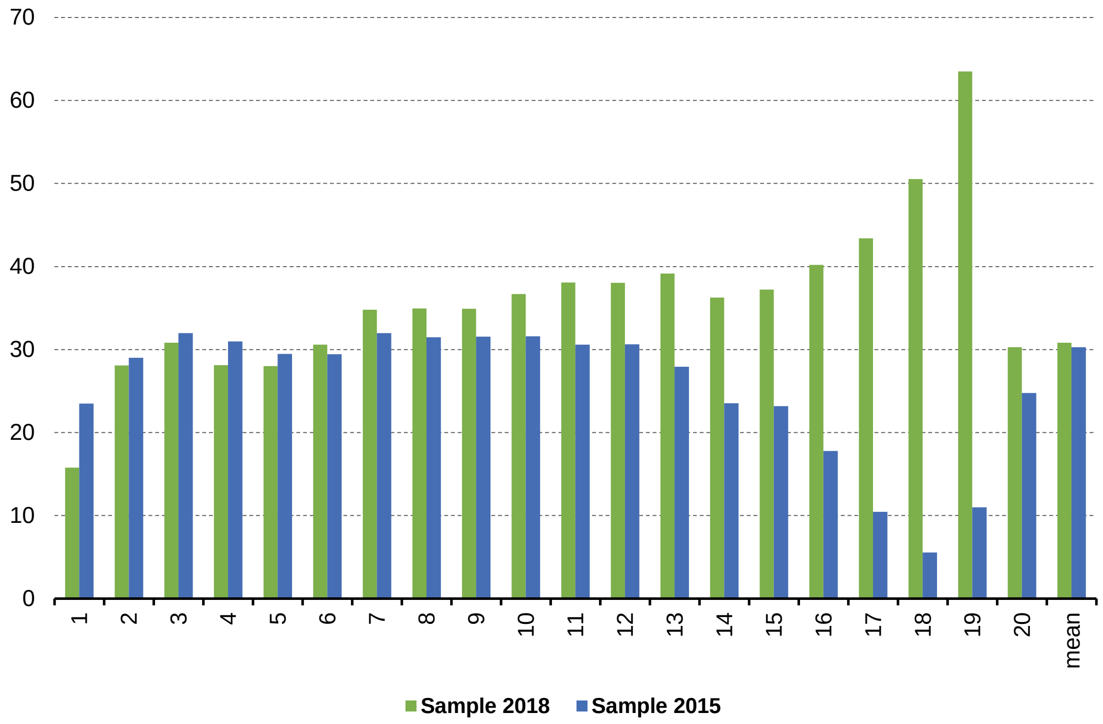
The distribution of points by class elevation is showed in figure 14. Because the different procedures in using the PI points and the introduction of the probability of change, in 2015 sample the percentages of points with an elevation more than 1 000 m are higher than in Master and 2018 sample while the points less than 100 m are more present in sample 2018. For elevation between 400 and 1 000 m the percentage allocation of two samples is the same as in Master while small differences are found in the intermediate elevation (100-400 m).

Figure 14: **Percentage distributions of points in Master and in 2015/2018 samples by elevation** (Percentage)



In the following figure 15 the ratios between the number of points in the two samples and the points in the Master belonging to the same class of probability of change (that has been estimated for all the point of the Master) is reported. The ratios have the meaning of the “coverage” of a sample with respect to its frame. The coverage of 2015 sample is slightly greater for the lower classes of probability to change while the points with higher probability to change are much more present in 2018 sample.

Figure 15: **Ratios between the number of points in 2015/2018 samples and in the Master**
(The probability to change classes are obtained dividing the interval 0-1 by 20)
(Percentage)



5 Conclusions

In coherence with the previous rounds of the LUCAS survey, the 2018 edition includes improvements on some aspects of the survey characteristics.

In this round, the survey is used as a sort of “multi-purpose” survey because it integrates different samples with different objectives:

(i) the estimates on land cover and land use,

(ii) an extended soil module where a topsoil sample is collected, for bulk density, soil biodiversity and organic horizon,

(iii) a test module for “grassland”, and

(iv) additional points for Copernicus programme.

The stratum variable STR18 has been updated, substantially ameliorating the uncertainty of classification and splitting two STR05 modalities, one of them being used to better define the reference population excluding the related points (transitional water).

Additionally to these improvements, deeper innovations in sampling design has been carried out in this round.

The innovation core was the estimation, for each point in the master, of the most probable land cover modality/modalities by a linear logistic regression model, on the basis of the actual data from the 2015 LUCAS survey, and considering about 16 covariates. This information is used (i) to calculated the CVs for the target variables at NUTS2 level and hence the sample size according to the desired sampling errors and (ii) to allocate the sample in the final strata.

The stratification adopted in 2018 sample design is not only finer than the one used for the previous surveys but it has been also a “dynamic” stratification inside the established study domains. The “atomic strata” (given by the Cartesian product of STR18, CLC and elevation classifications) are aggregated with an iterative algorithm that optimises the CVs of the target variables at NUTS2 level. The final strata produced in this way represent the most correlated combinations of modalities of the stratification “criteria” with the target variables and take into account the specificity of the country in the NUTS2 territories.

Stressing the use of PI, the survey has been considered a multi-mode survey. All the points are considered suitable to be selected and surveyed directly by surveyors or by photo interpretation. This last modality is used for the points of difficult access or associate to a low probability of change. In this way, the concept of “eligible” in the meaning of “eligible for field work” has been abandoned and no partition of the Master is done. The choice of which data collection mode has to be used is done after the sample selection on the basis of two information: a reachability and a propensity to change indexes. They are calculated for each point in the Master according to the physical characteristics of the point and by a statistical model of the status variations of the points belonging to the longitudinal structure of the previous surveys.

It is to be noted that while the reachability index is based more or less on the same parameters already used for the eligibility in the previous surveys, the propensity or probability of change index is quite new and it is worthy of further in-depth analysis.

Finally in 2018 LUCAS survey, because of the need to maintain the comparability between the different editions, an important element of continuity with the preceding rounds is given by considering as target information, in sampling and survey design, the variable “land cover” leaving “land use” as a sort of ancillary information. It may be relevant to promote a study on the implications of this choice and of an alternative vision of LUCAS.

6 References

http://ec.europa.eu/eurostat/web/lucas/overview

Baillargeon S. and Rivest L.-P. (2012). *The construction of stratified designs in R with the package stratification*. Survey Methodology, Vol. 37, No. 1, pp. 53-65

Baillargeon S. and Rivest L.-P. (2014). *Stratification: Univariate Stratification of Survey Populations*. R package version 2.2-5. https://CRAN.R-project.org/package=stratification

Barcaroli G., Pagliuca D., Willighagen E., Zardetto D. (2018). SamplingStrata: Optimal stratification of sampling frames for multipurpose sampling surveys. R package version 1.2. http://cran.r-project.org/web/packages/SamplingStrata/index.html

Ballin M., Barcaroli G. (2016). *Optimization of stratified sampling with the R package SamplingStrata: Applications to network data*. Computational Network Analysis with R: Applications in Biology, Medicine and Chemistry, Wiley

Ballin M., Barcaroli G., Catanese E, D'Orazio M. (2016). *Stratification in Business and Agriculture Surveys with R*. Romanian Statistical Review 2/2016, pp. 43-58

Barcaroli G. (2018). *Optimization of sampling strata with the SamplingStrata package*. https://barcaroli.github.io/SamplingStrata/articles/SamplingStrata.html

Barcaroli, G. (2014). *SamplingStrata: An R package for the optimization of stratified sampling*. Journal of Statistical Software 61 (4), 1-24.

Bethel J. (1985). *An Optimum Allocation Algorithm for Multivariate Surveys*. American Statistical Proceedings of the Survey Research Methods Section, pp. 209-212

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

Chromy J.B. (1987). *Design Optimization with Multiple Objectives*. Proceedings of the American Statistical Association Section on Survey Research Methods 1987, pp. 194-199

Cochran, W.G. (1977). *Sampling Techniques*. John Wiley & Sons, New York

Dalenius, T., Hodges, J.L. (1959). *Minimum Variance Stratification*. Journal of American Statistical Association, Vol. 54, pp. 88-101

Day C. D. (2006). *Application of an Evolutionary Algorithm to Multivariate Optimal Allocation in Stratified Sampling Designs*. Proceedings of the American Statistical Association Section on Survey Research Methods 2006 [CD-ROM]

Day C. D. (2010). *A Multi-Objective Evolutionary Algorithm for Multivariate Optimal Allocation*. Section on Survey Research Methods JSM 2010 pp.3351-3358

Gonzales J.M., Eltinge J.L. (2010). *Optimal Survey Design: a Review*. Section on Survey Research Methods - JSM, October 2010 https://www.bls.gov/osmr/abstract/st/st100270.htm

Gunning P., Horgan J.M. (2004). *A New Algorithm for the Construction of Stratum Boundaries in Skewed Populations*. Survey Methodology, Vol. 30, No. 2, pp. 159-166

Hartigan, J.A., Wong M.A. (1979). *A k-means clustering algorithm*. Applied Statistics, 28, pp. 100-108

Keskinturk T., Er S. (2007). *A Genetic Algorithm Approach to determine Stratum Boundaries and Sample Sizes of Each Stratum in Stratified Sampling*. Computational Statistics and Data Analysis, Vol. 52, Issue 1, 15 September 2007, s.53-67.

Kozak M., Wang H.Y. (2010). *On stochastic optimization in sample allocation among strata*. Metron – International Journal of Statistics 2010, vol. LXVIII, n.1, pp. 95-103

Lavallée P., Hidiroglou M.A. (1988). *On the stratification of skewed populations*. Survey Methodology, Vol.14, pp.33-43

Neyman, J. 1934. *On the two different aspects of the representative methods*. The method stratified sampling and the method of purposive selection. Journal of Royal Statistical Society, 97, 558-606.

Vose M. D. (1999). *The Simple Genetic Algorithm: Foundations and Theory*, MIT Press, Cambridge, MA

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: http://europa.eu/contact

On the phone or by e-mail
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 electronic mail via: http://europa.eu/contact

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: http://europa.eu

EU Publications
You can download or order free and priced EU publications from EU Bookshop at: http://bookshop.europa.eu. Multiple copies of free publications may be obtained by contacting Europe Direct or your local information centre (see http://europa.eu/contact)

EU law and related documents
For access to legal information from the EU, including all EU law since 1951 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/data) provides access to datasets from the EU. Data can be downloaded and reused for free, both for commercial and non-commercial purposes.

KS-TC-18-006-EN-N

Redesign sample for Land Use/Cover Area frame Survey (LUCAS) 2018

Eurostat's Land Use/Cover Area frame Survey (LUCAS) provides harmonised statistics on land use and land cover across the European Union. Land use shows the socio-economic usage of a given land: agriculture, commerce, industry, residential, etc. Meanwhile, Land cover refers to the bio-physical coverage of the land: crops, woodland, buildings, roads, etc. LUCAS is characterised by the unique, in-situ information that it provides. Since 2006, Eurostat has carried out this survey every 3 years. The latest LUCAS survey, covering all the 28 European Union (EU) countries, is taking place in 2018. As in the previous rounds, 2018 has been an opportunity to adapt, adjust and improve the LUCAS methodology in order to obtain the most accurate and harmonised data possible, while still allowing for comparison over time. This paper shows the characteristics of the new sampling design used for 2018, and the innovative methodology which is applied to it.

For more information

https://ec.europa.eu/eurostat/

Publications Office

ISBN 978-92-79-93921-1