Calibration in LUCAS 2018 Survey

MARCO BALLIN, GIULIO BARCAROLI, MAURO MASSELLI

2023 edition

STATISTICAL WORKING PAPERS

eurostat

Calibration in LUCAS 2018 survey

MARCO BALLIN, GIULIO BARCAROLI, MAURO MASSELLI

2023 edition

Manuscript completed in April 2023

This document should not be considered as representative of the European Commission’s official position.

Luxembourg: Publications Office of the European Union, 2023

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.

Theme: General and regional statistics
Collection: Statistical working paper

ISBN 978-92-68-03119-3 ISSN 2315-0807 doi: 10.2785/238299 KS-TC-23-002-EN-N

Abstract

Since the end of the pilot phase (2006), Eurostat has carried out Land Use/Cover Area frame Survey (LUCAS) every 3 years in 2009, 2012, 2015 and 2018. Harmonised statistics on Land Use and Land Cover are produced by the survey across the European Union. Land Use shows the socio-economic use of a given land (agriculture, commerce, industry, residence, etc.), while the Land Cover refers to its biophysical coverage such as crops, forest, buildings, roads, etc. The survey is based on a sample of points selected from the Master, an area frame of about one million points, each of them representing a square of 4 km² in a grid covering all the EU territory. The methodology has been improved in each round of the survey. In 2018, a deep revision had been carried out concerning the use of statistical model to predict the Land Cover modalities for every point in the master on which to base the stratification and the sample selection, the rules of assigning the observation method (directly in field or photo interpreted in office), a different specification of the eligibility concept and, finally, the estimation procedure.

Regarding the last topic, in the 2018 survey the estimation has been carried out, increasing the number of variables whose known totals are used in calibration. Other than the area by elevation classes at NUTS0 level (the only variable considered in the previous rounds of the survey) also CORINE Land Cover datasets of ‘artificial’, “woodland”, “agriculture”, “wetland”, “water” and High Resolution Layers “imperviousness” are considered at NUTS2 level.

In the paper, a brief description of the calibration technique and the way it had been applied in LUCAS 2018 as well as the rationale for this choice are reported. Some analysis are performed with the aim at studying the differences between calibrated and initial weights and their respective estimates using not only for the 2018 survey data but also the survey data of 2009, 2012 and 2015, calibrated with the same calibration model used as for the 2018 survey.

Keywords: calibration, sampling, land cover and land use survey

Authors: Marco Ballin, Giulio Barcaroli, Mauro Masselli

Acknowledgement:

This work has been carried out under the supervision of the Eurostat project team (Alessandra Palmieri, Savvas Zachariadis and Beatrice Eiselt), with the contributions of Flavio Lupia and Moritz Mercker, and with the support of GOPA project team (Elena Mettler and Christian Joebges).

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.

JEL classification: C83 Survey Methods • Sampling Methods

Table of contents

List of Tables

List of Figures

Abbreviations

CLC	CORINE land cover
CV	Coefficient of variation
EU	European Union
FAO	Food and Agricultural Organisation of United Nations
HRL	High resolution layer
HT	Horwitz-Thompson
LUCAS	Land Use/Cover Area frame Survey
LUE	Services and Residential area
LUD	Land Use with Heavy Environmental Impact
R	Software environment for statistical computing

Country codes

BE	Belgium
BG	Bulgaria
CZ	Czechia
DK	Denmark
DE	Germany
EE	Estonia
IE	Ireland
EL	Greece
ES	Spain
FR	France
HR	Croatia
IT	Italy
CY	Cyprus
LV	Latvia
LT	Lithuania
LU	Luxembourg
HU	Hungary
MT	Malta
NL	Netherlands
AT	Austria
PL	Poland
PT	Portugal
RO	Romania
SI	Slovenia
SK	Slovakia
FI	Finland
SE	Sweden

1 Introduction

Since 2006 Eurostat has implemented the LUCAS survey (Land Use/Cover Area Frame Survey) to improve the quality and the completeness of Land Cover and Land Use statistics, which contribute to some of the major EU policy areas ([1]). The LUCAS survey is used to monitor the land cover, the social and economic use of land, the biodiversity, and other environmental parameters. Sustainable Development Indicators and Agri-Environmental indicators on soil are some examples of LUCAS data use. The micro-data collected also serve to produce, verify and validate CORINE Land Cover (CLC) and Copernicus data.

The LUCAS survey is usually carried out at EU level every three years, considering a sample of geo-referenced points selected over the entire EU territory. The Core LUCAS survey gathers harmonized information on Land Cover (bio-physical coverage of land) and Land Use (socioeconomic use made of land) and their changes over time. It also provides territorial information facilitating the analysis of interactions between agriculture, the environment, and the countryside, such as water and land management (e.g. irrigation and grazing).

At a given geo-referenced point, other than collecting information on the point itself, LUCAS surveyors also take a series of photographs in all four cardinal directions. In the framework of each survey and in addition to the core LUCAS, specific information is also collected under so-called ‘ad hoc modules’, e.g. the Topsoil in 2009 and 2015, the Transects in 2009, 2012 and 2015, Grassland and Soil in 2018.

Over time, the survey’s methodology has changed. While retaining its initial coverage, its focus has shifted from that of an agricultural land survey to a broader Land Cover, Land Use and Landscape survey. Coverage was extended to 23 EU Member States in 2009 (Bulgaria, Cyprus, Malta and Romania were not included) to 27 Member States in 2012 and finally to all Member States in LUCAS 2015 and 2018 surveys. Sample sizes have increased accordingly.

The survey consists of a two-phase area sample. In the first phase, a frame of more than 1 million geo-referenced points is systematically selected from a 1 square km grid (corresponding to more than 4 million points) built over the entire EU territory. The frame also classifies each point of the Master in variables related to Land Cover classes that were obtained from photo interpretation implemented in 2006 and 2016. In the second phase, a sample of points is selected from the Master, at which the statistical information is collected in a circle with a diameter of 3 meter or in some cases 40 meters. Information is gathered on the field by surveyors or by photo-interpretation, whenever the point is too difficult to be directly surveyed for different reasons (e.g. military base, arduous site, refusals from the

[1] 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 for the improvement of the data collection system and analyses of Land Cover and Land Use 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. This set the beginning of the current three-yearly data collection frequency. LUCAS has become part of Eurostat's activities and budget since January 2008. Since 2012, the survey has been supported financially by other DGs of the Commission.

owner).

Data collection is carried out for blocks of countries by different private companies that are also in charge of checking data quality during fieldwork. A further step of editing and imputation is performed by Eurostat after data transmission. Finally, the estimates of the target variables are calculated by applying a system of weights to the microdata, which is obtained through a calibration procedure ([2]).

The present paper aims at describing that calibration procedure and to analyse the impact of calibration versus the Horwitz-Thompson (HT) estimator.

([2]) For Lucas survey methodology see (1) and (2) and the Eurostat site at https://ec.europa.eu/eurostat/web/lucas/methodology.

2 Definition of calibration

Generally speaking, calibration is a technique used to modify the sample weights so that the estimates of the totals of some variables in the sample equate the corresponding known values of the same variables in an external source.

Several definitions are available for ‘calibration’. Here, we adopt the one reported in Särndal (2007).

Definition: The calibration approach to estimation for finite populations consists of:

1. A computation of weights that incorporate specified auxiliary information and are restrained by calibration equation(s),
2. The use of these weights to compute linearly weighted estimates of totals and other finite population parameters: weight times variable value, summed over a set of observed units.
3. An objective to obtain nearly design unbiased estimates as long as non-response and other non-sampling errors are absent.

The calibration weights are obtained through solving the following system:

$$\begin{cases} \text{Min}\, \{\, G(w_k,\, d_k)\} \\ \sum_{k \in S} w_k\, x_k \,=\, T_x \end{cases} \qquad \text{(e.1)}$$

where:

- *S* is the sample drawn from the population U;
- *k* is the indicator of the generic unit in the sample (for instance the sampled points in LUCAS);
- d_k is the initial weight (the inverse of probability of inclusion) of unit *k*;
- $w_k \,=\, c_k d_k$ is the calibration weight for unit *k*, given as the product of a correction factor and of the initial weight.

Solving the system minimizes the distance between initial and calibrated weights by the function $G(w_k, d_k)$, subject to the constraint that the estimates of the variable/variables X so must equal their known totals T_x in the reference population U. X can be one or more auxiliary variables that are available for all units in U. They are assumed to be of good quality, to belong to external sources and to be correlated with the target variables (as is the case with the calibration, in LUCAS), or they can be one or more variables surveyed in the survey, and present in different sources for all the population units. In both the cases, the estimate of X making use of the calibrated weights will replicate the totals.

The calibration procedure is largely used by the National Statistical Institutes (NSIs) for its statistical properties and flexibility, generally for inflating the data of sample surveys, making use of statistical sources or registers containing strictly related information, which are available for the entire reference population. Calibration can moreover guarantee the consistency between data in different statistical sources, when the auxiliary variables coincide with some of the variables that are collected in the survey, given the fact that calibration imposes equality between the sample estimates and the ‘known totals’. This characteristic of the method increases the credibility of published statistical information by avoiding that the main totals of the same phenomenon are described by divergent numbers from different sources. Credibility is one of the most important issues for the NSIs.

Calibration is a practical method used to incorporate the available auxiliary information into the estimation process and to use that information, in order to improve the accuracy of estimates deriving from sample surveys. Its impact is particularly relevant, when missing-unit rates are high. The total non-response rate is negligible in the LUCAS survey, because those sample units that are not directly observable are transformed into photo-interpreted ones. Bias can nevertheless be introduced, mainly during field operations, for instance through the observation type, that is, the incidence of photo interpretation rather than field observation. Beyond differences in data-collection modalities, it may also be due to a difference in reference time ([3]). Bias may also arise from the high variability in the distances from which a point is observed. While many studies indicate the calibrated estimator’s adequacy in dealing with total non-response, its potential in the treatment of other non-sampling errors is still a matter of debate ([4]).

([3]) Because photos are kept t years before data collection for the survey at time T, the information by direct observation is referred to as being made at time T, while the information derived from photo interpretation is referred to as being made at time T-t.

([4]) For more information see (4) in References.

3 Calibration in producing LUCAS estimates

The estimates published by Eurostat in LUCAS survey rounds 2009, 2012 and 2015 were produced by calibrating the microdata by classes of elevation (and an *ad hoc* application was developed) because the variable 'Elevation' is correlated with most LUCAS variables, it does not change overtime and it is available for all Master points. The latter were categorised as being 'eligible' or 'not eligible' according to the possibility of directly observing them or not, with difficult-to-reach points mainly being located in mountains. In principle, only eligible points were considered in the sample, while estimates had to refer to the entire EU area. Estimates were thus affected by the attribution of averaged characteristics of eligible points to the non-eligible ones. So as to limit this kind of bias, a calibration by classes of altitude was applied that, for example, aimed to attribute to mountain areas the characteristics of points belonging to an elevation class close to theirs.

The calibration methodology used in the 2018 LUCAS survey was developed using the R package ReGenesees, which allows the calculation of sampling estimates by using calibration estimators ([5]) (Zardetto, 2015).

When using a calibration estimator, the weight attributed to each unit is obtained through a procedure consisting of the following distinct steps:

1. The initial weight of each sample unit, named 'direct weight', is calculated according to sampling design, as the reciprocal of the inclusion probability.
2. The direct weight is adjusted to account for missing units, if any, resulting in the 'base weight'.
3. Correction factors of the base weights are computed, taking into account equality constraints between some known totals from the population of reference and the corresponding sample estimates.
4. The 'final' or 'calibrated weight' is obtained by multiplying the base weights by their respective correction factors.

The second step proved not to be necessary, in our particular case, given that there are no total non-response cases in the LUCAS survey.

The third step is performed by solving a constrained optimization system (see (e.1) above) for which the known totals and the initial weights are taken as inputs, and the correction factors are the unknown quantities to be calculated.

Implicitly, using a calibration estimator requires the definition of a calibration 'model' wherein, conceptually, the dependent variables are the estimates of the target variables (e.g. Land Cover, Land Use, etc.), and the explicative variables are implicitly defined as being those, for which known totals in the reference population are available.

The known totals related to NUTS 2 and elevation (5 classes) were considered, together with other known totals derived from Copernicus estimates (High Resolution Layers) available for the same reference period as LUCAS (2018). Together with the areas by each class of elevation, the final model,

([5]) For more information see (5) in References.

at NUTS 2 level, includes:

- HRL-Imperviousness,
- CLC-Artificial,
- CLC-Agriculture,
- CLC-Forest,
- CLC-Wetland,
- CLC-Water.

The known totals are calculated from the Master, after including CLC and HRLs.

This standard calibration procedure was adopted for the generality of countries. During the period spanning between the release of the first batch of estimates and the final release, a detailed analysis of estimates (stocks and variations) was carried out, in order to detect critical situations. In particular, the focus was on the detection of non-plausible variations in some of the Land Cover aggregates (i.e. 'Artificial' and 'Water') and in 'Settlement'.

The known totals used to calibrate the initial weights were calculated from the Master. The choice of not using the totals directly obtained from the Copernicus website but rather the ones derived from the Master, is due to practical reasons.

It is not worthless to remember that the LUCAS variables, though have a similar label (e.g. LUCAS 'Crop' and CLC 'Agriculture'), do not correspond exactly with the CLC variables and then their estimated totals are not equal to the known totals used for calibration. Non-correspondence is explained by differences in definitions, the different observation methods (pixel versus point) and the sampling variance of LUCAS estimates.

The auxiliary variables are correlated with the corresponding Land Cover modalities in LUCAS.

The coefficients of correlation between the CORINE Land Cover variables used as known totals in calibration, and the LUCAS variables of Land Cover and Land Use are reported in Table 1. They are calculated for the above-cited dichotomous variables, using all the points in the Master updated with the information from Copernicus, together with the variables of the previous LUCAS surveys and estimates produced by a statistical model ([6]). In Table 1, the more significant correlations are those between 'Imperviousness' and CLC 'Artificial' with LUCAS 'Artificial' and 'U3-Transport, Utilities, Residential', between CLC 'Agriculture' and 'Forest- semi natural areas' with LUCAS 'Cropland' and 'Woodland', between CLC 'Wetland' with LUCAS 'Wetland' and finally between CLC 'Water' with LUCAS 'Water'.

([6]) For more extensive information on updating the Master, see (1) in References.

Table 1: Coefficients of correlation between the CORINE Land Cover variables used in calibration and LUCAS variables related to Land Cover and Land Use

LUCAS variables	HRL imperviousness - CORINE Land Cover variables					
	Impervious ness	Artificial	Agriculture	Forest-semi natural area	Wetland	Water
A - Artificial land	0.58	0.50	–0.09	–0.16	–0.03	–0.01
B - Cropland	–0.11	–0.12	0.54	0.46	–0.08	–0.04
C - Woodland	–0.13	–0.13	–0.55	0.65	–0.05	–0.02
D - Shrubland	–0.04	–0.04	–0.13	0.15	0.02	–0.01
E - Grassland	–0.04	0.03	0.23	–0.23	–0.03	–0.02
F - Bareland	–0.02	..	0.05	–0.04	–0.01	..
G - Water areas	–0.01	..	–0.04	–0.04	0.02	0.39
H - Wetlands	–0.03	–0.04	–0.11	–0.02	0.55	0.05
U1 - Agriculture, Mining, Fishing	–0.32	–0.36	0.26	–0.01	–0.16	–0.08
U2 - Manufacture, Energy	0.18	0.14	–0.04	–0.04	–0.01	0.06
U3 - Transport, Utilities, Residential	0.49	0.56	–0.12	–0.17	–0.03	0.04
U4 - Unused and Abandoned areas	–0.07	–0.06	–0.20	0.17	0.22	0.05

The relationship between LUCAS land cover and CLC land cover is also found by using the Chi square statistic (χ^2), the contingency coefficient and the V Cramer index. The independence hypothesis is rejected by a p-test result of 0.000. The important contributions to χ^2 are explained by the cells identified by LUCAS 'Artificial' with CLC 'Artificial'; LUCAS 'Cropland' with CLC 'Agriculture' and 'Forest'; LUCAS 'Woodland' with CLC 'Agriculture' and 'Woodland'; LUCAS 'Water' with CLC 'Water'; LUCAS 'Wetland' with CLC 'Wetland'. The association between the two variables is measured by a contingency coefficient equal to 0.724 versus a maximum of 0.89 while the V Cramer is equal to 0.525 versus a maximum of 1.

The correlations between LUCAS 'Cropland' and 'Woodland' and CLC 'Agriculture' and 'Forests' are explained by the overlap between relative definitions; the same holds for LUCAS 'Artificial' and HRL 'Imperviousness' and CLC 'Artificial'.

4 Comparing calibration estimates and HT estimates

The comparison between calibrated and non-calibrated estimates allows us to evaluate the impact calibration has on the estimates. For this purpose, Horvitz-Thompson (HT) estimates were produced. They are calculated, making use of the initial weights, that is, the multiplicative inverses of the inclusion probabilities of the sampled points. The analysis was performed, considering the main variables produced by the LUCAS survey at EU level:

- Land Cover (1 digit);
- Land Use (1 digit);
- Settlement;
- FAO classes (1, 2, 3);
- LUE (Services and Residential area);
- LUD (Land Use with heavy environmental impact).

Table 2 reports some of the statistics for the above variables, to describe the distributions of calibrated and initial weights in the 2018 survey. For most of the variables, the averages of calibrated weights are equivalent to the ones of initial weights; only the means of calibrated weights of Land Cover 'Water' and Land Use 'U2-Manifacture, Energy' are slightly higher than the ones of the initial weights. The medians also are very close to each other, except in the case of Land Cover 'Water'. Because the medians are lower than the averages, both for calibrated and for initial weights, their distributions display a positive asymmetry, that is, they are skewed to the left.

For the majority of the variable, the ranges (the distance from the minimum value to the maximum) and the standard deviation of the calibrated and initial weights are of the same level, except for the variables Land Cover 'Water', Land Use 'U2-Manifacture, Energy' and 'U3-Transport, Utilities, Residential', 'Settlement', and 'LUE', where the indicators of calibrated weights are greater than the initial ones, indicating a larger dispersion.

Table 2: Average, range, standard deviation and median of calibrated and initial weights – 2018 survey

LUCAS variables	averages		range		standard deviation		median	
	calibrated	initial	calibrated	initial	calibrated	initial	calibrated	Initial
Land Cover								
A - Artificial land	9.1	9.0	348.7	351.9	9.6	10.1	6.9	6.7
B - Cropland	12.0	12.0	2 709.7	2 785.1	54.2	54.2	8.8	9.0
C - Woodland	14.4	14.5	1 881.3	1 903.9	16.3	16.3	8.6	8.6
D - Shrubland	13.8	13.9	757.1	716.7	11.2	11.1	8.5	8.4
E - Grassland	11.3	11.3	1 925.0	1 842.8	34.3	31.6	8.7	8.7
F - Bareland, Lichens, Moss	13.8	13.8	1 197.2	1 220.5	17.1	17.2	8.6	8.5
G - Water areas	53.2	49.8	5 105.9	3 846.8	11.9	11.1	11.2	10.2
H - Wetlands	11.7	11.7	712.4	716.7	15.4	16.1	8.4	8.5
Land Use								
U1 - Agriculture, Mining, Fishing	13.0	13.1	2 712.0	2 785.1	24.9	25.1	8.8	8.9
U2 - Manufacture, Energy	11.6	10.8	1 133.1	433.9	35.5	18.2	7.3	7.4
U3 - Transport, Utilities, Residential	11.6	11.3	5 105.9	3 847.4	54.1	48.3	7.5	7.5
U4 - Unused and Abandoned areas	13.8	13.8	3 675.5	3 846.8	39.1	38.4	8.3	8.3
FAO forest								
FAO 1 - Forest	14.5	14.6	1 012.7	1 211.9	22.3	22.8	8.7	8.7
FAO 2 - Other Wooded Land	13.9	14.0	757.1	716.7	25.0	25.2	8.5	8.5
FAO 3 - Other Wooded Land no FAO	15.9	15.9	2 712.0	2 785.0	59.8	59.5	6.9	6.9
Settlement	9.9	9.7	1 133.2	822.0	14.2	13.2	7.4	7.4
LUE - Services and Residential area	12.2	11.9	5 105.9	3 846.8	67.2	59.6	7.6	7.6
LUD - Land Use with Heavy Environmental Impact	10.8	10.5	1 967.9	1 970.1	23.1	21.1	7.2	7.2

4.1 Evaluating calibration impact by comparing HT and calibration estimates

From the following equation establishing the relation between the calibrated estimate and the HT estimate for the generic variable Y (in our case each of the eight modalities of Land Cover, the four of Land Use, ‘settlement’, ‘LUE’, and ‘LUD’):

$$\widehat{Y_{cal}} = \sum_{k \in S} w_k \, y_k = \widehat{Y_{HT}} + \sum_{k \in S} (w_k - d_k) \, y_k \qquad \text{(e.2)}$$

we can obtain the following:

$$\widehat{Y_{cal}} - \widehat{Y_{HT}} = \sum_{k \in S} (w_k - d_k) \, y_k \qquad \text{(e.3)}$$

where $\widehat{Y_{cal}}$ is the estimate of the variable Y obtained by calibration, $\widehat{Y_{HT}}$ is the estimate of the same variable obtained by HT, w_k and d_k are respectively the calibration and the initial weights (d_k is the inverse of the probability of inclusion) assigned to the unit *k*, y_k is the value of the variable *Y* for the unit *k*, and *S* is the set of sampling units. Both the calibrated and the initial weights are expressed in terms of area and so they sum up to the total EU area.

The (e.3) is averaged over all possible samples, because the correctness of the HT estimator, E($\widehat{Y_{HT}}$) = Y, provides the bias of the calibration estimator:

$$E(\widehat{Y_{cal}} - \widehat{Y_{HT}}) = E(\hat{Y}_{\mathrm{cal}}) - Y = E(\textstyle\sum_{s} (w_k - d_k) y_k)$$

In case the function $G(w_k, d_k)$ is the Euclidean distance function (as used in ReGenesees for LUCAS calibration) a near-unbiasedness holds.

The difference (e.3) or the ratio between the two estimates can be used to evaluate the impact, and as a rough indicator of the possible bias of the calibration. It is worthwhile noting that (e.3) is different for each variable and it depends on the distribution of the variable itself among the sample units.

Table 3 shows the 2018 estimates for the total area of the above variables at EU level, together with the percentage ratios of calibrated to HT estimates that measure the impact of calibration. The table also presents the sampling errors (percentage coefficients of variation) calculated for calibrated and HT estimates, and their ratios; here, a ratio above 100 implies the higher sampling error of calibrated estimates.

For Land Cover variables the ratios range from 99.5 to 106.4, that is, the calibration has an impact ranging from -0.5% to +6.4%, in terms of the HT estimates. Calibrated estimates of ‘Woodland’ and ‘Shrubland’ are lower than those yielded by HT, while the opposite holds for ‘Artificial’, ‘Cropland’, ‘Grassland’ and ‘Water’, which are found to be higher than HT.

The calibrated estimates of Land Use are almost or higher to those HT produces except for ‘U1-Agriculture. Mining, Fishing’. In particular for the ’U2-Manifacture, Energy’ the ratio reaches 107.9. The FAO classes ‘Forest’ and ‘Other Wooded Land’ display a ratio of below 100, while ‘Other Wooded Land no FAO’ presents a slightly positive impact of 0.3 %. Finally, all ‘Settlement’, ‘LUE’ and ‘LUD’ display a positive impact for calibration estimates.

The sampling errors for calibrated estimates are slightly lower than those provided by HT. Nevertheless, one must note that the ratio accentuates the variation due to the small sizes of the values compared and, when one also considers the CV differences, the real difference in sampling error is in fact very small. Only for the modality ‘Water’ of Land Cover and ‘U2-Manufacture, Energy’ of Land Use we find a ratio greater than 160.

Table 3: Land Cover, Land Use, Settlement, FAO-Forest, LUE, LUD total areas estimated and coefficient of variation, by calibration and by HT estimator, and their ratios – 2018 survey

LUCAS Variable	Survey 2018					
	Calibrated		HT		ratios *100 (cal/HT)	
	estimated area	cv	estimated area	cv	estimated area	cv
Land Cover						
A - Artificial land	190 518	0.85	188 268	0.87	101.2	97.70
B - Cropland	1 041 286	0.53	1 041 314	0.64	100	82.81
C - Woodland	1 729 494	0.28	1 739 040	0.31	99.5	90.32
D - Shrubland	263 762	1.14	264 925	1.18	99.6	96.61
E - Grassland	820 493	0.53	820 021	0.53	100.1	100
F - Bareland, Lichens, Moss	110 111	2.29	110 213	2.23	99.9	102.69
G - Water areas	135 943	3.47	127 780	2.14	106.4	162.15
H - Wetlands	77 876	2.2	77 923	2.25	99.9	97.78
Land Use						
U1 - Agriculture, Mining, Fishing	3 284 320	0.23	3 294 175	0.26	99.7	88.46
U2 - Manufacture, Energy	13 667	9.15	12 665	5.46	107.9	167.58
U3 - Transport, Utilities, Residential	400 346	1.96	391 676	1.57	102.2	124.84
U4 - Unused and Abandoned areas	671 150	1.15	670 968	1.13	100	101.77
FAO forest						
FAO 1 - Forest	1 619 918	0.26	1 630 089	0.27	99.4	96.30
FAO 2 - Other Wooded Land	227 062	1.29	228 193	1.34	99.5	96.27
FAO 3 - Other Wooded Land no FAO	140 984	3.59	140 531	3.81	100.3	94.23
Settlement	342 080	3.09	336 529	2.51	101.6	123.11
LUE - Services and Residential areas	255 584	1.79	249 798	1.68	102.3	106.55
LUD - Land Use with Heavy Environmental Impact	171 639	0.84	167 438	0.81	102.5	103.70

The data reported in Table 3 can be visualised in Figure 1 and Figure 2.

Figure 1: Ratio of estimated areas and ratio of coefficients of variations, calculated by calibrated and by HT estimator, for the main LUCAS variables – 2018 survey

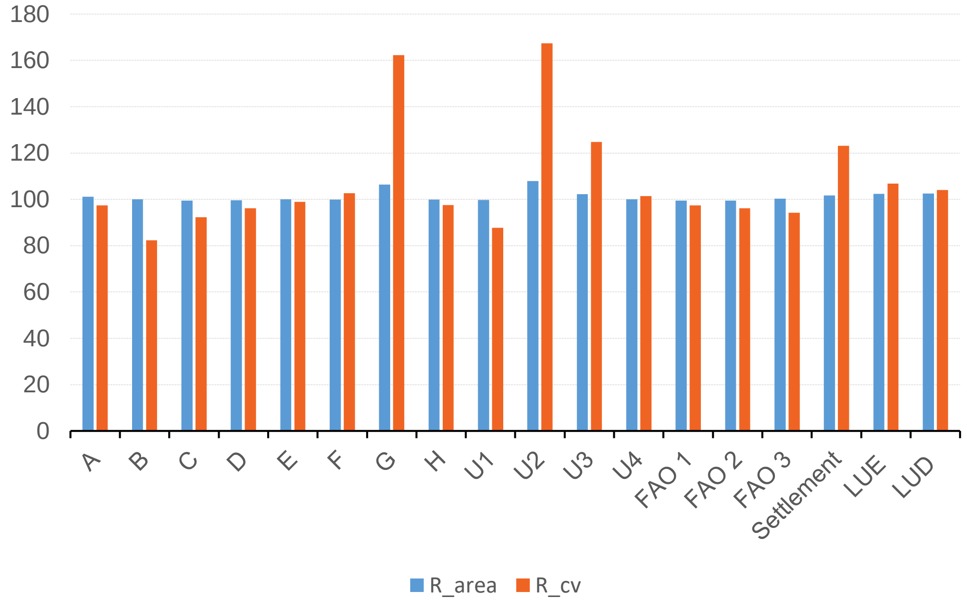
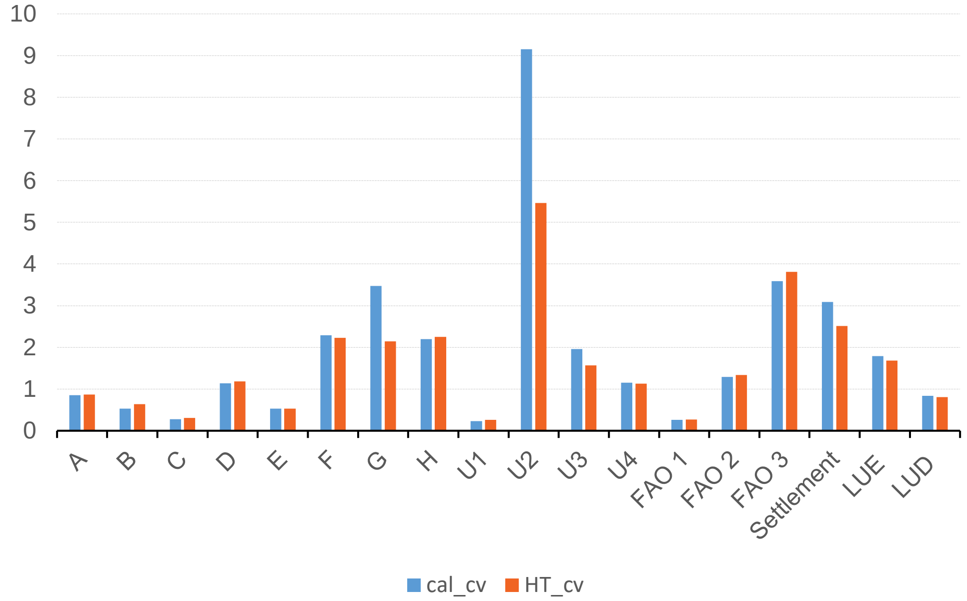


Figure 2: Coefficients of variation of the areas estimated by calibrated and by HT estimator, for the main LUCAS variables – Survey 2018



4.2 Evaluating calibration impact using the differences between weights

Equality (e.3) allows to calculate the impact of calibration by using the estimates of the total areas, as reported in Table 3, but also by using the weight differences at the level of sampling units. For this case (e.3) states that the differences between calibrated and HT estimates depend on the differences between initial and calibrated weights, as well as on the distribution of the variable *Y* in the sample units. For the quantitative variables, that means the level of Y, while for the qualitative variable the level is assigned as equal to one. For example, the estimate of total artificial area is obtained from the units that have the characteristic 'artificial' and hence only from a subset of the sample. In the LUCAS survey, the main estimates concern dichotomous variables corresponding to the modalities of Land Cover, Land Use, 'FAO Forest', 'Settlement', 'LUE', and 'LUD'. In those cases, the set of derived binary variables constitutes a transformation of the *mother* variable (e.g. Land Cover) into a *disjunctive normal form*, that is, a sample unit presents only one characteristic (e.g. 'Artificial') and the *union* of all the modalities (e.g. the characteristics of Land Cover) covers all the sampling units. Then, the sum over all sample units:

$$\sum_{k \in s}(w_k - d_k) \qquad \text{(e.4)}$$

indicates the impact on the entire *mother* variable, e.g. Land Cover.

Table 4 (below) reports the following indicators for the main 2018 LUCAS variables:

- Column (1) - the impact of calibration measured by the ratio between the calibrated estimates and the HT ones minus 100, that indicates how much the calibrated estimates are greater (or lower if the ratio is < 0) in percentage of HT estimates;
- Column (2) - for the units, for which the calibrated weights are greater than the initial weights ($w_k > d_k$), the average of the differences between calibrated and initial weights $\Sigma(w_k^+ - d_k^+)$ /n^+ , where n^+ is the number of units with (w_k - d_k) > 0 (or $w_k > d_k$), and where w_k^+ and d_k^+ are the calibrated and the initial weights of those units;
- Column (3) - for the units, for which the calibrated weights are smaller than the initial weights ($w_k < d_k$), the average of the differences between calibrated and initial weights $\Sigma(\mathrm{w}_k^- - \mathrm{d}_k^-)$ / n^- , where n^- is the number of units with (w_k - d_k) < 0 (or $w_k < d_k$), and where w_k^- and d_k^- are the calibrated and the initial weights of those units;
- Column (4) - the ratio of the difference between the number of units with $w_k > d_k$ (i.e. $w_k - d_k > 0$) and the number of units with $w_k < d_k$ (i.e. $w_k - d_k < 0$) to the total sample size expressed as a percentage: $(n^+ - n^-)/n *$ 100; a negative value indicates that the points with calibrated weights smaller than the initial weights outnumber the points with calibrated weights greater than the HT weights; a positive value indicates the opposite case;
- Column (5) - the percentage contribution to the impact of calibration of the units with $w_k > d_k$ on total area $(\sum_S w_k^+ - \sum_S d_k^+)/\sum_S w_k$ * 100;
- Column (6) - the percentage contribution to the impact of calibration of the units with $w_k < d_k$ on total area $(\sum_S w_k^- - \sum_S d_k^-)/\sum_S w_k$ * 100.

Using the above indicators represents a different way of evaluating the impact of calibration, which provides a deeper insight into the mechanism leading to a lower or to a higher estimation of calibrated weights, when comparing to HT weights, as reported in Column (1).

The total reported in Table 4 corresponds to the overall impact of the *mother* variables (as defined above, e.g. Land Cover). In Column (1), its ratio is equal to 100 by definition, because both the HT estimation and calibration must sum to the target EU overall area. This result has been achieved by a number of points with $w_k > d_k$ of 8.0 % lower than the corresponding points with $w_k < d_k$, but this condition is balanced by a greater difference between means of calibrated weights and HT weights: 0.55 versus –0.47.

The data reported in the table describe the mechanism leading to the calibration impact. The points' contribution to the calibration impact (Columns 5 and 6) depends on the differences between the averages of the calibrated and the initial weights, for units with $w_k > d_k$ and $w_k < d_k$ (Columns 2 and 3) and the differences between the number of sample points in the two states expressed as a percentage share of the total sample size (Column 4). The algebraic sum of contributions gives the calibration impact; a nil or negligible impact is found for variables, for which the two contributions are more or less equivalent.

A large part of the variables do display a relevant absolute percentage difference in numbers (Column 4) but the critical factor, with the calibration impact, appears to be the difference between the averages of calibrated weights and initial weights for the points with $w_k > d_k$ versus the same parameter for the points with $w_k < d_k$ (Columns 2 and 3).

Table 4: Ratio between calibrated and HT estimates, differences between averages of calibrated and initial weights, percentage difference of the number of points with $w_k > d_k$ and $w_k < d_k$, and contribution to the calibration impact of units with a calibrated weight higher/ lower than the initial one (2018 survey)

LUCAS variables	% ratio between calibrated and HT estimates	average of differences between calibrated and initial weights		% difference of the number of points with $w_k>d_k$ and the number of points with $w_k<d_k$ over the total sample size	% contribution to the calibration impact on total area	
		in sample units with $w_k>d_k$	in sample units with $w_k<d_k$		in sample units with $w_k>d_k$	in sample units with $w_k<d_k$
	(1)	(2)	(3)	(4)	(5)	(6)
Total	100	0.55	–0.47	–8.0	2.0	–2.0
Land Cover						
A - Artificial land	1.2	0.89	–0.88	11.5	5.5	–4.3
B - Cropland	0.0	0.28	–0.25	–5.8	1.1	–1.1
C - Woodland	-0.5	0.47	–0.48	–16.3	1.4	–1.9
D - Shrubland	-0.4	0.52	–0.51	–11.9	1.7	–2.1
E - Grassland	0.1	0.45	–0.43	–2.4	2.0	–1.9
F - Bareland	-0.1	0.42	–0.42	–3.3	1.5	–1.6
G - Water areas	6.4	11.86	–5.41	1.8	11.3	–5.0
H - Wetlands	-0.1	0.91	–0.81	–6.5	3.6	–3.7
Land Use						
U1 - Agriculture, Mining, Fishing	-0.3	0.37	–0.38	–9.6	1.3	–1.6
U2 - Manufacture, Energy	7.9	2.63	–1.51	14.3	12.9	–5.6
U3 - Transport, Utilities, Residential	2.2	1.35	–1.12	11.7	6.5	–4.3
U4 - Unused and Abandoned areas	0.0	0.75	–0.56	–13.9	2.3	–2.3
FAO forest						
FAO 1 - Forest	-0.6	0.45	–0.47	–17.4	1.3	–1.9
FAO 2 - Other Wooded Land	-0.5	0.51	–0.51	–13.7	1.6	–2.1
FAO 3 - Other Wooded Land no FAO	0.3	0.59	–0.52	2.4	1.9	–1.6
Settlement	1.6	1.02	–0.91	12.1	5.8	–4.0
LUE - Services and Residential area	2.3	1.27	–1.27	13.2	7.0	–4.5
LUD - Land Use with Heavy Environmental Impact	2.5	1.52	–1.0	9.4	6.4	–4.2

For example, in the case of the binary variable ‘Water’, for which a percentage difference in numbers of instances of 1.8 % is recorded, the calibration impact is 6.4 %, while the variable ‘Artificial’ displays a percentage difference of 11.5 %, and a calibration impact of 1.2 %. The distinction between the different impacts is due to the positive and negative contributions that, in turn, depend on the averages of calibrated and initial weights in the two sets of points with $w_k > d_k$ and $w_k < d_k$.

The variable ‘U2 - Manufacture, Energy’ presents a high percentage difference in sizes (14.3 %) that, together with a high average of calibrated and initial weights in the points with $w_k > d_k$ and $w_k < d_k$ (2,63 and –1,51) produces high contributions (12.9 % and –5.6 %) and hence a calibration impact of 7.9 %. The same mechanism holds for the variables ‘U3-Transport, Utilities, Residential’, ‘Settlement’, ‘LUE’, and ‘LUD’, leading to a moderate impact of about 2 %.

The nil or negligible calibration impact ranging between -0.6 % and +0.3 % of the remaining variables is due to a small differentiation in averages of differences between calibrated and initial weights making equivalent contributions to the impact.

Table 5 reports on a further analysis, carried out on the distributions of the differences between calibrated and initial weights. For all variables, those distributions display outliers identified by the distances (i) exceeding the 99th percentile (that is the value below which are contained 99 % of the differences) or (ii) falling below the 1st percentile (that is the value below which are contained 1 % of the differences). In particular, this applies for the variable ‘Water’, the Land Use derived binary variables ‘U2-U3’, ‘LUE’, ‘LUD’ and ‘Settlement’. The number of potential severe outliers can be easily calculated by keeping the 1 % of the frequency in both the tails of the distribution.

The same variables show a variability of differences that is much higher than that of others, when one considers the range and the standard deviation. Four variables (‘Artificial’, ‘Woodland’, ‘Grassland’ and ‘FAO - Forest’) have a negative skewness, that is, the distribution is skewed to the right, while the contrary holds for the remaining variables.

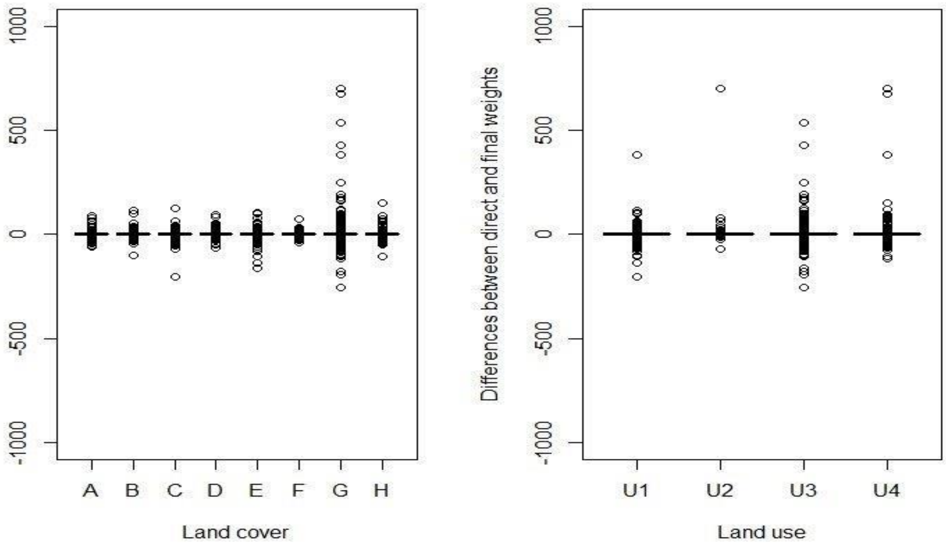
Summarizing, the variables can be divided into three groups, on the basis of the analysis of the distributions of differences between calibrated and initial weights at micro-level: a first group (‘Water’, ‘U2-Manufacture, Energy’ and ‘U3-Transport, Utilities, Residential’ ,‘LUE’) is characterized by high variability and skewness to the left; an intermediate group (‘Wetland’, ‘U4-Unused and Abandoned areas’, ‘Settlement’ and ‘LUD’) are also skewed to the left, and finally the remaining variables with a smaller variability that have both left and right skewness.

Table 5: Parameters of the distribution of the difference between calibrated and initial weights – Land Cover, Land Use, FAO, Settlement, LUE and LUD – 2018 survey

LUCAS Variables	mean	p99	p1	max	min	Skewness	standard deviation	Frequency
Land Cover								
A - Artificial land	0.1	4.9	–4.5	85	–62	–0.03	2.3	21 021
B - Cropland	0.0	1.4	–1.3	109	–120	6.4	1.0	86 528
C - Woodland	–0.08	2.9	–4.0	117	–203	–21.12	1.5	119 790
D - Shrubland	–0.06	3.1	–4.1	93	–64	10.8	2.1	19 054
E - Grassland	0.0	2.9	–2.6	101	–164	–20.62	1.8	72 661
F - Bareland	–0.01	2.8	–2.8	76	–37	12.7	1.7	7 964
G - Water areas	3.4	88.9	–59.0	2 860	–1107	24.9	71.8	2 567
H - Wetlands	–0.01	6.4	–5.8	151	–125	7.2	4.0	6 670
Land Use								
U1 - Agriculture, Mining, Fishing	–0.04	2.1	–2.7	374	–203	37.8	1.6	251 929
U2 - Manufacture, Energy	0.9	10.2	–9.0	696	–68	31.8	20.8	1 174
U3 - Transport, Utilities, Residential	0.3	6.3	–5.2	2 860	–1107	117.1	17.6	34 613
U4 - Unused and Abandoned areas	0.0	3.7	–4.4	696	–175	74.1	6.8	48 539
FAO forest								
FAO 1 - Forest	–0.09	2.6	–4.0	45	–203	–31.23	1.4	111 460
FAO 2 - Other Wooded Land	–0.07	3.0	–4.0	93	–64	10.4	2.2	16 307
FAO 3 - Other Wooded Land no FAO	0.1	3.9	–3.6	117	–120	8.2	2.7	8 861
Settlement	0.2	5.4	–4.8	696	–164	90.0	4.8	34 545
LUE Services and Residential area	0.3	6.5	–4.9	2 860	–1107	95.0	22.3	20 969
LUD Land Use with Heavy Environment Impact	0.2	6.9	–6.5	696	–97	59.5	7.3	15 887

The information from Table 5 can be visualized for the variables of the subsidiary categories of Land Cover and Land Use, in Figure 3. The y-axis reports the values of the differences, where values greater than zero are associated with sample units, for which $w_k > d_k$ while, below zero, one finds the units with $w_k < d_k$.

Figure 3: Distribution of the differences between calibrated and initial weights – 2018 survey



5 Comparing calibrations with an enlarged set of constraints and analysing the differences between calibrated and HT estimates in LUCAS 2009–2018

Two further studies were carried out in order to complete the analytical framework pertaining to the differences between the calibrated and the HT estimator: an analysis of the effects of an enlargement of the calibration model and the influence of calibration on the trend of estimates. For that purpose, the microdata collected by LUCAS 2009, 2012 and 2015 were calibrated using the same calibration model but the closest auxiliary data in terms of time: the CLC/HRL data of the year 2012 for the 2009 and 2012 surveys, and the CLC/HRL data of the year 2018 for the 2015 and 2018 surveys.

5.1 Comparing different calibrations in the 2012 and 2015 surveys

The first analysis takes into account the 2012 and 2015 surveys. For each, two sets of estimates are available: the published data calibrated only by classes of elevation and the data obtained using the same calibration model as for the 2018 estimation, but with CLC/HRL data referred to the year 2012, as introduced above.

Thus, for each survey, the two sets of estimates are obtained by the same sample design with the same sample units that have the same initial probabilities but different calibration weights.

Comparing the estimates derived from the two data sets, it is possible to isolate the effect of the two calibration procedures, which differ due to the constraint in the first calibration being contained in the set of constraints of the second calibration.

Through including rounds 2012 and 2015 of the survey in this analysis, and with the aim of ensuring a homogeneous comparison between the two years in terms of countries involved, the scope of coverage was slightly reduced to 27 Member States, rather than the 28 included in the 2018.

Table 6 reports the estimates of the variables Land Cover, Land Use, FAO Forest, and 'Settlement', which derive from the old and from the new calibration, together with their percentage ratios. Generally, for both years, a broader specification of constraints caused the same direction in higher/lower estimates, except in the case of Land Use 'U2-Manufacture,Energy', for which a high difference is found (a ratio of 109 % in 2015 compared to 99 % in 2012). The differences in the ratios, between the old and the new calibration, also show that, in a majority of cases, they remain at about the same level with few exceptions (Land Cover 'Bareland'([7]) and 'Wetland', FAO variables). This confirms that a larger specification of constraints produces higher/lower estimation profiles that are probably stable over the different rounds of the Lucas survey.

([7]) The definition of 'Bareland' has been modified between 2012 and 2015.

Table 6: Comparison between calibrations with different sets of constraints – 2012 and 2015 Lucas surveys

LUCAS Variables	old calibration estimates		new calibration estimates		total area ratio: new/old *100	
	2012	2015	2012	2015	2012	2015
Land Cover						
A - Artificial land	174 061	181 043	173 824	180 083	99.9	99.5
B - Cropland	981 777	960 223	1 056 448	1 061 582	107.6	110.6
C - Woodland	1 596 681	1 621 941	1 619 695	1 634 079	101.4	100.8
D - Shrubland	300 813	304 071	288 951	291 356	96.1	95.8
E - Grassland	932 779	895 664	867 832	827 330	93.0	92.4
F - Bareland	123 894	145 503	97 374	110 125	78.6	75.7
G - Water areas	131 928	131 650	134 951	134 841	102.3	102.4
H - Wetlands	70 893	72 730	73 815	73 495	104.1	101.1
Land Use						
U1 - Agriculture, Mining, Fishing	3 264 258	3 219 645	3 280 139	3 232 962	100.5	100.4
U2 - Manufacture, Energy	16 468	14 588	16 334	15 912	99.2	109.1
U3 - Transport, Utilities, Residential	392 779	409 053	368 172	388 814	93.7	95.1
U4 - Unused and Abandoned areas	639 322	669 540	648 245	675 203	101.4	100.9
FAO						
FAO 1 - Forest	1 511 675	1 558 127	1 412 701	1 420 904	93.5	91.2
FAO 2 - Other Wooded Land	226 067	236 573	294 917	321 384	130.5	135.9
FAO 3 - Other Wooded Land no FAO	144 891	145 912	135 576	139 655	93.6	95.7
Settlement	309 917	324 181	304 004	314 294	98.1	97.0

5.2 Differences between calibrated and HT estimates in LUCAS 2009–18

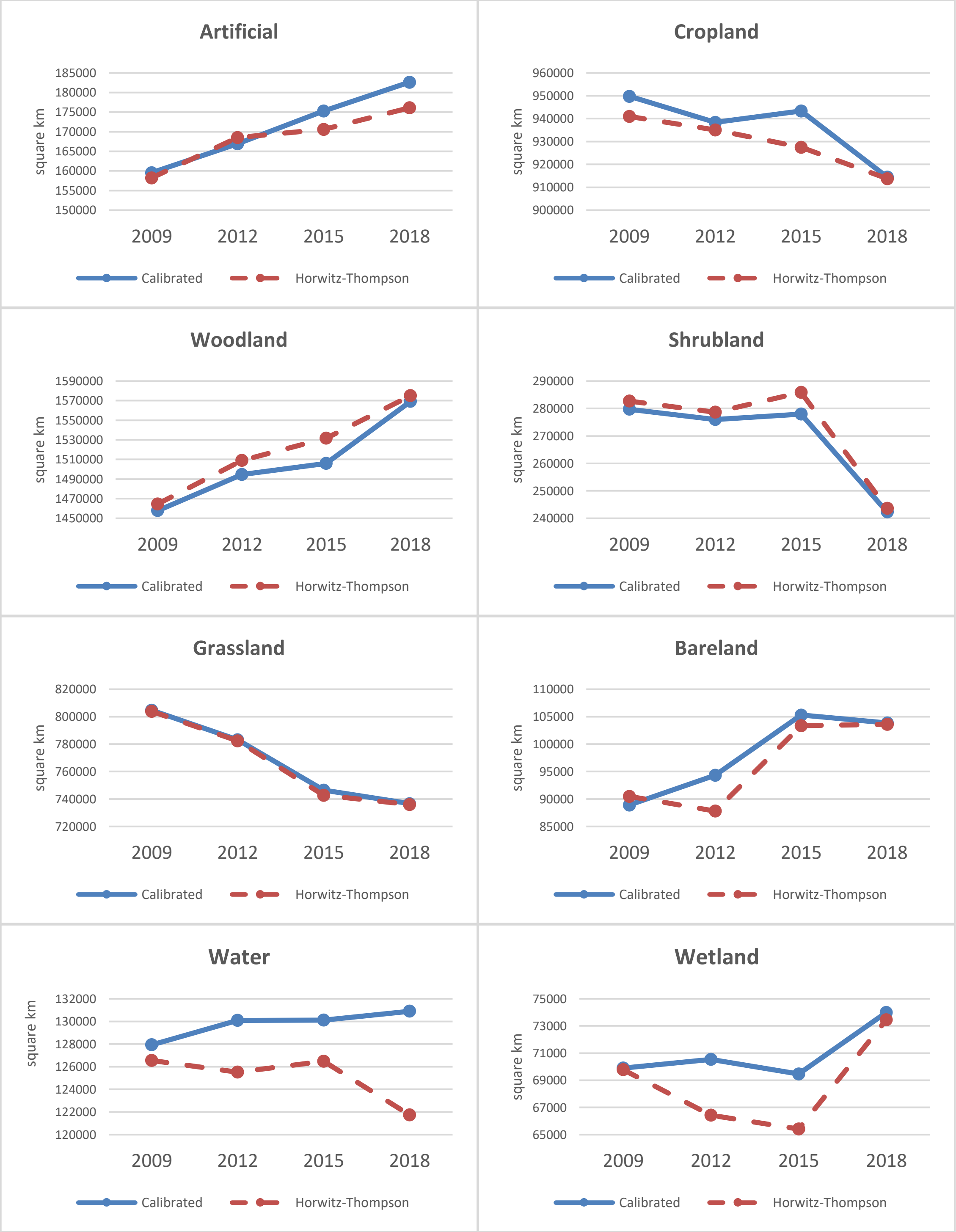
In terms of the impact of the calibration procedure on the LUCAS estimates, no clear evidence of differences could be found in the elaborations performed, between trends observed in EU estimates, depending on the number of Member States (23, 27 and 28). Therefore, the following considerations are valid in general for the three groups of estimates and they will refer to the 23 EU Member States that figure in all rounds of the LUCAS survey. The results are presented in the following plots depicting the trends over the period 2012–18 in Land Cover and Land Use estimates calculated by calibration and by the HT estimator.

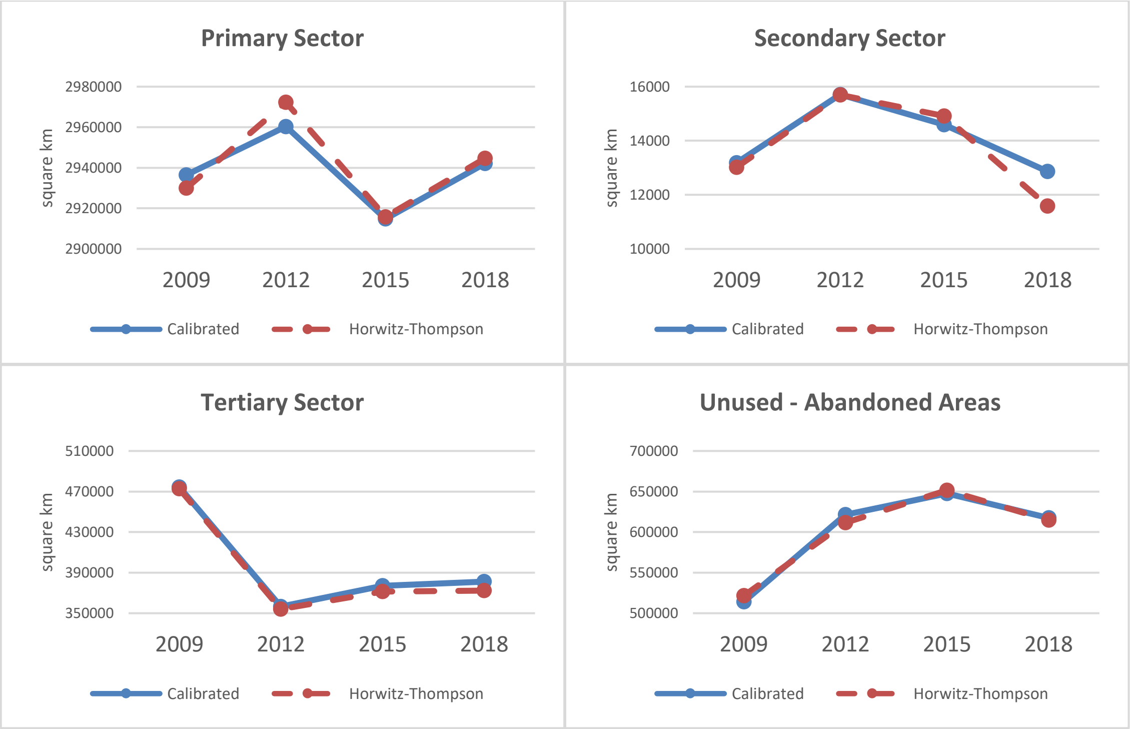
The estimates most affected by the calibration procedure are those related to Land Cover, rather than Land Use.

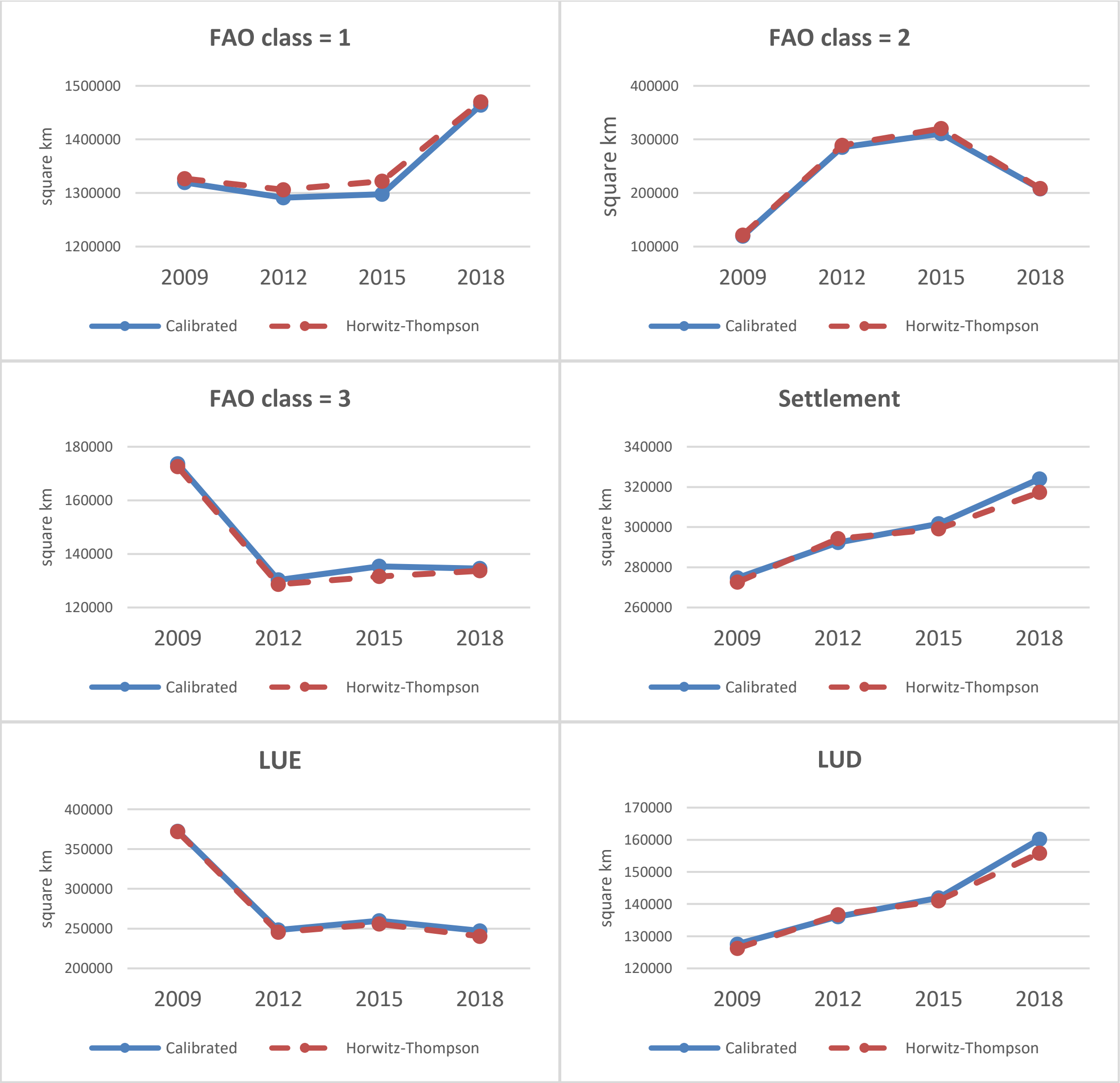
Under Land Cover, the value 'Grassland' is the one least affected by the calibration, whereas the HT and calibration estimates are more or less the same in each round of the survey. The case is different for 'Water areas', the estimates of which diverge in all rounds. An intermediate pattern is observed in the other cases:

- The estimates of 'Artificial' coincide in 2009 and in 2012, yet they diverge in 2015 and especially in 2018.
- The estimates of 'Cropland' diverge in the three rounds preceding 2018, when they then coincide.
- The estimates of 'Woodland', 'Shrubland', 'Bareland', and 'Wetlands' diverge in central rounds (2012 and 2015), the former two slightly, the latter two in a much more pronounced way.
- The estimates of Land Use only present two noticeable divergences: one in the case of 'U1-Agriculture. Mining, Fishing' in 2012 and one in that of 'U2-Manufacture, Energy' in 2018.
- The other groups of estimates related to 'Settlement', 'FAO classes', 'LUE', and 'LUD' do not display significant divergences between HT and calibration estimates, for the EU23.

Figure 4: trend of the estimates produced by calibrated and HT estimator in the years 2009-2018 for the variables Land Cover, Land Use, FAO Forest, Settlement, LUE and LUD – EU23




Primary Sector
square km
2980000
2960000
2940000
2920000
2900000
2009
2012
2015
2018
Calibrated
Horwitz-Thompson
Secondary Sector
square km
16000
14000
12000
10000
2009
2012
2015
2018
Calibrated
Horwitz-Thompson
Tertiary Sector
square km
510000
470000
430000
390000
350000
2009
2012
2015
2018
Calibrated
Horwitz-Thompson
Unused - Abandoned Areas
square km
700000
650000
600000
550000
500000
2009
2012
2015
2018
Calibrated
Horwitz-Thompson


FAO class = 1
square km
1500000
1400000
1300000
1200000
2009
2012
2015
2018
Calibrated
Horwitz-Thompson
FAO class = 2
square km
400000
300000
200000
100000
2009
2012
2015
2018
Calibrated
Horwitz-Thompson
FAO class = 3
square km
180000
160000
140000
120000
2009
2012
2015
2018
Calibrated
Horwitz-Thompson
Settlement
square km
340000
320000
300000
280000
260000
2009
2012
2015
2018
Calibrated
Horwitz-Thompson
LUE
square km
400000
350000
300000
250000
200000
2009
2012
2015
2018
Calibrated
Horwitz-Thompson
LUD
square km
170000
160000
150000
140000
130000
120000
2009
2012
2015
2018
Calibrated
Horwitz-Thompson

6 Conclusion

Calibration is a method to inflate the sampling data by incorporating available auxiliary information deriving from an external source into the estimation process in order to improve the accuracy of estimates. Calibration changes the initial weights in such a way as to equate the known totals of the auxiliary variables in the external source with the same totals calculated from the sampling data; the prerequisite of the method is the availability of the auxiliary information also for the sampling units.

Although the calibration methodology is well established, its implementation may consider possible alternatives. Such alternatives mainly concern the choice of the so-called ‘known totals’. The ‘known totals’ form a set of reliable statistics that are used to improve the accuracy of the estimates and/or to place the survey estimates in a ‘framework’ or ‘context’ facilitating their evaluation or their comparison.

In the case of LUCAS 2018, it seemed rational that the construction of this type of framework should be based on the statistics produced by the Copernicus project (in particular by the High Resolution Layers - HRLs) available for 2018, which is with the same reference period as that of the LUCAS data. This choice can be motivated both by the homogeneity in the process of constructing the HRLs across EU Member States, and because they are correlated with the main variables of Land cover and Land use observed by LUCAS, even if the corresponding definitions are slightly different.

Thus, in the LUCAS 2018 survey, the calibration methodology had been further developed, when compared to the previous survey rounds. While, in the 2009–2015 surveys, only classes of elevation were taken into account, in the latest survey, six other variables obtained from CLC and HRL data are also used, namely HRL ‘imperviousness’ and CLC ‘artificial’, ‘woodland’, ‘agriculture’, ‘wetland’, and ‘water’. In that way, the estimation of LUCAS variables profits from the correlations between those variables and the ones collected.

Another important reason for using Copernicus output is to lay the foundations for a stricter integration between the two data sources in the future, by specializing their functions and taking advantage of their particularities. The strength of LUCAS is certainly in the field observation, which is obviously not replaceable when material samples need to be collected (e.g. soil samples) or if the necessary details are not observable through the current capacity of remote observations. On the other hand, the accuracy of data produced by LUCAS could significantly be improved by the information deriving from Copernicus and its availability in the Master. Moreover, if and when it will be possible to align the definitions (e.g. for Land Cover) in both sources, the estimated LUCAS totals could be made to match those from Copernicus through a calibration that increases the reliability of the statistics produced.

In this report, following a brief introduction of the basic concept of calibration and the description of the actual calibration operated in the 2018 survey, a number of analyses are discussed. These compare estimation by calibration and by the HT estimator, based on the distributions of weights, the produced estimates, and the differences in weights calculated at the micro-level. Moreover, a simulation was performed, in which the data of LUCAS 2009–2015 rounds were calibrated by applying the same procedure as in 2018. Variable trends were analysed, as was also the enlargement of the set of known totals in the 2012 and 2015 surveys.

Regarding the 2018 calibration, its impact, compared with the HT estimator, is differentiated among the considered variables. More than half of the variables present a light or moderate impact of calibration, while a significant impact can be found for the modalities ‘Water’ of Land Cover and the ‘U2-Manifacture,energy’ of Land Use.

Similar effects are observed with gains in precision. For ten variables, the sampling errors of the calibrated estimates are lower than those of the HT estimates, while the remaining variables display a moderate or more significant increase in terms of absolute differences between the CVs.

The distributions of calibrated and initial weights are similar for most of the variables. Only in the case of Land Cover 'water', Land Use 'U2-manufacture, energy' and 'U3-transport, utilities, residential', 'settlement', and 'LUE', do the distributions of calibrated weights show a larger variability. Considering the distributions of the differences of the two sets of weights at micro-data level, which the impact of calibration on estimates depends on, the presence of outliers is detected for most of the variables. Almost half of them display strong variability, when looking at the range and the standard deviation ('water' in Land Cover, 'U2-manifacture,energy', 'U3-transport, utilities, residential' and 'unused and abandoned land' in Land Use, 'Settlement', 'LUE' and 'LUD').

The simulation carried out on the variations of LUCAS estimates over time (2009–18) highlights that calibrated estimates generally present a smoother trend and that they do not diverge significantly from the HT estimates.

Nevertheless, some variables, especially those related to Land Cover do diverge in some rounds of the survey, while the estimates for 'water' diverge in all rounds. The simulation aiming to measure the effects of a calibration procedure with an enlargement of known totals in the same survey (2012 and 2015 rounds) shows that the introduction of more known totals, between the two calibrations, causes observed divergence to progress in the same direction.

7 References

Ballin M., Barcaroli G., Masselli M., Scarnó M. (2018), ‘Redesign sample for Land Use/Cover Area frame Survey (LUCAS) 2018’, *Statistical Working Papers,* EUROSTAT.

D'Andrimont R., Verhegghen A., Meroni M., Lemoine G., Strobl P., Eiselt B., Yordanov M., Martinez-Sanchez L., and van der Velde M. (2021), ‘LUCAS Copernicus 2018: Earth-observation-relevant in situ data on land cover and use throughout the European Union’, *Earth Syst. Sci. Data*, Vol. 13, pp. 1119–1133, https://doi.org/10.5194/essd-13-1119-2021.

Guandalini A., Tillé Y. (2015), ‘Design-based Estimators Calibrated on Estimated Totals from Multiple Surveys’, *International Statistical Review*, Volume 85, Issue 2, pp. 250–269 https://doi:10.1111/insr.12160.

Särndal C. E. (2007), ‘The calibration approach in survey theory and practice’, *Survey Methodology*, December, Vol. 33 No. 2, Statistics Canada, pp 99–119.

Zardetto D. (2015), ‘ReGenesees: an Advanced R System for Calibration, Estimation and Sampling Error Assessment in Complex Sample Surveys’, *Journal of Official Statistics*, 31(2), pp. 177–203.

8 Annex: Selected tables by countries

Table A.1: percentage ratio between areas calculated by calibrated and initial weights – Land Cover, 2018 Survey

	A	B	C	D	E	F	G	H
BE	104,7	99,4	99,5	101,8	99	100,3	91,1	141,6
BG	98,6	99,8	99,8	99,6	100,9	96,5	108,5	98,2
CZ	105,9	99,1	100,5	99,7	99,6	99	98,7	99,3
DK	102,1	99,6	99,7	103,1	100,5	99,2	100,1	96,7
DE	100,3	99,6	99,7	103,9	100,5	99,5	100,6	116,9
EE	102,5	100,2	101,1	100,8	100,7	99,6	85,3	100,2
IE	98,7	99,2	98,1	102,4	99,5	102,7	105,5	105,6
EL	100,8	100,4	99,3	99	100,2	106,5	112,2	108,5
ES	98,7	100,4	99,7	99,6	100,2	100,2	104,7	101,4
FR	101,5	99,8	99	99,6	100,7	99,9	108,4	127,7
HR	103,1	100,9	99,5	99,8	99,9	101,2	100,2	104,1
IT	101	101,3	99	99	99,8	98,7	101,2	106
CY	98,7	100,1	99,9	100	100,3	100,6	102,3	99,9
LV	103,4	98,2	100,4	100,3	99,1	100,3	101,8	106
LT	99,2	99,3	101	100,5	99,1	99,8	105,3	94
LU	101	98,4	99,1	106,7	101,5	119	83,8	-
HU	104,4	100,1	99,1	99,9	100,4	100	88,8	115
MT	108,6	98,5	77,4	93,5	97,6	111,7	101,9	-
NL	99,4	95	98,2	98,5	97,1	92,2	139,5	86,2
AT	99,9	98,9	100,1	100,3	99,9	100,2	110,5	98,7
PL	99,6	100,3	99,7	99,9	99,8	98,8	100	106,1
PT	100,5	99,7	100,1	99,2	99,5	99,3	104,1	173,2
RO	101,4	100,2	100	99,9	99,9	99,8	97,2	98,2
SI	113,5	97,1	99,8	102,6	99,3	101,3	98	104,4
SK	98,8	99,5	101,1	99,5	99,8	99,6	81,1	101,3
FI	99,7	101,9	98,3	96,6	99	99,7	117,3	94,8
SE	115,5	99,2	99,4	101	102,2	99,9	99,6	100,5
UK	99,7	99,6	99,6	99,1	99,4	99,5	149,1	93,7
Total	101,2	100	99,4	99,6	100	99,9	106,8	99,9

Table A.2: percentage ratio between areas calculated by calibrated and initial weights – Land Use, 2018 Survey

	U1	U2	U3	U4
BE	99,3	106,6	101,9	102,6
BG	100,2	107	99,6	99,1
CZ	99,8	102,5	102,2	99,9
DK	99,5	102,6	102,8	98,6
DE	99,8	106	100,6	102,7
EE	100,8	101,8	91,6	101,3
IE	99,8	105,9	97,6	101,6
EL	99,9	108,2	102,1	99,8
ES	100,1	97,2	100,7	99,7
FR	99,7	99	101,6	100,5
HR	99,9	99,1	102	99,8
IT	100,1	104,5	100,6	99,4
CY	100,1	101,4	99,4	100
LV	99,6	112,3	102,5	101,9
LT	100	97,8	102	98
LU	99,1	86,5	108,2	97,2
HU	99,8	96,7	99	103,8
MT	100,6	-	100,8	98,4
NL	96	93,5	109,7	99,5
AT	99,7	116,2	101,9	100,2
PL	100	99,7	99,4	100,3
PT	99,9	105,6	100,8	99,7
RO	100	98,5	101,5	99,3
SI	99,2	112,7	112,7	100,2
SK	100,3	98,4	96,4	100
FI	98,7	210,1	108	99,7
SE	99,1	115,1	106,7	99,8
UK	99,3	101,9	101	101,2
Total	99,7	107,9	102,3	100

Table A.3: percentage ratio between areas calculated by calibrated and initial weights – FAO forest, settlement, LUE services and residential area and LUD Land Use with heavy environmental impact, 2018 Survey

	FAO - Forest	FAO - Other wooded land	FAO - Other wooded land no FAO	Settlement	LUE - services and residential area	LUD - Land Use with heavy environment impact
BE	99,9	102,6	97,3	102,4	102,8	99,8
BG	99,9	100,2	98,5	99,8	100,2	100,3
CZ	100,6	99,8	98,9	102,6	101,6	102,9
DK	99	100,2	103,2	102,5	103,8	100,7
DE	99,7	104	99,5	100,8	100,5	101,3
EE	101,2	100,7	100,3	100,7	89,5	100,5
IE	98,2	102,6	82,2	99,4	96,2	104,6
EL	99,2	99	99,8	102	101,6	102,8
ES	99,6	99,6	100,4	100,1	100,4	100,5
FR	98,8	99,5	100,5	101,2	101,5	102,5
FR	99,4	99,7	101	102	102	101,8
IT	98,9	99,2	100,4	101,3	100,8	100,6
CY	99,9	100	100	99,4	99	100
LV	100,2	102,1	100,3	102,8	102,8	103
LT	101,1	98,6	99,9	98,9	103,6	95,8
LU	98,6	98,3	96,9	107,1	115,5	101,6
HU	98,9	101,2	99,7	102,6	99,7	99,4
MT	69,8	94	85	104,2	102,3	95,6
NL	98,1	99,4	99,6	98,1	111	107
AT	100,1	100,3	101	101,2	102,2	102,8
PL	99,8	100,4	99,3	98,9	99,3	99,9
PT	100,3	99,1	99,2	101	100,8	102,9
RO	100	99,7	100,5	101,4	102,4	100,4
SI	99,8	102,1	100,7	113,7	111,8	112,9
SK	101	99,8	97,3	97,8	100,5	92,3
FI	98,2	96,1	102	105,9	108,8	110,6
SE	99,3	100,3	102	115,9	105,5	111
UK	99,2	99,2	106,3	99,6	101,9	98,6
Total	99,4	99,5	100,3	101,8	102,6	102,3

Table B.1: percentage differences between the averages of calibrated and initial weights $(ave_{cal}-ave_{ini})ave_{cal}*100$ – Land Cover, 2018 Survey

	A	B	C	D	E	F	G	H
BE	4,49	-0,58	-0,51	1,74	-0,99	0,34	-9,72	29,40
BG	-1,37	-0,17	-0,20	-0,36	0,86	-3,66	7,86	-1,79
CZ	5,60	-0,93	0,46	-0,30	-0,38	-1,03	-1,33	-0,70
DK	2,10	-0,42	-0,26	2,97	0,47	-0,79	0,07	-3,37
DE	0,30	-0,39	-0,34	3,80	0,47	-0,46	0,61	14,42
EE	2,46	0,18	1,10	0,75	0,69	-0,38	-17,20	0,23
IE	-1,34	-0,83	-1,95	2,32	-0,46	2,59	5,19	5,27
EL	0,75	0,43	-0,74	-1,01	0,24	6,07	10,90	7,81
ES	-1,28	0,42	-0,28	-0,36	0,19	0,18	4,48	1,39
FR	1,47	-0,25	-0,98	-0,45	0,68	-0,06	7,74	21,67
HR	2,97	0,91	-0,52	-0,21	-0,11	1,23	0,20	3,93
IT	0,96	1,24	-1,01	-0,98	-0,24	-1,30	1,16	5,65
CY	-1,30	0,09	-0,11	0,00	0,34	0,61	2,20	-0,13
LV	3,32	-1,79	0,36	0,27	-0,87	0,28	1,73	5,65
LT	-0,84	-0,75	1,04	0,46	-0,91	-0,24	5,04	-6,43
LU	1,03	-1,67	-0,87	6,29	1,49	15,95	-19,39	
HU	4,21	0,11	-0,93	-0,06	0,43	0,01	-12,58	13,05
MT	7,95	-1,54	-29,22	-6,99	-2,47	10,49	1,89	
NL	-0,63	-5,21	-1,84	-1,56	-2,95	-8,42	28,34	-16,03
AT	-0,11	-1,15	0,06	0,29	-0,12	0,17	9,48	-1,31
PL	-0,41	0,31	-0,25	-0,13	-0,16	-1,20	0,00	5,78
PT	0,50	-0,29	0,12	-0,80	-0,49	-0,72	3,90	42,27
RO	1,42	0,20	-0,01	-0,14	-0,09	-0,25	-2,87	-1,84
SI	11,88	-3,01	-0,22	2,55	-0,72	1,25	-2,04	4,17
SK	-1,19	-0,54	1,04	-0,55	-0,20	-0,40	-23,32	1,26
FI	-0,34	1,84	-1,76	-3,56	-0,98	-0,31	14,78	-5,45
SE	13,39	-0,82	-0,58	0,97	2,11	-0,13	-0,45	0,46
UK	-0,33	-0,44	-0,36	-0,87	-0,55	-0,55	32,93	-6,73
Total	1,19	0,00	-0,57	-0,41	0,02	-0,10	6,35	-0,06

Table B.2: percentage differences between the averages calculated by calibrated and initial weights (ave_{cal}-ave_{ini})ave_{cal}*100 – Land Use, 2018 Survey

	U1	U2	U3	U4
BE	-0,73	6,21	1,88	2,53
BG	0,15	6,56	-0,38	-0,89
CZ	-0,21	2,41	2,19	-0,15
DK	-0,47	2,55	2,72	-1,41
DE	-0,24	5,64	0,59	2,60
EE	0,83	1,75	-9,23	1,26
IE	-0,24	5,62	-2,42	1,60
EL	-0,13	7,59	2,05	-0,18
ES	0,06	-2,86	0,70	-0,26
FR	-0,25	-1,04	1,60	0,46
HR	-0,05	-0,88	1,96	-0,20
IT	0,12	4,31	0,59	-0,64
CY	0,08	1,42	-0,64	0,04
LV	-0,43	10,97	2,40	1,89
LT	-0,02	-2,22	1,93	-2,00
LU	-0,95	-15,63	7,61	-2,92
HU	-0,17	-3,36	-1,00	3,63
MT	0,61		0,80	-1,61
NL	-4,21	-6,92	8,81	-0,49
AT	-0,29	13,94	1,87	0,20
PL	0,04	-0,34	-0,58	0,27
PT	-0,07	5,27	0,83	-0,26
RO	-0,01	-1,50	1,44	-0,76
SI	-0,84	11,28	11,23	0,22
SK	0,31	-1,63	-3,69	-0,02
FI	-1,31	52,40	7,44	-0,29
SE	-0,94	13,10	6,28	-0,22
UK	-0,66	1,82	0,97	1,23
Total	-0,31	7,34	2,22	0,04

Table B.3: percentage differences between the averages calculated by calibrated and initial weights $(ave_{cal}-ave_{ini})ave_{cal}*100$ – FAO forest, settlement, LUE services and residential area and LUD Land Use with heavy environment impact, 2018 Survey

	FAO - forest	FAO - other wooded land	FAO - other wooded land not FAO	settlement	LUE - services and residential area	LUD - Land Use with heavy environment impact
BE	-0,12	2,54	-2,79	2,36	2,72	-0,20
BG	-0,12	0,20	-1,55	-0,20	0,25	0,27
CZ	0,55	-0,18	-1,07	2,52	1,60	2,85
DK	-0,99	0,22	3,11	2,48	3,68	0,73
DE	-0,27	3,80	-0,51	0,76	0,48	1,26
EE	1,17	0,74	0,33	0,66	-11,79	0,52
IE	-1,87	2,58	-21,67	-0,58	-3,90	4,44
EL	-0,77	-0,98	-0,21	2,01	1,59	2,75
ES	-0,39	-0,39	0,40	0,13	0,37	0,48
FR	-1,19	-0,50	0,47	1,18	1,43	2,47
HR	-0,57	-0,29	1,00	1,92	1,98	1,81
IT	-1,09	-0,79	0,35	1,25	0,78	0,58
CY	-0,12	0,03	-0,01	-0,65	-0,99	-0,01
LV	0,25	2,09	0,27	2,77	2,72	2,88
LT	1,09	-1,37	-0,08	-1,08	3,45	-4,38
LU	-1,38	-1,73	-3,20	6,65	13,44	1,61
HU	-1,10	1,18	-0,29	2,58	-0,30	-0,57
MT	-43,29	-6,34	-17,68	3,99	2,28	-4,57
NL	-1,93	-0,61	-0,39	-1,94	9,93	6,58
AT	0,07	0,31	0,95	1,15	2,16	2,70
PL	-0,19	0,42	-0,72	-1,14	-0,71	-0,06
PT	0,26	-0,91	-0,77	0,99	0,77	2,84
RO	-0,03	-0,31	0,50	1,43	2,37	0,45
SI	-0,22	2,02	0,70	12,08	10,58	11,41
SK	1,03	-0,21	-2,79	-2,28	0,52	-8,38
FI	-1,81	-4,02	2,01	5,57	8,06	9,62
SE	-0,66	0,29	1,94	13,70	5,20	9,91
UK	-0,81	-0,76	5,91	-0,41	1,85	-1,43
Total	-0,63	-0,50	0,32	1,76	2,50	2,25

GETTING IN TOUCH WITH THE EU

In person

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

On the phone or by email

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

FINDING INFORMATION ABOUT THE EU

Online

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

EU publications

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

EU law and related documents

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

Open data from the EU

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

KS-TC-23-002-EN-N

Calibration in LUCAS 2018 Survey

LUCAS survey provides harmonised statistics on Land Use and Land Cover across the European Union. The LUCAS survey is used to monitor the land cover, the social and economic use of land, the biodiversity and other environmental parameters. The 2018 survey is based on a sample of 336000 points selected from the LUCAS 2x2 km2 grid which is the list of geo-referenced of about one million points in the EU.

The paper describes the calibration technique, the way it has been applied in LUCAS 2018 and the impact of calibration on the estimates.

For more information

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

Publications Office
of the European Union

ISBN 978-92-68-03119-3