Use of auxiliary information in the sampling strategy of a European area frame agro-environmental survey

Laura Martino[1], Alessandra Palmieri[1] & Javier Gallego[2]
(1) European Commission: DG-ESTAT
(2) European Commission: DG -JRC
Email: laura.martino@ec.europa.eu

Keywords: area frame, auxiliary information, survey sampling, donor

Introduction

This paper addresses the use of auxiliary information for improving the efficiency of the sampling design. The case study is represented by an area frame survey carried out by Eurostat where auxiliary information is available only for some of the survey domains.

In the first section a short description of the European survey is provided and the motivations for the research activity are presented. In the second section the main pillars of the sampling design strategy and the reasons for using auxiliary information are described. The main activities foreseen for the future are mentioned at the end of the last section.

LUCAS: a European area frame agro-environmental sample survey

The European "Land Use/Cover Area frame statistical Survey" (LUCAS) is based on an area-frame sampling scheme. It aims to inform decision makers and the general public about coverage and management of the European territory and their changes. Agro-environmental parameters are investigated too.

During the pilot phase of the survey (started in 2001) the sampling strategy has moved from a one-phase cluster design to a two-phase sampling method with stratification (from 2006 onwards).

The LUCAS first phase sample is a systematic sample with points spaced 2 km in the four cardinal directions covering all European territory (EU) except Cyprus and Malta. It included a total of 958,325 points. Each point of the first phase sample was photo-interpreted and assigned to one of the 7 pre-defined land cover strata. The results of the stratification activity, conducted in 2005, are reported in Table 1.

Table 1: Stratification results

First phase sample		
	Area in %	**Variation Coefficient (%)**
Arable land	25.18	0.008
Permanent crops	2.94	0.01
Grassland	16.65	0.009
Woodland and shrubland	45.87	0.006
Bare land	2.06	0.01
Artificial land	4.09	0.01
Water	3.21	0.01

From the stratified first phase sample, a sub-sample of points (field sample) was extracted in 2006 to be classified during field visit according to the full land nomenclature (for a detailed description of the methodology adopted in the pilot phase of the survey see Martino & Fritz, 2008).

Moving from a single-phase clustered design (2001-2003 rounds) to a double-phase stratified design (2006 round) allowed a significant improvement in the efficiency of estimates (Gallego, 2007). Nonetheless in 2009 the need for a further enhancement in the sampling design came out for various reasons.

The main drawbacks of the sampling design adopted until 2006 were the imbalance of the strata size (the agricultural strata were over-represented) and the geographical detail focused only at EU level.

In 2009, being the first official round of the LUCAS survey covering all the EU countries (except Malta and Cyprus), the focus of the survey changed from a merely agricultural to a broadly agro-environmental one. Taking into account also the users needs for more geographically detailed figures, Eurostat was forced to further fine-tune the sampling strategy using as much as possible all the available auxiliary information.

Main pillars of the new sampling strategy

The new sampling strategy had to cope with the bond of providing: sufficiently precise estimates at NUTS1 (Eurostat, 2008) level and more, satisfactory precision for the main land cover classes and longitudinal data on land cover and land use (panel approach). These constraints affected both the design and the selection of the sample.

Sampling design

In compliance with the users' requests, a sample at NUTS2 level has been designed. The regions (NUTS2) were then split into two groups according to their total area:

1. group A: NUTS2 with total area below or equal to 500 km2;
2. group B: NUTS2 with total area above 500 km2.

Within the regions with an area above the threshold, those belonging to the 11 countries already surveyed in 2006 (group B1) were distinguished from those belonging to the remaining countries (group B2). For the first group of regions some auxiliary information, based on the previous round of the survey, was available, while missing for group B2.

A different sampling strategy has been adopted in the various groups.

For group A, due to the limited extension of the regions, no precision has been fixed and an allocation to strata proportional to their size has been adopted.

For regions belonging to group B1, auxiliary information from the 2006 round played an important role. A sampling scheme based on multivariate optimal allocation (Bethel, 1989) was devised taking into account a set of land cover classes; upper-bounds for the coefficient of variation (% values) were fixed based on the experience gained in 2006 (Table 2).

For group B2, no information was available from 2006 LUCAS survey, thus a different sampling strategy was adopted.

The land cover/use data collected within the Corine Land Cover (CLC) (EEA, 2006) were used as auxiliary information. CLC classes were grouped into 12 new ones. Based on this independent source, dissimilarity indexes were computed among regions belonging to group B1 and B2 according to the L1 distance:

$$dabs(r,r') = \sum_c \left| x_{cr} - x_{cr'} \right|$$

Where x_{cr} is the area of CLC land cover group c in region r.

The absolute differences were preferred to the percentage ones to avoid small regions to become donors for very big regions (and the other way round).

The main target of this exercise was identifying, for each region of the countries where LUCAS was not carried out in 2006, the "closest" region in group B1 to be used as donor. Sampling rates by strata taken from the donors were then applied to the 'recipient' regions and treated as sub-optimal.

Table 2: Upper-bound of expected error by Land Cover classes

Land Cover classes	Upper-bound of the expected error
Cereals	15%
Root crops, Vegetables, floriculture, ornamental plants and strawberries	25%
Fibre and oleaginous crops, non permanent industrial crops	25%
Fodder and temporary grassland	25%
Permanent crops and nursery	25%
Grassland	7.5%
Broadleaved woodland	20%
Coniferous woodland	20%
Mixed woodland	20%
Shrubland	20%
Bare land	20%
Artificial areas	15%
Water	20%

Sample drawing

As a sample selection method a scheme maximizing the distance of the points, both in the same and in different strata was designed. This scheme was maintained from 2006 round since it appeared to be effective (Jacques & Gallego, 2005). Points sampled in different strata, indeed, can be close to each other and give some redundant information, as spatial correlation happens also between strata.

To reduce the autocorrelation within and between strata, the basic sampling grid (2km*2km) has been divided into squared (9 by 9 that means 18 Km by 18 Km each) blocks of 81 points each (Figure1). The set of points with the same relative position in the block is named a replicate. The numbering of the replicates is done under the constraint that the distance with the previous ones is maximized.

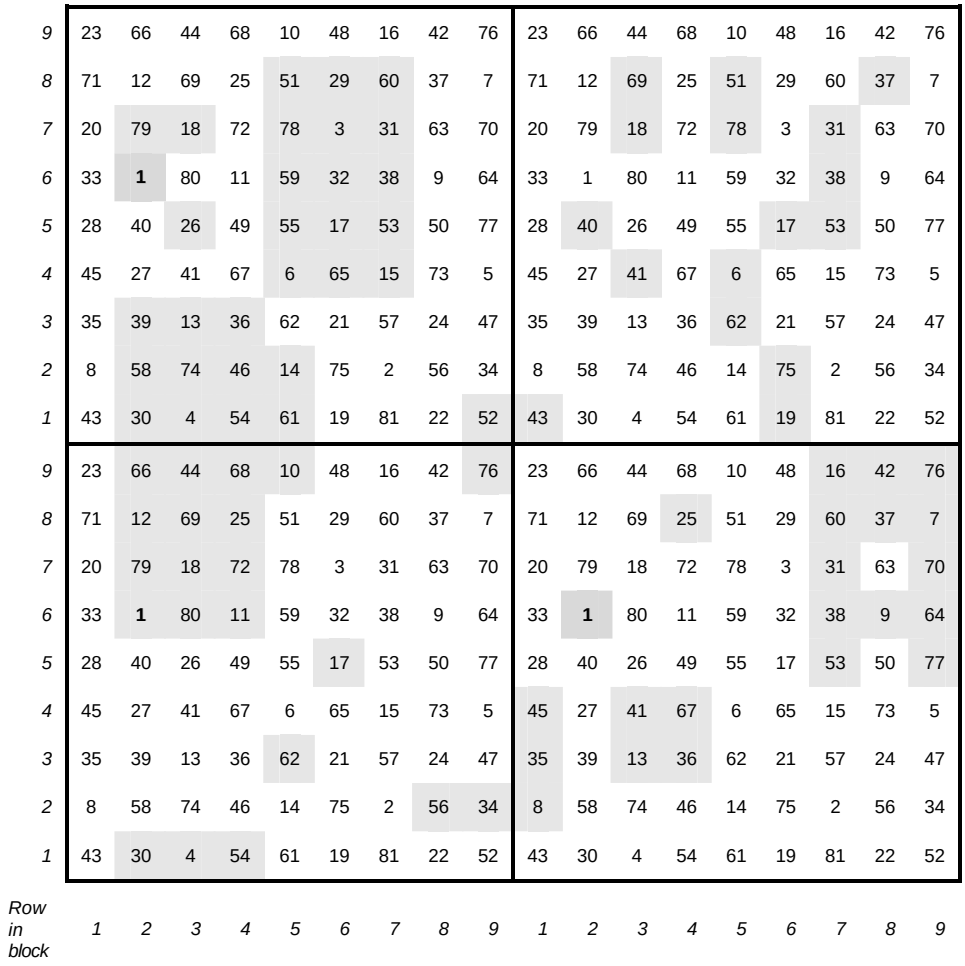
Replicates are then selected successively (starting with replicate 1) until the required sample size by domain is reached. From the replicate with the highest number, points are randomly selected.

The above described selection method was combined with a panel approach aiming to build time series of observations of the same points in different years and matrixes of transition for both Land Cover and Land Use over the time; in addition, a panel main sample allow targeted sub-samples for specific analysis. As a consequence LUCAS 2006 sample points were included as much as possible.

Finally points with an altitude above 1000m were considered inaccessible and excluded from the second phase sample.

The relative efficiency of the adopted sampling strategy versus a simple random sampling and a pure systematic sampling strategy will be computed as soon as results of the survey will be available.

Figure 1 – Sub-grid with replicates



9	23	66	44	68	10	48	16	42	76	23	66	44	68	10	48	16	42	76
8	71	12	69	25	51	29	60	37	7	71	12	69	25	51	29	60	37	7
7	20	79	18	72	78	3	31	63	70	20	79	18	72	78	3	31	63	70
6	33	**1**	80	11	59	32	38	9	64	33	1	80	11	59	32	38	9	64
5	28	40	26	49	55	17	53	50	77	28	40	26	49	55	17	53	50	77
4	45	27	41	67	6	65	15	73	5	45	27	41	67	6	65	15	73	5
3	35	39	13	36	62	21	57	24	47	35	39	13	36	62	21	57	24	47
2	8	58	74	46	14	75	2	56	34	8	58	74	46	14	75	2	56	34
1	43	30	4	54	61	19	81	22	52	43	30	4	54	61	19	81	22	52
9	23	66	44	68	10	48	16	42	76	23	66	44	68	10	48	16	42	76
8	71	12	69	25	51	29	60	37	7	71	12	69	25	51	29	60	37	7
7	20	79	18	72	78	3	31	63	70	20	79	18	72	78	3	31	63	70
6	33	**1**	80	11	59	32	38	9	64	33	**1**	80	11	59	32	38	9	64
5	28	40	26	49	55	17	53	50	77	28	40	26	49	55	17	53	50	77
4	45	27	41	67	6	65	15	73	5	45	27	41	67	6	65	15	73	5
3	35	39	13	36	62	21	57	24	47	35	39	13	36	62	21	57	24	47
2	8	58	74	46	14	75	2	56	34	8	58	74	46	14	75	2	56	34
1	43	30	4	54	61	19	81	22	52	43	30	4	54	61	19	81	22	52
Row in block	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*

References

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

EEA (2006) Global Land cover 2006 project.

Eurostat (2008) NUTS: Nomenclature of Territorial Units for Statistics.

Gallego F.J. (2007) Sampling efficiency of the EU point survey LUCAS 2006. *Paper presented at the 56th ISI session*, Lisbon , 22-29 September 2007.

Jacques, P. & Gallego, F.J. (2005) The LUCAS project – The new methodology in the 2005/2006 surveys. Agri-environment workshop: Belgirate, September 2005. http://forum.europa.eu.int/irc/dsis/landstat/info/data/index.htm

Martino L. & Fritz M. (2008) New insight into land cover and land use in Europe, *Statistics in Focus*, 33, Eurostat, Luxembourg