The use of a Point Sample as a Master Frame for Agricultural Statistics.

Francisco Javier Gallego
Joint Research Centre of the EC
Via Fermi 2749
I-21027 Ispra, Italy
Email: Javier.gallego@jrc.ec.europa.eu

ABSTRACT

Building an Area Sampling Frame for a specific survey has a cost that can be quite high. Resources can be optimised if a common sampling frame is used for several agricultural and environmental surveys. This may mean in particular re-using a multi-purpose stratification and graphical material, such as ortho-photographic documents that may have been produced.

In this paper we present the potential use of the Eurostat LUCAS survey (Land Use and Cover Area Frame Statistical survey) for additional purposes. In particular we assess the possible use for two purposes: sampling farms through points and surveys along the road.

Keywords: Area Sampling Frame; Master Frame; Multi-phase sampling

1. Introduction

Area sampling frames have been used for a long time by agricultural statistical services in a number of countries. The USDA June Agricultural Survey (JAS) of the US Department of Agriculture (USDA) is probably the longest and most important operational experience on area frame surveys (AFS). Several examples of agricultural area frame surveys around the world are reported in a two-volume report published by FAO (1996 1998); examples reported include: TER-UTI, run by the French Ministry of Agriculture, the first operational area frame survey in Europe, running since 1970 and fully operational since 1980, the ESYRCE survey in Spain (Ministerio de Agricultura, 2008), Morocco, the most stable area frame in Africa, and several examples in Latin America. The use of AFS for crop area estimation is well established and has important advantages

when list frames (e.g. agricultural censuses) are not properly updated or present a significant amount of missing farms (or households) or overlapping records.

Agricultural area frames may have two types of units: points or patches of land, often named segments. Segments may be delimited by physical bountaries or by a geometric pattern, such as a square grid, on a cartographic projection. The sample design of the USDA JAS is based on the delineation of Primary Sampling Units (PSUs) with physical boundaries that are subdivided into Secondary Sampling Units (SSUs) only if they are selected in the first sampling step (Cotter and Tomczac, 1994). The sampling method is well adapted to the US landscape, but might be too costly in areas with smaller irregular fields. Due to the high cost of setting up a sampling frame with physical boundaries, cheaper solutions have been adopted in the EU, including square sampling units and several types of point sampling methods (Gallego et al., 1994, Gallego and Delincé, 2010).

2. The LUCAS survey (Land Use/Cover Area-frame Survey)

In this paper we mainly focus on LUCAS (Land Use and Cover Area-Frame Survey), conducted by Eurostat, mainly for environmental monitoring.. LUCAS was first run in 2001 and 2003 with a systematic non-stratified sample based on clusters of 10 points (Gallego and Bamps, 2008). The sample design changed in 2006, moving from a clustered (two-stage) non-stratified sampling method to a two-phase sampling of unclustered points. The sampling scheme is inspired on the Italian AGRIT survey, with adaptations. The two phases of the sampling procedure are:

- A systematic grid of points is selected. One point every 2 km (around 1,100,000 points in the EU). The points are photo-interpreted with a symplified nomenclature on aerial ortho-photographs or the best available satellite images. This photo-interpretation provides the stratification, incomplete in the sense that not the whole population is stratified but only the grid that is the Master Sampling Frame of LUCAS (Gallego et al., 2010). Figure 1 shows the example of Austria. In this figure we have the visual impression of a complete stratification, but a zoomed image (Figure 2) shows that only one point every 2 km is stratified.

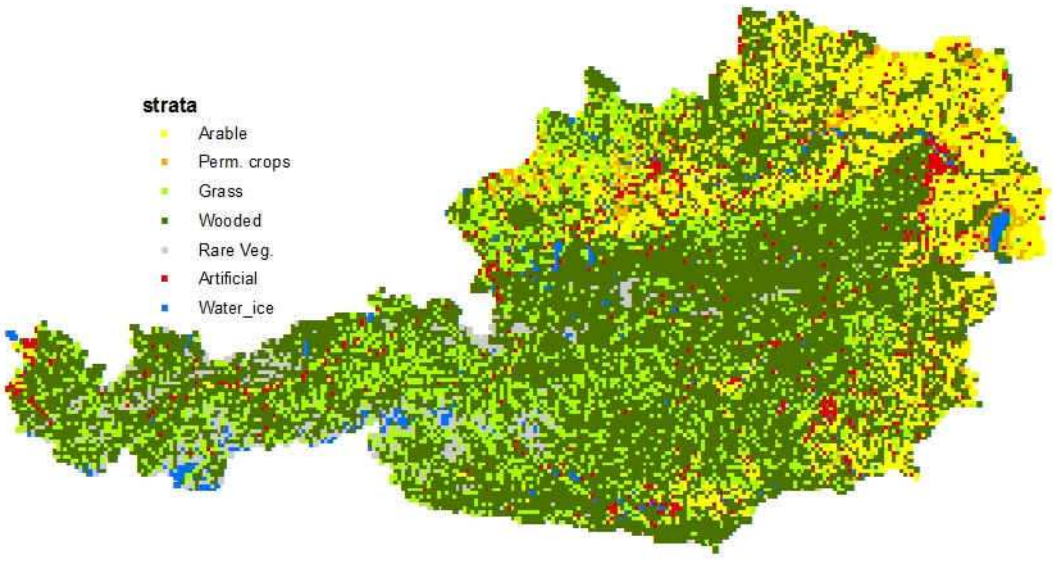

Figure 1: *Stratification by point photo-interpretation of Austria.*

Figure 2: *Stratification by point photo-interpretation in Western Austria. Zoomed view.*

- Points are subsampled in the second phase. The sampling rate in each stratum is tuned depending on the priority of the survey and the target accuracy for different levels of administrative units. In 2012 the second phase sample has nearly 300,000 points in the whole EU. The subsampling is again systematic with a variable number of replicates in a block of 18 km x 18 km. Figure 3 shows again the example of Austria, that illustrates the different sampling rates in different areas.

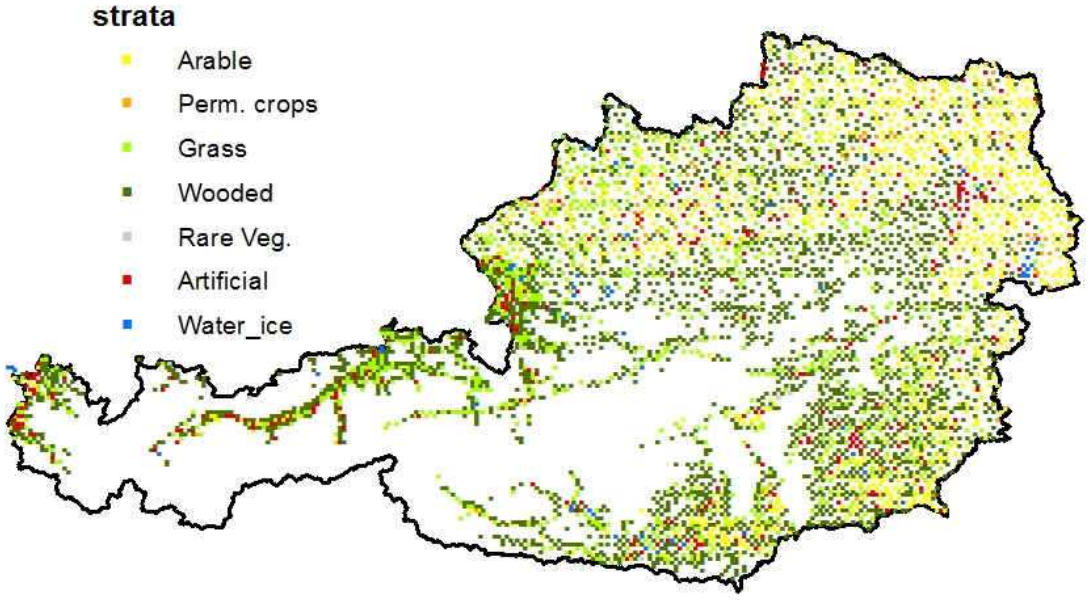

Figure 3: *Field survey sample in Austria. 2009.*

The choice of an approach that combines different types of systematic sampling is based on a wide literature showing that systematic sampling with a random starting point is superior to random sampling when the spatial correlation is a decreasing function of the distance (Cochran 1977, Bellhouse, 1988), and in particular for land cover data (Dunn and Harrison, 1993). The good behaviour of systematic sampling is explained because it ensures a good spatial distribution of the sample. Systematic sampling has a drawback that we consider minor: there is no sample-based unbiased estimator for the variance. Usual variance estimators heavily overestimate the variance when applied to systematic sampling. The advantages of systematic sampling are hidden if we estimate the variance with the usual estimators (Wolter, 1984).

In LUCAS we use a modified estimator that uses local measures of the variability: If we call *y* the proportion of area for a given land cover class *c*, the stratified estimator for the proportion $\bar{y}_{st}$ of a given land cover *c* is:

$$\bar{y}_{st} = \sum_h w_h \bar{y}_h \qquad (1)$$

where w_h is the weight of stratum *h* estimated from the first phase sample and $\bar{y}_h$ is the proportion of *c* from the survey in the stratum *h*:

$$\bar{y}_h = \frac{\sum_{i \in h} w_i y_i}{\sum_{i \in h} w_i} \qquad (2)$$

where w_i is the weight of each observation *i* in stratum *h*.

Variance estimators for two-phase sampling with stratification in the first phase can be found in classical books. For example Cochran (1977, chapter 12) gives:

$$v(\bar{y}_{st}) = \sum_h w_h s_h^2 \left(\frac{1}{n' v_h} - \frac{1}{N} \right) + \frac{N-n}{(N-1)n'} \sum_h w_h (\bar{y}_h - \bar{y}_{st})^2 \qquad (3)$$

where s_h^2 is an estimate of the variance of *y*. For LUCAS 2006/2012 we have adapted the formula using a local estimate of the variance:

$$s_h^2 = (1 - f_h) \frac{\sum_{i \neq j} w_i w_j \delta_{ij} (y_i - y_j)^2}{2 \sum_{i \neq j} w_i w_j \delta_{ij}} \qquad (4)$$

where δ_{ij} is a decreasing function of the distance between *i* and *j*:

$$\delta_{ij} = \begin{cases} 1 \big/ d(i,j) & \text{if } j \text{ is among the 8 closest points to } i \text{ in the stratum} \\ 0 & \textit{otherwise} \end{cases} \qquad (5)$$

In 2006 a priority on agriculture was decided for LUCAS. Therefore the agricultural strata were subsampled with a probability 5 times higher than the other strata. For the 2009 and 2012 editions the priority was focused on general land cover monitoring and the subsampling rates were more balanced among strata, with some geographic variability depending on the target accuracy per administrative unit. In LUCAS 2006 the same sampling rate was applied in each stratum across the 11 countries covered in that occasion, In 2009 and 2012 the sampling rate per stratum was tuned

separately for each “NUTS2” region. There are 272 NUTS2 regions in EU27, with a very variable size (http://epp.eurostat.ec.europa.eu/portal/page/portal/nuts_nomenclature/introduction).

3. Sampling farms with an area sampling frame.

Area frame surveys are particularly well adapted for land cover area estimation, in particular for crop area estimation, when direct observation of the crop is considered more reliable than reporting by farmers. Area frames can be also used as an intermediate step to sample farms whose conductors are interviewed to collect information on variables that cannot be directly observed on the field, such as fertilisers, socio-economic information, cropping intentions for next year, etc .

The so-called open, closed and weighted segment estimators are used to make the link between spatial units and farms (Hendricks et al., 1965). The main advantage of an area frame compared to list frames is that completeness of the frame and non-overlapping units are easy to ensure and extrapolation factors are reliable and not difficult to compute.

If a sample of segments is used as a basis to sample farms in a complex landscape with small plots and we select all the farms having a field totally or partially in the segment, the number of farms in the segment may be very large and the sample may become inefficient. To reduce the number of farms per segment, they can be subsampled through points. We can choose for example 5 points per segment. When a point falls on agricultural land, the farm operating the corresponding plot is selected and interviewed (Gallego et al., 1994). When compared to the above mentioned approaches (open, closed and weighted segment), sampling farms through points has the advantage that the “tract” (part of the segment that belongs to a particular farm) does not need to be measured. Some difficulties also appear, for example when the sampling frame is set up for the first time, linking points with farms may be a heavy operation unless a georefernced data are available for the cadaster or the agricultural census.

The sampling rule is as follows: points that fall on Utilised Agricultural Area (UAA) generate an element of the sample of farms. The definition of UAA may be flexible to some extent, but the concept of UAA used to decide if the point generates or not a sample unit should be the same considered by the farmer when they are asked about the total area of the farm and used for computation. Farm buildings and rough pastures may be included in the UAA if this is considered useful for the completeness of the frame.

If a point of the sample falls on UAA, the farmer managing that field is located and asked to provide global data for the farm, including total area and production of each crop or inputs such as fertilizers or pesticides. No question is asked about the production in each field. If several points fall on fields of the same farm, the farm will have in the computation a weight proportional to the number of points. The area of each crop can be estimated separately from the direct field observations and the farmers survey. This will provide a tool for cross-checking.

Locating the farmer may be a heavy task, depending on the livelihood structure. The task is heavier in regions in which people live in concentrated villages or towns rather than in scattered houses close to the fields. In some cases the owner or the manager of the farm may live in a city that is very far from the fields. However this is an investment to be made the first time the survey is run if the sample of points is kept constant: keeping a database with the link point-farm is much easier, although the amount of work depends on the communications structure or training level of the farmers, in particular whether or not the farmers have a telephone or access to internet.

Assume W_k is an additive variable for farm k, for example the area under a given crop, the production, amount of fertilizer, etc. (yield is not an additive variable). The UAA of the farm will be called A_k. The total area of the region under study is D and the area of each stratum D_h, in which we have selected a sample of n_h points. Farm k is selected with a probability proportional to the area A_k:

$$p_k = \frac{n_h \, A_k}{D_h} \qquad (6)$$

The Horwitz-Thompson estimator for unequal probability sampling, also called π-estimator (Särndal et al., 1992), for the total of W in stratum h is

$$\hat{W}_h = \frac{D_h}{n_h} \sum_k \frac{W_k}{A_k} \qquad (7)$$

Because the stratification is not perfect, there will be, even in agricultural strata, sample points that do not fall on UAA and therefore do not generate an element in the sample of farms. To deal with such points, we can define a fictitious farm that corresponds to all non-UAA and has a value $W_k = 0$.

The computation of the variance of a π -estimator requires the cross-probability of selecting pairs of units, and this is tricky when the sampling plan is systematic, as it happens in LUCAS. The question becomes easier if we introduce an instrumental variable

$$X = W \big/ A \qquad (8)$$

This corresponds to the idea of distributing W in a homogeneous way in the whole UAA of the farm. By definition the total of W is identical to the total of X. The instrumental variable X is now a variable defined in the geographical space and is not anymore sampled with unequal probability. X is also useful to deal with a problem that we have disregarded so far: what to do with farms that are shared by two or more strata in the AFS. If we had considered directly W the estimation of totals per stratum should have taken into account the area of farm k in stratum h. In exchange the variable X is a geographical variable linked to the point and we do not need it anymore. The totals of W and X per stratum do not coincide, but the overall total does. We are still disregarding the cases in which a farm is only partially included in the region or country under consideration.

Using X instead of W as target variable makes the computation of variance easier, but still needs more in-depth work. The sampling approach uses indeed three phases: systematic grid and field survey as carried out in LUCAS are the two first phases. For the third phase there is a stratification of the second-phase sample (LUCAS field sample) into UAA and non-UAA. This stratification is correlated with the stratification used for the second phase sampling (photo-interpretation of a systematic sample), but the concept of “photo-interpreted as agricultural” is not the same as identified as UAA in the field visit. Estimating the variance of a three-phase sample with simple random sample is complicated, but tractable (Fattorini et al., 2006). In our case, combining different types of systematic sampling in the first and second phase with another type of sampling (to be defined) in the third phase, there is certainly no unbiased variance estimator. Variance estimators with a moderate bias still need to be explored.

A major limitation of sampling farms through points is that covering farms that have livestock but no agricultural land becomes problematic. An option to be explored to overcome this issue is considering the subsample of points with artificial land cover and agricultural land use. In the LUCAS sample there are 700 points with such land cover/use combination. There is not enough information to estimate how many of them correspond to livestock farm facilities.

4. Field observations along the road.

We have exploited LUCAS data for an empirical assessment of the applicability of a field survey method that is not an ideal solution for unbiased crop area estimates, but may be practical when the available manpower for field observations is very limited or the territory is particularly difficult. To maximize the amount of data that can be collected within a given budget, the area estimation procedure is foreseen to limit the collection of field data with an approach that is called here "survey along the road".

The overall approach is divided into two steps: estimation of cropland based on photo-interpretation, and estimation of the proportion of a given crop compared with the total agricultural land.

$$X = Y \times Z \qquad (9)$$

Where

- X is the proportion of a given crop compared with the total geographical area
- Y is the proportion of a given crop compared with the total cropland. It would be estimated from a "survey along the road" possibly combined with classified images.
- Z is the proportion of the overall cropland on the total geographical area. It would be estimated by CAPI (Computer Assisted Photo-Interpretation) of a sample of unclustered points, possibly combined with classified images. The finest possible spatial resolution would be used for the photo-interpretation, while classified images of any resolution may be useful.

There are two major underlying assumptions for the validity of this approach:

a) The identification of cropland by photo-interpretation is reliable. The potential bias should be lower than the accuracy requirements of the crop area estimation.
b) While the distance to the closest road is usually correlated with the probability that a sampled point is cropland or not, it should not be strongly correlated with the probability of a particular crop given that the point is cropland. In other words in the above formula, the variable Y should be similar close to the roads and far away.

The concept of "cropland" remains undefined to some extent in this text. For the purpose of this this method cropland is an intermediate step in the estimation procedure. It can correspond to arable land including or excluding temporary fallow, it can correspond to other concepts of agricultural land (including permanent crops or different types of grassland). The more accurately cropland can be identified by photo-interpretation with a given definition of cropland, the better the concept will be adapted for this purpose.

An empirical test has been run with LUCAS data, considering the EU as simulation region. The results only have an indicative value, since it is not obvious that a particular property the spatial distribution of crops in the EU (assumption "b" above) may be extrapolated to other areas of the world. However the availability of data has determined the choice.

Beside the 234,700 points of the LUCAS 2009 survey, we have used a digital road network with scale 1:250,000 that roughly includes all paved roads and excludes dirt roads. The total length of the roads in this map is 2.6 million km. Bulgaria is still missing in this layer. The road density is not homogeneous, and this should be taken into account in the extrapolation weighting.

The distance from each LUCAS point to the closest road has been computed. Table 1 illustrates the proportion of major land cover classes for points close or far away from roads. For

artificial land we check an obvious increase of proportion when we look only close to the roads, while for shrubland it is the opposite. For cropland the difference is not dramatic. If we consider points less than 20m from the road we have 23.9% versus an average of 25.9%, while a distance range until 50 m gives a proportion of 26.7%. This probably means that the bias of a direct "along the road" survey is less dramatic than initially thought. Still this issue needs to be assessed for each specific landscape.

	<10 m	*10-20 m*	*20-50m*	*50-100m*	*100-200m*	*200-500m*	*500-3000m*	*>3000 m*	*Overall*
Artificial land	26.7	20.5	14.2	11.0	8.0	4.2	1.5	0.4	4.5
Cropland	22.7	25.1	28.7	31.9	32.9	32.5	22.5	3.2	25.9
Woodland	18.2	21.3	23.1	23.7	25.4	30.7	42.9	43.2	35.3
Shrubland	4.0	3.3	3.9	4.1	4.3	4.8	7.7	19.7	6.8
Grassland	26.0	27.4	27.5	26.8	26.0	23.4	17.6	9.2	20.7
Bare land	1.8	1.6	1.4	1.3	1.4	1.5	1.8	4.7	1.8
Water	0.5	0.7	0.9	1.1	1.5	2.3	4.0	10.2	3.3
Wetland	0.1	0.2	0.3	0.3	0.5	0.7	2.0	9.3	1.7
Total	3188	3160	8983	14452	26389	59428	104636	14473	234709

Table 1: *Proportion of LUCAS points per distance class to roads.*

For the purpose of area estimation with surveys along the road, the most meaningful indicator is the proportion of points within a given distance from the road. We should also remember that the field sample in LUCAS is not selected with equal probability. Each point i has a selection probability p_i and a weight for the extrapolation

$$w_i = \frac{1}{p_i} \qquad (10)$$

If we take the subsample of points that are "close" (within a given distance) to roads, the new weight is

$$w_i' = \frac{1}{p_i\, q_i} \qquad (11)$$

Where q_i is the probability that a sampled point is close to a road, in this case restricted to points with agricultural use.

Table 2 reports a similar table restricted to arable land with major crops considering a distance threshold of 100 m. If we use the weights w_i the differences between the proportions of some crops close and far from the roads appear as non-negligible. For example root crops or sunflower seem to appear much more often close to the roads. However the apparent difference when we use the corrected weights w'_i. appears moderate for major annual crops: within 1.5% for wheat, root crops, sunflower and rapeseed. Somewhat higher for barley and maize. This may be acceptable when we have a strong uncertainty. The situation is more difficult for olive trees and vineyards, with a positive bias, and for fodder and temporary grass, that seems to appear less often close to the roads. A more in-depth geographical analysis needs to be performed to better understand the bias for permanent crops and fodder.

	All LUCAS points	*Estimated % 100 m of a road Weight w_i*	*Estimated % 100 m of a road Weight w'_i*	*% bias*
Wheat	24.88	24.04	25.24	1.4
Barley	14.63	13.75	14.28	-2.4
Maize	12.31	12.34	11.94	-3.0
Root crops	3.62	9.01	3.65	0.8
Sunflower	2.50	3.73	2.53	1.2
rapeseed	6.18	1.89	6.24	1.0
fodder & temp grass	7.07	5.74	6.70	-5.2
olive trees	5.83	7.73	6.04	3.6
vineyards	3.75	6.29	3.97	5.9

Table 2*: Weighted estimation of the proportion of major crops from all LUCAS points and from LUCAS points within 100 m of a road.*

For a hypothetical area frame survey along the roads, a two-stage sampling scheme might be used with large cells (for example of 3 km x 3 km) as Primary Sampling Units. The size of the cells can be tuned in function of the road density in the available digital road map. Figure 4 below shows a 3 km segment with the roads that appear in a digital map 1:250,000. Inside the segment a systematic grid with 100 m step appears in grey (900 points). Only the 169 points at a distance < 100 m will be surveyed.

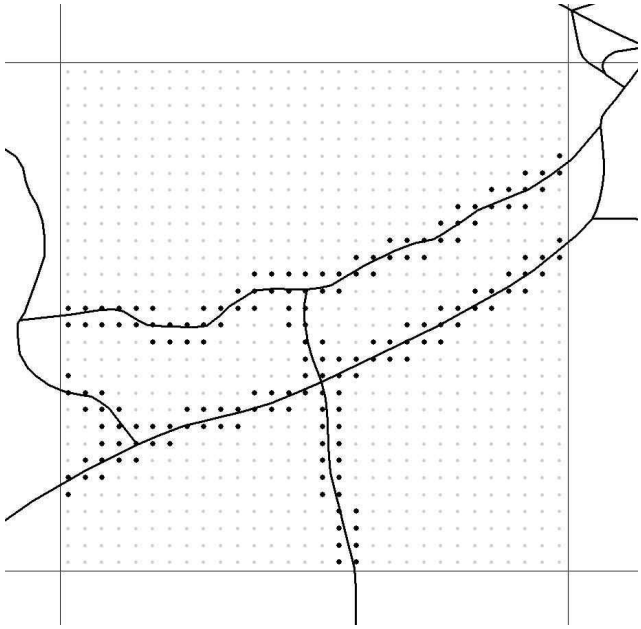

Figure 4: *Sample of points along the road in a segment of 3 x 3 km.*

REFERENCES

Bellhouse D.R., (1988) Systematic sampling, *Handbook of Statistics*, vol. 6, Krisnaiah P.R., Rao C.R., (eds) 125-146, North-Holland, Amsterdam

Cochran W., (1977) *Sampling Techniques*. New York: John Wiley and Sons

Cotter J., Tomczac C., (1994) An Image Analysis System to Develop Area Sampling Frames for Agricultural Surveys. *Photogrammetric Engineering and Remote Sensing*, 60(3) 299-306.

Dunn R., Harrison A.R., 1993, Two-dimensional systematic sampling of land use. *Journal of the Royal Statistical Society series*: Applied Statistics, vol. 42 n. 4, pp. 585-601.

FAO (1996) *Multiple frame agricultural surveys. Volume1: current surveys based on area and list sampling methods*. FAO statistical development series n.7, Rome.

FAO (1998) *Multiple frame agricultural surveys. Volume2: Agricultural surveys programmes based on area frame or dual frame (area and list) sample designs*. FAO statistical development series n.10, Rome.

Fattorini, L., Marcheselli, M., & Pisani, C. (2006). A three-phase sampling strategy for large-scale multiresource forest inventories. *Journal of Agricultural, Biological, and Environmental Statistics*, *11*(3), 296-316

Gallego F.J., Delincé J., Carfagna E., (1994) Two-Stage Area Frame Sampling on Square Segments for Farm Surveys. *Survey Methodology* , 20, 2, 107-115.

Gallego, F.J. Delincé, J., 2010, The European Land Use and Cover Area-frame statistical Survey (LUCAS). In *Agricultural Survey Methods*, R. Benedetti, M. Bee, G. Espa, F. Piersimoni. (Ed.), pp. 151-168 (New York: John Wiley & sons).

Hendricks W.A., Searls D.T., Horvitz D.G., (1965) A comparison of three rules for associating farms and farmland with sample area segments in agricultural surveys. *Estimation of areas in Agricultural Statistic*s, S.S. Zarkovich (ed.), 191-198, FAO, Rome

Ministerio de Agricultura (2008) *Encuesta de Superficies y Rendimientos de Cultivos. Resultados 2008*. Madrid, http://www.mapa.es/estadistica/pags/encuestacultivos/boletin2008.pdf.

Särndal C.E., Swenson B., Wretman J., 1992, *Model Assisted Survey Sampling*. Springer Verlag

Wolter K.M., (1984) An investigation of some estimators of variance for systematic sampling. *Journal of the American Statistical Association*, Vol. 79 No 388, pp. 781-790