European Commission

JRC TECHNICAL REPORTS

# Area Sampling Frames for Agricultural and Environmental Statistics

*Short guidelines for Developing Countries*

Javier Gallego

2015

EUR 27595 EN

Joint Research Centre

# Area Sampling Frames for Agricultural and Environmental Statistics

**JRC Science Hub**
https://ec.europa.eu/jrc

JRC98806

EUR 27595 EN

ISBN 978-92-79-54000-4 (PDF)

ISSN 1831-9424 (online)

doi:10.2788/88253 (online)



How to cite: Gallego F.J., 2015, Area Sampling Frames for Agricultural and Environmental Statistics, EUR 27595 EN, doi:10.2788/88253

# Table of contents

## CONTENTS

## Abstract

Agricultural statistics in many developing countries are based on expert opinions (subjective reports) for small administrative units, often Enumeration Areas (EA) of agricultural or popularion censuses. Other countries use list-frame surveys built on censuses or administrative registers. List frames are efficient if they are updated and quality-checked with a sufficient frequency. This is very difficult to achieve in developing countries, where the population of farms or agricultural households is generally very dynamic. If list frames are not sufficiently updated or have completeness problems, area frames provide a very useful alternative. The completeness of the frame can be guaranteed by just specifying the geographic boundaries of the country.

# 1. Introduction

The first element to define an area sampling frame is the geographic delimitation of the region of interest. If we know the boundaries of a country and we have a rule to divide it into non-overlapping units, we have a solid starting point to ensure completeness and non-redundancy of the frame. An area frame can be seen as a list of area units, even if the list is often implicit and can be infinite. The units may be defined by a geometric grid.

The specification of the sampling frame includes the cartographic projection, preferably an equal-area projection. Most cartographic projections are approximately equal-area and are perfectly acceptable to define an area sampling frame. In exchange latitude-longitude coordinates should be avoided as they may introduce a significant distortion, in particular in large countries. The main elements of an area sampling frame are the definition of the sampling and reporting units and the stratification. Stratification is not strictly indispensable to define a sampling frame, but an efficient stratification can make the difference between a strong and a weak sampling frame (this is true both for list frames and for area frames). Area frames are well protected against undercoverage, but this source of bias can appear if the perimeter of the region excludes some agricultural areas (minor islands) or more often if some strata are excluded from sampling because they are supposed to be non-agricultural, but they contain some agriculture.

The units of an area frame can be points, transects (lines of a certain length) or pieces of territory, often named segments. When the units of the frame are points, it may be called a "point frame". Points are in principle dimensionless, but in practice a certain size is attributed to them in the observation rules. For example several point surveys in Europe (LUCAS, AGRIT, TER-UTI, BANCIK) attribute a size of 3 m to points, coherent with the resolution and location accuracy of the images used as support to locate the point (FAO 1998, Gay and Porchier, 2000, Martino, 2003, Gallego and Delincé, 2010). Point frame surveys are very common in forestry, but less frequent in agriculture. In agricultural surveys the oldest operational area frame survey using points as sampling units is probably TER-UTI, run by the French Ministry of Agriculture.

For segment area frames the main choice to be made is between segments with physical boundaries (landscape elements such as roads, rivers or stable field boundaries) or segments with a regular geometric shape, such as squares, widely used by the MARS project in Europe (Gallego et al., 1994) and in several countries in southern Europe, in particular in Spain (MAPA, 2008). Area frames of segments with physical boundaries are used by the US Department of Agriculture (USDA) since the 30's.

## 2. Area frames of segments

The units of an area frame can be pieces of territory, often named segments. A frequent rule of thumb is that segments should have a size that allows carrying out the ground work in less than one working day, possibly two to four hours. This often corresponds to an average of 10-20 fields per segment. In landscapes with very small fields, such as the example in Rwanda shown in figure 1 it may happen that segments with a much larger number of fields are cost-efficient, but a careful analysis is recommended if segments with more than 30-40 fields appear frequently. The size of segments may be tuned separately in each stratum of the frame. If holders are to be interviewed, the segment size should consider the number of holders in addition to the area of the segment and the number of parcels. In developing countries, where surveyors do not have their own vehicle, the target of one segment per day may be more reasonable than several segments per day in order to optimize the logistics. For example several surveyors may be dropped in different areas by the same vehicle in the morning and picked up in the evening.

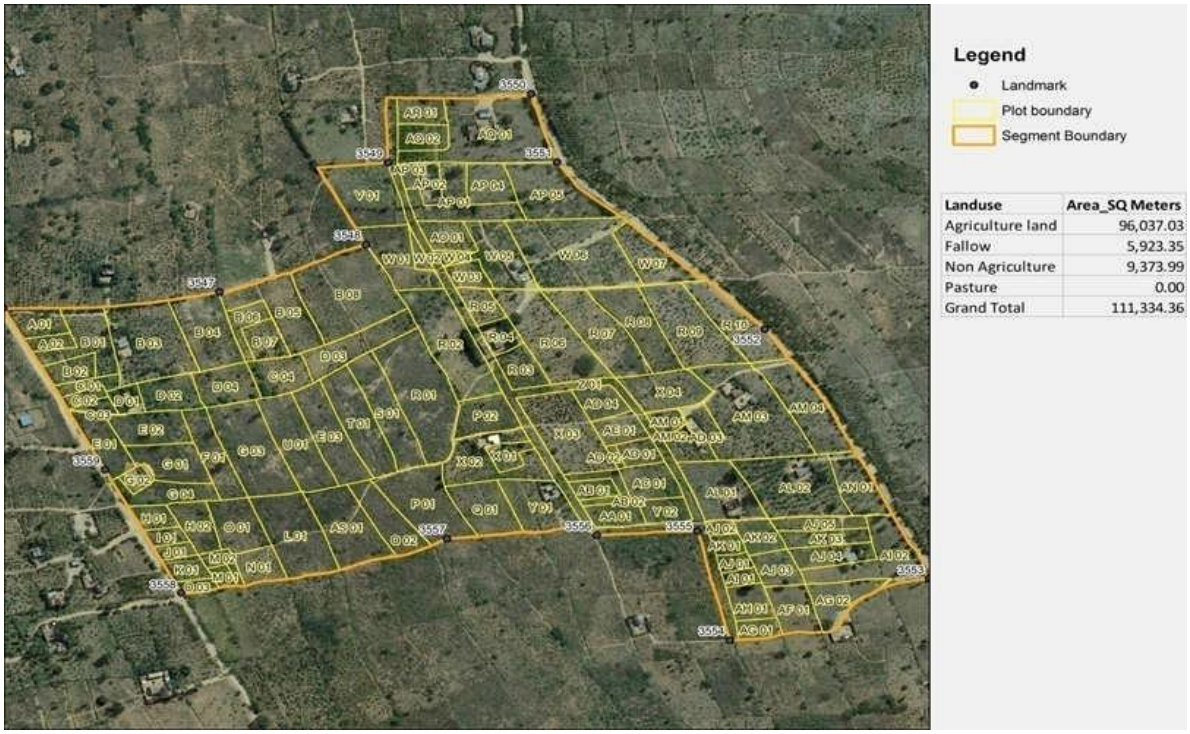



*Figure 1: Example of segment in Rwanda with a large number of fields.*

Segments can be delimited by physical elements, such as roads, rivers or permanent field boundaries. This has been the choice of the US Department of Agriculture (USDA) for the June Area-frame Survey (Cotter and Tomczac, 1994, Davies, 2009, Boryan and Yang, 2012). The same approach has been applied in a number of countries that have, in several cases, developed their area frames with the support of USDA. Several examples are described in volume 2 of FAO (1998) .. Defining an area frame with this approach requires an important initial investment, that is reduced by the use of Primary Sampling Units (PSUs), that are subdivided into Secondary Sampling Units (SSUs) or segments only if they are selected in the first sampling step. The concept of PSU in this type of area frames does not fully match with the usual concept of PSU in survey

sampling textbooks that generally refer to large units inside which a sample of several SSUs is selected. It is a particular case where only one SSU is chosen in each sampled PSU and the traditional variance computation formulas for two-stage sampling do not apply since the variance component for the second stage requires more than one SSU sampled in each PSU (Cochran, 1977). The Italian AGRIT project followed this approach between 1992 and 2001.

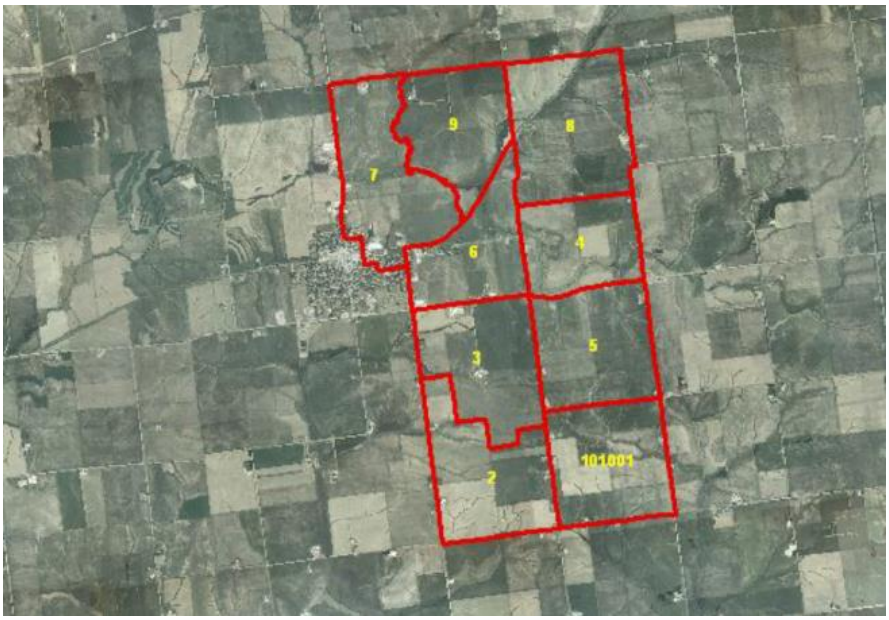



*Figure 2: PSU subdivided into several segments in the US*

*Figure 3: example of PSU with physical boundaries in Italy with delineated segments inside and segment (SSU) finally selected.*

Segments are cheaper to define with the help of a regular grid, often square, in the selected cartographic projection (Gallego, 1995). This was the choice of the Spanish MAST (Marco de Areas de Segmentos Territoriales), also described in the FAO 1998 volume. Some comparisons have been made between square segments and segments with physical boundaries: González et al. (1991) conclude that the standard errors are very similar, and therefore square segments are preferable because they are cheaper to define. Other statisticians consider that segments with physical boundaries reduce ground survey mistakes due to wrong location. However the concerns on wrong location are strongly reduced if GPS is used for the field survey.

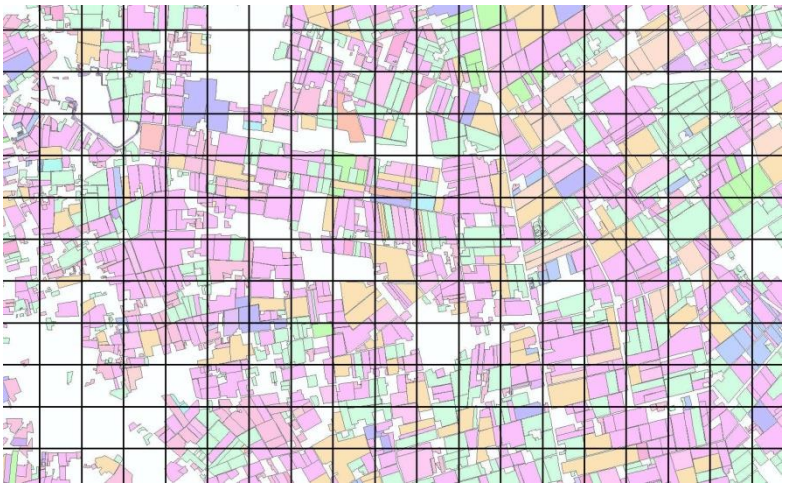


*Figure 4 Building an area frame with regular boundaries only requires defining a regular grid.*

# 3. Area frames of points

Area frames of points have been widely used for forest inventories. For agricultural and land cover surveys, there are several important examples in Europe, such as the French TER-UTI survey (FAO, 1998), operational since the 60's, BANCIK in Bulgaria (Hristoskova, 2003, Eiden et al., 2002), and the EU LUCAS survey with the sampling scheme applied in 2001 and 2003. All of them use a two-stage sampling scheme, with 10 to 36 points (SSUs) per PSU or cluster. The area covered by a PSU may be chosen as roughly corresponding to the size of a segment (a cluster of points can be seen as an incomplete observation of a segment). Nevertheless covering a larger area can be more efficient for clusters of points than for segments with complete observation, including delineating fields. For example the clusters of 36 points used in TER-UTI and BANCIK are an incomplete observation of a square segment of 324 ha with a distance of 300 m between points. The complete observation of the segment would have been extremely cumbersome, but the effort of walking 300 m between two points is reasonable in the average French or Bulgarian agricultural landscape.

The Italian Agrit project decided since 2002 to move from segments with physical boundaries to a sample setup of unclustered points (see TER-UTI, mentioned above, for an example cluster layout). The operational costs of the new Agrit suggested the need to review the cost functions that had been used to optimise the cluster size in previous studies (Gallego et al. 1999, Carfagna and Gallego, 1994). The comparison of results between the old and the new Agrit with similar cost (Martino, 2003) indicates that unclustered points can provide the cost-efficiency ratio for area frame surveys with the European conditions. This led to a new setup of the Eurostat LUCAS survey since 2006. The availability of cheap and accurate GPS has improved very much the feasibility of area frames, in particular when the sampling units are points.

In the European conditions, unclustered points in a two-phase sampling seem to be more efficient than clustered points in a two-stage sampling scheme where the first stage is defined by clusters of 5 x 2 points 300 m apart (Gallego and Delincé, 2010). However this does not necessarily apply to developing countries. In Europe, each surveyor has a personal vehicle and driving 4-6 km between a point and the next one is not a major problem. When surveyors do not have a personal vehicle and the road network has lower density and quality, clustered points are likely to be more efficient.

Data collection and processing is easier for points than for area segments and the cost-efficiency is generally better, although a rigorous comparison of operational costs is difficult: it would require the coexistence of both approaches in the same country, same institutional context and same degree of experience. Ground work with segments involve determining filed areas (reported by the holder, measured in the field or by delineation of plots on an ortho-image). This may require a certain time (a few weeks to several months) for large samples and consequently delays the production of estimates, although the introduction of GPS devices that record coordinates in the field reduces the processing time. Area segments give in exchange better information on the plot structure for landscape indicators or for co-registration if ground data are combined with satellite images for regression estimators (Gallego, 2004).

# 4. The use of transects

Transect surveys are often used for environmental purposes, for example to estimate the total length of linear landscape elements, such as hedges or stone walls; variables not usually measured in the agricultural domain. Some experiments have been carried out in the agricultural field with samples of long and relatively thin stripes on which aerial photographs are acquired (Jolly and Watson, 1979, Dancy et al., 1986, Reinecke et al, 1992). These sampling units should be considered thin-shaped segments or PSUs rather than real transects. The number of operational examples is limited, but the scheme deserves more attention for the estimation of cropland and nomadic livestock (Watson et al., 2007). A (possibly stratified) sample of stripes would be selected (figure 5) and observed with aerial photographs in which livestock can be identified and counted (figure 6).

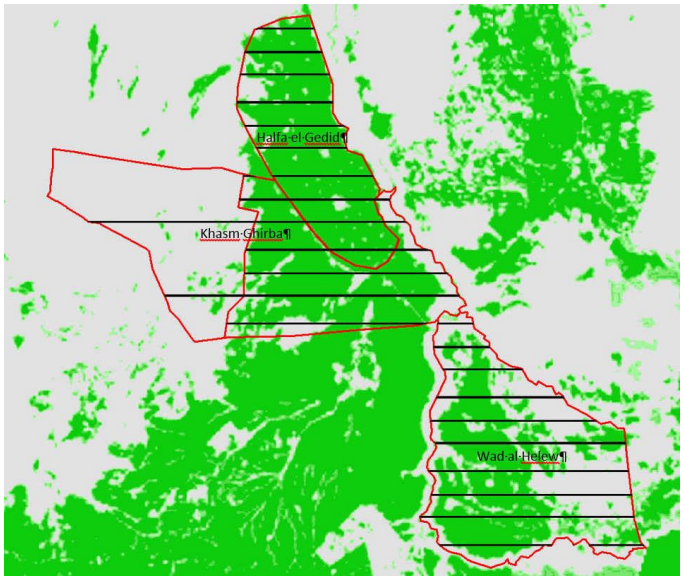



*Figure 5 Example of a possible sample of stripes in Sudan*

*Figure 6  aerial photo in which nomadic livestock can be counted.*

For specific landscapes with blocks of thin stripes, point sampling can be combined with transects perpendicular to the stripes for crop area estimation with direct observations (Kerdiles et al, 2013).

*Figure 7: Transects can be used for crop area estimation in landscapes with thin stripes.*

# 5. Stratification

The stratification of an area frame is generally based on geo-referenced features that can be observed in the land, possibly by image analysis or photo-interpretation. Typical definitions of strata in an agricultural area frame can be “cropland > 60%”, “cropland between 30% and 60%”, etc. Additional strata may be defined for specific crops or crop groups that are usually stable on the land. For example we may define a stratum with conditions of the type “irrigated crops > 50%”, “permanent crops dominant” or “grassland mixed with cropland”. The range of strata labels that we can imagine is very wide. Therefore if we are using a division of the territory into geographic units as a basis for a **master sampling frame**, **the unit should be characterized by a set of variables**, e.g. proportion of arable land, of irrigated land, permanent crops, grassland, etc. For each specific survey, the statisticians will define a suitable stratification on the basis of the available information. When EAs are used as PSUs a very efficient stratification can be defined on the basis of information from the most recent agricultural census. Each sampling unit must belong to one and only one stratum. Some tests have been conducted on splitting segments with strata boundaries. Part of a segment would belong to stratum h and another part to stratum h’ (Gallego et al., 1994). This approach has not given good results so far.

In practice the information on the variables that characterize the frame units is likely to be far from perfect, but this does not mean that the stratification will be inefficient. If a land cover map is available and we are building an area frame of segments, each sampling unit (segment) is usually not fully inside a single class of the land cover map. For each segment we will compute the area of each land cover class. In the example of Figure  we see a square segment overlaid on a land cover map. The segment contains two land cover classes according to the map: forest and rainfed arable land. In this example the map corresponds reasonably well to the background image, but this is not always true. In other cases the real number of classes may be higher and the proportions unknown. The area of each class will be computed from the land cover map and reported in the attribute table of the list of segments. The analyst will define the stratification for each specific survey on the basis of this table.

The situation is slightly different for point sampling: each point will belong to a land cover class and the stratification can be directly defined by simplifying the nomenclature of the land cover map. We can have strata, such as “rice”, “other irrigated arable land”, “rainfed arable land”, “permanent crops”, “heterogeneous agricultural areas” etc.

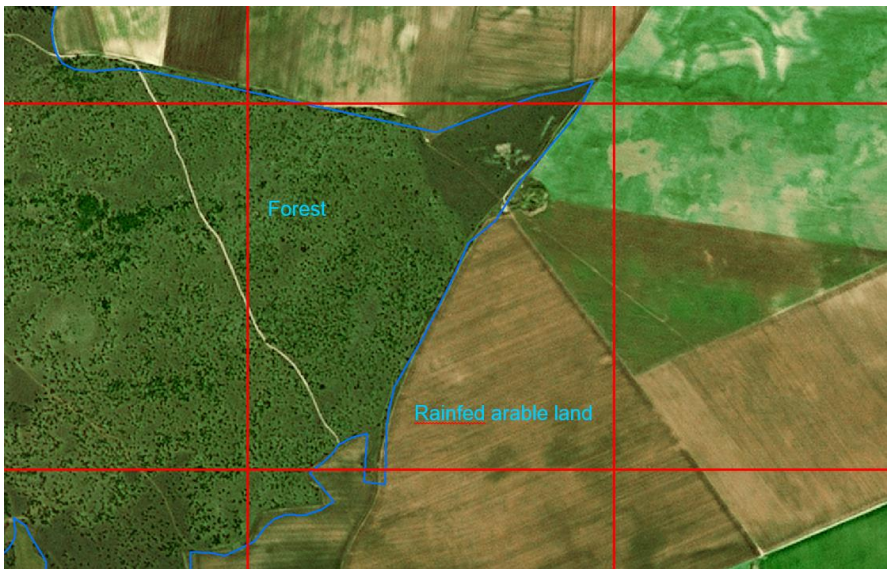



*Figure 8: example of a square segment on a land cover map (blue lines).*

# 6. Single and multi stage sampling

A two-stage point sample survey can be seen as a segment survey with an incomplete observation of segments. In the first stage a sample of area units (Primary Sample Units, PSU) is selected. In the second stage a sample of points is drawn in each PSU selected in the first stage. In this case segments will act as PSUs, but small administrative units (Enumeration Areas) can be also used as PSUs. A range of techniques may be used for the second sampling stage, including stratified random sampling and different types of (possibly stratified) systematic sampling.

As pointed out above, the term PSU is not always linked to a proper two-stage sampling scheme. In particular the segment sampling approach used by USDA-NASS is an improper two-stage sampling because only one secondary sampling unit (SSU) is selected in each PSU. In this case the PSUs are a tool to reduce the amount of GIS work and the two-stage sampling formulas for variance estimation do not need to be used.

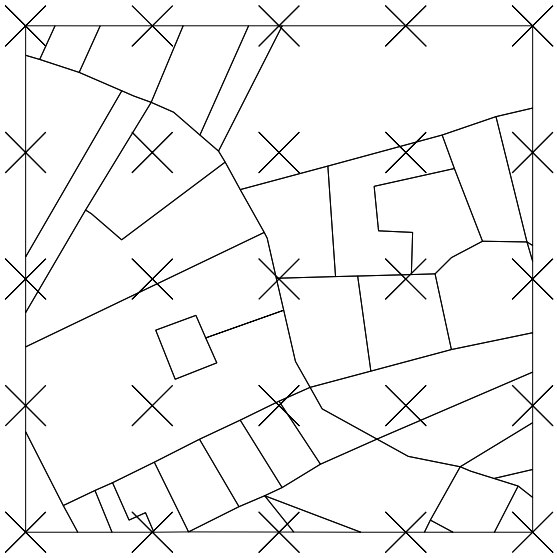
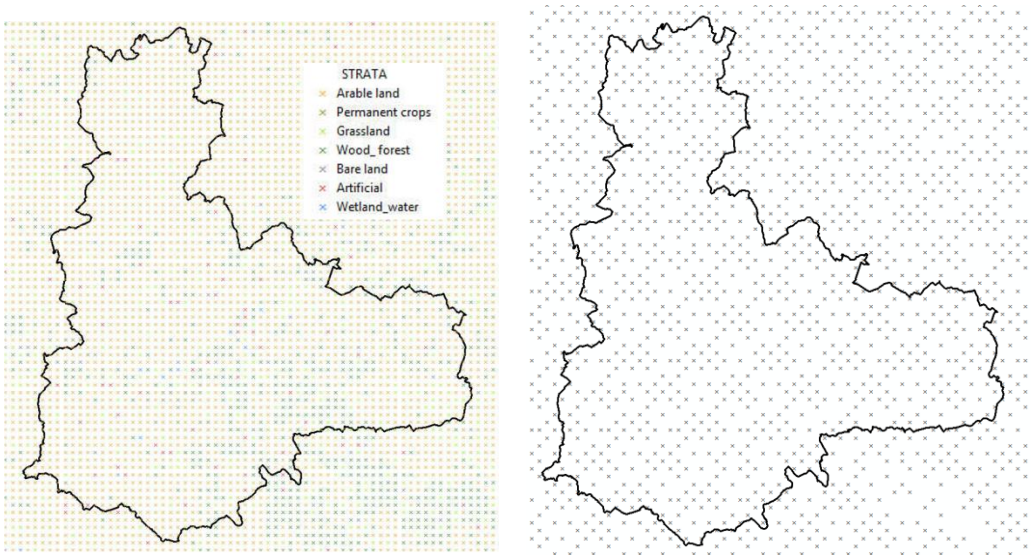


*Figure 9: example of second-stage sampling: a grid of points is sampled inside a square segment (first stage sampling unit).*

# 7. Multi-phase sampling

The idea of multi-phase (most often two-phase sampling) sampling is particularly linked with the principle of master sampling frame. In two-phase sampling a large sample is selected in the first phase, this first-phase sample or pre-sample is generally stratified with a procedure that allows a higher efficiency than a stratification system that can be applied to the full area frame. If this happens and the first-phase sample is large enough, this large sample can be a good master sample or a basis for a master sample.

Two-phase sampling is particularly interesting for non-clustered point sampling. Typical examples are the Italian AGRIT survey since 2002 and the Eurostat LUCAS survey since 2006. The method has become operational also in Haiti and pilot tests are being conducted in several Sub-Saharan countries. In both cases, AGRIT and LUCAS, the first phase sample is a systematic grid on the whole area of the targeted region. The grid has a 500 m step in AGRIT and 2 km in LUCAS (figure 10). The points in the grid are photo-interpreted on aerial ortho-photos or VHR satellite images with a simplified nomenclature, for example "arable land", "permanent crops", "pastures" and "non-agricultural". Photo-interpretation of a point is usually more accurate than polygon photo-interpretation to produce a land cover map because the photo-interpreter focuses their attention into a single point. The better photo-interpretation quality largely makes up for the loss of efficiency due to the incompleteness of the stratification (only the first-phase sample of points is stratified). The first phase sample (usually a very large number of points) is subsampled in the second phase with a rate that depends on the strata so that most of the final sample is concentrated on the strata with the highest priority.




First phase sample with a 2-km step and stratification

Second phase sample (unclustered points)

*Figure 10 Two-phase sample of points with incomplete stratification.*

# 8. Systematic sampling

Systematic sampling is often applied in list frames: the elements of the frame are sorted with some criterion and the sample contains elements *i, i+k, i+2k*, etc., where *i* is random and the step *k* is adjusted to get the targeted sample size. The efficiency of systematic sampling depends on the degree of auto-correlation between neighbouring elements: it avoids elements that are too close to each other and the corresponding redundancy of information, measured by auto-correlation.

In area frames of segments with physical boundaries systematic sampling is sometimes applied by ordering PSUs in a serpentine arrangement, but this does not necessarily avoid neighbouring elements in the sample. Systematic sampling in an area frame uses jumps of amplitude $k_x$ and $k_y$ along both axes (possibly $k_x=k_y$) and gives a better guarantee of homogeneous geographic distribution.

The main disadvantage of systematic sampling is that there is no unbiased estimator of the variance, but the practical implications of this fact are limited. The usual formulas for random sampling overestimate the variance under systematic sampling, but alternative formulas based on local variances have been proposed to substantially reduce the bias (Wolter, 1984). A more significant drawback of the straightforward systematic sampling is the difficulty of accommodating the sample size to the available budget without rerunning the whole process (Stehman, 2009). Systematic sampling with multiple replicates keeps the good spatial distribution and subsequent standard error reduction and is quite flexible to accommodate to sample size changes and unbiased estimation of sampling errors (Gallego and Delincé, 2010).

When the results of a sampling survey are politically sensitive and there is a risk of unacceptance by stakeholders who may wish to check every step, remote sensing has the advantage of being more easily traceable. For random sampling it is more difficult to prove that the sample is really the outcome of the first attempt of random numbers extraction.

# 9. Direct observations

Part of the information collected in a survey may be collected by direct observation in the field, in particular the crop area and yield, while other information items, such as the amount of fertilizers or pesticides used, require a personal interview. Direct observations protect against subjectivity but limit the types of observation that can be collected.

Direct observations are relatively easy for crop area estimation if appropriate graphic material (ortho-rectified aerial photographs or satellite images) and GPS devices are provided to the enumerators. However it may be necessary to visit the points or segments several times in the year if the crop calendar is complex.

Several studies have witnessed that plot area reported by holders often give a very strong positive bias (overestimation), in particular for very small fields. A recent study by the World Bank in Zanzibar observes that holders usually report ¼ of an acre or ½ acre for cassava fields that have less than 450 m2, with an average overestimation of more than 300%. This is an extreme example, but the conclusion is generally valid that area estimates based on self-reporting have a strong risk of overestimation. This means that measuring the area of plots is strongly recommended for list frame surveys.

Area frame surveys with a sample of segments require plot area data. It can be the area reported by the holder if it is collected and deemed reliable. In the case of direct observation mode, the survey usually involves delineating fields using an ortho-photo or satellite image as background, digitizing the plot in a GIS environment is the most obvious way to measure the plot. If the background images are not very recent or the plot boundaries are not clearly visible, GPS provides a precious tool to determine the boundaries and thus to measure the plot. Walking around the plot with a GPS is an alternative way to delineate the plots in a segment, but this approach requires a further editing process to ensure a coherent (seamless and non-overlapping) set of boundaries of the components of the segment: fields, roads or other landscape elements.

In the case of point sampling, measurement of the field or plot area is not necessary for crop area estimation. For this purpose only the area of the stratum and the proportion of points that fall in a certain crop are needed. However if the direct observations are combined with an interview with the holder for the estimation of other parameters, measurement of the plot in which the point falls is recommended to compare the area with self-reported data.

Direct observation is more problematic for yield estimation, in particular when the yield is heterogeneous inside a given field. Authorization by the farmer may be necessary to collect a crop sample and this involves an investment of time to locate the farmer. Upwards bias in yield estimation with crop cutting experiments (CCE) has been described by several authors (Murphy, 1991, Casley and Kumar, 1988, Poate, 1988). A source of bias that has been observed sometimes is that enumerators tend to avoid points for the crop cutting sample in which the state of the crop is very poor. Taking pictures and recording coordinates can be a good practice to ensure that the point where the crop sample has been collected is the point that has been sampled. Another source of overestimation is the way surveyors tend to use the frames used to delimit the area to be harvested in the crop cutting experiment (CCE): they tend to include the plants that are on the border more often than excluding them.

In some countries direct observations without previous contact with the farmer may be badly perceived by the population. Such observations may be even unwise for the security of the enumerator. In this case the advantages of direct observations may be debatable, but combining and comparing results from direct observations and farm surveys improves the quality control of both surveys.

Direct observations in an area frame substantially eliminate some sources of bias related to the reliability or distortion of farmer's replies on cultivated area or yield, but some care is needed on the application of coefficients that take into account yield losses from the moment of the observation until harvest and stocking.

For the estimation of livestock units, direct observations are problematic because of two reasons at least:

- Double counting or failure to count units can happen because livestock moves from one area to another. Compensation may be hoped, but is not sure.
- Livestock is often concentrated in herds or in stalls and there will be a large number of zeroes and a small number of sampling units with very large values. This will result in large variances.

For crop area, other potential risks of bias that remain with direct observation and require attention in the survey design or execution are:

- Wrong location of the enumerators on the ground. This can happen for example when the boundaries between cultivated fields do not coincide with the boundaries that can be seen in the support image (from a previous year). Location error can introduce a bias if it is correlated with the land cover/use, but otherwise location error does not introduce a bias in the area estimates. The location error has become a less important issue with the accuracy improvement of GPS.
- Wrong identification of crops. This may be a serious problem for unusual crops that enumerators are not able to identify, but is a minor issue for major crops if the dates for the survey are properly organized.
- Unsuitable observation date: the enumerator visits the field too early (when the crop has not yet emerged or is in a very early development stage) or too late (the crop has been harvested and no visible traces are left on the ground). When the crop calendar has a strong variability, several visits may be necessary.
- An important bias can also appear if only one visit per year is made in areas where the proportion of double or multiple cropping is significant.
- In some countries enumerators face difficulties to reach the fields or even a point from which the field can be seen (large properties with restricted access).
- Width of linear elements. In segment surveys enumerators are generally asked to delineate fields or other landscape elements in the segment. To avoid large inaccuracies in the delineation of thin elements (like a narrow road), a width threshold is applied. For example elements thinner than 10 m are delineated as a single line with no area. Thus the area of the thin element is systematically attributed to the contiguous patches (fields) and generates an overestimation of crop area. However this bias can be estimated (and thus corrected) by measuring the area of linear elements in a small subsample of segments. In Western Europe it is often of the order of 2-3% and therefore it is wise to apply a coefficient 0.97-0.98 to the crop area estimations to compensate this source of over-estimation. For other landscapes the proportion needs to be estimated. For point surveys the threshold to consider that linear elements have no area is often around 3 m and the proportion of neglected area to compensate is much smaller.
- Improper modifications of the sampling scheme. Some examples of risky strategies are described below in this chapter. When possible a sampling strategy should be tested on a pseudo-population, i.e. a set of data that behaves approximately like the real world.

# 10. Sampling farms using an area frame.

Farms can be sampled through an area frame (FAO, 1996, 1998, Gallego et al., 1994), but the identification of farmers linked with the sampled points or segments may be costly. Therefore this approach is only recommended when there are doubts on the completeness and reliability of a list frame (e.g. agricultural census). Farm surveys through area frames offer different alternative approaches and estimators. Hendricks et al, (1965) and FAO (1996) analyse three classical options linked to segment sampling: the so-called open, closed and weighted segments. The farm sampling approach through points described by Gallego et al. (1994, 2013) can be seen as a variant of the weighted segment, although there are some significant differences, in particular there is no need to compute the area of the "tract", part of the segment that belongs to a particular farm. 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.

Farms that have livestock and do not have their own land are difficult to pick with an area frame, but this is not the only weakness of area frames to sample farms. In most cases, for types of farms for which available lists are reliable and updated, list frames are more efficient because the stratification captures better the different production typologies. This happens in particular for large or specialized commercial farms (FAO, 1996, 1998). A dual frame combining a list frame of commercial farms and an area frame is often a good solution.

The sampling units may be segments or points. The estimation units may be farms or tracts, i.e. the part of the cropland (or the utilized agricultural land if grassland is included) of the farm inside the segment. Let us briefly remind the basic ideas of different approaches

## The open segment.

In this approach a farm is selected if the headquarters are inside a sampled segment. We need to define very clearly what we mean by headquarters. For agricultural households the dwelling is a possible criterion, but the question is more difficult for agricultural enterprises or farms that are managed by several families. The selection probability of the farm is the selection probability of the associated segment. Using the open segment obliges to include urban strata in the sampling frame and the number of farms per segment can be very heterogeneous. For many types of landscapes a large number of segments may have no farms associated (Figure 11). When area frames are based on a geometric grid, the open segment approach requires a particularly precise rule on the reference point to locate the dwelling (main entrance door?). Ohterwise cases may appear (farm 1 in figure 11) in which the headquarters are split by the limit of two segments. A modified version of the open segment would be to subsample farms inside each sampled segment. We would be in a two-stage sampling scheme and the sampling probability would be the probability of the segment multiplied by the subsampling rate inside the segment.


*Figure 11: Example of agricultural landscape with a few farm headquarters. Many segments do not contain any farm headquarters. For farm 1 it is difficult to decide with which segment it is associated with the open segment approach.*

## The closed segment.

In the closed segment the reporting unit is the tract, part of the farm's fields inside the segment. This approach is problematic because the farmer does not always have precise information on the target variables referred to the field(s) inside the segment. If the segment is defined by a geometric grid, the farmer's answers become more difficult. The closed segment is not recommended for surveys involving an interview with the farmer. In the example of figure 12 the square segment that has been sampled contains 5 tracts, parts of farms identified by colours. The sampling probability of each tract coincides with the sampling probability of the segment. The area of each tract and the area of each crop in each tract are easily computed in a GIS environment if the boundaries inside the segment have been delineated during the segment enumeration, but the farmer may find it difficult to report the average yield or the amount of fertilizer used inside the tract.

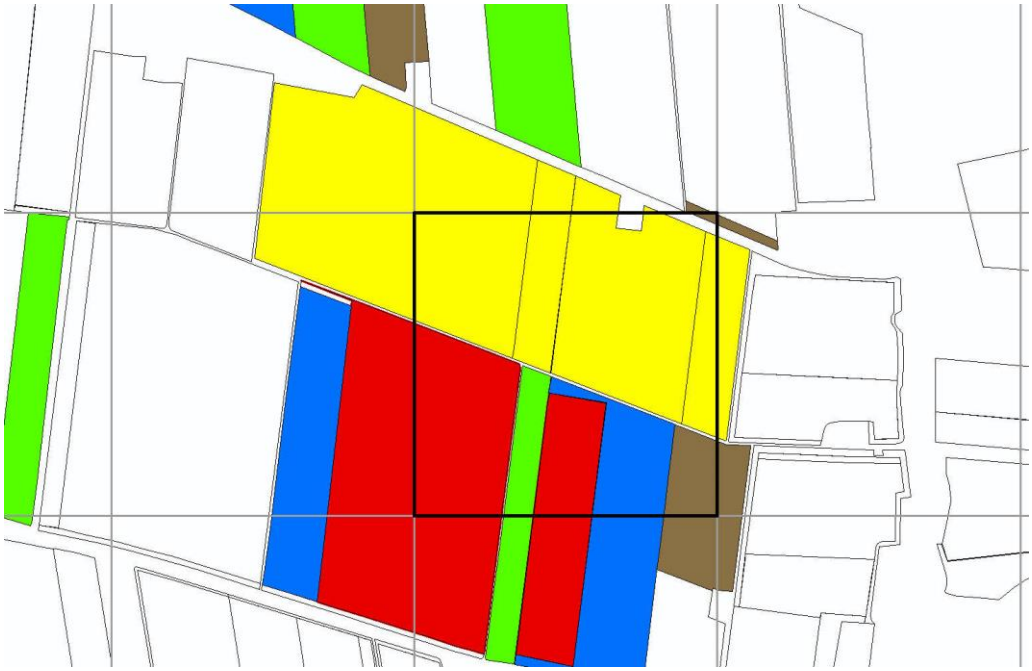


*Figure 12: Tracts inside a square segment.*

## The weighted segment.

When we do not have sufficient information about how an additive variable *Y* is distributed inside a farm, the difficulty of reporting *Y* for a tract may be bypassed using the weighted segment approach. The area under a given crop, the production, amount of fertilizer, etc. are examples of additive variables, while the yield is not an additive variable.

We call $T_{jk}$. the area of the tract corresponding to farm k in the segment j that has been sampled. The weighted segment attributes to the tract a part of the additive variable proportional to the area:

$$x_{jk} = \frac{T_{jk}}{A_k} y_k$$

This approach creates a fictitious variable *X* that is uniformly distributed in the farm area $A_k$, and that has, by definition, the same total value as *Y* for each farm. The total values of X and Y also coincide for an administrative region if we accept that each farm only has fields in one administrative region. This assumption is usually not exactly true. The existence of fields of a given farm in different administrative regions may introduce a distortion in any farm survey (area or list frame), but the impact is generally minor.

## Subsampling farms by points inside a segment.

If there is a large number of farms with fields in each sampled segment, the weighted segment approach may be inefficient because of a too heavy workload per segment (too many interviews). In this case the scheme may be modified subsampling farms (or rather tracts) inside the segment. If the farms are subsampled by points the surveyor will not need to produce the list of farms with fields in the segment. For regular shape segments the sampling procedure will be more traceable if a fixed pattern of points, for example in Figure 13 a pattern of 5 points is used with the central point and 4 points close to the corners. The template is the same for all the segments in a stratum. Data are obtained only for farms corresponding to points falling on Utilised Agricultural Area (UAA). The definition of UAA used in the field work may be adapted to the specific characteristics of the territory. Farm buildings and rough pastures may be included or excluded in the chosen definition of UAA. The crucial point is that the definition used to decide if a point falls on UAA must be consistent with the definition UAA considered for the interview with the farmer.

In the example of figure 13, point 3 falls on woodland and point 2 on a built area. They will generate two zero-valued records in the farm file. The enumerator will have to locate the farmers for the other three points. The farm corresponding to point 1 has other fields in the segment, that will be implicitly included in the survey as the farmer will be asked about the overall data for the farm. Points 4 and 5 belong to the same farm that will appear twice in the farm file (table 1).

*Table 1: Observations generated by points sampled in the segment of figure 13. the data refer to the entire farm*

| Segment | Point | Cropland | Perma-nent. Crops | Wheat | | Barley | | ..... | |
|---|---|---|---|---|---|---|---|---|---|
| | | | | Area | Production | Area | Production | ..... | |
| 1 | 1 | 19 | 5 | 10 | 62 | 0 | 0 | ..... | |
| 1 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | ..... | |
| 1 | 3 | 0 | 0 | 0 | 0 | 0 | 0 | ..... | |
| 1 | 4 | 33 | 0 | 19 | 121 | 3 | 12 | ..... | |
| 1 | 5 | 33 | 0 | 19 | 121 | 3 | 12 | ..... | |
| 2 | ..... | ..... | ..... | ..... | ..... | ..... | ..... | ..... | |

Farmers are located and asked to provide global data for the farm, including total area and production of each target crop. No question is asked about the production of each field or the part of fields inside the segment. This is not necessary because the final formulae to compute the estimates do not use the tract area or the value of Y in the tract. Figure 13 illustrates a version of the method that uses both farm and non-farm points.

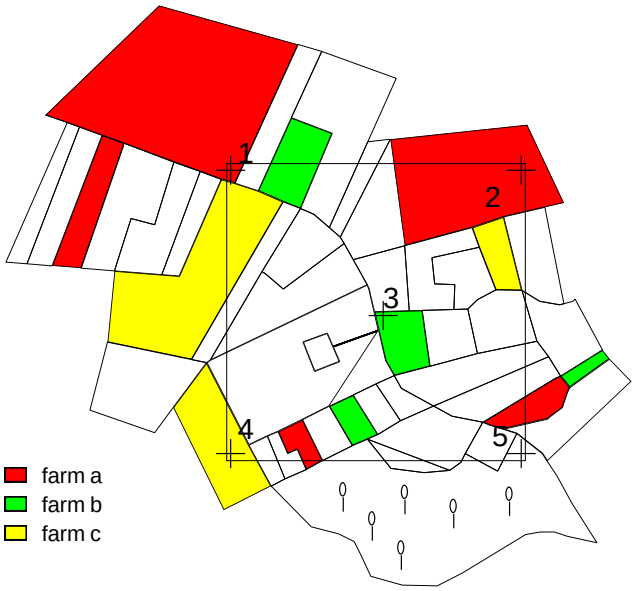



*Figure 13 sampling farms (tracts) inside a square segment*

The Horvitz-Thompson estimator for the total of *X* in segment j is

$$\hat{X}_j = \frac{1}{F_j}\sum_{k=1}^{F_j}\frac{x_{jk}}{\pi_{jk}} = \frac{1}{F_j}\sum_{k=1}^{F_j}\frac{T_{jk}}{A_k}y_k\frac{U_j}{T_{jk}} = \frac{1}{F_j}\sum_{k=1}^{F_j}\frac{U_j}{A_k}y_k$$

where $T_{jk}$ is the area of farm *k* in segment *j* (tract), $U_j$ is the total area of segemtn *j* and $A_k$ is the total area of farm *k*. The result is later extrapolated to estimate the total for the stratum and for the region of interest. According to the formula above the area of the tract does not need to be computed, although this approach requires the total UAA in the segment, easy to compute in a GIS environment, in particular if the field observations are recorded with a tablet with GPS facility. A variant of this approach can be defined that uses only points that fall on UAA instead of all points in the segment. For this approach and additional issues, such as computation of variances, see Gallego et al. (1994).

## Sampling farms directly by points (single stage)

Farms can be sampled with an area frame without using segments, but on the basis of a sample of points. The sampling rule is as follows: points that fall on UAA generate an element of the sample of farms. Farm *k* is selected with a probability proportional to its area $A_k$.

As in the previous section, the definition of UAA may be flexible to adapt to the specific characteristics of the local agriculture, but it has to be consistent along the process: the concept of UAA used to decide if the point generates or not a sample unit should be the same considered when the total area of the farm is recorded. Farm buildings and rough pastures may (or may not) be included in the UAA if this is considered useful for the completeness of the frame, but the choice has to be consistent in all the steps. An additional question mark is whether the farmer has a good knowledge of the farm area. In many countries objective measurement of plots is necessary because the area declared by farmers are often unreliable

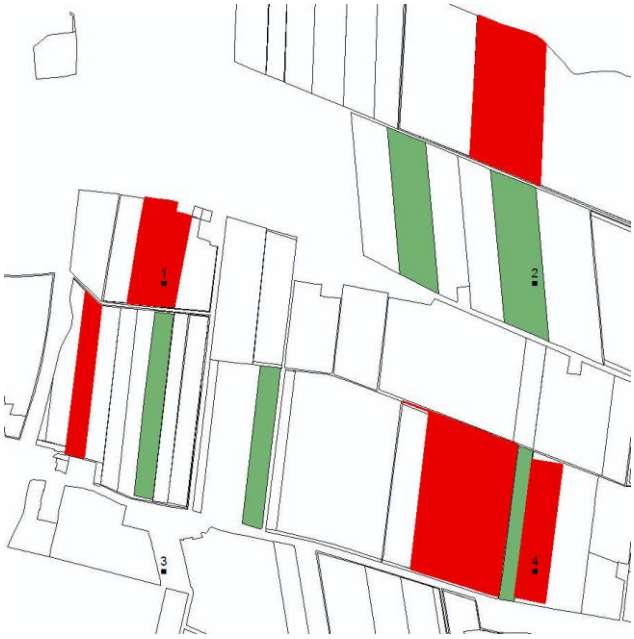



*Figure 14: sampling farms by points. Plots highlighted with the same colour correspond to the same farm.*

Other additive variables, such as production of each crop or inputs such as fertilizers or pesticides are recorded, generally using CAPI (Computed Assisted Personal Interview) . There is no need to put questions 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 direct field observations and compared with the data from the interview with the farmer. This will provide a tool for cross-checking in a quality control process to be run on a subsample (for example a 5-10% of the total sample).

Locating the farmer may be a heavy task, depending on the livelihood structure. The involvement of staff with a good knowledge of the local population (extension workers) is essential. The task is heavier in regions in which people live in concentrated 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. Drafting a look-up table that links points with farmers is an investment to be made the first time the survey is run. Keeping a mainly constant sample of points (with a rotation of a small proportion) is recommended: updating a database with the link point-farm is lighter than updating a census.

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 $A_k$:

$$\pi_k = \frac{n_h A_k}{D_h}$$

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

$$\hat{X}_h = \frac{D_h}{n_h} \sum_k \frac{y_k}{A_k}$$

Sample points that do not fall on UAA do not generate an element in the sample of farms (for example points). To deal with such points, we can define a fictitious farm that corresponds to all non-UAA and has a value $y_k = 0$. A simple stratification into two strata (agriculture and non-agriculture) minimizes this type of event. The stratification is more precise if it is carried out on a sample of points. This will involve a multi-phase sampling scheme: a large first-phase sample that can be looked at as a master sample is stratified and a subsample is selected later to build a sample of farms (Gallego, 2013).

A major limitation of sampling farms through points is that covering farms that have livestock but no agricultural land becomes problematic. An option to overcome this issue is considering a dense sample of points in built areas of rural regions in search of dwellings of livestock owners without agricultural land.

## 11. Non-sampling errors in an area sampling frame.

As stated above, an area sampling frame usually has a negligible undercover referring to farms that cultivate land or manage pastures (owned or rented). Double counting is also easy to prevent in a GIS environment. The main source of bias is linked to farms with livestock that use common pastures, in particular nomadic livestock, but this is not the only source of bias.

However an insufficiently careful management of the frame can introduce a significant bias. The most important source of bias is probably the exclusion from the sampling process of strata that have been labelled as purely non-agricultural. For example in the CORINE Land Cover Map of the European Union, although it is a good quality land cover map, between 2% and 3% of the cropland is located in polygons that have been labelled as purely non-agricultural. This percentage is likely to be higher for comparable maps (Africover) in developing countries with a more complex landscape and smaller fields. Excluding theoretically pure non-agricultural areas will lead to an underestimation, that may be considered moderate, but that should be taken into account. The bias may be more important if we apply criteria such as excluding segments in which the agricultural proportion is minor according to the land cover map.

Another example of risky management practice that can introduce a significant bias is illustrated in Figure 15 We can call it "extended segment". A regular grid (square segments in this case) used to define sampling units, but when a segment is sampled the observation unit is defined considering the full parcels for all the plots that intersect the square segment. The overestimation is obvious when an expansion factor of the type (N/n) is used. The bias is less obvious, but still significant, if correction factors are applied to compensate the increase of the segment size. This type of modifications should be avoided, or at least their impact should be assessed by simulation if sufficient real data are available.

*Figure 15: Example of an "extended segment"*

## 12. Linking area frames with census or administrative information: the use of enumeration areas (EA)

We have discussed the use of census enumeration areas or small administrative units in chapter 4 but it may be meaningful to discuss it here focusing more specifically on the sampling issues. EAs are generally associated to a delimited territory and therefore can be looked at as defining an area sampling frame, but if we do not use the specific characteristics of an area frame, we can consider the set of EAs as a list frame, discussed in chapter 5. In this chapter we shall only deal with the use of EAs with some type of area frame technique. This presumes that EAs have well defined geographic boundaries are not only defined as a list of dwellings or built areas.

A sample of n EAs is selected in the first stage, usually with some type of probability proportional to size (pps), where size does not necessarily refer to the geographic area of the EA. It may refer for example to the number of farms or to the agricultural area. If we want to sample EAs with a probability proportional to the geographic area or to the UAA, we can use an area frame approach. For example we can select a systematic sample of points (a square grid with a 10 km or a 20 km step may be an option). This will ensure a relatively homogeneous geographic distribution. We can use several sampling rules:

- We select all the EAs on which a grid point falls. In this case we are sampling with a probability proportional to the geographic area.
- A more reasonable criterion may be sampling with a probability proportional to the UAA. We shall only select the EA if the grid point falls on UAA.
- We may wish to give a higher probability to EAs with a stronger share of high-value agricultural products, for example irrigated land or permanent crops. In this case we can define two replicates as systematic grids suitably shifted from each other. A point of the first replicate will generate a sample EA if it falls on any UAA, while a point of the second replicate will only select the corresponding EA if it falls on high-value cropland. This type of area frame approach can be tuned with several types of land, but it requires that the different typologies of cropland can be recognized in an efficient way, ideally by photo-interpretation of available images.

An EA can be looked at as a large segment of an area frame, but being large, the traditional open, closed and weighted segment without subsampling of farms are likely to be very inefficient. Subsampling farms becomes necessary. This can be performed by building a list of farms having most of their activity in the EA or sampling farms by points, as described above in this chapter.

EAs can be also used as sampling units in a survey with direct observation (no farmer interview) focused only on area and production estimation or on environmental issues. In this case the exhaustive observation of all the fields may be very inefficient. Subsampling points for direct observation should provide a way of improving the efficiency. However to our knowledge there little or no experience on this type of technique.

One of the advantages of using EAs as PSUs both as a list frame or as an area frame is the frequent existence of relevant data for all EAs in the country, such as expert reports on crop area. Such figures may be subjective to a certain extent, but they are a valuable source of covariates to be used with more objective data estimated for a sample of EAs.

## References


Boryan, C. G., and Yang, Z., "A new land cover classification based stratification method for area sampling frame construction," *Proc. in First Intl. Conf. on Agro-Geoinformatics, Shanghai, China, August 2-4th, 2012*. http://ieeexplore.ieee.org/xpl/articleDetails.jsp?tp=&arnumber=6311727&queryText%3DBoryan

Carfagna, E., Gallego, F.J. 1994, On the optimum size of segments in a two-step survey on area frame: Possible improvements for rapid european estimates. Conference on the MARS Project: overview and perspectives. EUR Publication n° 15599 EN Office for Publications of the E.C. Luxembourg. pp. 139-143

Casley, D.J. and Kumar, K. 1988. "*The collection, analysis and use of monitoring and evaluation data*". Baltimore, MD: Johns Hopkins University Press for the World Bank.

Dancy, K. J., Webster, R., & Abel, N. O. J. (1986). Estimating and mapping grass cover and biomass from low-level photographic sampling†. *International Journal of Remote Sensing*, *7*(12), 1679-1704.

Davies c., 2009, Area frame design for agricultural surveys. USDA-NASS, http://www.nass.usda.gov/Publications/Methodology_and_Data_Quality/Advanced_Topics/AREA%20FRAME%20DESIGN.pdf

Delincé J., 2001, A European approach to area frame survey. *Proceedings of the Conference on Agricultural and Environmental Statistical Applications in Rome (CAESAR)*, June 5-7, Vol. 2 pp. 463-472 http://www.ec-gis.org:/

Eiden G., Vidal C., Georgieva N., 2002**, Land Cover/Land Use change detection using point area frame survey data; Application of TERUTI, BANCIK and LUCAS Data** . In "*Building agri-environmental indicators: focussing on the European area frame survey LUCAS*". European Communities, EUR Report 20521 EN ISBN 92-894-4633-1, pp 55-74. http://agrienv.jrc.it/publications/ECpubs/agri-ind/

FAO, 1996, *Multiple frame agricultural surveys*, FAO statistical development series, n.7, 119 pp.

FAO, 1998, *Multiple frame agricultural surveys, Volume2: Agricultural survey programmes based on area frame or dual frame sample designs*. FAO statistical development series, n. 10, 242 pp.

Gallego F.J., 2004, Remote sensing and land cover area estimation . *International Journal of Remote Sensing*. Vol. 25, n. 15, pp. 3019-3047 .

Gallego F.J., Delincé J.,Carfagna E., 1994, Two-Stage Area Frame Sampling on Square Segments for Farm Surveys. *Survey Methodology* , vol 20, No. 2 , pp. 107-115.



Gallego F.J., Delincé J.,Carfagna E., 1994, Two-Stage Area Frame Sampling on Square Segments for Farm Surveys. *Survey Methodology* , vol 20, No. 2 , pp. 107-115.



Gallego F.J., Feunette I., Carfagna E., 1999, Optimising the size of sampling units in an area frame. *GeoENV II - Geostatistics for Environmental applications*. ed: Gómez - Hernández J. et al. Kluwer, Series: Quantitative geology and geostatistics, vol.10, pp. 393-404

Gallego, F.J., Delincé J., 2010, The European Land Use and Cover Area-frame statistical Survey (LUCAS), in *Agricultural Survey Methods*, ed: R. Benedetti, M. Bee, G. Espa, F. Piersimoni. Wiley, Ch. 10. pp. 151-168, John Wiley & Sons.

Gay Ch, Porchier, J.P., 1998, Land Cover and Land Use Classification using TER-UTI , *Agricultural Statistics 2000, Washington*, March, 18-20. pp. 193-201. http://www.nass.usda.gov/as2000/proceedings/page-193.pdf.

González F., López S., Cuevas J.M., 1991, Comparing Two Methodologies for Crop Area Estimation in Spain Using Landsat TM Images and Ground Gathered Data. *Remote sensing of the environment*. no 35, pp. 29-36.

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, ed. S.S. Zarkovich, pp. 191-198, FAO, Rome.

Hristoskova N., 2003, Bancik : Bulgarian Area Frame Survey. *Polish Seminar : Information Systems in Agriculture*. Krakow, July 9-11. 2003

Jacques, P., Gallego, F.J., 2005, The LUCAS project – The new methodology in the 2005/2006 surveys. Workshop on Integrating agriculture and environment: CAP driven land use scenarios, Belgirate, September 26-27, 2005. http://forum.europa.eu.int/irc/dsis/landstat/info/data/index.htm

Jolly, G. M., & Watson, R. M. (1979). Aerial sample survey methods in the quantitative assessment of ecological resources. *Sampling Biological Populations'.(Eds RM Cormack, GP Patil and DS Robson.) pp*, 203-216.

Kerdiles, H., Spyratos, S., Gallego, J., & Dong, Q. (2013). Assessing the crop acreage in Mengcheng county on the North China plain using adapted regression estimator method. *2nd International Conference on Agro-Geoinformatics: Information for Sustainable Agriculture, Agro-Geoinformatics 2013,* 577-582 http://iopscience.iop.org/1755-1315/17/1/012057

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

Martino L., 2003, The Agrit system for short-term estimates in agriculture: A project for 2004. *Polish Seminar : Information Systems in Agriculture*. Krakow, July 9-11. 2003.

Murphy, J., Casley, D.J. and Curry, J.J. 1991. "Farmers' estimations as a source of production data". World Bank Technical Paper 132. Washington, DC: World Bank

Poate, C.D. 1988. "A review of methods for measuring crop production from smallholder producers". Experimental Agriculture, 24, 1–14

Reinecke, K.J., Brown, M.W. & Nassar, J.R. 1992, "Evaluation of aerial transects for counting wintering mallards", *Journal of Wildlife Management,* vol. 56, no. 3, pp. 515-525.

Stehman, S. V. (2009). Sampling designs for accuracy assessment of land cover. *International Journal of Remote Sensing, 30*(20), 5243-5272.

United Nations, 2005. *Designing Household Survey Samples: Practical Guidelines.* s.l.:United Nations Secretariat. Statistics Division

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.

## List of abbreviations and definitions

| | |
|---|---|
| AFS: | Area Frame Sampling |
| EA | Enumeration Area |
| EU | European Union |
| FAO | Food and Agriculture Organization of the United Nations |
| LUCAS | Land Use/Cover Area-frame Survey |
| PPS | Probability Proportional To Size |
| PSU | Primary Sampling Unit |
| GSARS | Global Strategy to improve Agricultural and Rural Statistics |

## List of definitions

**Area Sampling Frame** An area frame is a listing of land areas. The listing can be done in single or multiple stages. An area sampling frame where the ultimate sampling unit is either a parcel (segment of land) or a point to which a reporting unit of land associated with a holding is defined**.**

**Closed Segment:** A method of defining a reporting unit when the sampling unit is a segment of land selected from an area sampling frame. The reporting unit is a tract of land within the segment boundaries comprising all or part of a holding. Data are collected only for the land within the segment boundaries.

**Cluster Sampling:** A method of sampling to reduce frame development and data collection costs. The population is partitioned into primary units (clusters); each compromised of secondary units that may be listings of farms, segments of land units, or points. Clusters are land areas defined either administratively (villages, counties, etc.), geographically using natural boundaries, or using geo-referenced boundaries. A sample of clusters is selected using any sampling method and surveyed in its entirety or subsampled using two stage sampling.

**Frame:** A frame is the set of source materials from which the sample is selected (UN, 2005). It is the basis used for identifying all the statistical units to be enumerated in a statistical collection

**List Sampling Frame.** The ultimate sampling units are lists of names of holders or households. In the context of this Handbook, list frames are lists of farms and/or households obtained from agricultural or population censuses and/or administrative data.

**Master Sampling Frame:** In the context of the Global Strategy to Improve Agriculture and Rural Statistics, a Master Sampling Frame is a frame or combination of frames that covers the population of interest in its entirety and allows the linkage of the farm as an economic unit to the household as an social unit and both to the land as an environmental unit. Sampling frame designed to reach the integration of agriculture into the national statistical system by means of establishing a closer link between results from different

statistical processes and different statistical units. Such frame allows the selection of different samples (including samples selected from different sampling designs) for specific purposes: agricultural surveys, household surveys, management of farms surveys. The specific feature of a MSF is that it makes possible to draw samples for several different surveys or different rounds of the same survey, as opposed to building an ad-hoc frame for each survey.

**Multiple Frame Surveys** A sample survey based on multiple sampling frames. For agriculture this includes the joint use of area and list sampling frames. The frames are usually not independent; some of the frame units in one of the frames are present in the other frame.

**Multistage Sampling** A method of sampling which for agriculture uses large geographic areas or clusters as the first stage. The final sample frame is then only developed within the selected clusters in one or stages of sampling. A two stage sampling design is when the clusters are sub-sampled and the sampled secondary units are the reporting units. If sampled selected units are sub-sampled again, it is a three-stage sampling design. In general a multistage sampling design is the sub-sampling (in two or more stages) of primary sampling units (clusters).

**Non Sampling Errors** Any error in the entire process from frame development to data to analysis that is systematic and not a random error. These errors include over or under coverage of the sample frame, poorly worded questionnaires, etc.

**Open Segment** A method of defining a reporting unit when the sampling unit is a segment of land selected from an area sampling frame. The reporting unit depends on the location of the headquarters or household of the holder. If it falls within a sample segment, data are collected for the holding's entire operation whether or not it is included in the segment. No data are collected for holdings with land in the segment but whose headquarters is outside the segment.

**Point Sampling:** Final sampling unit is a point. The reporting unit is the holding associated with the land covering the point

**Primary Sampling Unit:** Same as a cluster defined above

**Replicated Sampling** A method of sampling used to simplify the estimation of sampling errors or to facilitate rotating panel surveys. Sampling procedure that instead of drawing one large sample of size "n", m independent samples of equal size n/m are selected.

**Sampling errors:** the errors in statistics obtained from a sample survey because data are collected from only sample units.

**Sampling frame:** *(see frame)*.

**Segments** Final land unit selected from an area sampling frame.

**Single-stage sampling:** sampling scheme in which the sample is selected directly from a list of units covered by the survey.

**Weighted Segment Estimator**: A method of defining a reporting unit when the sampling unit is a segment of land selected from an area sampling frame. The reporting unit is all land operated by every holding that also has land within the sample segment. The estimator is based on the ratio of the holders land in the segment to the land area in the entire operation.

## List of figures

Europe Direct is a service to help you find answers to your questions about the European Union
Free phone number (*): 00 800 6 7 8 9 10 11
(*) Certain mobile telephone operators do not allow access to 00 800 numbers or these calls may be billed.

A great deal of additional information on the European Union is available on the Internet.
It can be accessed through the Europa server http://europa.eu

**How to obtain EU publications**

Our publications are available from EU Bookshop (http://bookshop.europa.eu),
where you can place an order with the sales agent of your choice.

The Publications Office has a worldwide network of sales agents.
You can obtain their contact details by sending a fax to (352) 29 29-42758.

LB-NA-27595-EN-N

## JRC Mission

As the Commission's in-house science service, the Joint Research Centre's mission is to provide EU policies with independent, evidence-based scientific and technical support throughout the whole policy cycle.

Working in close cooperation with policy Directorates-General, the JRC addresses key societal challenges while stimulating innovation through developing new methods, tools and standards, and sharing its know-how with the Member States, the scientific community and international partners.

*Serving society*
*Stimulating innovation*
*Supporting legislation*


doi:10.2788/88253

ISBN 978-92-79-54000-4