Area Frame Sample Designs: A Comparison with the MARS Project

Elisabetta Carfagna

Department of Statistics, University of Bologna, Italy

ABSTRACT: In almost every country in the world, annual estimates for most types of agricultural commodities are available. In some countries, estimates are based on subjective evaluation and there is no way to judge their accuracy. In some other countries, non-probability samples are used. Finally, in many countries, probability samples from a list or from an area frame or from both are used. This paper briefly reviews the main characteristics of various area frame designs and presents a comparison with the sample designs adopted by the MARS (Monitoring Agriculture with Remote Sensing) project. This project has the aim of elaborating methods to combine multiple frame agricultural surveys and remote sensing, and evaluating their performance in Europe. Some results concerning an analysis carried out to improve MARS project area sample designs are also given.

1. List Frame and Area Frame

In the European Union (EU), Council Regulation 837/90 obliges all Member States to provide, on an annual basis, information on area under cultivation and production of cereals obtained by surveys implemented according to statistical methods and meeting specific requirements with regard to quality, objectivity and reliability. These requirements are becoming more and more important since the European Commission needs reliable and comparable estimates of main commodities of the member states to delineate its agricultural policy. For this reason, experiments with new cost-effective methodologies have begun.

To meet the requirements specified above, Statistical Offices of many countries send postal questionnaires to a sample of farms or organize personal interviews of farmers. The frame generally used is a list of the farms, from which a stratified random sample is selected. The main problems of this procedure are the quality of the list, which is not, in general, updated and complete, and the number of non-respondents, which is often very high when postal questionnaires are sent.

The list of farms is created from agricultural censuses, which are normally carried out every five or ten years. In only a few countries are more frequent censuses made. Whatever the frequency, the problem is that the lists do not represent exactly the true population of farms. Some farms included in the list do not exist at the moment of the survey, while new ones have not been identified for inclusion. The expansion factors, after some time, are no longer adequate. In addition, the stratification based on certain characteristics (i.e. arable land) of farms may be not very efficient because of changes in the farm structure.

1.1 A Comparison

An interesting example of the consequences of the obsolescence of a list is given by an experience in the Region Emilia Romagna in Italy. Two different surveys were carried out. The first one was by the Regional Authority on behalf of ISTAT (Italian Statistical Office) in 1989. The second one was carried out by ITA Consortium in 1990, on behalf of the European Commission, in the framework of an action of the MARS project named “Regional Inventories“. The aim of the latter was to evaluate the performance of the combined use of multiple frames and remote sensing data to produce agricultural

estimates in Europe. High resolution satellite data were used for stratification as well as for improving estimate efficiency through the regression estimator.

The Regional Authority interviewed a probability sample of 1700 farmers selected from a list stratified on the basis of farm size. While the method adopted by ITA was very similar to the one developed by the National Agricultural Statistics Service (NASS) of the U.S. Department of Agriculture [Cotter and Nealon 1987], the sampling population was constituted by about 50 hectare area units (segments) with permanent and well recognizable boundaries (such as roads, rivers, etc.) The stratification was based on a land use map created by photo-interpretation of Landsat 5 satellite pictures, scale 1:100,000. A total of 617 farms were selected by points in sampled segments. Although the sample size was much lower than in the project organized by ISTAT, estimate precision was comparable for main crop areas, as shown in Table 1 [Carfagna et al. 1992a].

Table 1. Area Estimates of Major Crops based on Various Sources of Information (1000 hectares)

	MARS 1990 (segments)		MARS 1990 (farms)		ISTAT	Region 1989	
	Area	CV%	Area	CV%	Area	Area	CV%
Soft wheat	212	5.7	208	6.9	212	217	3.8
Durum wheat	46	14.9	48	15.2	72	60	9.8
Barley	43	11.2	50	17.7	38	38	9.2
Maize	85	*8.5	63	12.8	50	58	23.9
Rice	0	-	4	59.1	6	7	42.8
Sugar beet	111	*7.1	96	9.6	97	119	16.9
Soybean	76	*6.0	55	11.6	53	47	22.7
Vineyards	78	*13.3	76	18.7	70	75	17.3
Orchards	91	*13.1	96	19.7	98	85	19.8

* Estimates obtained using regression estimators.

In the framework of the MARS project, some experiments have been carried out in eight EU countries, as well as in some Eastern European countries. It has been demonstrated that the methodology is operational and very useful, particularly in former communist countries where the rapid transformation of ownership and farm structure does not allow to create and maintain up-to-date lists [Gallego et al. 1994].

2. Multiple Frames and Area Frame Surveys

"The dual frame sample designs can benefit from the advantages of both types of samples: the list sample is extremely efficient for estimating data of large holdings and holdings that produce rare items; and the area sample, on the other hand, gives a better probability model that ensures a complete coverage and provides unbiased estimates." [Gonzalez-Villalobos and Wallace 1998] The main advantage of a multiple frame sample design is that it does not require an up-to-date and complete list of holdings, but simply the list of biggest farms, which is generally available [Carfagna et al. 1992a,b]. The basic idea is that a complete list of very large agricultural holdings and of holdings that produce rare items is easier to create and update than the list of all agricultural holdings.

The use of an area frame sample design can, however, create specific problems for some types of agricultural items. For example, in some European countries, holdings with cattle breeding quite often have no Utilized Agricultural Area (UAA). Thus, cattle tend to be underestimated if selection

probabilities of the sampling design are proportional to the UAA of the holding, unless a multiple frame sample design is adopted. Thus, the importance of a multiple frame approach is basically due to the fact that it reduces a possible bias for some items. Furthermore, multiple frame sample designs are more efficient than simple area frame designs for all commodities that are highly concentrated in a small number of farms.

In a multiple frame survey process, a very crucial step is the detection of the overlap between the list frame and the area sample. In fact, the record linkage is not easy and sometimes it is even impossible due to laws on confidentiality.

2.1 Some Multiple Frame Experiences

Since the mid-1950s, NASS has carried out research activities on development of area and multiple frame agricultural surveys in the United States, where a dual sampling frame is currently used. A sample from a list is combined with an area sample design, in which sampling units are segments with permanent physical boundaries. Similar multiple frame agricultural surveys are carried out every year in many countries in the world, such as Brazil and Honduras.

Indeed, in the U.S., the list component is large, since out of the approximately 2 million farms, the list frame covers around 1.1 million farms and the list frame sample includes approximately 70,000 farms. Instead, the 11,000 area sample segments include approximately 37,000 farms.

In Canada, a multiple frame sample design has been used since 1983 for the annual National Farm Survey. The area frame component, however, was conducted for the last time in 1995. In 1996 and 1997, the non-overlapping farms found through the 1995 area survey have been used to account for new births (to 1995) and the 1991 census undercoverage.

Statistics Canada has access to the Census of Agriculture data, whose coverage is about 97-98 percent complete in the census year, and is carried out every 5 years. Farmer interviews are conducted by telephone using computer-assisted telephone interviewing (CATI) technology and the response rates range from 89 to 99 percent. For these reasons and due to the low number of non-overlapping farms and budgetary limitations, the area frame component has been eliminated. It is important to note that the list frame is considered frozen for the five year period between two successive censuses; thus, the list is not complete if the area frame survey is not performed to compensate for births and changes to existing farms.

"Whatever limitations existed with the Area Farm Survey could have likely been solved by an increase in the sample size while at the same time decreasing the segment size. However, the current costs were starting to be prohibitive and any increase was deemed to be out of the question. Furthermore, the increased availability of administrative files combined with advances in record linkage were seen as an opportunity to steer the Agricultural Survey Program away from a dual frame approach using a list and area component to a dual frame approach using lists only. At the moment, the income tax files are being studied as a potential source to identify new entrants to agriculture and to account for census undercoverage, by comparing the list of tax files on two successive years." [Mayda and Chinnappa 1994].

Statistics Canada assumes that dual frame approach using an area component has a main limitation — new farms constitute a rare population and are more likely to be small in terms of land area, thus difficult to find. These might, however, be important in terms of sales (e.g. pig farms, poultry farms

or feedlots) but missed by the area sample. Moreover, errors in overlap detection entail an overestimation of variables of interest [Denis and Morabito 1997].

2.2 Some Area Frame Experiments

Many countries carry out general purpose agricultural surveys concerning almost all agricultural commodities on the basis of area frame only. The main reason for this is the difficulty to create and update a good list of largest farms. Examples of such countries are Albania, Argentina, Morocco and Pakistan, which use an area frame very similar to the one adopted in the U.S. Some other countries, such as Italy, France and Spain, perform special purpose surveys based on area frames without farmer interviews, since the main interest is to produce timely area estimates for the main land uses.

3. Square Segments versus Segments with Physical Boundaries

In the EU, the first experiences with area frame surveys, carried out by Regional Inventory action of the MARS project, were in some regions in Italy, Spain, Germany, France and Greece [Taylor et al. 1996] (later also in other countries such as the Czech Republic). They were based on square segments without physical boundaries. The usual sample design superimposed a square grid on the region of interest. Then, from each block covering an area of 125 km^2, a subgroup of square segments of about 50 hectares was selected through systematic sampling with three replicates for each block in intensively cultivated areas, and only one replicate in areas of low agricultural interest. In the following experiments, some aspects of the area frame sample design were modified, but the main point was still the use of regular square segments.

With a regular grid approach, the area frame is defined as soon as the limits of the region of interest are known in some cartographic coordinates (UTM, Lambert or any other system) and a grid is overlaid on this area. The cells of the grid are the segments. An area frame with square segments can be carried out using a manual or an automated procedure. Most of the realization phases of a regular grid area frame can be automated. The regular grid approach is much faster than the physical boundary approach if an automated process can be adopted. Automated procedures can be adopted if computers (at least PCs), trained people, and specific software — mainly digitization software and GIS (Geographic Information Systems) — are available.

3.1 Advantages and Disadvantages of Area Frames Based on Square Segments

In an area sample survey using segments with recognizable physical boundaries, an enormous effort in work, time, costs and support materials is necessary for the area frame construction and sample selection. NASS has developed computer-assisted area frame construction and sample selection that cuts by over one-half the time required for the stratification of an average U.S. state on paper-based cartographic material. Costs are still considerable and the digitalization of linear elements (roads, railways and so on) has to be performed.

Area frame construction and sample selection are much simpler, faster and cheaper when the regular grid approach is adopted. Satellite data can also be used for stratification if square segments are adopted, but the photo-interpretation can be much faster, since the polygons do not need to have physical boundaries.

On the other hand, the method of square segments presents some disadvantages, mainly due to the difficulty of locating the segments on the ground. In fact, identifying theoretical limits on the ground is

more difficult than recognizing permanent physical boundaries. Particularly in the case of segments without physical boundaries, there is an increased risk of location shift and of modifying the shape of the segment that is actually surveyed, compared to the segment that had been theoretically selected. However, this will not introduce a bias if location and shape errors are independent of the land cover and the estimates are based on the percentage of the land cover rather than on the area itself.

In areas with few stable and recognizable physical limits, it is difficult to find good boundaries for the area frame, and the regular grid approach could be more appropriate. In areas with few roads, as in some developing countries or in forested areas, it is sometimes impossible to find permanent physical boundaries. In such cases, non-sampling errors concerning data collection can be reduced if square segments are used.

Use of a GPS (Global Positioning System) could help to reduce location errors.

3.2 Special Purpose Surveys based on Area Frames

If a special purpose survey is organized for estimating areas of main crops, an important task in EU countries for determining subsidies based on acreage, the most reliable method is drawing the field limits on an aerial photograph during the ground survey and digitizing those fields to measure their size in the office. With this approach, no farmer is interviewed.

Some experiments were done in Emilia Romagna, Italy and in Navarra, Spain [Gonzalez Alonso et al. 1991] to compare the performance of square segments and segments with physical boundaries of the same target size. The comparison has shown that estimates' coefficients of variation are similar.

One problem is that the territory has no square limits, so the borders of the area of interest and of the strata do not match with the limits of the cells of the grid. The common solution is to decide that cells on the border are completely included in the area of interest if more than half of their surface is inside the region and completely excluded otherwise. Following this procedure, the region actually studied is approximated following the square grid.

Indeed this approximation can produce a bias on the estimates if the behavior on either side of the border is very different, such as on a coast line where pieces of land are removed and are replaced by pieces of sea. Another option would be to include only the piece of the segment in the study area and discard other pieces. In this case, many segments on the border are not squares and their size is smaller than the target size [Gallego 1995].

Concerning the approximation of the stratification to the square grid, the problem is less serious than in the case of the region border. In fact, modifying the border between two strata to follow the grid can reduce the efficiency of the stratification, but it will not introduce a bias in the estimates.

Theoretically, the land use stratification carried out for an area frame with physical boundaries is more efficient than using square segments, since the latter are attributed to the different strata on the basis of majority, which results in higher variability within strata. However, the median value of the stratification efficiency for main crops in different countries in the frame of the MARS project's "Regional Inventories" ranges between 1.4 and 1.5, which is similar to the values obtained by the Italian Ministry of Agriculture through photo-interpretation of SPOT satellite images at a scale 1:25,000 and the usual procedure of identifying primary sampling units with permanent physical boundaries.

4. Farm Selection in Square Segments

The Regional Inventory action of the MARS project has also tested the validity of an area frame with square segments as an instrument to gather information on farms in general purpose surveys. The usual closed and weighted segment estimators can be applied with many difficulties in the case of square segments, since the enumerator has to describe the theoretical limits of the segments to the farmer and ask him which part of the farm is included in the segment.

The relatively small size of farms in western Europe and the positive low distance spatial autocorrelation of surveyed variables have suggested the strategy of sampling farms by points inside segments. Some of these points fall on UAA and identify some farms, while some points may fall on non-UAA and identify non-farms with zero values for all the variables. In the first stage of this sample design, segments are selected without replacement. In the second stage, the unit is the portion of the UAA of a farm included in the selected segment. Since the weighted segment estimator is used, we do not need to know its area, just global information about the farm [Carfagna et al. 1992a,b].

If a farmer does not cooperate or cannot be found, we have missing values. The exclusion of non-respondents could produce a serious bias because the zero values corresponding to non-UAA are never missing. The solution adopted is to attribute to the missing value the average values of responding farms in the segment, if there are any; otherwise, the average of the responding farms for all segments in the current stratum. There is still a risk of bias if farmers who cannot be located or refuse to cooperate have a peculiar behavior, e.g. if they are smaller or less efficient than the average farm. A different approach would be to eliminate both missing values and a proportional number of zero values corresponding to non-UAA points. Both approaches give the same estimate for the total.

Some experiments have been carried out in Italy, Spain, Germany, Greece and the Czech Republic. These have shown that the procedure works well. The biggest problem, however, is identifying the location of points on the ground by the enumerator.

5. Unclustered Point Sampling

When few physical boundaries are present or the survey has to be conducted in a very short time, a general purpose agricultural survey based on unclustered point sampling is sometimes preferred. "There are several methods of defining the segment once the point has been located. One possibility is delineating a circle with some fixed radius around the point, then enumerating all farming operations that have land within the circle. …The main advantage to this method is that detailed, current materials are not needed to define the segment. The enumerator can use what is present around the point to define the segment. In addition, aerial photography is usually not required. The main disadvantage is the responsibility placed on the enumerator. The enumerator not only has to collect the information correctly, but also has to properly define the segment. Extensive training is needed to reduce the probability of non sampling errors. The second type of point sampling from an area frame can be considered a special case of the method just described. In this method, random points are selected within the boundaries of each stratum. Here, instead of defining a segment around the point, the enumerator determines who operates the land where the point falls, then enumerates only the one operator to determine the agricultural activity on the entire farming operation." [Garibay et al. 1996].

The first type of point sampling is not very different from an area frame based on a square grid. A difference is that square segments are non-overlapping, while circles are overlapping. Advantages and disadvantages are the same if the circle is drawn on an aerial photograph in the office. If aerial

photography is not used or the circle is drawn directly by the enumerator, then non-sampling errors are probably much higher.

The second type of point sampling corresponds to farm selection with probability proportional to the size of the farm and has been adopted in Nicaragua. Location errors of the points on the ground can be relevant if aerial photography is not used. This sample design is efficient if the spatial autocorrelation between farms in a stratum is very high. On the other hand, traveling costs are high (unless enumerators live near the points) and, if a farmer cannot be found or refuses to cooperate, the travel is useless since missing data correspond to the point. Thus, if a fixed budget is considered, this sample design does not maximize estimate precision. A cluster of points is probably a better sampling strategy. If a cost function can be estimated, the optimum cluster size can be evaluated.

5.1 Unclustered Point Sampling for Land Use Statistics

An experiment of unclustered point sampling was carried out by the Italian Ministry of Agriculture in the Region Emilia Romagna in 1995. The aim was estimating areas of the main crops. Points were systematically distributed on the territory. Then a so-called dynamic stratification was carried out after sample selection on the basis of photo-interpretation of SPOT satellite images. The stratification concerned the presence of winter crops, summer crops or permanent crops. In June, a first ground survey was conducted only on points with winter crops, permanent crops and doubtful points. Only points with summer crops were visited during the second ground survey. Points were located on the ground using a GPS.

Results appeared very encouraging since the stratification process is fast and cost-effective when compared with ground survey. Estimates and their variances, however, were calculated treating photo-interpreted points as if they were surveyed on the ground. A more accurate analysis is needed to evaluate possible bias due to photo-interpretation and a model which takes into account measurement errors should be adopted.

6. A Multinational Area Frame Sample Design in Europe

A project named "Rapid Estimates in the European Union" has been developed by the European Commission in the framework of the MARS project to get, independently of member states, rapid estimates of inter-annual change of area and production of main crops at the European level [Gallego et al. 1994]. Rapid crop area inter-annual change estimates are being produced in the European Union by analysis of high resolution satellite images on a stratified sample of 53 sites of 40 km×40 km. With the inclusion of three new member states (Austria, Finland and Sweden), the number of sites has become 60.

Image classification for each site is performed on 1 to 4 high resolution satellite images per site during the agricultural season (March to October). "Ground truth" for training the classifier is obtained by photo-interpretation of approximately 16 segments sized 700 m×700 m per site. Segments are selected at random by stratified sampling in some sites, and located on a regular grid in others where no stratification is needed.

Image analysis requires a good knowledge base built on ground information coming mainly from ground surveys made in the preceding years. On the same segments enlarged to 1400 m×1400 m, 40 points are visited after systematic selection out of a grid overlaid on the segments. Ground visits are also used as ground truth for an *a posteriori* check of photo-interpretation and for independent area

estimates. In each segment, 32 out of the 40 points are used to select farmers for interviews to collect data on production and cropping intentions for the coming year.

6.1 Optimum Segment Size

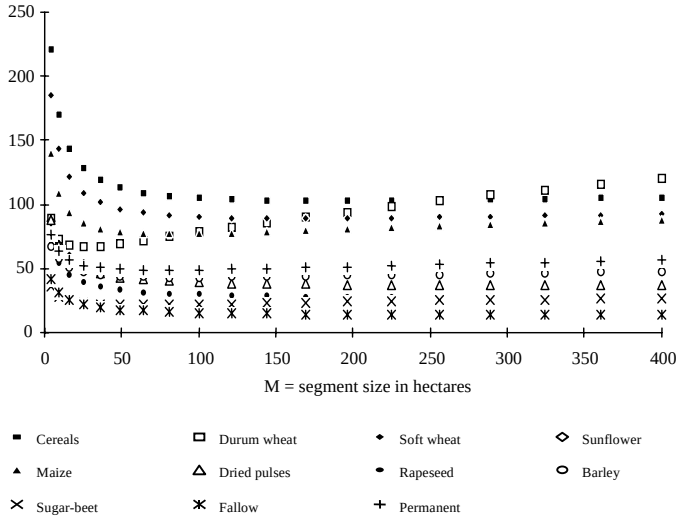
Until 1992, ground visits were performed on segments sized 700 m×700 m. This size seemed reasonable to the organizers of the project. Indeed, sample size is a crucial aspect for area frame sampling, considerably influencing estimate precision. Besides, if segments without physical boundaries are used, the target segment size can be easily changed. For these reasons, we carried out a study to find out the optimum segment size for this project, that is, the segment size that minimizes the estimator variance with a fixed budget and a specific cost function.

To determine the optimum segment size, it is necessary to create a link between the estimator variance and the segment size. For this purpose, each segment of a previous survey is considered as a cluster of squares in cluster sampling. These squares are small areas in which each segment can be decomposed [Carfagna et al. 1994]. The link between estimator variance and segment size allows to identify the combination of segment size and corresponding sample size that minimizes the estimator variance.

The result of the analysis showed the sample size that minimizes crop area variance is about 50 hectares for most crops when we make estimates for a particular year. The optimum segment size is influenced by the characteristics of the territory, average size of fields, agricultural practices and so on. Other studies carried out on many segments in Italy gave the same result concerning segment size.

If we consider inter-annual area change, the optimum segment size is much larger. Figure 1 describes the standard error corresponding to different segment sizes. It shows that the optimum is about 20 hectares for durum wheat, 65 hectares for maize, 80 hectares for sunflowers, sugar beets and permanent grassland, 120 hectares for barley, 150 hectares for soft wheat and cereals, and 230 hectares for fallow (arable land under repose in the crop rotation). The analysis of optimum segment sizes and of increases in variance due to different segment sizes has suggested to enlarge the segments to 196 hectares from the 1994 survey onwards [Carfagna and Gallego 1995a,b].

Figure 1. Standard Error for Inter-annual Area Change for Different Segment Sizes

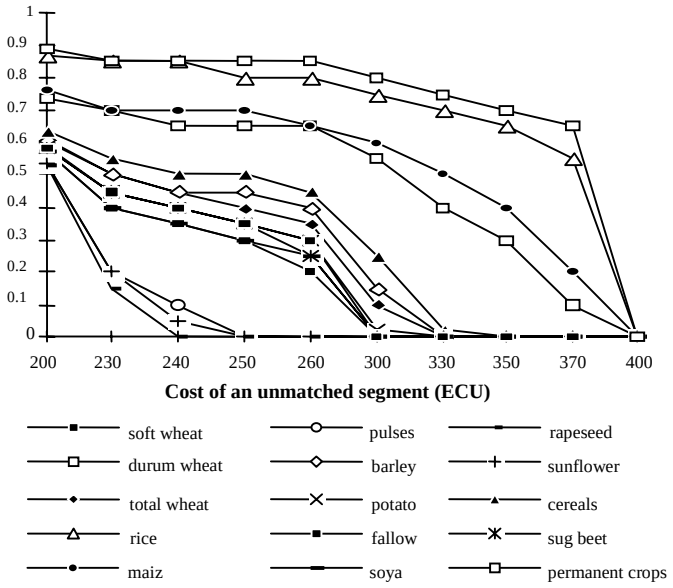


7. Repeated Sampling of the Same Population

If rapid area estimates are to evolve in the sense of estimating crop areas as well as crop area change, a good strategy would be combining current survey data with information derived from the previous year, disregarding information collected two or three years before to avoid inertia in estimate series. The best combined estimator of the population mean of crop area for the current year is found by weighting the two independent estimates deriving from matched and unmatched portions of the sample inversely as their variances [Cochran 1977, chapter 12].

Since survey costs are different for matched and unmatched segments, the unmatched proportion of the sample that minimizes the variance of the population mean (or total) of area for the current year is influenced by the ratio of matched and unmatched costs. We have considered different unmatched costs depending on the kind of technical support used for ground survey (aerial photographs or SPOT images). Some other unmatched costs are relative to photo-interpretation of segments. Different unmatched costs determine different optimum unmatched portions of the sample for each crop, as shown in Figure 2.

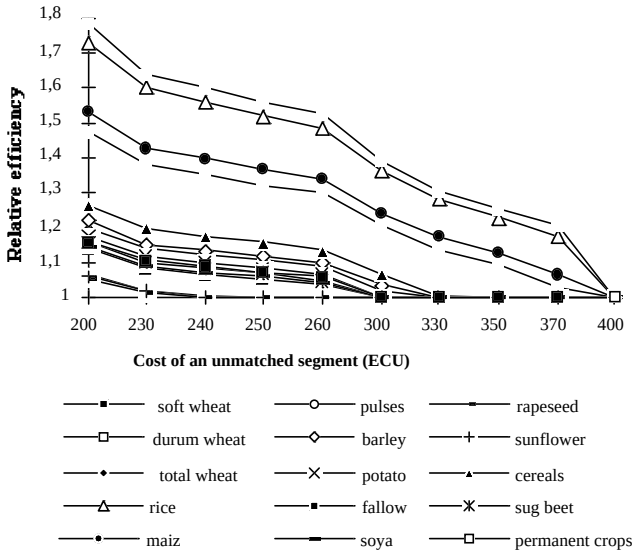
Figure 2. Optimum Unmatched Portion for Various Crops and Different Unmatched Segment Costs



Obviously, the smaller the unmatched cost, the larger the optimum unmatched portion of the sample. It is interesting to notice that for durum wheat, permanent crops, rice and maize, the optimum unmatched portion tends to be much higher than for other crops because of different levels of linear correlation between area per segment in the two years.

Figure 3 shows the relative efficiency associated with different unmatched portions for the various crops. An increase of efficiency of about 1.4 is present for unmatched costs less than or equal to 240 ECU for durum wheat and maize, and 310 ECU for rice and permanent crops. Low efficiency increases are reached for other crops.

Figure 3. Relative Efficiency for Various Crops and Different Unmatched Segment Costs



Replacement of part of the sample on each occasion also allows to take into account possible changes in the population. Then, since farmers selected by points inside the segments are interviewed, the portion of unmatched segments allows to reduce respondent burden and consequently to improve data quality.

8. Contribution of the First Sampling Stage to the Estimator Variance

The approach based on square sites of 40 km×40 km was mainly determined by the size and shape of currently marketed images, approximately square. This shape of sites, although useful for remotely sensed data, could be inefficient for ground survey. In order to decide if modifications to the sampling plan were needed, an analysis of the contribution of first sampling stage to the estimator variance was performed. The evaluation of the sampling strategy has been carried out comparing the estimator variance with the contribution to variance due to the first sampling stage. The aim was to evaluate the effect of clustering of segments in sites through the comparison of the contribution to the estimator variance due to the first sampling stage with the estimator variance itself.

For some crops, like durum and soft wheat, maize and fallow, the first stage contribution is high compared to the total variance. For these crops, it is not important to collect more information inside the sites. It would be more efficient to have a larger number of smaller sites with fewer segments inside.

For other crops, like barley, cereals, rice, potatoes, sugar beets and permanent crops, most of the variance is due to sampling inside sites. In these cases, the analysis suggests to adopt a smaller number of larger sites and a larger number of segments per site.

The main item of the survey is wheat area inter-annual change, for which a larger number of sites would give smaller variance. On the other hand, the opposite strategy could reduce the coefficients of variation for crops whose estimate precision is low. An analysis of priorities allows to decide if it is more useful to further increase precision for the most important crops or to get more reliable estimates for other crops.

Although crop area change is the target variable, area estimation is also important, and for this variable, results are very different. Most of the variance is due to first stage sampling for almost all the crops. For this variable, two stage sampling gives redundant information for all crops and a better sampling strategy could be many smaller sites, maybe so small that one site coincides with one segment (one stage sampling).

9. Correlograms and Sample Design

An analysis of spatial characteristics of area for the different crops substantially confirms the suggestion for sampling design improvement derived from comparison of first stage and total variance. Correlograms (graphs of autocorrelation at different distances) [Cressie 1991, Haining 1990, Cliff and Ord 1973, 1981] for area in 1995 are positive and show high values for some crops like cereals and permanent crops, and low values for others like sunflowers, rapeseed, soybeans and sugar beets. The presence of positive spatial autocorrelation suggests that, for area estimation, clustering selected segments in sites is less efficient than sampling unclustered segments, more or less, for all crops. Thus, the use of sites as primary sampling units is a strategy that can be justified only by lower costs for ground work, image acquisition and analysis.

Figure 4. **Correlogram for Area 1995 at EU level**

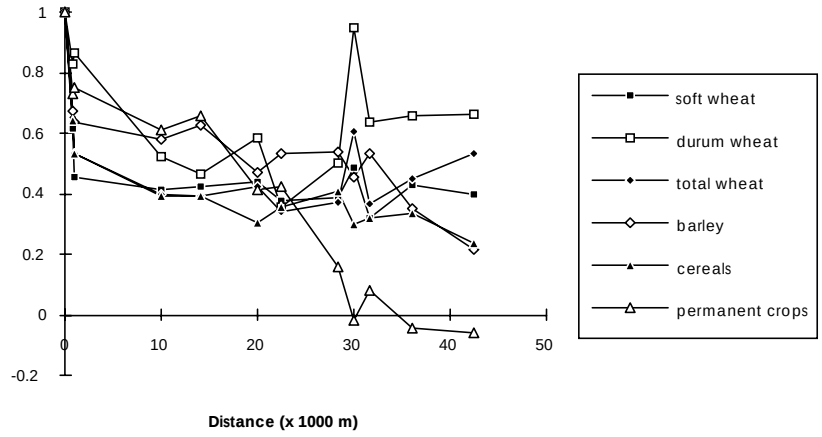


Figure 4 shows that durum wheat and permanent crops have a strong spatial autocorrelation at short distances, but at larger Euclidean distances, the correlograms of the two crops have different behaviors. For permanent crops, the correlogram assumes high values up to 15000 m, then decreases towards zero and even small negative values for larger distances. The correlogram for durum wheat decreases to a value of 0.4 at about 23000 m and increases for larger distances. For these two crops, stratified random sampling of segments should be preferred to systematic sampling. In fact, for permanent crops, we can notice a strong low order autocorrelation and an autocorrelation function not concave upwards. For durum wheat, the shape of the correlogram suggests a risk of periodicity.

The soft wheat correlogram shows a fast descent to 0.45 and a certain stability around this value for larger distances. The trend of the correlogram for total wheat is noticeably influenced by soft wheat. The correlogram for cereals is similar to the one for soft wheat, except for large distances where it is influenced by the decrease of autocorrelation values of barley.

The maize autocorrelation function decreases very slowly from 0.5 for 700 m to 0.4 for 3000 m. Rice has a specific behavior — its correlogram assumes a value of about 0.5 for 700 m distance and descends to about zero from 1000 m distance onwards. For the other crops, the descent is less rapid.

9.1 Correlogram for Inter-Annual Area Change

Since we have data collected in 1994 and 1995 on the same segments, we could treat these data through a space-time series analysis [Zani 1993]. In this context, we prefer to transform the data into a simple spatial series by subtracting 1994 crop area data from 1995 ones.

Figure 5. Correlogram for Inter-annual Area Change at EU level

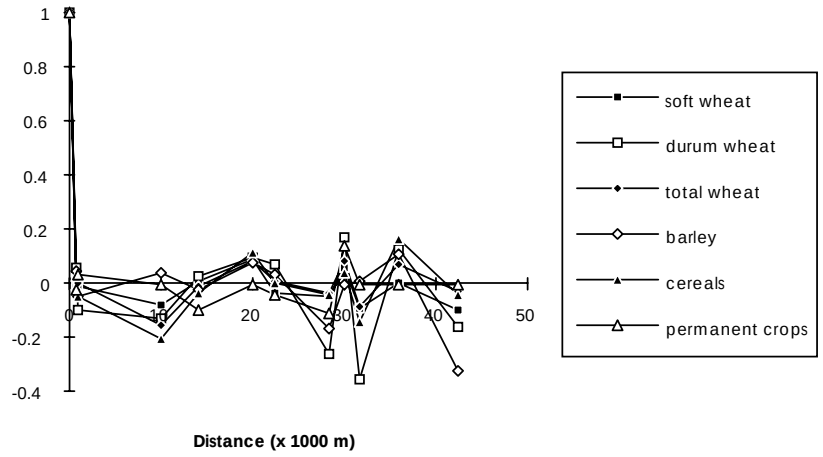


Figure 5 shows that correlograms for area change estimation are very different from the ones described for area estimation in one year. Their values are almost zero for short distances and tend to be negative in some cases. Since autocorrelation is near zero for almost all crops from short distances onwards, clustering of segments determines a small loss of efficiency if compared with one stage systematic sampling for most crops, and an increase of efficiency for some others, like sugar beets whose spatial correlation values are almost zero for short distances and tend to be negative for distances larger than 3500 m. Although permanent crops show high spatial autocorrelation for area, for inter-annual change the spatial correlogram tends to assume small negative values from 700 m onwards. Thus, clustering segments in sites could be a more efficient approach than sampling unclustered segments. Finally, rice area change has an important negative autocorrelation for a distance of 700 m and about zero autocorrelation onwards. For this variable, sites are a very efficient strategy.

10. Autocorrelation and Homogeneity Measure

In order to evaluate the optimum site size and segment number per site, since in our case results of a previous survey are available, one approach could be to simulate smaller sites than the actual ones and to calculate corresponding homogeneity indexes [see Hansen et al. 1953].

If a very simple cost function is specified,

$$C = C_1 m + C_2 m \bar{n} \ , \tag{1}$$

where m is the number of sites and $\bar{n}$ is the average number of segments per site, then the optimum number of segments per site can be given by

$$opt(\bar{n}) = \sqrt{\frac{C_1}{C_2} \frac{1-\delta}{\delta}} \ , \tag{2}$$

where δ is the measure of homogeneity, and the optimum number of sites is

$$opt(m) = \frac{C}{C_1 + C_2 opt(\bar{n})} \ . \tag{3}$$

If the homogeneity measure, δ, assumes a high value, the optimum number of segments per site is small. Consequently, the sampling plan should present a large number of small sites. However, if the homogeneity measure presents a low value, then the optimum number of segments per site is large and a small number of large sites should be included in the sample.

Indeed, in the case of a two stage area frame sample design, the number of sites (primary sampling units - PSUs) and segments (secondary sampling units - SSUs), their size, and the distance between PSUs and between SSUs can vary. Thus, to adopt the usual formulas for the optimum two stage sampling design [Hansen et al. 1953], some assumptions have to be made, such as the segment size and the distance between segments in each site.

For sites of a size smaller than 40 km×40 km, δ could be calculated, while for larger sites, δ should be extrapolated to evaluate the optimum sampling design for each crop. Obviously, a compromise solution should be reached.

We have formulated the homogeneity measure as a function of the Moran autocorrelation index and, consequently, we have expressed the optimum site and segment numbers as a function of spatial autocorrelations analyzed in previous sections.

First, recall that the definition of δ is:

$$\delta = \frac{2\sum_{i=1}^{M}\sum_{j<k}(x_{ij}-\mu)(x_{ik}-\mu)}{(N-1)\sum_{i=1}^{M}\sum_{j=1}^{N}(x_{ij}-\mu)^2} = \frac{\sum_{i=1}^{M}\sum_{j=1}^{N}\sum_{k=1}^{N}(x_{ij}-\mu)(x_{ik}-\mu)}{(N-1)\sum_{i=1}^{M}\sum_{j=1}^{N}(x_{ij}-\mu)^2} \ , \ j \neq k \ . \tag{4}$$

δ can also be expressed as [D'Orazio 1993]:

$$\delta = \frac{\sum_{h=1}^{MN}\sum_{l=1}^{MN}(x_h-\mu)(x_l-\mu)w_{hl}}{(N-1)\sum_{h=1}^{MN}(x_h-\mu)^2} \ , \ h \neq l \ , \tag{5}$$

where $w_{hl} = 1$ if segments h and l are in the same site, $= 0$ otherwise.

δ may also be written as:

$$\delta = \frac{\sum_{h=1}^{MN}\sum_{l=1}^{MN}(x_h-\mu)(x_l-\mu)w_{hl}\ \sum_{h=1}^{MN}\sum_{l=1}^{MN}w_{hl}}{(N-1)\sum_{h=1}^{MN}(x_h-\mu)^2\ \sum_{h=1}^{MN}\sum_{l=1}^{MN}w_{hl}}\ ,\ h\neq l\ , \tag{6}$$

and since $\sum_{h=1}^{MN}\sum_{l=1}^{MN}w_{hl} = MN(N-1)$, then

$$\delta = \frac{MN\sum_{h=1}^{MN}\sum_{l=1}^{MN}(x_h-\mu)(x_l-\mu)w_{hl}}{\sum_{h=1}^{MN}\sum_{l=1}^{MN}w_{hl}\ \sum_{h=1}^{MN}(x_h-\mu)^2}\ ,\ h\neq l\ . \tag{7}$$

Then, following Arbia [1986]:

$$\sum_{h=1}^{MN}\sum_{l=1}^{MN}(x_h-\mu)(x_l-\mu)w_{hl}=\sum_{h=1}^{MN}\sum_{l=1}^{MN}(x_h-\mu)(x_l-\mu)z_{hl}^{(1)}+\sum_{h=1}^{MN}\sum_{l=1}^{MN}(x_h-\mu)(x_l-\mu)z_{hl}^{(2)}+..+\sum_{h=1}^{MN}\sum_{l=1}^{MN}(x_h-\mu)(x_l-\mu)z_{hl}^{(D)} \tag{8}$$

where $z_{hl}^{(d)} = 1$ if the distance between segments h and l in the same site is d, $= 0$ otherwise, and D is the maximum distance between segments in a site.

Substituting expression (8) in formula (7) we get:

$$\delta = \frac{MN}{\sum_{h=1}^{MN}\sum_{l=1}^{MN}w_{hl}\ \sum_{h=1}^{MN}(x_h-\mu)^2}\sum_{d=1}^{D}\sum_{h=1}^{MN}\sum_{l=1}^{MN}(x_h-\mu)(x_l-\mu)z_{hl}^{(d)} \tag{9}$$

$$= \sum_{d=1}^{D}\left[\frac{S_0^{(d)}}{MN(N-1)}\ \frac{MN\sum_{h=1}^{MN}\sum_{l=1}^{MN}(x_h-\mu)(x_l-\mu)z_{hl}^{(d)}}{S_0^{(d)}\sum_{h=1}^{MN}(x_h-\mu)^2}\right],\ h\neq l\ . \tag{10}$$

where $S_0^{(d)}$ is the number of segments at a distance d.

Then, since

$$I^{(d)} = \frac{MN\sum_{h=1}^{MN}\sum_{l=1}^{MN}(x_h-\mu)(x_l-\mu)z_{hl}^{(d)}}{S_0^{(d)}\sum_{h=1}^{MN}(x_h-\mu)^2}\ ,\ h\neq l\ , \tag{11}$$

is the Moran autocorrelation index at contiguity level d,

$$\delta = \frac{1}{(N-1)}\sum_{d=1}^{D}I^{(d)}\ \frac{S_0^{(d)}}{MN}\ . \tag{12}$$

Thus, we can estimate δ corresponding to different site sizes summing the values of the Moran autocorrelation index [Cliff and Ord 1973] for distances between the segments less than or equal to d, weighted by the ratio between the number of segments at distance d and the total number of segments, and dividing the result by the number of segments in a site minus one.

We have calculated the optimum number of segments per site and the corresponding optimum number of sites for values 1, 2 and 5 of the ratio $\frac{C_1}{C_2}$. Considering the value 5 for the ratio, the optimum number of sites for crop area estimation is the actual one for rice, soybeans and rapeseed, while it is double for all the cereals.

If the variable of interest is inter-annual area change, then a small number of large sites is the best strategy for almost all the crops. In fact, the correlogram assumes values very near zero and often negative. The result is that the homogeneity index is generally negative and almost zero.

11. Conclusions

In the near future, the need of reliable estimates for agricultural commodities will increase. Currently, list frames are more often used than area or multiple frames, but a comparison between list and multiple frame sample designs can give different results in different situations depending on the aims of the projects and on the information available in a country. The assumption that list frames are more efficient than multiple frames is not always confirmed by experience. The effects from the use of obsolete and incomplete lists can be considerable if a description of the agricultural sector and not only the production of figures for main agricultural items is the aim. In fact, the largest farms generally produce the largest amount of most commodities, but the behavior and the economic and social conditions of the other farms are generally very different.

When land use data are to be estimated, area frames based on square segments can be a very good and cost-effective alternative to area frames based on segments with physical boundaries, particularly where it is difficult to find many permanent physical boundaries. If the survey also concerns data that can only be collected through farmer interviews, it is very difficult to identify the portion of a farm included in a square segment. On the other hand, if farms are selected by points and the weighted segment estimator is used, then the only problem is to locate the points on the ground.

Selecting unclustered points is an interesting strategy, but it is cost-efficient only in particular circumstances. Furthermore, non-sampling errors can be relevant. A very good strategy to take into account missing data is needed.

Area frames based on regular grids can be easily modified with low costs, and spatial characteristics of variables of interest observed during a previous survey can be very useful in order to improve the sample design. Particularly, previous data allow the determination of the optimum segment size, the estimation of the contribution to the estimator variance due to each stage of sampling, and the evaluation of correlograms and of homogeneity measures. This information is essential to review a sampling plan. We have developed this kind of analysis on the Rapid Estimates action of the MARS project carried out by the European Commission.

An interesting result is the link between the autocorrelation function and the homogeneity measure, which allows to estimate homogeneity for different site sizes, and to evaluate the optimum combination

of segment and site numbers and consequently their optimum size. Another interesting aspect of the analysis is the noticeable difference between the correlograms for crop area and for area change estimation. On the basis of observed data, we can conclude that, if change is the main variable of interest, the wide range of sampling methods based on the assumption that spatial variables are generally positively spatially autocorrelated [Ripley 1981, Dunn and Harrison 1993] are not appropriate.

Area frame sample designs are typically multipurpose. Since the behavior of the different variables is often quite different, a compromise solution should be determined when evaluating the optimum site size, number of sites, number of sampling stages, number of segments per site, and so on. Thus, we can say that the analysis that can be done on spatial characteristics of the variables is the basic information for deciding how to improve a sampling plan.

References

Arbia, G. (1986), "The Modifiable Area Unit Problem and the Spatial Autocorrelation Problem: Towards a Joint Approach," *Metron*, 44, pp. 391-407.

Carfagna, E., Delincé, J., Minelli, U. and Orlandi, G. (1992a), "Farm survey based on area frame sampling: The case of Emilia Romagna in 1990," in *the Application of Remote Sensing to Agricultural Statistics*, Luxembourg: Office for Official Publications of the European Communities.

Carfagna, E. and Gallego, F. (1995a), "Sulla dimensione ottimale del segmento per rilievi a terra e fotointerpretazione," *Statistica e Telerilevamento*, AIT Quad. n. 3 1995, CNR Pisa.

Carfagna, E. and Gallego, F. (1995b), "Extrapolating Intracluster Correlation to Optimize the Size of Segments in an Area Frame," Conference on Applied Statistics in Agriculture, Manhattan, Kansas, 24-26 April 1994.

Carfagna, E., Gallego, F. and Tarsi, L. (1994), "On the Optimum Size of Segments in a Two-stage Survey on Area Frame. Possible Improvements for Rapid European Estimates," in *The Mars Project Overview and Perspectives*, Luxembourg: Office for Official Publications of the European Communities, pp. 139-143.

Carfagna, E., Ragni, P., Rossi, L. and Terpessi, C. (1992b), "Area frame: Un nuovo strumento per la realizzazione delle statistiche agricole in Italia," in *Contributi alla statistica spaziale*, ed. S. Zani, Parma: Istituto di Statistica Università di Parma.

Cliff, A.D. and Ord, J.K. (1973), *Spatial Autocorrelation*, London: Pion Limited.

Cliff, A.D. and Ord, J.K. (1981), *Spatial Processes. Models and Applications*, London: Pion Limited.

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

Cotter, J. and Nealon, J. (1987), *Area Frame Design for Agricultural Survey*, Washington, D.C.: National Agricultural Statistics Service, USDA.

Cressie N. (1991), *Statistics for Spatial Data*, New York: John Wiley & Sons.

Denis, J. and Morabito, J. (1997), "An Assessment of the Use of Administrative Data in the Identification of New Farms," to appear in *1997 Proceedings of the American Statistical Association, Survey Research Methods Section*.

D'Orazio (1993), *Campionamento di blocchi di unità territoriali in presenza di autocorrelazione spaziale,* in STATCHEM93, Venezia, Cafoscarina, pp.67-73.

Dunn, R. and Harrison, A.R. (1993), *Two-dimensional Systematic Sampling of Land Use*, JRSS: Applied Statistics, Vol. 42 No. 4, pp. 585-601.

Gallego, F.J. (1995), *Sampling Frames of Square Segments*, Report EUR 16317, Luxembourg: Office for Official Publications of the European Communities, 68 pp., ISBN 92-827-5106-6.

Gallego, F.J., Delincé, J. and Carfagna, E. (1994), "Two Stage Area Frame on Squared Segments for Farm Surveys," *Survey Methodology*, 1994, Vol. 20, No. 2, pp. 107-115.

Garibay, R., Steiner, M. and Gonzalez, M. (1996), *Area Frame Point Sampling: An Exploratory Study to Measure Nicaragua Agricultural Product*, Washington, D.C.: National Agricultural Statistics Service, USDA.

Gonzalez Alonzo, F., Lopez Soria, S. and Cuevas Gozalo, J.M. (1991), "Comparing Two Methodologies for Crop Area Estimation in Spain Using Landsat TM Images and Ground-Gathered Data," *Remote Sensing of Environment*, 1991, Vol. 35, pp. 29-35.

Gonzalez-Villalobos, A. and Wallace, M.A. (1998), *Multiple Frame Agricultural Surveys,* Volume II, FAO Statistical Development Series No. 7, in press.

Haining, R.P. (1990), *Spatial Data Analysis in the Social and Environmental Sciences*, Cambridge University Press.

Hansen, M., Hurwitz, W. and Madow W. (1953), *Sample Survey Methods and Theory*, New York: John Wiley & Sons.

Mayda, J.E. and Chinnappa, N. (1994), *A Multiple Frame Agricultural Survey Design: A Case Study from Statistics Canada*, National Case Study prepared for FAO.

Ripley, B.D. (1981), *Spatial Statistics*, New York: John Wiley & Sons.

Taylor, J., Sannier, C., Delincé, J. and Gallego, F.J. (1996), *Regional Crop Inventories in Europe Assisted by Remote Sensing: 1988-1993. Synthesis Report* (in press), Luxembourg: Office for Official Publications of the European Communities.

Zani, S. (1993), *Metodi statistici per le analisi territoriali*, ed., Milano: Franco Angeli.