IDENTIFYING THE MOST APPROPRIATE MASTER FRAME FOR AN INTEGRATED SURVEY

IDENTIFYING THE MOST APPROPRIATE AREA FRAME, LIST FRAME AND MULTIPLE FRAME FOR SPECIFIC LANDSCAPE TYPES, TAKING INTO ACCOUNT THE KIND OF DATA SOURCES AVAILABLE

Luis Ambrosio Flores
luis.ambrosio@upm.es
July, 2013

SUMMARY

The general features of a sampling frame are the number, size and type of sampling units, as well as the materials used to localize sampling units and statistical units. We review the literature on the features of the sampling frames for agricultural surveys, on the one hand, and on the features of sampling frames for household surveys, on the other hand. Generally, both of them use multiple frames: in household surveys the frame unit is the enumeration area (EA), as defined in censuses (100 households, on average), while in agricultural surveys it is a parcel of land (field). The number of sampling stages is usually three in household surveys (sampling units are clusters of EAs in the first stage, EAs in the second stage and blocks (10 households) in the third stage) and two in agricultural surveys (Clusters of segments and segments).

We propose a conventional master sampling frame for integrated surveys, which combines the features found on this review. It is a multiple frame, where the frame units are parcels of land in the area frame and farms and households in the list frames. In urban areas, the number of sampling stages is three (sampling units are clusters of 2- 4 EAs in the first stage, EAs with 10 blocks of 10 households in the second stage and households in the third stage. In rural areas the number of sampling stages is two (sampling units are EAs in the first stage and segments in the second stage.) The statistical units are households and persons, farms and farmers, and fields. EA provides a link between the area frame component (based on cartographic material) and the list frame component (based on censuses).

This master sampling frame is the most appropriate for an integrated survey, once the features capable of improvement, such as the sampling unit size or the number of sampling stages, have been optimized. To identify the optimum sampling unit size for specific landscape types, a cost-efficiency approach is proposed. This approach is developed for the lattice case. The importance of the lattice case derives from the fact that so much data is now available through remote sensing, which comes in the form of a lattice grid (pixels). The landscape effect on sampling error is modeled using correlogram functions, and the optimum sampling unit size is found by minimizing the sampling error subject to a cost function, taking into account the data gathering difficulty in each landscape type.

CONTENTS

1. INTRODUCTION

The Master Sample of Agriculture focuses on the economic dimension: crop areas, crop yields, livestock production, aquaculture, fisheries, forestry and inputs and outputs in the agriculture holdings. The Master Sample of Household focuses on the social dimension: the social well-being of the farm and rural household. The aim here is to develop a Master Frame for surveys integrating these two dimensions, together with land cover and other environmental issues (tillage methods and use of chemical fertilizers, pesticides and herbicides), which is the third dimension to take into account for analyzing the sustainability of production systems.

The term "integrated" refers here to the use of the same sampling frame and related materials in multiple surveys, as well as the same concepts, survey personnel, and facilities. The importance, to a national statistical organization, of planning a program of integrated surveys, as opposed to ad hoc design of individual surveys has been highlighted by the United Nations Statistical Office (UN): the development of a high quality Master Sampling Frame is expensive and the costs could not be justified if the frame were to be used in only one survey. Similarly, the use of a Master Sample may make it possible to decrease the costs of sample selection, including the preparation of frames for the second and subsequent stages of selection, attributable to each survey [UN (1986)].

The Master Frame to be developed has to be highly accessible to developing countries. Kish (1996) distinguished between transferability and accessibility: the former denote the ease with which a method or technique can be adapted from a "donor" country to a "receiving" country; the later refers to the low cost of the access to the method. The methods of survey sampling are highly transferable, but are not highly accessible since "the application of sampling methods requires intimate knowledge of the local situations, of available sampling frames, and local resources". This is the reason that the type of available data sources must be taken into account.

As noted Kireyera (1982) a few years ago, the appropriate sampling frames were invariably absent in many African countries and the cost of construction and maintenance was prohibitive. Noting the importance of map material for the construction of a sampling frame in rural and urban areas from India, Murthy (1969) realized that this material was not available at that time. Today, the remote sensing-based material is available at low cost and therefore suitable sampling frames can be accessible to many countries around the world [FAO (2012)], including developing countries [Hannerrza and Lotsch (2008)].

To identify the most appropriate Master Frame for an integrated survey, we will proceed as follow. Firstly, we will try to define a conventional Master Sampling Frame by identifying the common features of current Master Sampling Frames of Agriculture and Master Sampling Frames of Household (sections 2, 3 and 4). Secondly, we will try to concretize the features of a conventional Master Sample for an integrated survey based on this conventional Master Sampling Frame (section 5). Finally, we will try to identify the aspects of this conventional Master Sample that are susceptible of being optimized and we will review the approaches proposed in the literature to optimize sampling design (section 6). The most appropriate Master Frame has the same features

than the conventional one, once the aspects susceptible of being optimized have been optimized.

2. MASTER SAMPLING FRAMES OF AGRICULTURE

King (1945) credited the idea of a Master Sample to Rensis Likert, from the Bureau of Agricultural Economics (BAE) in the United States Department of Agriculture. The idea was to have a large sample from which subsamples of farmers could be selected. In later usage, the term "Master Sample" has come to be applied to the materials used in the creation of the first sample. That is, the term is often applied to the frame rather than to the sample itself.

While large-scale area samples had been used in India and in Europe in the 20s and 30s (King, 1945; Stephan, 1948), the Master Sample of Agriculture developed by the BAE represented a major step forward in sampling technique. The Master Frame should be as complete, accurate and current as practicable and only an area frame can provide these three characteristics. Hence, use of cartographical materials is required to build the area frame and to obtain measure of size for the sampling units, and an extensive use of cartographical materials distinguished the BAE approach from previous approaches.

The sampling units of the BAE Master Sample were relatively small and the units were designed by the statisticians, specifically for the purpose of the sample. An attempt was made to optimize sampling units with respect to travel time and interviewing costs [Jessen (1942)]. These innovations set standards for area sampling that continue to this day [Fuller (1984, 2010)] as can be seen in FAO (1996, 1998).

In UN (1986) one can find examples of countries where surveys were conducted comprising agricultural and household surveys, using list frames (Ethiopia 1980-1983). The problem with lists is that frames usually are incomplete, contain duplications and are not updated. Sampling areas is an improvement over other designs and therefore an area frame should be a basic component of the master frame.

2.1 Stratification

In the Master Sample of Agriculture developed by the BAE and the Bureau of the Census at the Statistical Laboratory of the Iowa State College, three primary strata were defined on the basis of incorporation and density of population [King (1945)]. The total area of the United States was classified into: (i) incorporated areas (such as a city, town or village), (ii) unincorporated areas (which resembles a city, town or village but lacks its own government) relatively densely populated and (iii) open country strata (agricultural areas relatively sparsely populated) [Jessen (1945)].

The stratum (i) includes all incorporated cities and towns (it represented 1% of the 1940 census land area, 63.6% of the 1940 census population and 3.4% of 1940 census farms), including unincorporated places regarded as "urban" by the Bureau of the Census. The stratum (ii) was further stratified into "rural" and "urban" groups according to Census rule (those places having a population of less than 2500 persons were classed "rural" in 1940 while those of 2500 or more population are classed "urban") and the areas classified as urban were included in (i) (it represents 3% of the 1940 census land area, 10.5% of the 1940 census population and 5.9% of the 1940 census farms). The

remaining areas, after the designation of the incorporated and un-incorporated areas, were included in (iii) and were labeled by "open country" (contains 96 per cent of the 1940 census land area, 25.9% of the 1940 census population and about 91 per cent of the farms of 1940 census definition) [Jessen (1945)].

2.2 Segmentation into sampling units

The strata were assigned a number of small areas (segments) which average about 2.5 square miles (700 hectares) in size but which vary according to strata, location and other circumstances. Nevada has the largest average size: 108 square miles (27972 hectares); Indiana the smallest: 0.71 square miles (184 hectares). The sample areas have been selected from every one of the 3070 counties' of the United States and contain within their boundaries approximately 1/18 of the land area of the United States, 1/18 of the farms (about 300000) and 1/18 of the rural population.

By design, these small areas were formed to contain, within reasonable limits, a given desired number of farms or persons within reasonably identifiable boundaries. A big procedural job was the setting up of criteria for use in determining the size, shape, and boundaries of these areas.

2.2.1. Within the open country stratum

Giving consideration to the problem of finding on the maps sufficient identifiable boundaries for small areas (segments), three general regions were considered within each of which a different average size of sampling unit would be sought. In region number one (the Corn Belt) the segment varied from two to six farms in size, but the goal was set at four farms per sampling unit. In region two (the areas outside the Corn Belt but within the Public Land Survey System), an average of three farms per sampling unit was thought feasible. In region three (all other areas of the country), the goal was set at six farms per sampling unit.

These sizes were the results of an attempt to strike a suitable balance between the smaller area which is usually statistically more efficient and the larger area which is less costly to enumerate (per farm) and prepare for, and which generally has fewer problems of boundary identification. Some study has been made of these problems [Jessen (1942); Jessen and Houseman (1944)] which are reviewed and are extended in section 5.3.

2.2.2. Within the unincorporated stratum.

In the unincorporated stratum some rather arbitrary criteria for assigning the number of sampling units to each of the unincorporated places were applied: (i) a number of sampling units was so assigned that the expected number of farms per sampling unit was six; (ii) a minimum of two sampling units was required for each unincorporated place regardless of how few farms were indicated; (iii) the number of sampling units assigned by rules (i) and (ii) was modified when necessary to utilize most effectively identifiable boundaries shown by the map.

Lists were prepared which grouped the unincorporated places into three size groups: (i) from 100 to 499 population, (ii) from 500 to 999 populations, and (iii) 1,000 population

and over. In addition to the name of the place, this listing showed its population, its expected number of farms, the number of sampling units into which it would be divided, and the cumulative total of sampling units through the three size groupings.

2.2.3. Within the incorporated stratum

As in the case of the un-incorporated stratum, a listing was made of all incorporated places within each county, grouped into the two size groups-those under 2500 population (rural according to the 1940 census) and those 2500 and over (urban). From special tabulations provided by the Bureau of the Census the number of 1940 census farms for each incorporated place was obtained. Each incorporated place was assigned a number of sampling units so that the average number of farms per sampling unit would be approximately six. With the sampling units so determined, sampling followed the same procedure as stated for the un- incorporated stratum.

2.3 Sampling scheme

For sample selection, clusters of sampling units (called count units) were considered. A random number between one and eighteen was chosen as the starting point and the cluster containing every eighteenth sampling unit was designated on the cumulative total column on the tabulation sheet. The scheme of numbering clusters of sampling units within minor civil divisions and minor civil divisions within the county made possible the ordering of clusters on the tabulation in such a way that geographic distribution of the sampling units would be maximized by this kind of systematic selection.

The cluster selected in this way was then located on the map, and its subdivisions into the appropriate number of sampling units were so indicated that each sampling unit could be identified. Only those clusters in the sample required actual subdivision into sampling units. This saved considerable clerical time. The sampling unit for the sample was chosen by selecting a random number from one to the number of sampling units in the cluster.

2.4 Master Sample of City Areas

The described Master Sample is an adequate sample of both farms and population in the open country stratum only. The areas set up for agriculture were not efficient for population samples in the cities and villages where the size of the sampling areas should be determined by the number of persons rather than the number of farms.

In the village and town stratum the Master Sample was redrawn by means of a supplemental selection of smaller areas with an average of 8 to 10 dwelling units per area, an adequate sample of persons in the unincorporated stratum. The extension follow the same general method of sampling, that is, the use of small areas clearly designated on maps, but the size of the unit will be measured in terms of total dwelling units rather than farms.

Note that the area sample is versatile: with appropriate rules of association it could be used to select samples of farms, fields, dwelling units or households. The two-stage design used to select the master sample substantially reduces the amount of preparatory

work needed, since only the selected count units had to be subdivided into sampling units.

3. MASTER SAMPLING FRAME OF HOUSEHOLD SURVEYS

The United Nations Statistical Office has carried out a series of publications designed to assist countries in planning and implementing household surveys in the context of the National Household Survey Capability Program [(UN (1986)]. The central topic of this Program is the development and maintenance of Master Sampling Frames and Master Samples of Household Surveys.

The principles that govern the establishment of a Master Sample Frame of Household, according to UN's guidelines, are little different from those for sampling frames in general, and particularly from that of agriculture. Some of these principles are: (i) the Master Frame should be as complete, accurate and current as practicable, (ii) define Primary Sampling Units (*PSU*s) in frame from area units such as census Enumeration Areas (*EA*s) with mapped, well-delineated boundaries and for which population figures are available, (iii) define Master Sample *PSU*s that are large enough or numerous enough to sustain many surveys, or repeat survey rounds, during intercensal period, (iv) use census list of households as frame at last stage only if very recent – usually no more than one-year old, (v) use dual or multiple frames with caution by ensuring procedures are in place to deal with duplications, (vi) employ system of sample rotation – either households or *PSU*s – in repeat surveys that use master samples.

To follow the aforementioned recommendations, particularly (i) and (ii), extensive use of cartographic material such as maps, aerial photos and satellite images is required, since only area frames assure completeness, accuracy and up-to-datedness of a Master Frame. However, the usual starting point is not an area frame but a country's population census, were units are administratively labeled but seldom geo-referenced in cartographic material. Labels differs from one country to another but typically include such terms, in descending order, as province or county; district; tract; ward and village (rural areas) or block (urban areas). For census purposes, administrative sub-divisions are further classified into enumeration areas or *EA*s [UN (1986)].

3.1 Frame units

The census Enumeration Area (*EA*) is the most appropriate area frame unit because it is often the smallest geographical unit that is defined and (sometimes) is delineated on maps in a country. In most countries *EA*s are purposely constructed to contain roughly equal numbers of households – often about 100 – in order to provide comparable workloads for census-takers.

3.2 Sampling units. Country examples

Depending on the surveys, *EA*s frame units turn out to be smaller or larger than the appropriate sampling unit. The size of the Primary Sampling Units (*PSU*s) must be sufficiently large to accommodate multiple surveys without interviewing the same respondents repeatedly. *EA*s turn out to be usually small and this is why the *PSU*s are defined aggregating several *EA*s in a cluster. In this case, *EA*s play usually the role of Secondary Sampling Units (*SSU*s).

Country examples can be found in UN (2005), where *PSU*s were defined by aggregating *EA*s. In Vietnam, for instance, the Master Sample is based on the 1999 population census. *PSU*s were defined as communes, in rural areas, and wards, in urban areas. They were defined in this way because it was decided that a minimum of 300 households would be necessary in each *PSU* to serve the Master Sample. Alternatively, *EA*s were considered as *PSU*s but they were too small and would have had to be combined with adjacent *EA*s in order to qualify satisfactorily as *PSU*s.

Each sample *PSU* contained, on average, 25 *EA*s in urban areas and 14 in rural areas. For the second stage of selection three *EA*s were selected in each sample *PSU*, using probabilities proportional to size. The *SSU*s were the *EA*s, which contain an average of approximately 100 households according to the 1999 census – 105 in urban areas and 99 in rural areas. For survey applications, a third stage of selection is administered in which a fixed number of households is selected from each sample *EA*. That number may vary by survey and by urban-rural. For example, 20 households per *EA* might be chosen for rural *EA*s and 10 per *EA* for urban ones.

Mozambique is another country example. Master Sample *PSU*s were constructed from the 1997 Population Census. They consist of geographical groupings of, generally, 3-7 census *EA*s, the latter of which contain about 100 households on average. The second stage of selection was a sub-sample of the *PSU*s. At the third stage a sample of one *EA* was selected in each of the *PSU*s. The *EA* were selected with equal probability because their sizes are roughly the same. The final stage of selection occurred following field work in which a fresh list of households was compiled to bring the 1997 sample frame up to date. From the lists so compiled, a systematic sample of 20 households in rural areas and 25 in urban was selected for interviews.

Finally, Cambodia's National Institute of Statistics developed a Master Sample in 1999 from its 1997 population census. It was decided to use villages as the *PSU*s because they are large enough (on average, 245 households in urban areas and 155 in rural), to have enough households to accommodate several surveys during the intercensal period. The alternative of using census *EA*s was considered but discarded because they are only half the size of villages, on average.

Within each selected Master Sample *PSU*, area segments of size 10 households (on average) were formed. The number of segments created within each *PSU* was computed as the number of census households divided by 10 and rounded to the nearest integer. The typical *PSU* contains about 18-30 segments, providing an ample number of segments in each *PSU* to sustain all surveys.

A limitation of the Master Sample of Households design is the use of compact clusters (all the households in the sample segment are adjacent to each other). This increases the design effect somewhat over non-compact segments, that is, a systematic sample of households within a larger cluster. Note that, to reduce the design effect, *EA*s are portioned into segments of only size 10 households.

In a conventional household survey, the last stage of selection is based on a list frame. Thus the *EA*s come from an area frame but a list frame defining the sample households

within the *EA*s or a sample of segments (blocks in urban areas) into which *EA*s are usually divided is used in the last sampling stage.

4. CONVENTIONAL MASTER SAMPLING FRAME

A conventional Master Sampling Frames is defined here by combining the features (frame unit, sampling unit and observation unit) found in the literature on Master Sampling Frames of Agriculture (reviewed in section 2) and on Master Sampling Frames of Households (reviewed in section 2). The target population is a country with its territory, its citizens, its natural resources and its economical resources. The Master Sampling Frame that we propose here to design integrated surveys of this target population is a multiple frame, made up of area frames and list frames. The required basic materials are cartography for the construction of the area frames, including aerial photography and satellite imagery; and censuses for the construction of the list frames. Additional lists are required to improve the estimation of specific characteristics.

Primary stratification

A primary stratification consists in dividing the country's territory into two strata: urban areas and rural areas. The OCDE criterion is proposed for this stratification.

The limits between these two strata have to appear on the cartographical material

Frame unit

A sampling frame is made up of frame units. The frame unit considered here is the Enumeration Area (*EA*), as defined in the censuses, where the objective in establishing *EA*s size is to limit and more or less equalize the workloads of individual census enumerators: this size is, on average, 100 households for urban areas and 100 farms for rural areas. If neither population census nor agricultural census is available, then an *EA* having the features of those described in sections 2 and 3 is defined.

The *EA*s limits have to appear on the cartographical material. In this way, they are the link between area frames (based on cartographic material) and list frames (based on censuses).

Secondary stratification

The *EA*s are stratified in a number of strata, using as criterion land use intensity. According to this criterion, *EA*s from urban areas are classified in the first stratum and *EA*s from rural areas are stratified in a number of strata (non more than five) according to the percentage of cultivated land.

Sampling unit

The sample selection procedure includes more than one stage and a sampling unit has to be defined for each stage. In the conventional Master Sample of Households, three stages are considered: in the first stage, the Primary Sampling Unit (*PSU*) is a cluster of *EA*s, in the second stage, the Secondary Sampling Unit (*SSU*) is the *EA*, and in the third stage, the sampling unit is a small area called "block". In the conventional Master

Sample of Agriculture, the number of stage is reduced to two. The *PSU*s are *EA*s and in the last stage, the sampling units are the segments.

The size of the sampling unit is the result of an attempt to strike a suitable balance between the smaller area which is usually statistically more efficient and the larger area which is less costly to enumerate. In urban areas, *PSU* size is usually between 2 and 4 *EA*s. The last stage sampling unit size is fixed at 10 households per block (in urban areas) or 10 farms per segment (in rural areas), on average, that is 10 blocks (in urban areas) or 10 segments (in rural areas) per *EA*, on average.

The area of the block and the area of the segment to achieve its target size change between landscape types. A target size is defined for each stratum, as a function of the average building size (in urban areas) or the average field size (in rural areas) in the stratum. For instance, if the target number of fields per segment is 20, then the target segment size is 20 multiplied by the average field size in the stratum.

Observation units

Four observation units are considered: the person, the household, the farm (subsistence, small holder or commercial farms) and the field. In urban areas, the main observation units are persons and households. In rural areas, the main observation units are farmers, farms and fields. In the Multiple Frames framework, person, farm and household observation units are selected from a list frame. This list has to be elaborated but only in those segments and blocks included in the sample. Segments and blocks are area sampling units and are selected from an area frame, as is also selected the field.

5. CONVENTIONAL MASTER SAMPLE

A systematic sampling scheme is used to select a sample of *PSU*s within each stratum, with probabilities proportional to size. To maximize geographic distribution of the sampling units, *PSU*s are numbered in serpentine fashion. In urban areas, a *PSU* is a cluster of *EA*s and a sample of them is selected with equal probability, within each *PSU* included in the sample. *EA*s included in the sample are subdivided into the target number of blocks and a sample of them is selected with equal probability. In rural areas, *EA*s are the *PSU*s and those *EA*s included in the sample are subdivided into the target number of segments and a sample of them is selected with equal probability. Note that only those *EA*s included in the sample are actually subdivided into the target number of blocks (urban areas) or segments (rural areas).

6. THE MOST APPROPRIATE MASTER SAMPLING FRAME AND MASTER SAMPLE FOR AN INTEGRATED SURVEY. EXAMPLES

The most appropriate Master Frame has the same features than the conventional one. However, population census or agricultural census not always is available and two main scenarios regarding the kind of data sources available are considered bellow: in one it is assumed that an updated agricultural census is available and in the other scenario it is assumed that satellite imagery is the only material available.

Satellite imagery is a data source on land cover and land use usually available, and hence it is a good basis for integrating agricultural and household surveys with

environmental issues. A Geographical Information System is a useful tool for land use stratification using satellite data, as well as for geo-referencing the administrative units of a country, particularly the enumeration areas (*EA*s).

Once the *EA*s have been geo-referenced, both the farms registered in the agricultural census and the farm and nonfarm households registered in the population census can be linked with each other and with land use data. That can be achieved defining an appropriate criterion of association between farms, households and *EA*s.

6.1 Master sampling frame with an agricultural census. Examples

The agricultural census should include all kind of farms: subsistence farming households, familiar farms and corporate farms. Subsistence farming and familiar farming are invariably associated to a households and the Master Sampling Frame has to take into account this association: each farm should be associated with a household unless it is a corporate or institutional farm.

Just as the link between households and *EA*s is established through the place of residence, the link between farms and *EA*s can be established through the land of each farm. Once the *EA*s are geo-referenced to the satellite imagery, the farms registered in the agricultural census, together with the households associated with them, and together with the nonfarm households registered in the population census, can be linked with the appropriate *EA* and, as a result, the Master Sampling Frame integrates the population census and the agricultural census.

As pointed out in section 2, this Master Sampling is adequate to select samples of farms and persons in the open country stratum only. In this stratum, the *EA*s are partitioned into segments (sampling units), which size is determined as a function of the average number of farms per segment. The sampling units set up for agriculture are not efficient for population samples in the cities and villages, where the size of the sampling areas has to be determined by the number of persons, rather than the number of farms.

As pointed out is section 3, in the village and town stratum the Master Sampling has to be redrawn by means of aggregating *EA*s to form *PSU*s sufficiently large to accommodate multiple household surveys. The *EA*s plays then the role of *SSU*s and a third sampling stage is required for a supplemental selection of smaller areas (segments) with an average of 8 to 10 dwelling units per area.

Country example: Chile´s Master Sample of Agriculture

In Chile, a Master Sample of Agriculture was designed, using an Area Frame based on the 2007 Agricultural Census [Ambrosio (2012)]. The total area of Chile was classified into two primary strata: (i) Urban areas (such as city, town or village) and (ii) non-urban areas (the remaining areas). The areas in stratum (i) were delimited using the digital limits available from the 2011 Population Census. The non-urban areas were stratified into three secondary strata, according to land use: agriculture (cultivated land), livestock (meadow and extensive grasslands) and forest and bushes.

Stratification

Satellite imagery were considered as a basis for land use stratification but was discarded because the Chilean landscape is a mosaic of agricultural fields, dispersed between non-agricultural areas, and finding sufficient identifiable boundaries for small areas (segments) on the images is a big problem. Instead, the agricultural and livestock areas were stratified using the 2007 Agricultural Census [INE (2007)]. In this Census the frame units are the *EA*s and their limits are available on digital form. Land use data aggregated at the *EA*s level is available and these data were used to classify *EA*s into strata, using multivariate statistical methods. A similar approach can be found in Abaye (2010).

Sampling units

Giving consideration to the problem of finding on the maps and satellite images sufficient identifiable boundaries for sampling units, it was decided to divide the *EA*s into segments of geometrical limits, instead of identifiable boundaries. A square grid of side 500 meters was superimposed on the stratified *EA*s to define segments of size 25 hectares. Each sampling unit (segment) is assigned to the strata where lie the highest part of its area and the strata limits were adjusted to those of square segments.

The *EA* average size is 100 farms

Multiple frames

This is an adequate Master Sample to estimate areas under the main crops, using the closed segment estimator. It is also a good basis for selecting a sample of farms, and particularly livestock farms to estimate livestock characteristics. However, to improve estimates of very specific crops (industrial crops and vegetable crops), complementary samples selected from list frames are required.

Eliminating overlaps between area and list frames were considered but was discarded because it is very laborious. Instead, we used estimators that take into account overlaps. A review of these estimators can be found in Ambrosio (2012).

6.2 Master sample frame when there is not a recent agricultural census. Examples

As pointed out in section 4, only area frames assure completeness, accuracy and updating of a Master Frame. Hence, the Master Sampling Frame is based on an area frame.

Stratification

If there is not cartographic material available, satellite images can be used to build the area frame [Gallego (1995)]. The starting point is to classify satellite imagery by land-use categories. Then, the sampling units have to be referred to satellite imagery.

Sampling units

The *EA* is chosen as frame unit since it allows to link population and agricultural censuses and to refer both to satellite imagery. In urban areas, the *EA*s are grouped in *PSU*s and sampled in the second sampling stage. In non-urban areas, a target sampling

unit size is chosen for the third sampling stage and then a number of sampling units are assigned to each *EA*. A criterion to associate farms and households to sampling units is established in order to geo-reference the farms and the households as well as to link the area frame with the population census.

Multiple frames

In non-urban areas, to improve estimates of very specific crops (industrial crops and vegetable crops), as well as to estimate characteristics of especial farms, complementary samples selected from list frames are required. These list frames can be elaborated with the help of professional organizations and the agro-industrial sector. In urban areas these list are usually required for sampling rare populations.

Eliminating overlaps between area and list frames is very laborious. Estimators that take into account overlaps are usually more cost-efficient.

Country example: Guatemala's Master Sample of Agriculture

In Guatemala, the last agricultural census is dated 2003, but it will be updated this year. The frame unit in the agricultural census is the *EA* coming from the population census. There are 12546 geo-referenced *EA*s. They are aggregated at the following administrative levels, in ascending order: 3219 tracts, 334 counties, 22 departments and 8 regions. Thus the average number of *EA*s per administrative level (to the nearest entire) is 4 *EA*s/tract, 38 *EA*s/county, 570 *EA*s/department and 1568 *EA*s/region.

Stratification

A Master Sample of Agriculture is being built using a map of land cover and land use dated 2005 [Ambrosio (2013)]. More recent satellite imagery were considered but they were discarded because the Directorate General of Strategic Geographical Information and Risk Management of the Ministry of Agriculture, Livestock and Food, which has built the land use map, is carrying out updating.

The total area of Guatemala was classified into two primary strata: (i) Non-agricultural areas (including cities, towns or villages, as well as permanent water areas, forest areas protected by the law and other non-agricultural land) and (ii) agricultural areas (the remaining areas in the country, including agro-urban areas ("traspatio")).

An area frame was built for agricultural areas. Agricultural areas were stratified into five strata, according to land use. Areas where the percentage of cultivated land is more than 60% were divided into two strata according to the field size: the stratum A is that of big fields and the stratum B is that of small fields. The stratum C is defined by areas where the percentage of cultivated land is between 20% and 60%. The areas fewer than 20% of cultivated land are classified as stratum D. A specific stratum for vegetable crops, coded as E, was delimited [FAO (1996,1998)].

Sampling units

A target segment size was defined for each stratum, as a function of the average field size in the stratum. In order to keep non-sampling errors within tolerable limits, it was

considered that the average number of fields per segment should be between 15 and 25 [Taylor et al (1997)]. In the stratum A, the average field size is estimated to be 1.25 hectares and, as a result, the target segment size is 25 hectares [((15+25)/2) x 1.25=25]. In the same way, target segment sizes for the remaining strata were defined: 6.25 hectares for stratum B; 50 hectares for stratum C and 100 hectares for stratum D.

Initially, it was decided to use segments with physical boundaries in the specific stratum for vegetable crops and segments with geometrical boundaries in the remaining strata. Finally, segments with geometrical boundaries were used in all the strata, due to budget and calendar reasons. The target segment size in the specific stratum for vegetable crops was 6.25 hectares.

Land stratification was carried out using the digitized land cover and land use map. The starting point is to compute the percentage of cultivated land in a square lattice of 1000 meters side: 100 hectares, which is the target segment size in the biggest stratum, D. Then, each square of the grid is classified in a stratum according to its percentage of cultivated land. Squares in the stratum more than 60% of cultivated land are sub-stratified according to the average field size in the stratum. Squares in the stratum A were divided into four squares of size 25 hectares each (which is the target segment size in this stratum) and those in stratum B were divided into 16 squares of size 6.25 hectares each one. The squares of side 1000 meters that were classified in stratum C, were divided into two parts of size 50 hectares each one (which is the target segment size in this stratum).

The total vegetable crops area in the land use map was partitioned into segments of 6.25 hectares.

To link the area frame with the agricultural census and with the population census, each segment is associated with an *EA*

Observation units

Within each segment, three observation units are considered: the field, the farm and the household. Each field is associated with a farm and each farm is associated with a household, unless it is a corporate or institutional farm. Since each segment is associated with an *EA*, the area frame is linked to the agricultural census as well as to the population census.

Multiple frames

To improve estimates, complementary samples from list frames will be used together with the area sample. List frames are being elaborated with the help of professional organizations and the agro-industrial sector.

Master Sample of City Areas

The Master Sample of Agriculture is an adequate sample of farms and persons in the "agricultural areas" stratum, including households with “traspatio” activities. However, it is not adequate for urban areas. In urban areas, census tracts are being considered as *PSU*s. An average *PSU* contains 4 *EA*s and 400 household. *EA*s are being considered as

*SSU*s. An average *EA* contains 100 household and it is being considered that a sample of 2 *EA*s will provide enough households (between 100 and 300) to accommodate the several required surveys, included food security surveys.

The *EA*s that are included in the sample of the urban areas stratum will be divided into blocks, with an average of 8 to 10 dwelling units per block. A sample of blocks will be selected with equal probability and the households within the limits of these blocks will be included in the sample.

Within each *EA*, two observation units are considered: the farm and the household. Households with "traspatio" activities are a third observation unit category, which is relatively important in agro-urban areas, particularly for some livestock species (poultry house).

7. OPTIMIZING THE SAMPLING DESIGN

Desirable frame properties are mainly of three categories: quality-related, efficiency-related and cost-related. Quality-related are those properties which make it possible to minimize non-sampling errors: mainly coverage and measurement errors. Efficiency-related properties of frames are those qualities that make possible to minimize sampling errors. Properties that favor quality and efficiency in the use of sampling frames usually have costs attached to them. Cost-efficiency in this context refers to the relationship between sampling error and the cost of producing survey estimates: the most efficient survey design is the one that produces the desired level of precision at the lowest possible cost or the one that minimizes the sampling variance, conditionally to unitary costs and total budget [UN (1986, 2005)].

Most of the literature on survey design, both theoretical and applied, addresses the question of how to optimize the design of a single survey. However, a large sample is needed for master samples in order to provide enough farms and households to support multiple surveys over several years without having to interview the same respondents repeatedly. Furthermore, even single surveys are multipurpose in nature and the optimum for a purpose it is not optimum for other purposes. Some aspects of the multipurpose single surveys are well developed in the literature, such as multipurpose stratification [Jessen (1978)] or the multiple probability proportional to size sample selection [Steiner (2005)].

However, as pointed out by Groves (1989), optimizing would not be useful if the sole purpose is to solve for the optimal design. The cost-error approach is interesting because it offers a systematic way to evaluate alternative sampling designs. When faced to the problem of designing a large multipurpose sampling, the designer can use this approach to compare the cost-efficiency of the set of single-purpose sampling designs, corresponding to the set of purposes. "This can only benefit the design relative to the blind acceptance of some standard solution" [Groves (1989, p.56)].

Optimizing sampling unit size

The features of a Master Sample deal with the number, size and type of sampling units. The most cost-efficient sampling unit size can be theoretically identified, according to

statistical criteria. Once it has been identified, it can be used not for optimizing but "proximizing" [according to Kish, quoted by Groves (1989, p. 53)] the Master Sampling Frame. The sampling variance depends on the variability of the frame units within and between sampling units. A review of the literature on models of this variability can be found in Jessen (1978). More recently, Gallego and Carfagna (1995), Carfagna (1997), Gallego et al (1999), and Ambrosio et al (2003) propose to assess this variability using variogram functions.

Jessen (1978) differentiates between frame units and sampling units (a sampling unit is a cluster of frame units). This author approaches the problem as follow. Consider two sampling frames, a and b, of a given population. Sampling units are assumed to be of the same size and shape within a given frame but differs between frames.

Frame	Nº of sampling units	Size of the sampling unit (number of frame units)	Total	Mean	Variance	Sample size
a	M_a	$N_a = N/M_a$	Y	$\bar{Y}_a = Y/M_a$	S_a^2	m_a
b	M_b	$N_b = N/M_b$	Y	$\bar{Y}_b = Y/M_b$	S_b^2	m_b

If the sampling scheme is simple random, then the total unbiased estimators and their variances in each frame are as follow:

Frame	Estimator of the Total	Variance Estimator
a	$\hat{Y}_a = M_a \hat{\bar{Y}}_a$	$V(\hat{Y}_a) = M_a^2 (1 - f_a) \frac{S_a^2}{m_a}$
b	$\hat{Y}_b = M_b \hat{\bar{Y}}_b$	$V(\hat{Y}_b) = M_b^2 (1 - f_b) \frac{S_b^2}{m_b}$

The most efficient sampling unit is that minimizing the sampling variance for a given cost or makes minimum the cost for given variance.

7.1 Relative efficiency

The relative efficiency of the frame b with respect to the frame a is

$$RE_{b/a} = \frac{V(\hat{Y}_a)}{V(\hat{Y}_b)} = \frac{M_a^2 (1 - f_a) \frac{S_a^2}{m_a}}{M_b^2 (1 - f_b) \frac{S_b^2}{m_b}}$$

Since sampling units, a and b, can be very different in size, it is convenient comparing their efficiency under the constraint that the total number of frame units included in the sample be equal:

$$m_a N_a = m_b N_b \Leftrightarrow \frac{m_a}{N_a} = \frac{m_b}{N_b} \Leftrightarrow f_a = f_b$$

So that,

$$m_a = m_b \frac{M_a}{M_b} = m_b \frac{N_b}{N_a}$$

Thus, the relative efficiency of the frame b with respect to the frame a is:

$$RE_{b/a} = \frac{V(\hat{Y}_a)}{V(\hat{Y}_b)} = \frac{M_a^2(1-f_a)\frac{S_a^2}{m_a}}{M_b^2(1-f_b)\frac{S_b^2}{m_b}} = \frac{M_a S_a^2}{M_b S_b^2}$$

If a is the frame unit and b is a cluster of N_0 frame units, then:

$$N_a = 1 = N/M_a \Rightarrow M_a = N$$
$$N_b = N_0 = N/M_b \Rightarrow M_b = N/N_0$$

and ,

$$RE_{b/a} = \frac{S_a^2}{S_b^2 / N_0}$$

7.2 Net relative efficiency

Let C_a and C_b be the cost of observing a sampling unit a and b, respectively. If C is the total available budget, then the sample size that it is possible to observe is $m_a = C/C_a$ using sampling unit a and $m_b = C/C_b$ using sampling unit b. Suppose, to simplify, that the sampling fractions are small enough to be $f_a \cong 0$ and $f_b \cong 0$.

Thus,

$$RE_{b/a} = \frac{V(\hat{Y}_a)}{V(\hat{Y}_b)} = \cong \frac{M_a S_a^2 C_a}{M_b S_b^2 C_b}$$

If a is the frame unit and b is a cluster of N_0 frame units, then:

$M_a = N; M_b = N/N_0$ and $C_b = N_0 C_a$, so that, $RE_{b/a} = \frac{S_a^2}{S_b^2/N_0}$

7.3 Variance functions

The optimum sampling unit size - N_0 - is that which makes minimum the sampling variance for given C, or that which makes minimum the total estimation cost for a given sampling variance level. If both sampling variance as costs were known for every relevant value of N_0, then the net relative efficiency could be used as criterion for choosing N_0. Furthermore, if it would possible to express both, sampling variance as costs , as N_0 functions, then it would be possible to determine the optimum N_0, by minimizing the variance function constrained to a given cost or minimizing the cost function constrained to a given variance level. This point will be considered later (see section 7.6). Now a more simple procedure will be considered

The sampling variance using cluster random sampling is (Jessen (1978), p 107):

$$V(\hat{\bar{Y}}) = M^2(1-f)\frac{S^2}{m}\left[1+(N_0-1)\rho\right]$$

where ρ is the intraclass correlation coefficient. If M is large, then

$$\rho = \frac{S_E^2 - \bar{\bar{S}}_D^2}{S^2} = \frac{S_E^2 - \bar{S}_D^2/N_0}{S^2}$$

and

$$V(\hat{\bar{Y}}) = M^2(1-f)\frac{S^2}{m}\left[1+(N_0-1)\frac{S_E^2 - \bar{\bar{S}}_D^2}{S^2}\right]$$

where

$$S^2 \cong N_0 S_E^2 + \bar{S}_D^2$$

and

$$V(\hat{\bar{Y}}) = M^2(1-f)\frac{1}{m}\left[N_0^2 S_B^2 + N_0 \bar{S}_W^2\right] \quad [1]$$

where,

$$S_B^2 = \frac{S_b^2/N_0 - \bar{S}_D^2}{N_0}$$

The sampling variance of the mean by frame unit is:

$$V(\hat{\bar{Y}}) = (1-f)\frac{1}{m}\frac{(N_0 S_B^2 + \bar{S}_D^2)}{N_0} \quad [2]$$

In a given population, the variance between individual elements -S^2- is fixed and independent of the sampling unit (cluster) size and shape. To each partition of the population in clusters corresponds a decomposition of S^2 into two components $-S_B^2, \bar{S}_D^2-$. Assuming that this decomposition is independent of N_0 and using a simple cost function:

$$C = mC_B + mN_0C_D \quad [3]$$

where C_B are C_D the observation costs per sampling unit (cluster) and frame unit, respectively, then the N_0 optimum, e.g. that minimizing [1] or [2] constrained to [3] is,

$$N_{0Optimo} = \sqrt{\frac{C_B}{C_D}\frac{\bar{S}_D^2}{S_B^2}}$$

Or, in terms of the intraclass correlation coefficient,

$$N_{0Optimo} = \sqrt{\frac{C_B}{C_D}\frac{1-\rho}{\rho}}$$

Smith (1938) considers that if the sampling units (clusters) were formed choosing at random N_0 frame units from among the population, then $\bar{S}_D^2 = 0$ and $S^2 \cong S_b^2/N_0^2 = S_E^2$, so that $S_b^2/N_0 = S^2N_0$. However, clustering usually implies $\bar{S}_D^2 > 0$ so that $S_b^2/N_0^2 \cong S^2 - \bar{S}_D^2 < S^2$ and $S_b^2/N_0 < S^2N_0$.

Smith (1938) proposed a variance function, $S_E^2 = S^2/N_0^g$, where g is a constant less than 1, which can be estimated form the sample data. Mahalanobis (1940) and Jessen (1942) consider the variance function $\bar{S}_D^2 = AN_0^g$, where $g > 0$ an d A are constants, which can be estimated from the sample data. If $N_0 = N$, then $S^2 = AN^g = A(N_0M)^g$. If M is large, then $S_b^2/N_0 \cong N_0S^2 - (N_0 - 1)\bar{S}_D^2$, so that the expressions given S^2y $\bar{S}_D^2$ as a function of N_0, we have: $S_b^2/N_0 \cong AN_0^g\left[N_0M - N_0 + 1\right]$. Hansen et al (1953) propose an intraclass correlation function, $\rho = AN_0^h$, where h is usually positive and less than 1. See Pennigton and Volstad (1991) for other variance functions.

Using these functions, the variance estimator, $V(\hat{Y})$, given in [1] or [2], can be expressed as a N_0 function. Thus, the optimum N_0 is found by minimizing $V(\hat{Y})$ subject to C. We will follow later this approach for the lattice model, but using the correlogram function, instead of the variance function.

7.4 Sampling on a lattice

The importance of the lattice case derives from the fact that so much data on land use is now available through remote sensing, which comes in the form of pixels (lattice grid). The population is a finite number of frame units arranged in a lattice. Following Das (1950) and Bellhouse (1977), the *MN* frame units of the population are considered

arranged in $M=ml$ rows and $N=nk$ columns. The frame units are grouped in mn sampling units with $N_0 = lk$ elements in each.

This kind of arrangement is often the basis for spatial sampling and is found or can be constructed from many cartographic or other representations of space, including remote sensing. In the UTM (Universal Transverse Mercator) system, the territory is divided into square units of size "lxl" by lines parallel to the axis of a Cartesian system of reference. These units are arranged in blocks of mxm units. The UTM co-ordinates identify the row and the column of the square unit.

Theoretical results in this area provide some evidence on the relative efficiencies of the different sampling schemes. For the one-dimensional case, Cochran (1946) shows that when the correlation function is non-negative and non-increasing convex, systematic sampling of frame units is more efficient than either random or stratified sampling. Das (1950) points out that this is true even if the correlation function takes negative values. Hajek (1959) has extended Cochran's (1946) results and has shown that systematic sampling is optimum among all designs with the same probabilities of inclusion (see also Iachan (1985)).

For the two-dimensional case, Bellhouse (1977) has shown there is no uniformly optimum sampling scheme for a general class of correlograms. He does show that when the correlation function is such that $\Delta_u^2\Delta_v^2\rho(u,v) \geq 0; \forall u,v \geq 0$ then a systematic sample will always be preferred to a stratified random sampling scheme. Das (1950) identifies situations when stratified sampling is more efficient than random and when systematic sampling is more efficient than stratified random sampling.

However the problem with theoretical findings to date is that it can be difficult to know when the different conditions hold and hence what sampling strategy should be selected in any practical situation. The purpose here is to propose a more practical basis from which to choose a sampling frame and a sampling scheme. As pointed out by Gallego and Stibig (2013) and Carfagna (1998), the correlogram is a useful tool for optimizing an area sampling design. Here, we will focus on this tool following Ambrosio et al. (2003).

7.4.1 Random sampling

A random sampling scheme of sampling units is analysed first: (i) ϑ sampling units are chosen with equal probability from the mn sampling units of the population; (ii) the $N_0 = lk$ frame units in each of the sampling unit is observed. So, the sample size is ϑlk frame units, grouped in ϑ sampling units of $N_0 = lk$ frame units each.

A design-based unbiased estimator of the population mean per sampling unit is the sample mean:

$$\hat{\bar{Y}}_{su(r)} = \frac{1}{\vartheta}\sum_{i=1}^{\vartheta} y_{i.} \qquad [4]$$

The design-based variance of this estimator is:

$$V(\hat{\bar{Y}}_{su(r)}) = (1-\frac{\vartheta}{mn})\frac{1}{\vartheta}\frac{1}{mn-1}\sum_{i=1}^{mn}(y_{i.}-\bar{y}_{..})^2$$

where:

$$y_{i.} = \sum_{j=1}^{lk} y_{ij}$$

$$\bar{y}_{..} = \frac{1}{mn} \sum_{i=1}^{mn} y_{i.}$$

The expected value of $V(\hat{\bar{Y}}_{su(r)})$, assuming that the data have been generated by a second – order stationary process, is [Das(1950)]:

$$EV(\hat{\bar{Y}}_{su(r)}) = (1 - \frac{\vartheta}{mn}) \frac{1}{\vartheta lk} \sigma^2 [1 - \frac{mnlk - 1}{mn - 1} \Phi(ml, nk) + \frac{mn(lk-1)}{mn-1} \Phi(l,k)] \qquad [5]$$

where:

$$\Phi(ml, nk) = \Phi(M, N) = \frac{2}{M(MN-1)} \sum_{v=1}^{M-1} (M - v) \rho(0, v) + \frac{2}{N(MN-1)} \sum_{u=1}^{N-1} (M - u) \rho(u, 0) + \frac{2}{MN(MN-1)} \sum_{u=1}^{N-1} \sum_{v=1}^{M-1} (N - u)(M - v) \Psi(u, v)$$

$$\Phi(l,k) = \frac{2}{l(lk-1)} \sum_{v=1}^{l-1} (l - v) \rho(0, v) + \frac{2}{k(lk-1)} \sum_{u=1}^{k-1} (k - u) \rho(u, 0) + \frac{2}{lk(lk-1)} \sum_{u=1}^{k-1} \sum_{v=1}^{l-1} (k - u)(l - v) \Psi(u, v)$$

$$\Psi(u,v) = \rho(u,v) + \rho(-u,v)$$

The term $\rho(u,v)$ is the value of the correlogram function in a point of the frame located a number u of frame units from the origin in the rows direction and a number v of frame units from the origin in the columns direction. The term $\Phi(M,N)$ is the average correlation between all pairs of frame units of the population. The term $\Phi(l,k)$ is the average correlation between all pairs of frame units of the same sampling unit.

7.4.2. Random sampling and stratified sampling of frame units

In random sampling of frame units, a sample of ϑlk frame units is selected with equal probability among the $mnlk$ elements of the population. A design-based unbiased estimator of the population mean per frame unit is the sample mean:

$$\hat{\bar{Y}}_{fu(r)} = \frac{1}{\vartheta lk} \sum_{i=1}^{\vartheta lk} y_i \qquad [6]$$

The design-based variance of this estimator and its expected value assuming that the data have been generated by a second – order stationary process are [Das (1950)]:

$$V(\hat{\bar{Y}}_{fu(r)}) = (1 - \frac{\vartheta}{mn}) \frac{1}{\vartheta lk} \frac{1}{mnlk - 1} \sum_{i=1}^{mn} \sum_{j=1}^{lk} (y_{ij} - \bar{Y})^2 \ ,$$

where $\bar{Y} = \frac{1}{mnlk} \sum_{i=1}^{nm} \sum_{j=1}^{lk} y_{ij}$

$$EV(\hat{\bar{Y}}_{fu(r)}) = (1 - \frac{\vartheta}{mn}) \frac{1}{\vartheta lk} \sigma^2 [1 - \Phi(M,N)] \qquad [7]$$

In stratified random sampling, each sampling unit is considered as a stratum of frame units and a random sample of frame units of size ϑ is selected in each. A design-based unbiased estimator of the population mean per frame unit is:

$$\hat{\bar{Y}}_{fu(str)} = \frac{1}{mn}\sum_{i=1}^{mn}\frac{1}{\vartheta}\sum_{j=1}^{\vartheta} y_{ij} \qquad [8]$$

The design-based variance of this estimator and its expected value assuming that the data have been generated by a second – order stationary process are [Das (1950)]:

$$V(\hat{\bar{Y}}_{fu(str)}) = \frac{1}{(mn)^2}\sum_{i=1}^{mn}(1-\frac{\vartheta}{lk})\frac{1}{\vartheta}\frac{1}{lk-1}\sum_{j=1}^{lk}(y_{ij}-\bar{Y}_i)^2$$

where $\bar{Y}_i = \frac{1}{lk}\sum_{j=1}^{lk} y_{ij}$

$$EV_{fu(str)}(\hat{\bar{Y}}) = \frac{1}{(mn)^2}\sum_{i=1}^{mn}(1-\frac{\vartheta}{lk})\frac{1}{\vartheta}\sigma^2[(1-\Phi(l,k)] \qquad [9]$$

7.4.3. Relative efficiency

We focus on the relative efficiency of using sampling units (clusters), instead of frame units

Random sampling of sampling units vs. random sampling of frame units

The relative efficiency of random sampling of sampling units with respect to random sampling of frame units, $RE_{su(r)/fu(r)}$, is defined as the quotient between the expected values of the sampling variances,

$$RE_{su(r)/fu(r)} = \frac{EV(\hat{\bar{Y}}_{fu(r)})}{EV(\hat{\bar{Y}}_{su(r)})}$$

The expected values $EV(\hat{\bar{Y}}_{fu(r)})$ and $EV(\hat{\bar{Y}}_{su(r)})$ are averages of the design-based variances $V(\hat{\bar{Y}}_{fu(r)})$ and $V(\hat{\bar{Y}}_{su(r)})$, respectively, over all finite populations with a correlogram function $\rho(u,v)$.

The relative efficiency, $RE_{su(r)/fu(r)}$, basically depends on the correlation structure of the population:

$$RE_{su(r)/fu(r)} = \frac{EV(\hat{\bar{Y}}_{fu(r)})}{EV(\hat{\bar{Y}}_{su(r)})} = \frac{[1-\Phi(M,N)]}{[1-\frac{mnlk-1}{mn-1}\Phi(M,N)+\frac{mn(lk-1)}{mn-1}\Phi(l,k)]} \qquad [10]$$

Thus, the basic requirements is the identification of the correlogram function, $\rho(u,v)$.

It also depends on the size and shape (M and N) of the frame, through MN and $\Phi(M,N)$, and on the dimensions of the sampling unit, through lk and $\Phi(l,k)$. The relative efficiency is independent of the sample size and of the population distribution of the studied variable.

Note that if the size of the sampling unit, $l \times k$, is big enough for $\frac{1}{lk} \square 0$, then

$\left[1+\frac{MN-mn}{mn-1}\Phi(l,k)-\frac{MN-1}{lk-1}\Phi(M,N)\right] \square 1+(lk-1)\Phi(l,k)-mn\Phi(M,N)$ and, in order for random sampling of sampling units to be more efficient than random sampling of frame units, $RE_{su(r)/fu(r)}>1$, it is enough that $\Phi(l,k)<\Phi(M,N)$.

In this case, the design-effect is the inverse of the relative efficiency:

$$DE_{su(r)/fu(r)}=\frac{1}{RE_{su(r)/fu(r)}}=\frac{EV(\hat{\bar{Y}}_{su(r)})}{EV(\hat{\bar{Y}}_{fu(r)})}=\frac{[1-\frac{mnlk-1}{mn-1}\Phi(M,N)+\frac{mn(lk-1)}{mn-1}\Phi(l,k)]}{[1-\Phi(M,N)]}$$

Stratified sampling of frame units vs. random sampling of frame units

The relative efficiency of stratified sampling of frame units with respect to random sampling of frame units, $RE_{fu(str)/fu(r)}$, is defined as the quotient between the expected values of the sampling variances,

$$RE_{fu(str)/fu(r)}=\frac{EV(\hat{\bar{Y}}_{fu(r)})}{EV(\hat{\bar{Y}}_{fu(str)})}=\frac{[1-\Phi(M,N)]}{[1-\Phi(l,k)]} \qquad [11]$$

Note that for stratified random sampling of frame units to be more efficient than random sampling of frame units, $RE_{fu(str)/fu(r)}>1$, it is enough that $\Phi(M,N)<\Phi(l,k)$.

In this case, the design-effect is the inverse of the relative efficiency:

$$DE_{fu(str)/fu(r)}=\frac{1}{RE_{fu(str)/fu(r)}}=\frac{EV(\hat{\bar{Y}}_{fu(str)})}{EV(\hat{\bar{Y}}_{fu(r)})}=\frac{[1-\Phi(l,k)]}{[1-\Phi(M,N)]}$$

7.4.4 Modeling the spatial correlation structure

To assess the correlation structure, a theoretical correlogram is usually considered in the literature, which is defined as a continuous of the Euclidean distance between sampling points.

Two theoretical correlograms are considered here:

(i) Exponential:

$$\rho(u,v)=\rho\left(h=\sqrt{u^2+v^2}\right)=(1-\tau)e^{-h/a_0} \text{ and}$$

(ii) Spherical:

$$\rho(u,v)=\begin{cases}\rho\left(h=\sqrt{u^2+v^2}\right)=(1-\tau)\left[1-\frac{3}{2}\frac{h}{a_0}+\frac{h^3}{2a_0^3}\right]; h\le a_0\\ \rho\left(h=\sqrt{u^2+v^2}\right)=0; h>a_0\end{cases}$$

where h denotes the Euclidean distance, $dist(s,s')$, between a frame units located at the point, s, of coordinates (u_s, v_s), and a frame unit located at the point, s', of coordinates $(u_{s'}, v_{s'})$, so that $dist(s,s') = h = \sqrt{u^2 + v^2}$ where $u = (u_{s'} - u_s)$, $v = (v_{s'} - v_s)$.

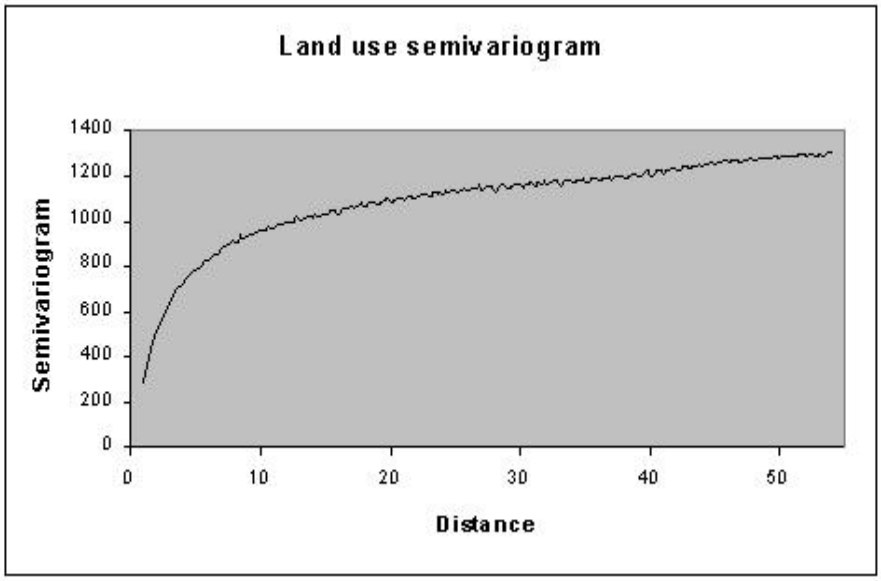
These theoretical correlograms decrease with the distance, h, and depends on two parameters: the range rate , a_0, and the ratio $\tau = C_0 / (C_0 + C_h)$, where C_0 is the nugget effect, i.e., the variation at the origin or near the origin (independent of the distance), C_h is the partial sill (a function of the distance h between sampling points) and $(C_0 + C_h)$ is the sill, i.e. the maximum variation far from the origin.

Note that, as $\tau \to 1$, in such a way that the partial sill tends towards zero $(C_h \to 0)$, the variogram tends to be a pure nugget effect, independent of distance, and the spatial correlation tends to be zero. When $C_h > 0$, the variogram is equal to the sum of two components (both positive), one of them, C_0, is independent of the distance but the other, C_h, depends on the distance and the spatial correlation increases when C_h increases with respect to C_0 in such a way that $\tau \to 0$.

7.4.5 Simulations

Many factors such as topography, climate, soil fertility and land ownership structure, determine differences in a_0 and τ between countries and between regions within a given country and even at different points in time. Here, several values of a_0 and τ will be considered, ranging from minimum to maximum values found in the literature, and their effect on the relative efficiency of sampling units size, as well as on the design effect, will be analyzed by simulation, using the two correlogram functions.

Ambrosio et al. (2003) consider an area of 200 km by 200 km in Castilla y Leon, Spain, as a case study. The frame unit is a 1 km by 1 km square. Agriculture in the area includes irrigated and non-irrigated herbaceous and woody crops. Only non-irrigated herbaceous crops are considered here. The data source is a digitized land use map, from which the area of each land use is calculated for each frame unit. Since the full data set is available in this case, the full data set is used to calculate the empirical variogram. Figure 1 shows the empirical semivariogram of non-irrigated herbaceous crops.



Graphically, the empirical values of the parameters can be set equal to $\tau = \frac{C_0}{C_0 + C_h} = \frac{300}{1300} = 0.23$, and $a_0 = 50Km$.

These empirical values are useful to simulate a set of likely correlation structures. A set of five values is assigned to a_0 : 60, 300, 600, 900 and 1200. For each of these values a set of four values is assigned to τ : 0.00, 0.25, 0.50 and 0.75. Twenty spherical and twenty exponential correlogram models are defined, using this data set.

7.4.5.1 Average correlation in the population $\Phi(M,N)$

Square lattices (M=N) of side 100, 200, 300, 400 and 500 are used as populations. Table 1 shows the average correlation between all pairs of frame units in the population, $\Phi(M,N)$, computed for each one of the considered populations, using each one of the twenty spherical models and each one of the twenty exponential models of correlogram.

Table 1 Average correlation in the population $\Phi(M,N)$

Spherical correlation function

Average correlations in the population	**Range**	τ **value**	**M-values=N-values**				
			100	200	300	400	500
$\Phi(M,N)$	60	0.00	0.159678	0.047874	0.022524	0.013028	0.008477
		0.25	0.119759	0.035905	0.016893	0.009771	0.006358
		0.50	0.079839	0.023937	0.011262	0.006514	0.004238
		0.75	0.039920	0.011968	0.005631	0.003257	0.002119

300	0.00	0.743741	0.514249	0.337837	0.226361	0.159746
	0.25	0.557806	0.385687	0.253378	0.169770	0.119810
	0.50	0.371871	0.257125	0.168919	0.113180	0.079873
	0.75	0.185935	0.128562	0.084459	0.056590	0.039937
600	0.00	0.870199	0.743752	0.623988	0.514254	0.417889
	0.25	0,652649	0,557814	0,467991	0,385691	0,313417
	0.50	0,435100	0,371876	0,311994	0,257127	0,208944
	0.75	0,217550	0,185938	0,155997	0,128564	0,104472
900	0.00	0,913260	0,827517	0,743754	0,662962	0,586132
	0.25	0,684945	0,620638	0,557815	0,497221	0,439599
	0.50	0,456630	0,413758	0,371877	0,331481	0,293066
	0.75	0,228315	0,206879	0,185938	0,165740	0,146533
1200	0.00	0,934891	0,870204	0,806352	0,743754	0,682827
	0.25	0,701168	0,652653	0,604764	0,557816	0,512121
	0.50	0,467445	0,435102	0,403176	0,371877	0,341414
	0.75	0,233723	0,217551	0,201588	0,185939	0,170707

Exponential correlation function

Average correlations in the population	**Range**	τ **value**	**M-values=N-values**				
			100	200	300	400	500
$\Phi(M,N)$	60	0.00	0.455378	0.240666	0.143346	0.093524	0.065324
		0.25	0.341534	0.180499	0.107509	0.070143	0.048993
		0.50	0.227689	0.120333	0.071673	0.046762	0.032662
		0.75	0.113845	0.060166	0.035836	0.023381	0.016331
	300	0.00	0.843314	0.715976	0.611865	0.526244	0.455416
		0.25	0.632485	0.536982	0.458899	0.394683	0.341562
		0.50	0.421657	0.357988	0.305933	0.263122	0.227708
		0.75	0.210829	0.178994	0.152966	0.131561	0.113854
	600	0.00	0.917544	0.843321	0.776401	0.715979	0.661344
		0.25	0,688158	0,632490	0,582301	0,536984	0,496008
		0.50	0,458772	0,421660	0,388201	0,357990	0,330672
		0.75	0,229386	0,210830	0,194100	0,178995	0,165336
		0.00	0,944067	0,891939	0,843322	0,797949	0,755576

900	0.25	0,708050	0,668955	0,632491	0,598462	0,566682
	0.50	0,472033	0,445970	0,421661	0,398974	0,377788
	0.75	0,236017	0,222985	0,210831	0,199487	0,188894
1200	0.00	0,957682	0,917548	0,879467	0,843322	0,809002
	0.25	0,718261	0,688161	0,659600	0,632492	0,606752
	0.50	0,478841	0,458774	0,439734	0,421661	0,404501
	0.75	0,239421	0,229387	0,219867	0,210831	0,202251

The average correlation in the population, $\Phi(M,N)$, increases when the range increases and, given the range, it increases when the partial sill, C_h, increases with respect to the nugget effect, C_0, in such a way that $\tau \to 0$ and the correlation tends to its maximum value. This average correlation decreases when the population size increases.

7.4.5.2 Average correlation within sampling units $\Phi(l,k)$

Square sampling units of sides (measured in number of frame units) 2, 4, 6, 8 and 10 are considered. Table 2 shows the average correlation between all pairs of frame units within sampling units, $\Phi(l,k)$, computed for each one of the considered populations, using each one of the twenty spherical models and each one of the twenty exponential models of correlogram.

Table 2 Average correlations within sampling units $\Phi(l,k)$

Spherical correlation function

Average correlations within sampling units	**Range**	τ **value**	**l-values=k-values**				
			2	4	6	8	10
$\Phi(l,k)$	60	0.00	0.9715519	0.9464872	0.9209101	0.895229	0.8695721
		0.25	0.728664	0.7098654	0.6906826	0.6714217	0.6521791
		0.50	0.485776	0.4732436	0.4604551	0.4476145	0.4347861
		0.75	0.242888	0.2366218	0.2302275	0.2238072	0.217393
	300	0.00	0.9943097	0.9892909	0.9841593	0.9943097	0.9892909
		0.25	0.7457323	0.7419681	0.7381195	0.7457323	0.7419681
		0.50	0.4971548	0.4946454	0.4920797	0.4971548	0.4946454
		0.75	0.2485774	0.2473227	0.2460398	0.2485774	0.2473227
	600	0.00	0.9971548	0.9946453	0.9920793	0.9971548	0.9946453
		0.25	0.7478661	0.745984	0.7440595	0.7478661	0.745984
		0.50	0.4985774	0.4973227	0.4960397	0.4985774	0.4973227

	0.75	0.2492887	0.2486613	0.2480198	0.2492887	0.2486613
900	0.00	0.9981032	0.9964302	0.9947195	0.9981032	0.9964302
	0.25	0.7485774	0.7473227	0.7460396	0.7485774	0.7473227
	0.50	0.4990516	0.4982151	0.4973597	0.4990516	0.4982151
	0.75	0.2495258	0.2491076	0.2486799	0.2495258	0.2491076
1200	0.00	0.9985774	0.9973227	0.9960396	0.9985774	0.9973227
	0.25	0.7489331	0.747992	0.7470297	0.7489331	0.747992
	0.50	0.4992887	0.4986613	0.4980198	0.4992887	0.4986613
	0.75	0.2496444	0.2493307	0.24999	0.2496444	0.2493307

Exponential correlation function

Average correlations within sampling units	**Range**	τ **value**	**l-values=k-values**				
			2	4	6	8	10
$\Phi(l,k)$	60	0.00	0.9812161	0.9650314	0.9488226	0.9328355	0.9171298
		0.25	0.7359121	0.7237736	0.7116169	0.6996266	0.6878474
		0.50	0.490608	0.4825157	0.4744113	0.4664177	0.4585649
		0.75	0.245304	0.2412579	0.2372056	0.2332089	0.2292825
	300	0.00	0.9962138	0.9928899	0.9895053	0.9962138	0.9928899
		0.25	0.7471604	0.7446674	0.7421289	0.7471604	0.7446674
		0.50	0.4981069	0.496445	0.4947526	0.4981069	0.496445
		0.75	0.2490535	0.2482225	0.2473763	0.2490535	0.2482225
	600	0.00	0.9981051	0.9964376	0.9947361	0.9981051	0.9964376
		0.25	0.7485788	0.7473282	0.7460521	0.7485788	0.7473282
		0.50	0.4990525	0.4982188	0.497368	0.4990525	0.4982188
		0.75	0.2495263	0.2491094	0.248684	0.2495263	0.2491094
	900	0.00	0.9987363	0.9976234	0.996487	0.9987363	0.9976234
		0.25	0.7490522	0.7482176	0.7473653	0.7490522	0.7482176
		0.50	0.4993681	0.4988117	0.4982435	0.4993681	0.4988117
		0.75	0.2496841	0.2494059	0.2491218	0.2496841	0.2494059
	1200	0.00	0.9990521	0.9982169	0.9973639	0.9990521	0.9982169
		0.25	0.7492891	0.7486627	0.7480229	0.7492891	0.7486627
		0.50	0.499526	0.4991085	0.4986819	0.499526	0.4991085
		0.75	0.249763	0.2495542	0.249341	0.249763	0.2495542

The average correlation within sampling units follows the same pattern than in the population: it increases when the range increases and, given the range, it increases when the partial sill, C_h, increases with respect to the nugget effect, C_0, in such a way that $\tau \to 0$ and the correlation tends to its maximum value.

7.4.5.3 Relative efficiencies of random sampling of sampling units with respect to random sampling of frame units $RE_{su(r)/fu(r)}$

The relative efficiency of sampling units with respect to frame units is computed, using [10]. Results are in Table 3. As can be seen, the efficiency loss when sampling units are used instead of frame units (design effect), is high and it increases when the sampling unit size increases.

Conventionally, the sampling unit size is limited to a few frame units: in the BAE Agricultural Frame, the target sampling unit size ranges from three to six farms (see section 2) and in the UN Household Master Frame the target sampling unit size is ten households (see section 3). However, according to our simulations results, the efficiency loss (design effect) corresponding to these target sizes could be very high, even if the correlation is low. The use of sampling units is justified because the cost of using frame units instead of sampling units is very high. In section 7.5 we deal with the optimum.

Table 3 Relative efficiencies of random sampling of sampling units with respect to random sampling of frame units $RE_{su(r)/fu(r)}$

Spherical correlation function

Range=60	**l-value=k-value**				
	2	4	6	8	10
τ **value**	M=N=100	200	300	400	500
0.00	0.2564366	0.0659517	0.0301378	0.0174414	0.0114927
0.25	0.3250896	0.0870315	0.0400041	0.0231953	0.0152973
0.50	0.4302915	0.1264599	0.0591464	0.0345033	0.0228192
0.75	0.6118168	0.2266551	0.1122516	0.0669091	0.0447109
Range=300	**l-value=k-value**				
	2	4	6	8	10
τ **value**	M=N=100	200	300	400	500
0.00	0.2541581	0.0637951	0.0284282	0.0160478	0.0103143
0.25	0.4394699	0.1030611	0.0421359	0.0228046	0.0143474
0.50	0.6255449	0.1724766	0.068423	0.0360434	0.0223157
0.75	0.8123854	0.3284032	0.1392854	0.0736918	0.0454659

Range=600	**l-value=k-value**				
	2	4	6	8	10
τ **value**	M=N=100	200	300	400	500
0.00	0.2541028	0.063725	0.0283475	0.0159586	0.0102238
0.25	0.5486366	0.1353963	0.0521666	0.0266184	0.0159846
0.50	0.7478074	0.250191	0.096465	0.0472597	0.0273074
0.75	0.8914734	0.4637769	0.2075712	0.1042457	0.0597645

Spherical correlation function

Range=900	**l-value=k-value**				
	2	4	6	8	10
τ **value**	M=N=100	200	300	400	500
0.00	0.2540927	0.0637128	0.0283347	0.015945	0.0102093
0.25	0.6226031	0.1663582	0.0628757	0.0312223	0.0182816
0.50	0.8101708	0.3162712	0.1250799	0.0603969	0.0340376
0.75	0.9237943	0.5558726	0.270371	0.1382504	0.0784105
Range=1200	**l-value=k-value**				
	2	4	6	8	10
τ **value**	M=N=100	200	300	400	500
0.00	0.2540892	0.0637086	0.0283303	0.0159405	0.0102046
0.25	0.6758081	0.1953578	0.0735114	0.0359314	0.0207072
0.50	0.8478496	0.3719691	0.152341	0.0735716	0.0410575
0.75	0.9413024	0.6213191	0.320241	0.1707073	0.0973316

Exponential correlation function

Range=60	**l-value=k-value**				
	2	4	6	8	10
τ **value**	M=N=100	200	300	400	500
0.00	0.2565622	0.0652956	0.0294791	0.0168476	0.0109578
0.25	0.3574590	0.0913410	0.0404854	0.0229010	0.0148079
0.50	0.4946324	0.1393072	0.0617657	0.0347870	0.0224187
0.75	0.6919357	0.2569748	0.1202961	0.0687708	0.0445617

Range=300	**l-value=k-value**				
	2	4	6	8	10
τ **value**	M=N=100	200	300	400	500
0.00	0.2545385	0.0639780	0.0285166	0.016083	0.0103202
0.25	0.5164056	0.1293510	0.0517398	0.0270923	0.0165327
0.50	0.7159620	0.2360600	0.0950074	0.0483881	0.0287265
0.75	0.8730826	0.4414369	0.2039710	0.1070265	0.0635588
Range=600	**l-value=k-value**				
	2	4	6	8	10
τ **value**	M=N=100	200	300	400	500
0.00	0.2543085	0.0638373	0.0284173	0.0160057	0.0102564
0.25	0.6323126	0.1757771	0.0679043	0.0341488	0.0201483
0.50	0.8174194	0.3348461	0.1379736	0.0684996	0.0393505
0.75	0.9272674	0.5787440	0.2966025	0.1583049	0.0926921

Exponential correlation function

Range=900	**l-value=k-value**				
	2	4	6	8	10
τ **value**	M=N=100	200	300	400	500
0.00	0.2542333	0.0637918	0.0283857	0.0159813	0.0102365
0.25	0.7034841	0.2177269	0.0837204	0.0412588	0.0238635
0.50	0.8655124	0.4113149	0.1774155	0.0881077	0.0500219
0.75	0.9490463	0.6621419	0.3705221	0.2046958	0.1207111
Range=1200	**l-value=k-value**				
	2	4	6	8	10
τ **value**	M=N=100	200	300	400	500
0.00	0.2541958	0.0637694	0.0283702	0.0159694	0.0102268
0.25	0.7515834	0.2556628	0.099056	0.0483035	0.0275827
0.50	0.8935601	0.4720739	0.2134941	0.1069897	0.0605302
0.75	0.9607911	0.7180234	0.4305047	0.2464039	0.1472139

The efficiency loss (the design effect) increases when the range decreases and, given the range, it increases when the partial sill, C_h, increases with respect to the nugget effect, C_0, in such a way that $\tau \to 0$ and the correlation tends to its maximum value.

The efficiency loss (the design effect) increases when the sampling unit size increases.

7.4.6 Modeling the cost

Consider a simple cost function $C_{fu} = c_1 n + c_2 d$, where n is the number of frame units in the sample, d is the distance between the n sampling points (the number of frame units between sampling points), c_1 is the unitary cost of observing frame units and c_2 is the travel cost per frame unit. Once the sample size has been fixed at n frame units, the first component, $c_1 n$, of the total cost, C_{fu} , is fixed.

However, the second component, $c_2 d$, depends on the spatial pattern of the n sampling points. This second component can be reduced by reducing the length, d , using sampling units instead of frame units. If sampling unit size is l^2 frame units, then the number of sampling points reduces to $\vartheta = \frac{n}{l^2}$. Note that if the sampling unit size increases until $l^2 = n$, then the number of sampling points reduces to its minimum, $\vartheta = 1$.

The length d between n random sampling points is approximately $\sqrt{n}$, so that the total cost using frame units is $C_{fu} = c_1 n + c_2 \sqrt{n}$. Using sampling units of size equal to l^2 frame units, the number of sampling points reduces to $\vartheta = \frac{n}{l^2}$ and the total cost is $C_{su} = c_1 n + c_2 \frac{\sqrt{n}}{l}$. The cost reduction using sampling units instead of frame units is $C_{su} - C_{fu} = c_2 \left(\sqrt{\vartheta} - \sqrt{n} \right) = c_2 \sqrt{n} \left(\frac{1}{l} - 1 \right)$ and there is interest on increasing the sampling unit size l in order to reduce the costs. However, as Table 3 shows, when the sampling unit size increases, the efficiency loss (the design effect) increases and this is the cost-efficiency problem.

7.5. Landscape effect

A landscape is an area of the Earth's surface, with differentiable visible features regarding living elements of land cover (indigenous vegetation) and human elements (land use). The fields size and shape are elements of the landscape, as well as their spatial distribution, forming a mosaic of pieces of land with different uses. The sampling design cost-efficiency depends heavily on the adaptation of the design to the specific features of the landscape that supports the target population. To gain cost-efficiency, the sampling designer has to identify the correlation structure among the observation units.

7.5.1 Landscape effect on the sampling error

Once the sample size has been fixed, the expected value of the sampling error of a given sampling scheme depends only on the correlation structure [that is, on the correlogram function, $\rho(u,v)$] as can be seen on [5], [7] and [9]. The value of $\rho(u,v)$ in a point of the frame depends on the number u of frame units from the origin to this point in the rows direction and of the number v of frame units from the origin to this point in the columns direction, but it is independent of size of the fields.

As a result, the relative efficiency between any two sampling schemes is independent of fields' size. That is why the conventional approach, where a target sampling unit size is defined in terms of the number of frame units (farm, household or segment), lk, instead of in terms of surface, is correct. The surface of the target sampling unit is usually calculated later, by multiplying the target number of frame units per sampling unit by the average area of frame units.

7.5.2 Landscape effect on the cost

The landscape has an effect on the unitary cost of observing frame units, c_1, and the travel cost, c_2. As a simple model of this effect, assume J types of landscape that differ on the difficulty of data gathering. Let be $c_{1j} = c_{10}t_{1j}$ and $c_{2j} = c_{20}t_{2j}$, where c_{10} and c_{20} are unitary observation costs and travel cost, respectively, and t_{1j} and t_{2j} are measures of difficulty of data gathering associated to the j^{th} landscape type.

Thus, the total cost using frame units is $C_{jfu} = c_{10}t_{1j}n + c_{20}t_{2j}d = c_{10}t_{1j}n + c_{20}t_{2j}\sqrt{n}$. The total cost using sampling units of size l^2 frame units is $C_{jsu} = c_{10}t_{1j}n + c_{20}t_{2j}\frac{\sqrt{n}}{l}$ and the landscape has an effect on the solution to the cost-efficiency problem, trough the measures of difficulty of data gathering, $\left(t_{1j}, t_{2j}\right)$. The cost reduction using sampling units instead of frame units is $C_{jsu} - C_{jfu} = t_{2j}c_{20}\sqrt{n}\left(\frac{1}{l} - 1\right)$.

7.6. Optimum sampling unit size

The cost-efficiency problem is to find a survey design that produces the desired level of precision at the lowest possible cost or that minimizes the sampling variance, conditionally to unitary costs and total budget.

The expected sampling variance,
$EV(\hat{\bar{Y}}_{su(r)}) = (1 - \frac{\vartheta}{mn})\frac{\sigma^2}{\vartheta lk}[1 - \frac{mnlk - 1}{mn - 1}\Phi(ml, nk) + \frac{mn(lk-1)}{mn-1}\Phi(l,k)]$, can be expressed as a function of the sampling unit size, $N_0 = lk$, using the correlogram function. Hereon, we limit ourselves to the $l = k$ case, so that $N_0 = l^2$ and $l = k = \sqrt{N_0}$ and

$$EV(\hat{\bar{Y}}_{su(r)}) = (1 - \frac{\vartheta N_0}{MN})\frac{\sigma^2}{\vartheta\sqrt{N_0}}[1 - \frac{MN - 1}{(MN/N_0) - 1}\Phi(M,N) + \frac{(MN/N_0)(N_0 - 1)}{(MN/N_0) - 1}\Phi(\sqrt{N_0}, \sqrt{N_0})] \qquad [12].$$

Thus, the optimum N_0 can be found by minimizing $EV(\hat{\bar{Y}}_{su(r)})$ subject to

$$C_{jsu} = c_{10}t_{1j}\vartheta N_0 + c_{20}t_{2j}\sqrt{\vartheta} \qquad [13]$$

where $n = \vartheta N_0$ is the sample size in terms of number of frame units and ϑ is the sample size in terms of number of sampling units.

We form the Lagrange function, $\psi(\boldsymbol{\theta}) = EV(\hat{\bar{Y}}_{su(r)}) + \lambda\left(C_{jsu} - c_{10}t_{1j}\vartheta N_0 - c_{20}t_{2j}\sqrt{\vartheta}\right)$, where $\boldsymbol{\theta}=\left[N_0\ \vartheta\ \lambda\right]^T$ is a unknown vector . The optimum $\boldsymbol{\theta}$ is the solution of $\frac{\partial\psi(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}} = \mathbf{0}$, that is the solution of the equations system:

$$\frac{\partial\psi}{\partial N_0} = \frac{\partial EV(\hat{\bar{Y}}_{su(r)})}{\partial N_0} - \lambda c_{10}t_{1j}\vartheta = 0$$

$$\frac{\partial\psi}{\partial\vartheta} = \frac{\partial EV(\hat{\bar{Y}}_{su(r)})}{\partial\vartheta} - \lambda c_{10}t_{1j}N_0 - \lambda c_{20}t_{2j}\frac{1}{2\sqrt{\vartheta}} = 0$$

$$\frac{\partial\psi}{\partial\lambda} = C_{su} - c_{10}t_{1j}\vartheta N_0 - c_{20}t_{2j}\sqrt{\vartheta} = 0$$

To find the optimum, $\boldsymbol{\theta}=\left[N_0\ \vartheta\ \lambda\right]^T$, we use a Newton-Raphson approach. The starting point is to develop $\frac{\partial\psi(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}$ as a Taylor's serie around a initial value, $\boldsymbol{\theta}^{(0)}=\left[N_0^{(0)}\ \vartheta^{(0)}\ \lambda^{(0)}\right]^T$, so that, $\frac{\partial\psi(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}} \cong \frac{\partial\psi(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}^{(0)}} + \frac{\partial^2\psi(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T}\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}^{(0)}}\left(\boldsymbol{\theta}\text{-}\boldsymbol{\theta}^{(0)}\right)$.

Then, $\frac{\partial\psi(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}^{(0)}} + \frac{\partial^2\psi(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T}\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}^{(0)}}\left(\boldsymbol{\theta}\text{-}\boldsymbol{\theta}^{(0)}\right) = \mathbf{0}$, is solved iteratively:

$$\boldsymbol{\theta}^{(m+1)} = \boldsymbol{\theta}^{(m)} + \left[-\frac{\partial^2\psi(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T}\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}^{(m)}}\right]^{-1}\frac{\partial\psi(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}^{(m)}}$$

Simulations

As a first approximation, we replace $\Phi(l,l)$ in $EV(\hat{\bar{Y}}_{su(r)})$ by an empirical function of l. Using data on Table 2, we fit the regression function, $\Phi_{a_0\tau}(l,l) = \alpha_0 a_0 + \alpha_1\tau + \beta l^2 + \varepsilon$ (note that only β is relevant to find the optimum and the fitted β, corresponding to the spherical correlation function, is $\hat{\beta} = -0.00075$), and we assume in [12] that $\Phi(\sqrt{N_0},\sqrt{N_0}) \cong \beta N_0$. In addition, it is required to assign values to (M,N), σ^2, $\Phi(M,N)$, C_{su}, (c_{10},c_{20}) and (t_{1j},t_{2j}).

A square lattice of 90000 frame units (M=N=300) is considered for simulations. We assign to σ^2 the value 1300, which is the sill of the empirical variogram found by Ambrosio et al. (2003) in a case study (see figure 1). We assign the value 0.17 to $\Phi(M,N)$, which corresponds to $a_0 = 300$ and $\tau = 0.50$ on Table 1.

We assign a value of 15000 € to C_{su}. To assign values to (c_{10}, c_{20}), we use the structure of the total cost of a survey as found in the literature [Groves (1989, Table 11.4): materials, 4%; salaries 38%; travel costs, 20%; and other (including, clerical work, pre-study, training and communication), 38%]. In Spain, the cost of materials, salaries and travel is 50€ per segment of size 700 meters by 700 meters. The estimated total cost per segment would be 80€ (50/0.62) and we consider $c_{10} = (0.04 + 0.38) \times 80€ = 34€$ per segment and $c_{20} = 0.2 \times 80€ = 16€$ per segment. The average number of fields (frame units) per segment is 15, so that $c_{10} = \frac{34}{15} = 2.3€$ per frame unit (field) and $c_{20} = \frac{16}{15} = 1.1€$ per frame unit (field). Finally, a set of landscape types could be considered for simulations and a set of values could be assigned to (t_{1j}, t_{2j}), according to the difficulty of data gathering in each landscape type. Here we limit ourselves to the value $t_{1j} = t_{2j} = 1.5$.

Thus, the cost-efficiency problem that we consider is finding the value $\boldsymbol{\theta}=[N_0\ \vartheta\ \lambda]^T$ that minimizes

$$EV(\hat{\bar{Y}}_{su(r)}) = (1 - \frac{\vartheta N_0}{90000}) \frac{1300}{\vartheta\sqrt{N_0}} [1 - \frac{90000-1}{(90000/N_0)-1} 0.17 + \frac{(90000/N_0)(N_0-1)}{(90000/N_0)-1} (-0.00075 N_0)]$$

subject to:

$$15000 = 2.3 \times 1.5 \times \vartheta N_0 + 1.1 \times 1.5 \sqrt{\vartheta}$$

The solution is $\theta=[2.2358\text{e+}003\ 1.9346\text{e+}000\ -1.8715\text{e-}005]^T$, and by approximating to the nearest entire, is $\vartheta = 2236$ and $N_0 = 2$.

8. REFERENCES

Abaye , A.T (2010): Combining enumeration area maps and satellite images (land cover) for the development of area frame (multiple frames) in an African country: Preliminary lessons from the experience of Ethiopia. National Statistical Data Quality Directorate. Director Central Statistical Agency. Addis Ababa. Ethiopia

Ambrosio L., Iglesias L. and Marin C. (2003). Systematic sampling design for the estimation of spatial means. *Environmetrics*, **14**: 45-61

Ambrosio L (2012): Marcos múltiples de áreas y listas. Diseño de encuestas sobre la estructura y la producción agrícola de Chile. Informe Técnico. Ministerio de Agricultura de Chile. Instituto nacional de Estadística de Chile. Universidad Politécnica de Madrid.

Ambrosio L (2013): Marco de muestreo y diseño de la Encuesta Nacional Agropecuaria 2013. Informe Técnico. Ministerio de Agricultura, Ganadería y Alimentación de Guatemala. Instituto Nacional de Estadística de Guatemala. FAO. Universidad Politécnica de Madrid.

Bellhouse DR. 1977. Some optimum designs for sampling in two dimensions. *Biometrics* **64** : 605-611.

Bureau of Agricultural Economics, U. S. Department of Agriculture (1936). *Proceedings of Conference on Statistical Methods of Sampling Agricultural Data.* Conference held July 14-17, 1936, Iowa State College, Ames, Iowa.

Carfagna E. (1997). Correlograms and sample design in Agriculture. *Bulletin of the International Statistical Institute, Book 1, Contributed papers*, Istanbul, pp. 5-6

Carfagna E. (1998). Area Frame Sample Designs: A Comparison with the MARS Project. *Proceedings of Agricultural Statistics 2000,* ISI, Voorburg, NL, 1998, pp.261-277 http://www.nass.usda.gov/as2000/proceedings/page-261.pdf

Cochran, W. G. (1942). Sampling theory when the sampling units are of unequal sizes. *Journal of the American Statistical Association* 37, 199-212.

Cochran W.G. 1946. Relative accuracy of systematic and stratified random samples for a certain class of populations. *Ann. Math. Statistic*. **17 :** 164-177.

Cochran, W.G. (1977), *Sampling Techniques*, third edition, Wiley, New York.

Das AC. 1950. Two-dimensional systematic sampling and associated stratified and random sampling. *Sankhaya* **10:** 95-108.

FAO (1996): *Multiple frame agricultural surveys*. Vol.1. Current surveys based on area and list sampling methods. Statistical Development Series, 7. Rome

FAO (1998): *Multiple frame agricultural surveys*. Vol.2. Agricultural survey programmes based on area frame or dual frame (area and list) sample designs. Statistical Development Series, 10. Rome.

FAO (2012): *Global Strategy. Improving AG-Statistics*. Expert meeting 3. Master sampling frames for agricultural and rural statistics.

Fuller W.A. (1984) *The master sample of agriculture*. In Statistics: An Appraisal, David, H.A. and David, H.T., eds. Ames: Iowa State University Press.

Fuller W.A. (2010). *The master sample of agriculture*. Special Collections Department. Iowa State University. http://www.lib.iastate.edu/arch/rgrp/13-24-00-05_report.html

Gallego F.J. (1995) *Sampling Frames of Square Segments*, Report. EUR 16317 EN, Office for Publications of the E.C. Luxembourg. 68 pp.

Gallego F.J.and Carfagna E. (1995) *Extrapolating Intracluster Correlation to Optimize the Size of Segments in an Area Frame*. In Applied Statistics in Agriculture. Manhattan. Kansas State University. pp. 261-270

Gallego F.J., Carfagna E., Fuenette I. (1998). *Geographic Sampling Strategies and Remote Sensing.* Report to EUROSTAT, F2, Agricultural Products and Fisheries.

Gallego F. J., Feunette I. and Carfagna E (1999). *Optimising the size of sampling units in an area frame*. In Gómez J., Soares A. and Froidevaux R. (Editors). geoENV II-Geostatistics for Environmental Applications. Kluwer Academic Publisher.

Gallego, F.J. and Stibig, H.J. (2013). Area estimation from a sample of satellite images: The impact of stratification on the clustering efficiency. *International Journal of Applied Earth Observation and Geoinformation*, **22**: 139-146

Groves, R. M. (1989): Surveys errors and surveys cost. Wiley.

Hájek J. 1961. Concerning relative accuracy of stratified and systematic sampling in plane. *Colloquium Mathematicum*, Vol VIII (1) : 133-134.

Hannerrza F. and Lotsch A. (2008). Assessment of remotely sensed and statistical inventories of African agricultural fields. *International Journal of Remote Sensing*, **29:** 3787–3804

Hansen, M.H., Hurwitz, W.N. and Madow, W.G. (1953), *Sample Survey Methods and Theory*, Wiley, New York.

Iachan R. 1985. Plane sampling. *Statistics & Probability Letters* **3 :** 151-159.

International Statistical Institute (1975), *Manual on Sample Design*. World Fertility Survey Basic Documentation, Voorburg, Netherlands.

Jessen, R. J. (1942). Statistical investigation of a sample survey for obtaining farm facts. *Research Bulletin No. 304,* Iowa Agricultural Experiment Station. Iowa State College, Ames, Iowa.

Jessen, R. J. (1945). The Master Sample of Agriculture: II. Design. *Journal of the American Statistical Association* **40**: 46-56.

Jessen, R. J. (1947). The master sample project and its use in agricultural economics. *Journal of Farm Economics* 29, 531-540.

Jessen, R. J. (1978). *Statistical Survey Techniques.* Wiley, New York.

Jessen, R. J. and Houseman, E. E. (1944). Statistical investigations of farm sample surveys taken in Iowa, Florida, and California. *Research Bulletin No. 324,* Iowa Agricultural Experiment Station. Iowa State College, Ames, Iowa.

King, A. J. (1945). The Master Sample of Agriculture: I. Development and use. *Journal of the American Statistical Association* **40:** 38-45.

King, A. J. and Simpson, G. D. (1940). New developments in agricultural sampling. *Journal of Farm Economics* **22:** 341-349.

Kireyera B. (1982). On sampling frames in African censuses and surveys. *Journal of the Royal Statistical Society*. Series D (The Statistician), **31** : 153-167.

Kish, L. (1965), *Survey Sampling*, Wiley, New York.

Kish L. (1996). Developing samplers for developing countries. *International Statistical Review*, **64**: 143-152

Murthy M.N. (1969). Population Census as the Source of Sampling Frame in India. Sankhyā: *The Indian Journal of Statistics*, Series B (1960-2002), **31**: 1-12.

Pennington M. and Volstad J.H. (1991). Optimum size of sampling unit for estimating the density of marine populations. *Biometrics* **47**: 717-723.

Statistical Laboratory (1944-1965). *Annual Reports (various issues)*. Iowa State University, Ames, Iowa.

Steiner, M. (2005). Sample Design for Agricultural Surveys in China, *Proceedings of the 55th Conference of the International Statistical Institute*, Sydney, Australia.

Stephan, F. F. (1948). History of the uses of modern sampling procedures. *Journal of the American Statistical Association* 43, 12-39.

Taylor J., Sannier C., Delincé J., and Gallego F.J. (1997) *Regional Crop Inventories in Europe Assisted by Remote Sensing: 1988-1993*. Synthesis Report. EUR 17319 EN, Office for Pubbications of the EC. Luxembourg. 71pp.

UN (1986). *National Household Survey Capability Programme. Sampling Frames and Sample Designs for Integrated Household Survey Programmes*. Department of Technical Co-Operation for Development and Statistical Office. United Nations. New York

UN (2005). *Designing Household Survey Samples: Practical Guides*. Department of Economic and Social Affairs. Statistics Division. Studies in Methods. Series F Nº 98. United Nations. New York