Using Satellite Imagery and geo-referencing technology for building a master sampling frame

Elisabetta Carfagna
Food and Agriculture Organization of the United Nations and University of Bologna
e-mail: elisabetta.carfagna@fao.org

Abstract

A master sampling frame is a sampling frame that provides the basis for all data collections based on sample surveys and censuses in a certain sector. Current technologies, in particular the availability of remote sensing, the ability of geographic information systems to overlay and integrate efficiently different layers of geographic information, have completely transformed the way of building master sampling frames for the agricultural sector and considerably reduced the cost and the time needed.
The evolution of global positioning systems has substantially changed the field work, influencing also the definition of master sampling frames. In this paper, we analyze how the use of satellite imagery and geo-referencing technology has influenced the building of master sampling frames for agricultural statistics.

Keywords: Remote sensing data, master sampling frame, area frame, list frame, Geographic information systems, Global Positioning Systems.

1. Introduction

A master sampling frame is a sampling frame that provides the basis for all data collections through sample surveys and censuses in a certain sector, allowing to select samples for several different surveys or different rounds of the same survey, as opposed to building an ad-hoc sampling frame for each survey. The aims of the development of a master sampling frame are: avoiding duplication of efforts, reducing statistics discrepancies, connecting various aspects of the sector, allowing the analysis of the sampling units from the different viewpoints, and having a better understanding of the sector.
In the case of agricultural sector, if both economical and social characteristics are relevant for a country, surveys have to collect information on the agricultural characteristics of the farm, including information on land area, and on the socio-economic characteristics; thus, the master sampling frame should allows linking the farm characteristics with the household.
By definition, a sampling frame, must cover the entire survey population exhaustively and without overlaps. According to the type of information available in a country, different kinds of frames are used for selecting sample units for sample surveys covering the various relevant aspects of agricultural statistics.
When the master sampling frame includes the geographic dimension of the statistical units, farms and households can be connected to the land cover and use dimensions. This generates a series of benefits. The link of the farm with its geo-referenced plots, which can be observed on the ground and measured, allows the assessment of the quality of self reported responses of farmers and the use of these measurements for benchmarking. Moreover, this link facilitates agro-environmental analysis.
High relevance of the geographic dimension is typical of area frames. The traditional approach to set up an area sampling frame was based on collections of printed maps and aerial photographs and involved a large amount of manual work.
Current technologies, mainly the ability of Geographic Information Systems (GIS) to efficiently handle different layers of geographic information, including remote sensing based thematic maps, have made this task much lighter. The stratification of sampling

units, for example, can be performed in a more efficient way through remote sensing data in a GIS.
More recently, the evolution of Global Positioning Systems (GPS), with sufficiently accurate devices at affordable prices, has substantially changed the field work, influencing also the definition of sampling frames, since the choices in the definition of the sampling frame need to take into account the field survey aspects.
In this paper, we analyze how the use of satellite imagery and geo-referencing technology has influenced the process for building master sampling frames.

2. Different kinds of master sampling frame

The Global Strategy to Improve Agricultural and Rural Statistics (World Bank *et al.* 2011 and FAO *et al.* 2012) aims at improving the agricultural statistics of developing countries through its main pillars:

i. The establishment of a minimum set of core data that countries will collect to meet current and emerging demands;
ii. The integration of agriculture into national statistical systems in order to satisfy the demands of policy makers and other users who rely on comparable data across locations and over time;
iii. The sustainability of the agricultural statistics system through governance and statistical capacity building.

The integration will allow avoiding duplication of efforts, connecting economic, social, physical (land cover and use data) characteristics of the sample units, analyzing the sampling units from the different viewpoints (economical, social and physical) and reducing statistics discrepancies. The integration can be achieved by implementing:

a) A set of methodologies that includes the development of a master sample frame for agriculture;
b) An integrated survey framework;
c) A data management system which makes data and results available.

In this paper, we shall focus on one component of this integration: the master sampling frame and particularly, on the use of satellite imagery and geo-referencing technology for building it.
A master sampling frame is a sampling frame that provides the basis for all data collections based on sample surveys and censuses in a certain sector. In the case of agricultural sector, the master sampling frame allows linking the farm characteristics with the household and thus having a better understanding of the rural dimension.
Given the structural characteristics of the agricultural sector and the level of development of the national statistical system, different kinds of master sampling frames are currently adopted:

1) *Population census enumeration areas*: the population census is usually conducted using an administrative structure in which cartographic or other mapping materials are used to divide the country into enumeration areas. The sampling frame is the list of enumeration areas. In agricultural censuses and surveys, a sample of enumeration areas is selected, the list of households in selected enumeration areas is created and a sample is extracted from each of these lists, following a two stages sample design.
2) *List frame based on the population census*: the list of farms or agricultural households identified on the basis of specific agricultural questions included in the population census questionnaire. This approach has been recently proposed by FAO and UNFPA (2012) for avoiding to face the cost of the agricultural census; for an analysis of advantages, disadvantages and requirements see also Keita and Gennari, 2013 and Carfagna *et at.*, 2013.
3) *Agricultural census enumeration areas:* In many countries, a sample agricultural census is conducted: some enumeration areas are randomly selected and screened for farms. The resulting sampling frame consists of the agricultural census enumeration areas.

4) *List frame based on the agricultural census*: the list of farms is created through the census of agriculture. Information collected through the census can be used for efficient sample designs and, where possible, for interviews through mail, email, etc. A major weakness is that the list rapidly becomes out-of-date. An out-of-date list of farms erodes all of the data quality dimensions because the completeness of coverage decreases over time, thus affecting the comparability and accuracy of the resulting estimates.
5) *List of farms based on administrative sources*, such as business registrations or tax collections. This sampling frame offers the advantages of the lists created through agricultural census, but it needs to be updated regularly. Moreover, a big disadvantage of the administrative sources is that they may not include the total population, especially units below a threshold required to be registered or pay taxes. In other words, while they will be inclusive of commercial farms, they are not likely to include small-scale farms and subsistence farming units (see Carfagna and Carfagna, 2010 and Carfagna et al. 2013).
6) *Area frame*: there are two meanings of an area sample survey, a restricted and a general meaning, as stated in FAO 1996 and 1988. An area sample survey designates, in the general meaning, a probability sample survey in which, at least for one sampling stage, the sampling units are land areas. In a more restricted meaning, an area sample survey designates a probability sample survey in which the final stage sampling units are land areas called segments and the selection probabilities are proportional to their area measures. Both approaches foresee the subdivision of the analysed territory into non-overlapping pieces of land, according to specific criteria, to create the area sampling frame. The population or agricultural census enumeration areas can be considered as an area frame only in the general meaning.
7) *Multiple frame*: a list of large, commercial farms (easy to update) and, in case, of other kinds of farms, is combination with the area frame, in order to take advantage of the strengths of the area frame (complete coverage also of small and subsistence farms and link with the land) and of the list frame (possibility to use characteristics of the farm -like size and type- in the sample design, easy identification of selected farms through their addresses, in some cases telephone or mail or email can be used instead of personal interviews, etc.).

For more details see World Bank, FAO, UNSC, 2011, Annex B.

3. Use of satellite imagery for building a master sampling frame

The sampling frames from 1) to 4) of previous section allow linking households and farms but generate a very vague link with the land (only at enumeration area level), unless the parcels of the households and farms are digitized. Digitizing all the parcels of the statistical units constituting the sampling frame is unaffordable from the cost and time viewpoints and could be even unfeasible, since farmers tend to omit field far from their households (see Kilik *et al*., 2013). Moreover, this geographic information becomes out of date as fast as the list of farms, since it refers to it.

Concerning the sampling frame 5), some kinds of administrative data are geo-referenced, for example some subsidies are linked to the fields and request digital information, allowing a partial link with the land, only for some of the parcels linked to the subsidies (see Carfagna and Carfagna, 2010).

The link with the land is important because agriculture statistics mostly refer to variables associated with land such as crops, livestock, forests, water and aquaculture and the most reliable way for estimating main agricultural variables is through collecting data on land parcels. Moreover, the land is the basis for collecting physical information for producing agro-environmental statistics.

Remote sensing data add the geographical dimension to the sampling frames:1) to 3)

of the previous section, in fact they provide land cover, vegetation indexes and physical boundaries. Since remote sensing data are already in digital format, the digitized enumeration areas can be overlaid to remote sensing data in order to associate information concerning the land cover to the enumeration areas.
The land cover of the enumeration areas is particularly useful for stratification, when the sampling frame is constituted by the population census enumeration areas (no agricultural auxiliary information can be derived from the population census) and when the sampling frame is constituted by the list of farms or agricultural household identified on the basis of specific agricultural questions included in the questionnaire for the population census. In fact, in the latter case, only a limited number of very focused questions related to agriculture can be added to the population census questionnaire, in order to avoid respondent burden and collect reliable information; thus almost no auxiliary information is associated to the units of the sampling frame to be used for sample designs, including for stratification.

The simplest way for associating the spatial information of remote sensing data to the enumeration areas and administrative units is through classification of remote sensing imagery into major categories, such as cultivated land, woodlands, grasslands, bare soil and urban areas. This classification allows stratifying the enumeration areas and the administrative units in order to improve the efficiency of the sample design of the sample surveys to be carried out for producing the agricultural and rural statistics.
Unless land cover/use is changing rapidly, this classification does not need to be updated frequently (every 10 years in relatively stable conditions). Only the borders between urban, agricultural and non agriculture areas change more frequently, but this change has a very limited effect on this kind of stratification.

The spatial, spectral, and temporal, resolutions of the sensors are important factors to take in account for building or updating a master sampling frame. Several kinds of satellites are available, with different resolutions; however, the most commonly used are Landsat and SPOT. New very promising satellites will be shortly accessible, with no-cost, to remote sensing users: Landsat 8 from USA-NASA, recently launched (pixel size: OLI Multispectral bands 30 meters, OLI panchromatic band 15 meters and TIRS Thermal bands 100 meters) and Sentinel by EU, ESA, that is due to be launched in early 2014.

When an area or multiple frame is adopted, the sampling frame is constituted by parcels of land; thus the link with the land cover is implicit in the definition of area frame. The stratification of the sampling units of an area frame according to their land cover, using remote sensing data, is more detailed and efficient than the stratification of enumeration areas. The use of remote sensing data has considerably reduced the cost and time for building area sampling frames.

4. Use of geo-referencing technology for building a master sampling frame
The development of a master sampling frame has changed completely with the use of Geographic Information Systems (GIS) which allow overlaying and integrating different geographic information layers (borders of administrative areas, enumeration areas, fields, land cover databases, coordinates of headquarters of farms and households) and Global Positioning Systems (GPS) which allows geo-referencing observations and data collected, which can then be overlaid to the other geographic information layers through GIS.
The time and cost needed for building all kinds of master sampling frame have decreased dramatically.

For area frames, the need to collect information on the ground on area units with physical boundaries has become less relevant, since segments with regular, theoretical

boundaries, like squares, rectangles etc. can be easily overlaid to ortho-photos or very high resolution satellite images for data collection on the ground. The use of segments with regular theoretical boundaries further reduces the cost for building the muster sampling frame, since this approach eliminates the need to draw the primary sampling units with permanent physical boundaries and then to break down the selected primary sampling units into segments. Moreover, experiments conducted in Europe (Carfagna, 1998) showed that the kind of segment (with or without physical boundaries) does not affect the accuracy of data collected on the ground and the efficiency of the land cover stratification.
When a Personal Digital Assistants (PDA) is used for data collection, the border of the fields derived from photo-interpretation of an aerial photo or from a previous survey can be showed on the screen of the PDA and the delineation of the field limits reduces to the delineation of the changes. Moreover, data can be directly downloaded and imported in a GIS, for more details see Keita *et at.* 2010).

When the master sampling frame is an area or multiple frame, during the data collection process, farmers operating the parcels included in the segment have to be identified and rules of association have to be used to connect farms or households to selected segments, in order to collect data on variables which cannot be directly observed on the ground, like socio-economic variables. Most commonly used rules are the so called closed, open and weighted segment estimators. Satellite maps and aerial photos make the research of farms and households easier and faster, particularly where farmers live in villages.

Since master sampling frames are multipurpose by definition, the optimal size of the sample units has to be a compromise and the optimum compromise for variables which can be observed on the ground can reveal to be too large for collecting socio-economic data, since the number of farmers operating fields on a segment can be large and related work too long and cumbersome. In these cases, a two stage sampling of farms can be implemented: a grid of points can be overlaid to the selected segments and farmers operating the fields under the points are selected (Gallego *et al.* 1994). This approach allows optimizing both the sample and segment size for collecting data on physical variables (land use, area and yield of crops, agro-environmental variables, etc.) and the sample size for estimating socio-economic parameters.
The use of GPS makes this approach much simpler and offers the possibility to carry out panel surveys identifying the same field in the subsequent surveys.

Other types of master sampling frame have become easy to implement with the support of GPS for data collection, like clustered and un-clustered point sampling, since identifying a point on the ground with good approximation has become much easier with mapping grade accuracy GPS (error less than 1 m – 5 m) and, in countries where the field size is not very small, even with recreational grade accuracy GPS (error 5-20 m), for more details see Keita *et al.*, 2010 and Keita, 2013.

5. Conclusions

We have analyzed how the use or remote sensing data, GIS and GPS have influenced the building of most commonly used master sampling frames for agricultural statistics, focusing on the process to build them. We have also discussed how the availability of low cost GPS with acceptable accuracy, combined with the use of remote sensing data and GIS has facilitated the adoption of specific kinds of master sampling frame which require less preparatory work, like area frames without physical boundaries, clustered and un-clusters point sampling and increased the accuracy of data collected through the survey. More extensive use of remote sensing data, GIS and GPS, associated to more efficient and accurate methodologies will facilitate the improvement of quality

and timeliness of agricultural and rural statistics in developing countries, as advocated by the Global Strategy to Improve Agricultural and Rural Statistics.

References

Carfagna, E. (1998). Area frame sample designs: a comparison with the MARS project, Proceedings of Agricultural Statistics 2000, International Statistical Institute, Voorburg. pp. 261-277.

Carfagna, E. and Carfagna, A. (2010) Alternative sampling frames and administrative data; which is the best data source for agricultural statistics?, in R. Benedetti, M. Bee, R. Espa & F. Piersimoni, eds. Agricultural Survey Methods. Chichester, UK, Wiley. 434 pp.

Carfagna E. Pratesi M. and Carfagna A. (2013) Methodological developments for improving the reliability and cost-effectiveness of agricultural statistics in developing countries, *the 59th World Statistical Congress, Special Topic Session* (STS043) "Using geospatial information in area sampling and estimation for agricultural and environmental surveys", Hong Kong, 25-30 August 2013.

FAO (1996) Multiple Frame Agricultural Surveys, vol. I Current surveys based on area and list sampling methods ("FAO statistical development series", n. 7, 119 pp., FAO, Rome, 1996

FAO (1998) Multiple Frame Agricultural Survey, vol. 2, Agricultural Survey Programs Based on Area Frame or Dual Frame (Area and List) Sample Designs, Food and Agricultural Organization, Rome, Italy.

FAO, World Bank and United Nations Statistical Commission (2012) *Action Plan of the Global Strategy to Improve Agricultural and Rural Statistics*, FAO, Rome.

FAO, UNFPA (2012) *Linking Population and Housing Censuses with Agricultural Censuses*, Food and Agriculture Organization of the United Nations, 2012, http://www.fao.org/docrep/015/i2680e/i2680e.pdf

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

Keita N. (2013) Assessing the effect of slope and weather conditions on area measurement using GPS, *the 59th World Statistical Congress, Invited Paper Session* (IPS007) "Improving agricultural statistics through methodological validation and research", Hong Kong, 25-30 August 2013.

Keita N, Carfagna E. and Mu'Ammar G. (2010) Issues and guidelines for the emerging use of GPS and PDAs in agricultural statistics in developing countries, *Proceeding of ICAS-V, Fifth International Conference on Agricultural Statistics, Integrating Agriculture into National Statistical Systems Kampala*, Uganda 13-15 October 2010, pp. 1-14. Conference organized by FAO, ISI, UNSD, World Bank, Eurostat, AFDB, USDA. http://isi-web.org/news/icas-v
http://www.fao.org/fileadmin/templates/ess/documents/meetings_and_workshops/ICAS5/PDF/CONFERENCE_PROCEEDINGS.html

Keita N., Gennari P. (2013) Building a Master Sampling Frame by Linking the Population and Housing Census with the Agricultural Census, *the 59th World Statistical Congress, Special Topic Session* (STS063) "Role of population and housing and agricultural censuses in the national statistical systems", Hong Kong, 25-30 August 2013.

Kilic T., Zezza A., Carletto C., Savastano S. (2013) Missing(ness) in Action: Selectivity Bias in GPS-Based Land Area Measurements, *the 59th World Statistical Congress, Invited Paper Session* (IPS007) "Improving agricultural statistics through methodological validation and research", Hong Kong, 25-30 August 2013.

World Bank, FAO and United Nations Statistical Commission (2011) *Global Strategy to Improve Agricultural and Rural Statistics*, World Bank, Washington, DC.