Appropriate Survey Methods for Different Country Profiles - Key Challenges, Gaps and Remaining Methodological Issues

Elisabetta Carfagna[1] Andrea Carfagna[2]
[1]FAO, Statistics Division and University of Bologna, Department of Statistical Sciences
Via delle Terme di Caracalla, Rome, Italy
E-mail Address: elisabetta.carfagna@fao.org , elisabetta.carfagna@unibo.it

[2]Independent consultant
Via G. Leopardi 43, Offida, AP, 63035, Italy
E-mail Address: andreacarfagna@virgilio.it

ABSTRACT

Several kinds of approaches are followed for producing agricultural and rural statistics. The aggregation of the data collected by the extension workers is often used as agricultural statistics. In other countries, agricultural statistics are given by subjective estimates of experts. Other countries rely mainly on sample surveys, based on list or area frames or both.

Various possible methods for producing agricultural and rural statistics are analysed, in order to identify the most appropriate approach, given the characteristics of the country.

Then, the key challenges, the gaps and remaining methodological issues are discussed.

Keywords: Survey methods; Agricultural statistics; Administrative data

1. Introduction

In many developing countries, the quality and quantity of agricultural statistics are low and, in the last decades, have undergone a serious decline; see World Bank *et al.* (2011) and FAO et al. (2012). On the other side, allocating high shares of public resources to data collection for producing agricultural statistics is a difficult choice for most developing countries.

Statistically sound methods, based on probabilistic samples selected from complete and updated lists of farmers allow producing accurate and timely agricultural statistics if good quality

data are collected through the interviews. These statistics are essential for knowledge based planning, in order to facilitate rural development and reduce poverty and food insecurity. However, traditional statistical methods are very costly. Consequently, there is a strong need to review the methods adopted in developing and developed countries, in order to assess how their cost efficiency can be improved. Moreover, countries have very different level of statistical information, statistical capacity, farm size, farmers literacy, availability and quality of administrative data, and so on. These differences should be taken into account when developing a survey designs for agricultural statistics. Some considerations on this topic are presented in this paper.

2. Different kinds of statistical information

Several countries follow the traditional approach for producing agricultural statistics, (see Benedetti *et al*., editors, 2010): a complete enumeration census of farms is carried out every 5-10 years. The census is carried out by interviewing all the farmers in the country. The census list is used for sample surveys of farms and is updated by integrating different kinds of registers or other kinds of administrative data. Data are collected through computer assisted personal interviews or computer assisted telephone interviews, email or mail.

Some developing countries follow the same approach and carry out complete enumeration censuses of agriculture and carry out sample surveys in order to produce annual estimates of the main variables. However, this is a very costly approach that cannot be followed by several developing countries. Moreover, updating the census list properly requests availability of large and updated registers; this is not always the case in developing countries.

Several developing countries follow different approaches. Particularly, in some of them, only the population census is conducted; generally, using an administrative structure in which cartographic or other mapping materials are used to divide the country into enumeration areas. Some of these countries include few specific agricultural questions in the population census questionnaire, in order to build up the list of farms to be used for sample surveys.

In many developing countries, a sample agricultural census is conducted: some enumeration areas of the population census are randomly selected and screened for farms. These enumeration areas are considered as the agricultural enumeration areas and all or a subset of the listed farms are surveyed.

Finally, few countries generate the list of farms on the basis of administrative sources, such as business registrations or tax collections and do not conduct agricultural censuses.

3. Censuses and administrative data

Several North European countries are using registers more and more extensively, in order to reduce the cost and the respondent burden due to data collection (see for example Wallgren and Wallgren, 2007 and 2010). It is important to clarify that, in these countries, the registers are not used for direct tabulation, they replace the censuses, not the sample surveys. Registers, and not censuses, are used for building the list frame for sample surveys.

Even in Sweden, a country which initiated to make an extensive use of registers for statistics decades ago, the annual agricultural statistics are produced through sample surveys, based on a list frame built through registers, mainly tax files.

Subsidies are an important source of data in European countries; however, their use for direct tabulation is not feasible, as explained in Carfagna and Carfagna, 2010. In Sweden, for crops with subsidies based on surface and for other crops which are generally cultivated by the same farms, the bias is low, but for other crops the downwards bias can be about 20%; moreover, the subsidies in Europe are progressively less linked to the surface of cultivated crops.

Updating the list frame, generated by a census, through the subsidies register is not an easy task, consider that, in 2009, the business register and the farm register at Statistics Sweden were not harmonized yet (Wallgren and Wallgren, 2010).

The census list frame updated through the integration with registers can have a very low coverage for some categories of farms, as showed by a study conducted in Campania Region, in Italy, in 2002, two years after the census of agriculture (Giovacchini 2012). An area frame sample survey of farms cultivating flowers was conducted and the comparison was done with the census list updated with registers, like the register of farmers for the use of pesticides, not the subsidies register, since this kind of farms does not receive subsidies. The under-coverage, came out to be 48%; 54% if only farms with a surface smaller than or equal to half an hectare is taken into account, note that farms of this size account for 74% of farms cultivating flowers detected by the area sample survey. In this study, farms were selected through a grid of points located on the selected square segments; this means that farms were selected with probability proportional to size, thus larger farms are more likely included in the sample.

The cost and the respondent burden of a census of agriculture can be somewhat reduced through the generation of a pre-census list by integrating different kinds of registers or other kinds of administrative data. In fact, the census enumerators can be provided with the pre-census list of farms and a pre-compiled part of the questionnaire.

However, the quality of the pre-census list can be low also with good administrative data, very sophisticated record linkage procedures and geo-location of administrative information.

In several countries, agricultural statistics are computed by aggregating administrative data, like the declarations of extension workers and experts' guesses. The main task of extension workers is facilitating agricultural development by supporting farmers; thus they have a conflict of interest; moreover, generally they are note requested to follow specific rules (like taking into consideration specific fields and farms selected by the National Statistical Office or the Ministry of Agriculture) in order to come up with an estimate. These kinds of statistics show problems in terms of definitions, objectivity, timeliness, reliability and generally are not able to detect relevant changes in the time series, for a detailed analysis see Carfagna and Carfagna (2010).

4. Different surveys for different typologies of countries

Given the kind of information available in the country, the structural characteristics of the agricultural sector and the level of development of the national statistical system, different approaches should be adopted for collecting data for producing agricultural statistics.

Where a recent complete enumeration census of agriculture is available and the quality of this census is high, also from the coverage viewpoint, the list of farms created by this census should be

used for sample surveys. In fact, the information at farm level collected through the census can be used for efficient sample designs and, where possible, for interviews through mail, email, etc. (indeed, data collection through emails is still not widespread even in developed countries).

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. If the quality of the agricultural census is poor, these problems are faced since the first round of the sample survey. Thus, the assessment of the quality of censuses is extremely important and the quality of a recently carried out census is not necessarily high. Despite of this, most countries do not test the coverage of agricultural censuses.

Where the agricultural census is old or its coverage is not complete or other aspects of its quality are poor, an area frame should be conducted. 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.

Data for variables which cannot be directly observed in the fields, particularly socio-economic variables, are collected through interviews of farmers which operate the fields included in the selected area units (some estimators have been developed). When designing the sample, in order to prefix the number of farmers to be interviewed, thus the coefficients of variation for main parameters and increasing the efficiency of the sample design, taking into consideration the spatial autocorrelation, the farms can be selected by points in the sample area units, as described in Gallego *et al*., (1994).

In case a list of large, commercial farms (easy to update) and, in case, of other kinds of farms, can be created and multiple frame approach should be adopted, in order to reduce instability of estimates and increase their efficiency. A multiple frame is a combination of an area frame with one or more list frames, 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 technical details see Carfagna (1998) and Carfagna and Carfagna (2010).

If the agricultural census is not a complete enumeration census and some enumeration areas are randomly selected and screened for farms, the sampling frame for sample surveys consists of the agricultural census enumeration areas. This implies a two stage sample design based on the selection of the enumeration areas made before conducting the agricultural census, generally with very poor information related to agriculture. The main alternative to the use of this sampling frame is the multiple frame described above.

If a recent population census has been conducted using an administrative structure in which cartographic or other mapping materials are used to divide the country into enumeration areas, a traditional option is using the list of enumeration areas as sampling frame, although it is not efficient. 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. Also in this case, it is worthwhile evaluating if a multiple frame approach is more efficient.

A recent proposed comes from FAO and UNFPA (2012), for avoiding to face the cost of the agricultural census, and has been adopted in some countries like Mozambique and Burkina Faso: a list frame is created 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 is promising for countries where agriculture is not an important economic sector, like small islands. More work is needed for testing the quality of data collected using long questionnaires and the coverage of the list of farms generated from the populations census; particularly, the entity of under and over coverage in different categories of countries should be assessed. Finally, it should be taken into account that the list frame of farms generated through the module on agriculture submitted to the households presents very few auxiliary variables; thus the efficiency of the sample designs for annual sample surveys is very low and this may have a strong impact on annual survey costs. For more details and an analysis of advantages, disadvantages and requirements see Keita and Gennari (2013) and Carfagna et at. (2013).

When a sample survey has to be designed in countries where the list of farms is based on the integration of various administrative sources, such as business registrations, tax collections, and subsidies registers, much attention has to be devoted to the risk of under-coverage, especially units below a threshold required to be registered or pay taxes are generally excluded, as well as those which do not apply for subsidies. The under-coverage and over-coverage have to be carefully checked before using such a list. In fact, while this kind of list generally include 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).

Concerning the over-coverage, an interesting test has been carried out in Italy on a sample of 15,682 farms included in the pre-census list generated through a very careful integration of registers, including subsidies register, and using sophisticated record linkage procedures. It showed that only 39.15% of the farms in the pre-census list existed and were active at the census date; 44.74% of these farms were not active (over-coverage) and the pre-census test was not able to assess the existence of 16.11% of these farms (Berntsen and Viviano, 2011). This level of over-coverage suggests that the integration of different kinds of administrative data exposes to a high risk of over coverage and should be made only where the reliability of these data is very high and the definitions adopted are compatible with the ones of the census.

In case the over-coverage and under-coverage are high, it is advisable to use the registers for creating the list of commercial or large farms, reducing the size of the list and the risk of over-coverage, and combine it with an area frame for accounting for the under-coverage.

The kind of area frame to be adopted - area frame with or without physical boundaries, clustered or un-clustered points, transepts - depends on the characteristics of the country; for a detailed description see Carfagna (1998 and 2007).

5. Reliability of data collected through surveys

In this session we focus on one aspect of non sampling errors, that is the reliability of farmers declarations.

In principle, this reliability tends to be higher for larger farms and higher education level of the farmer. Indeed, assessing this reliability is quite difficult for variables which cannot be observed on the ground by the enumerator, typically socio-economic variables.

Let us focus on the area of fields and on possible ways for increasing its reliability.

In the framework of the project GCP/INT/903/FRA, the Statistics Division of FAO conducted pilot surveys in Cameroon, Niger, Madagascar and Senegal. For each field, the self reports of the farmers and the estimates of the enumerators where collected and the area was measured with the traditional method (compass and rope) and with a standard GPS (about 250 USD). Unfortunately, the kind of crop cultivated was not reported; thus we can draw just general conclusions; in fact the reliability of self declaration for market crops should be higher.

If we consider compass and rope as the gold standard, we can notice that the self reports of farmers tend to overestimate the area of fields, see Carfagna *et al.* (2013).

The compass and rope measurements minus corresponding self reported measures against compass and rope measurements confirms the tendency of self reported measures to overestimate the area of fields, particularly for small fields, as already noticed by several authors, see De Groote and Traoré (2005) and Carletto et al. (2013).

Self reported field measures are not a good proxy for estimating the area of fields measured with compass and rope (Stock and Watson, 2003); in fact, the experiment conducted by FAO shows that the R-squared is only 0.5919, the slope is 0.5694091 and the constant 1,022.847.

Additional research is needed for assessing if the self-declarations can be improved if the fields are showed on a map to the farmers during the interview. This is generally the case when an area frame is adopted.

This experiment suggests to be very careful when farmers declarations are used. Indeed, for most of the variables collected through an interview, only some consistency checks can be done and no real comparison with a reliable measurement instrument.

The farmers tend to overestimate the area of the fields, instead the enumerators tend to underestimate the area, in fact the median of the paired difference, on each observation, between the area measured with rope and compass and the estimate of the enumerator is 50 square meters and the median of the relative difference is 0.0555556 %. However, the parametric and non parametric tests (Stock J. H., Watson M. W.,2003 and Student, K. Pearson,1931) we have made have showed that the distribution of measurements with compass and rope and enumerators guesses are not significantly different.

Using the guess of the enumerators as a proxy for estimating the area of fields measured with compass and rope is less risky than using the self reported field measures. In fact, the R-squared is 0.7859, the slope is 0.8293134 and the constant 382.6549.

The objective measurement of the area of fields can be used for benchmarking self reports also for other kinds of information, at least as an alert, in case the discrepancy is high. So, the measurement of the area of the fields should be made even if self declarations of the farmers are collected. Unfortunately, measuring the area of the fields of a farm with the compass and rope proven to be time consuming and very cumbersome.

A standard GPS allows reducing the time needed for the measurements in the average of one third. The same experiment proved that that the measurements with GPS receivers are generally accurate, although the accuracy tends to be lower for very small fields, particularly under dense and partial tree canopy cover (due to the low quality of the signal), for more details, see Keita and Carfagna (2009).

Therefore, we would suggest measuring the area of fields with a GPS, when the quality of the signal is good and the fields are not too small (for very small fields the use of compass and rope is suggested) for two aims: collecting accurate measurements of the area of the fields and using the area of the fields as a warning for the reliability of socio-economic information when the discrepancy concerning the area of fields is high.

When an area frame is conducted, the enumerator sees at least some of the fields of the farm and can play the role of the warning tool even if the area of the fields is not measured with compass and rope or GPS.

6. Conclusions

Different kinds of approaches for producing reliable agricultural statistics have been analyzed, and most appropriate ones are identified, according to the kind of information available in the country and the agricultural characteristics of the countries. Sampling frames based on agricultural census, population census, administrative registers and area and multiple frames have been taken into consideration.

Our conclusions are that the most appropriate approach depends on the specific characteristics of the country and that some aspects of the implementation of some approaches need further research, like the over and under coverage of list frames created integrating various kinds of administrative data or when a module with a few questions concerning agricultural variables is included in the questionnaire for a population census.

Another aspect to be further analysed, in the contest of developing countries, is the identification of the farmers when they are selected through an area frame and they live far from the fields they operate. The average time needed for identifying the farmers and the risk of missing data, in the different typologies of developing countries should be assessed.

Finally, we have highlighted the risk of collecting unreliable data through farmers interviews and proposed some preliminary ways for addressing this problem.

REFERENCES

Benedetti R., Bee M., Espa R., Piersimoni F., eds. (2010) Agricultural Survey Methods. Chichester, UK, Wiley. 434 pp.

Berntsen E., Viviano C. (2011) La progettazione dei censimenti generali 2010-2011: la rilevazione di controllo della copertura e qualità del prototipo di registro statistico delle aziende agricole (Clag) e la riconciliazione con la Base integrata delle fonti amministrative (Bifa), Istat working papers, n.1 2011

http://www.istat.it/it/files/2011/06/Istat_Working_Papers_1_2011.pdf

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. (2004) List frames, area frames and administrative data, are they complementary or in competition, invited paper to the Meeting MEXSAI Conference organised by Eurostat, FAO, OCSE, UN/ECE, NASS/USDA, JRC, ISI, Istat, SAGARPA, Cancun (Mexico), 2-4 November 2004, Proceeding on the web site: http://www.nass.usda.gov/mexsai/;

Carfagna, E. (2007) A comparison of area frame sample designs for agricultural statistics, Bulletin of the International Statistical Institute, the 56th Session, 2007, Proceedings, Special Topics Contributed Paper Meetings (STCPM11) "Agricultural and Rural Statistics" organized by Michael A. Steiner (NASS-USDA), 1470, Lisboa, August 22-29, 2007 F:\papers\1470.pdf, pp. 1-4.

Carfagna E. (2013) Using satellite imagery and geo-referencing technology for building a master sampling frame, the 59th World Statistical Congress, Invited Paper Session (IPS110) Developing a master sampling frame for integrated survey, Hong Kong, 25-30 August 2013 http://www.statistics.gov.hk/wsc/IPS110-P1-S.pdf

Carfagna, E., 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., 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 http://www.statistics.gov.hk/wsc/STS043-P1-S.pdf

Cotter J, Davies C., Nealon J., Roberts R. (2009) Area Frame Design for Agricultural Surveys in Benedetti, Bee, Espa, Piersimoni (Editors), Agricultural Survey Methods, Wiley, New York.

Carletto C., Savastano S., Zezza A. (2013) "Fact or artifact: The impact of measurement errors on the farm size–productivity relationship", Journal of Development Economics, 103 (2013), pp. 254-261

De Groote, H., Traoré, O. (2005) The cost of accuracy in crop area estimation, Agricultural Systems 84, 21–38.

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

Giovacchini A. (2012) Area and point sampling frames for agricultural statistics, presentation at the High Level Stakeholders Meeting on the Global Strategy - From Plan to Action, FAO, December 2012

Keita N., Carfagna E. (2009) Use of modern geo-positioning devices in agricultural censuses and surveys, Bulletin of the International Statistical Institute, the 57th Session, 2009, Proceedings, Special Topics Contributed Paper Meetings (STCPM22) "Using advanced data collection methods and modern tools to improve agricultural statistics data quality", Durban, August 16-22, 2009

http://www.statssa.gov.za/isi2009/ScientificProgramme/IPMS/0617.pdf, pp. 1-23.

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.

Stock J. H., Watson M. W. (2003), Introduction to Econometrics, Pearson Education.

Student, K. Pearson (1931), On the “z” Test, Biometrika, Vol. 23, No. 3/4 (Dec., 1931), pp. 407-415

Wallgren, A., Wallgren B. (2007) Register-based Statistics – Administrative Data for Statistical Purposes. John Wiley & Sons Ltd.

Wallgren A., Wallgren B. (2010) “Using Administrative Registers for Agricultural Statistics” in Benedetti, Bee, Espa, Piersimoni (Editors), Agricultural Survey Methods,

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

Wilcoxon F. (1945), Individual comparisons by ranking methods, Biometrics Bulletin, vol. 1, pp. 80-83.