Methodological developments for improving the reliability and cost-effectiveness of agricultural statistics in developing countries

Elisabetta Carfagna[1] Monica Pratesi[2] Andrea Carfagna
[1] FAO and University of Bologna, Bologna, ITALY
[2] University of Pisa, Pisa, ITALY

[1] Contact author: Elisabetta Carfagna, e-mail: elisabetta.carfagna@unibo.it

Abstract

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 and the reliability of information collected through interviews is sometimes debated. 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. In some cases, new methods should be developed, which allow for more efficient use the new technologies, mainly Geographic Information Systems (GISs), GPS Global Positioning Systems (GPSs) and remote sensing. Some considerations on this topic are presented in this paper.

Keywords: Agricultural statistics, cost-effectiveness, technological and methodological development.

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). Because three out of four poor people in developing countries live in rural areas, accurate and timely agricultural statistics are essential for knowledge based planning, in order to facilitate rural development and reduce poverty and food insecurity.

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.

In most countries, agricultural statistics are computed by aggregating administrative data, like the declarations of extension workers (whose main task is facilitating agricultural development by supporting farmers), experts' guesses and sometimes registers. These kinds of statistics show problems in terms of definitions, objectivity, timeliness, reliability and so on (for a detailed analysis see Carfagna and Carfagna, 2010).

Statistically sound methods, based on probabilistic sampling allow overcoming these problems when agricultural statistics have to be produced; 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. In some cases, new methods should be developed, which allow for more efficient use the new technologies, mainly Geographic Information Systems (GISs), GPS Global Positioning Systems (GPSs) and remote sensing. In this paper, we propose some considerations on these topics.

2. Censuses, sample surveys and registers

A traditional approach for producing agricultural statistics, adopted in most developed

countries is the following (see Benedetti et al. eds. 2010): a complete enumeration census is carried out every 5-10 years. The census is carried out by interviewing the farm. 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 or mail.

Some developing countries follow the same approach and carry out complete enumeration censuses of agriculture and carry out sample surveys in order produce annual estimates of the main variables. However, this is a very costly approach that cannot be followed by many 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 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 (see Selander et al., 1998 and Wallgren and Wallgren, 1999) 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. Consider that, 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.

Geographic Information Systems (GIS) can contribute to reduction of complete enumeration costs and respondents' burden through the generation of a pre-census list by integrating different kinds of registers or other kinds of administrative data. In fact, geo-locating the headquarter of farms facilitates the integration of different registers and other kinds of administrative data, in order to create the pre-census list. 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. For example, a test carried out in Italy on a sample of 15,682 farms included in the pre-census list showed that only 39.15% of these farms existed and were active at the census date; 44.74% of these farms were not active and 16.11% of

these farms were not identified through the pre-census test (Berntsen and Viviano, 2011). This level of over-coverage suggests to adopt this approach only where the reliability of administrative data is very high and the definitions adopted are compatible with the ones of the census. This approach should be advocated only in specific kinds of developing countries.

An alternative solution recently proposed by FAO and UNFPA is eliminating censuses of agriculture and adding a few questions on agriculture in the questionnaire for the population census. The aim is creating the list of farms without facing the cost of the agricultural census (Linking Population and Housing Censuses with Agricultural Censuses, FAO and UNFPA, 2012 http://www.fao.org/docrep/015/i2680e/i2680e.pdf). This approach is promising particularly for countries where agriculture is not an important economic sector, like small islands. Mozambique conducted its population and housing census in 2007 and included an agricultural module in questionnaire. Burkina Faso, in 2007, carried out the survey with the core module jointly with the population and housing census, and the supplementary modules completed as a separate operation soon afterwards. In other countries, like Canada, the population and the agricultural censuses are carried out jointly, allowing linkage of the records concerning households and farms. For other examples see FAO and UNFPA, 2012.

As said before, this approach is promising, although 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.

3. Area and multiple frame sample surveys

A completely different approach is using area or multiple frame sample surveys for producing annual estimates of main agricultural variables (Carfagna 1998 and 2004; Cotter et al., 2009). Due to its completeness, an area frame should be preferred in some cases (Carfagna and Carfagna, 2010): if other complete frame is not available, or an existing list of sampling units changes very rapidly, an existing frame is out of date, if an existing frame was obtained from a census with low coverage or if a multiple purpose frame is needed for estimating many different variables linked to the territory (agricultural, environmental etc.).

The most relevant disadvantage of an area frame sample designs has been for decades the cost of implementing the survey program, due to the need of producing detailed cartographic material. This disadvantage have been widely overcome by the availability of GIS, Global Positioning System (GPS) and remote sensing data. There is no more need to delineate primary sampling units and secondary sampling units with physical boundaries in order to identify the boundaries on the ground during the survey. Alternative kinds of area frames have been adopted in several developed and developing countries (segments with theoretical boundaries like squares, or rectangles and clustered or un-clustered points).

Another disadvantage of area frames is the sensitivity to outliers and the instability of estimates. The most widespread way to avoid instability of estimates and to improve their precision is adopting a multiple frame sample survey design. For agricultural surveys, a list of very large operators and of operators that produce rare items is combined with the area frame. If this list is short, it is generally easy to construct and update.

Some studies are still needed for assessing the difficulty in identifying the farmers through an area frame, when interviews have to be conducted for collecting data concerning socio-economic variables and where respondents live far from the selected

area units. The computation of the average time needed for identifying the farmers and the risk of missing data, under the different conditions, request additional research.

4. Reliability of data collected through surveys

After all the considerations concerning the sampling frame, in the next paragraphs we focus on reliability of data collected through surveys, focusing on areas of plots measured with GPS, self-reported estimated by enumerators.

4.1 Reliability of GPS and self-reported data

In the framework of the project GCP/INT/903/FRA, the Statistics Division of FAO conducted pilot surveys in Cameroon, Niger, Madagascar and Senegal to assess if a standard GPS (about 250 USD) can be used for measuring areas on the ground. For each plot, the area was measured with the traditional method (compass and tape) and GPS. The result was that the measurements with GPS receivers are much faster and generally accurate, although the accuracy tends to be lower for very small plots, particularly under dense and partial tree canopy cover (Keita and Carfagna, 2009).

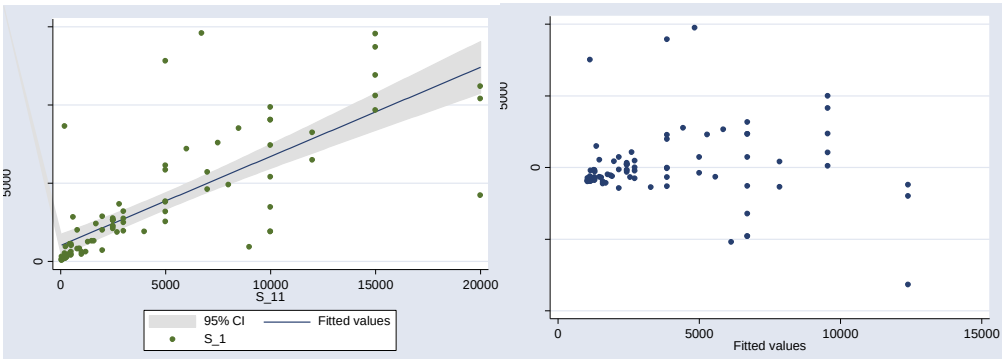
The project, for the same plots, collected also the self reports of the farmers and the estimates of the enumerators. Unfortunately, the kind of crop cultivated was not reported; thus we can draw just a general conclusion; in fact the reliability of self declaration for market crops should be higher.

If we consider compass and tape as the gold standard, we can notice that the self reports of farmers tend to overestimate the area of plots. Eliminating the outliers (using 87 plots) the median of the paired difference, on each observation, between the area measured with tape and compass and the self reported area is -50 square meters and the median of the relative difference is -9.577033 %,

The Student's paired test for two variables observed on one sample (Student and Pearson, 1931) showed that the paired difference between the two kinds of measurements does not have a mean equal to zero. Student's test assumes that the paired difference is Normally distributed and that the variances of the two variables on the sample are equal, although unknown. Thus we perform also the non parametric Wilcoxon matched-pairs signed-ranks test (Wilcoxon, 1945) for testing the equality of matched pairs of observations and relaxing the mentioned hypotheses. The null hypothesis is that both distributions are the same. The Wilcoxon signed-ranks test confirms the result that the two distributions are significantly different.

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

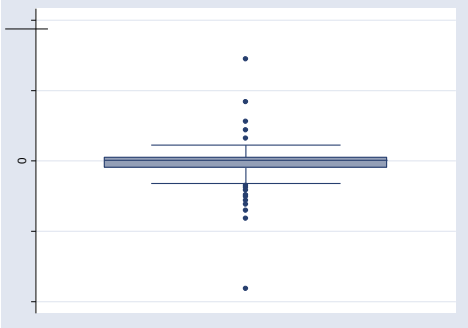
Self reported plot measures are not a good proxy for estimating the area of plots measured with compass and tape (Stock and Watson, 2003); in fact, the R-squared is only 0.5919, the slope is 0.5694091 and the constant 1,022.847. See the regression fit and the residuals:

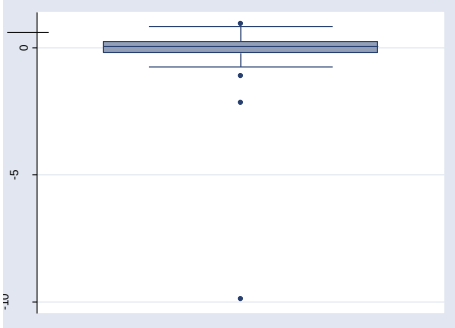


Additional research is needed for assessing if the self-declarations can be improved if the plots are showed to the farmers when during the interview. This is generally the case when an area frame is adopted. The objective measurement of the area of plots could be used for benchmarking self reports also for other kinds of information.

4.2 Reliability of enumerators

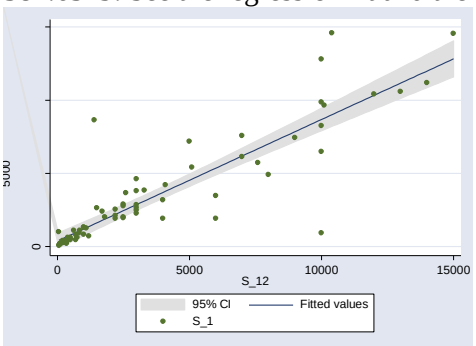
The estimates of the enumerators tends to underestimate the area of plots, in fact the median of the paired difference, on each observation, between the area measured with tape and compass and the estimate of the enumerator is 50 square meters and the median of the relative difference is 0.0555556 %, see the box plots below.

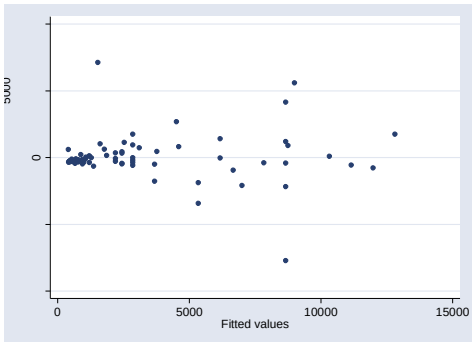




However, the Student's paired test for two variables observed on one sample shows that the paired difference between the two kinds of measurements has not a mean significantly different from zero. According to the non parametric Wilcoxon matched-pairs signed-ranks test, null hypothesis that both distributions are the same is not rejected. Thus, the two distributions are not significantly different.

Using the guess of the enumerators as a proxy for estimating the area of plots measured with compass and tape is less risky than using the self reported plot measures. In fact, the R-squared is 0.7859, the slope is 0.8293134 and the constant 382.6549. See the regression fit and the residuals:





5. Conclusions

Different kinds of approaches for producing reliable agricultural statistics have been analyzed, in order to assess if and how the costs can be reduced, taking advantage of the new technologies, mainly Geographic Information Systems, Global Positioning Systems and remote sensing, namely integration of various kinds of administrative data for creating a pre-census list, adding a module on agricultural variables to the questionnaire for a population census, using an area frame combined with a list of large farms. Our conclusion is that most appropriate approach depends on the specific situation of the country and 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 in included in the questionnaire for a population census. Another topic is the difficulty to identify the farmers when are selected through an area frame and leave far from the fields they operate, the corresponding cost and risk of missing data. Finally, we have highlighted the risk of

collecting unreliable data through farmers interviews.

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. 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.

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, 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

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.

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.