Methodological developments for improving agricultural statistics for different typologies of countries

Elisabetta Carfagna*
University of Bologna, Bologna, Italy – elisabetta.carfagna@unibo.it

Andrea Carfagna
Independent consultant, Offida, Italy – andreacarfagna@virgilio.it

Abstract

The need to reduce costs for producing accurate statistics forces countries to increase the efficiency of their statistical systems, exploring innovative approaches for sampling frame construction, sample design and estimation. Several countries are using more extensively low cost information for improving the accuracy of estimates, such as administrative data (Carfagna and Carfagna 2010 and Carfagna *et al*. 2013) and georeferenced information. The levels of statistical capacity, farm size, farmers literacy, availability and quality of administrative data are crucial for the identification of most appropriate approaches for the different typologies of countries. In this paper, we focus on advantages, disadvantages and requirements of some methodological developments, taking into consideration different typologies of countries, and devoting particular attention to the use of georeferenced information.

Keywords: methodological developments, agricultural statistics, georeferenced data, GIS operations.

1. Introduction

Statistics are essential for knowledge based planning, in order to facilitate rural development and reduce poverty and food insecurity. 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 interviews. 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.

Using administrative data, Geographic Information Systems (GIS), Global Positioning Systems (GPS) and remote sensing, easily accessible at decreasing prices, allows increasing the cost efficiency of the statistical systems, and requires changes in the methodological approach. However, countries have very different levels of statistical information, statistical capacity, farm size, farmers' literacy, availability and quality of administrative data, and so on. These differences are relevant for increasing the cost-efficiency of statistical systems of countries.

In this paper, we focus on advantages, disadvantages and requirements of some methodological developments, taking into consideration different typologies of countries, and devoting particular attention to the use of Geographic Information Systems (GIS), Global Positioning Systems (GPS) and remote sensing data.

In sections 2 and 3, the most appropriate sampling frames and sample designs for different typologies of countries are analysed. In section 4, the impact of GIS, PDA, GPS and remote sensing in the sampling frame construction is discussed. In section 5 and 6, attention is devoted to some statistical issues related to the use of georeferenced information at the estimator level.

2. Different sampling frames 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 is the most efficient sampling frame. 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).

Where the list generated by the agricultural census is old, or its coverage is not complete or other aspects of its quality are poor, an area frame should be preferred. An area sample survey design, in the general meaning, a probability sample survey in which, at least for one sampling stage, the sampling units are land areas. This approach foresees the subdivision of the analysed territory into non-overlapping pieces of land, according to specific criteria, to create the area sampling frame. Point frames are generally considered area frames because points are small circles on the ground.

Sometimes, a list of large, commercial farms (easy to update) and, in case, of other kinds of farms, is combined 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.). The multiple frame approach also allows improving the efficiency of estimates and reducing their instability (Carfagna, 2001; Carfagna and Carfagna, 2010).

A crucial aspect of this approach is the identification of the area sample units included in the list frame; the two different supports increase the difficulty of this kind of record matching. When units in the area frame and in the list sample are not detected, the estimators of the population totals are upwards biased.

3. Area sampling frame and sample design tailored to the characteristics of the country

Points on systematic grids overlaid on satellite images and stratified according to the land cover are being adopted in some countries like Haiti and some regions in Malawi, mainly because they are very simple to implement. However, their efficiency depends on the travel cost, which is related to the road system in the country. Moreover, where farms and fields are very small, the fields surrounding the one identified by the sampled point generally present different characteristics and crops; thus this information is not redundant. In countries presenting these characteristics, a clustered point sampling or an area frame based on segment are more cost-efficient.

The kind of area frame to be adopted - area frame with or without physical boundaries, clustered or un-clustered points, transepts – should be selected on the basis on a careful analysis of the characteristics of the country; for a detailed description see Carfagna (1998). Correlograms can be used to optimise the area sampling unit (segment) size under a fixed budget and a given cost function (Carfagna 1998); in fact, area sampling units can be considered as clusters of elementary units. The optimal size is studied through the intra-cluster correlation, which is computed as a weighted average of correlogram values (Carfagna, 1998, Gallego *et al*., 1999). In some landscapes, an analysis of correlograms can suggest the use of a two-stage sample design and give the basic data for computing the optimum combination of number and size of primary and secondary sampling units (Carfagna *et al*. 2008).

Classified satellite images provide a proxy variable for the spatial structure of land cover, which can be used to optimise a sampling frame when available ground data is not sufficient to estimate correlograms (the graph of the spatial autocorrelation function at increasing distances). Moreover, correlograms based on remote sensing data can be used for feeding some sequential selection techniques that require autocorrelation at short distances, generally difficult to estimate from previous ground surveys. This happens for example for the DUST sampling technique (Dependent area Units Sequential Technique), that modifies the sampling selection probabilities, once a first set of segments has been sampled, according to the autocorrelation for contiguous segments (Arbia, 1993).

We have to highlight that the spatial autocorrelation estimated through remote sensing data can be used for feeding the procedures described above only if the spatial resolution of remote sensing data is not too far from ground data, particularly where the field size is small. In fact, the Moran coefficient, the Geary ratio and Cliff-Ord statistic are scale dependent: the spatial correlation values decline with the scale; moreover, they are dependent on the zoning system used in the aggregation, as noted by Qi and Wu (1996). Thus, the average field size in the country or parts of the country has a strong impact on the kind of remote sensing data that should be used.

4. Use of GIS, PDA, GPS and remote sensing for sampling frame construction and data collection

The development of a sampling frame has changed completely with the use of Geographic Information Systems (GIS), which allow overlapping 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 allow geo-referencing data collected on the ground, which can then be overlaid to the other geographic information layers through a GIS. The time and cost needed for building all kinds of 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 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.

When the 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.

Since sampling frames for agricultural statistics are generally multipurpose, 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 facilitates this approach.

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). However, this approach is risky in countries where the field size is small, particularly when recreational grade accuracy GPS (error 5-20 m) are used, for more details see Keita, 2013. In addition, the possibility to carry out panel surveys of farms identifying the same field in the subsequent surveys depends on the field size compared to the GPS accuracy.

The most common way for increasing the efficiency of an area frame is through classification of remote sensing imagery into major categories, such as cultivated land, woodlands, grasslands, bare soil and urban areas and the computation of percentages of agriculture. Unless land cover/use changes rapidly, this classification does not need to be updated frequently (every 10 years in relatively stable conditions). In some cases, strata are associated to the prevalence in an area of specific crops or groups of crops (summer or winter crops for example).

The spatial, spectral, and temporal, resolutions of the sensors are important factors to take in account for building, updating or stratifying an area sampling frame. In case the spatial resolution of remote sensing data is low, compared to the spatial variability of agriculture or the co-registration of the information layers combined in the GIS is weak, the efficiency of the stratification is low.

If a non-surveyed stratum is defined in areas presumed to be non-agricultural, low spatial resolution or weak co-registration introduce a bias if this stratum has some marginal agriculture. A test made by Gallego *et al*. (1999), based on CORINE Land Cover, showed that the stratum defined as “non agricultural” contained approximately 4% of the agricultural land.

Similar results were obtained comparing the ground survey carried out in 2005 in the Italian project named Agrit, with the photo-interpretation, in 2006, of the points of the first phase sampling. 37 per cent of the points attributed to non-agricultural strata in the photo-interpretation belong to agricultural strata according to the ground survey and particularly, 6 per cent to arable land and 18 per cent to permanent crops. Since the project did not foresee to select points in non-agricultural strata, the risk of severe bias is high for area estimates of crops.

5. Use of remote sensing data as auxiliary variable at the estimator level

Calibration and regression estimators are the main approaches for combining accurate and objective observations on a sample (e.g. ground observations) with the exhaustive knowledge of a less accurate or less objective source of information, or co-variable (classified images). There are two main types of calibration estimators, often named “direct” and ”inverse” (for a discussion see Gallego, 2004). The regression estimator has been used for crop area estimation since the early times of satellite earth observation (Hanuschak *et al*., 1980).

The appropriate spatial resolution of remote sensing data to be used, at the estimator level, for producing agricultural statistics depends on the characteristics of the country, particularly on the size and heterogeneity of parcels. A useful rule of thumb is using images for which most pixels are fully inside a plot and only a minority of pixels is shared by several plots. In fact, many mixed pixels (shared by several land cover types) reduce the linear correlation between ground observations and the image classification, which plays a crucial role in reducing the variance of the estimate produced using only ground observations. Moreover, a big amount of mixed pixels influence the skewness of the distribution of the image classification and inflate the variance of the regression estimator, particularly for crops cultivated in small plots. Finally, many mixed pixels disturb the linear relationship that should hold between ground observations and the image classification introducing a bias in the regression estimator when the sample size is small.

A common practice in remote sensing is excluding mixed pixels from the training set for image classification. This can improve the quality of the discrimination between classes. However, excluding mixed pixels to compute the area corresponding to a crop type on the sample is not coherent with the computation of the area classified to a crop type in the whole study area (or stratum), for which mixed pixels cannot be identified. Ignoring the existence of mixed pixels in the classification or photointerpretation of satellite images generates an overestimate of the relationship between remote sensing and ground data and, consequently, an underestimate of the variance of the estimators. The entity of the underestimation of the estimator variance is proportional to the amount of mixed pixels, which is related to the pixel and field size, and to the classification algorithm.

In order to overcome above-mentioned problems, sub-pixel analysis techniques are available, but they have not proved yet to be operational. Usual image classification attributes one class to each pixel; this is often known as sharp or hard approach. Alternative soft or sub-pixel methods are not new but they are receiving a growing attention. Soft classifications can have at least three different conceptual bases: probabilistic, fuzzy or area-share (Pontius and Cheuk, 2006). In the probabilistic conception, each pixel belongs to a class with a certain probability. The fuzzy conception corresponds to a vague relationship between the class and the pixel; it is very attractive for classes with an unclear definition, but difficult to use for area estimation. In the area-share conception, classes have a sharp definition and the classification algorithm estimates the part x_{ik} of pixel i that belongs to class k.

The relative efficiency of estimators combining ground and remote sensing data is lower when the ground survey is performed on a sample of points rather than on a sample of segments. In Italy, with segments, the relative efficiency of the regression estimator ranged from 1.4 (colza) to 8.6 (soy-been);

instead, in 2002, with points, it ranged from 0.9 (colza) to 1.2 (soy-been and soft wheat). Consequently, in years 2004, 2005 and 2006, the regression estimator was replaced by the Empirical Best Linear Unbiased Predictor (EBLUP). The relative efficiencies for 2005 and 2006 were higher than the ones obtained with points in 2002 but considerably lower than the ones obtained in 2000 with segments. Moreover, EBLUP had a strange effect on the area estimates: in 2005 the area estimate for durum wheat increased and for all other crops decreases, up to - 4.42 % (tomato). In 2006, all EBLUP estimates were lower than direct expansion estimates (from - 0.56 to - 7.29). These results suggested that, probably, the model introduced a bias. Consider that the coefficients of variation of the direct expansion estimator applied to the ground data were very low for the area of main crops; in 2005, the area reduction per cent due to the EBLUP was higher than the CV per cent for maize, sunflower, soy-bean, sugar beet and tomato and in 2006 for soft wheat, barley, maize, sunflower, sugar beet and tomato.

Due to the low efficiency and the risk to introduce a bias, we would not recommend the use of remote sensing data at the estimator level for improving agricultural estimates produced with point sampling.

6. Small area estimation

The need of statistics for small geographical domains has fostered the use of small area estimators based on spatial auxiliary variables. Already in 1988, Battese *et al*. proposed small area estimators for improving the estimate in an area with a very small sample exploiting the link between ground surveys (variable) and classified images (co-variable) in a large area.

Several small area estimators have been developed, e.g. EBLUP (Rao, 2003), Generalized Regression estimator (Rao, 2003), SEBLUP (Petrucci and Salvati, 2006; Pratesi and Salvati, 2009), Model Based Direct Estimator (Chandra and Chambers, 2005), Spatial MBDE (Chandra *et al*., 2007), M-quantile regression small area estimator (Chambers and Tzavidis, 2006). As for the EBLUP analysed before, the risk to introduce a bias can be high, particularly when the amount of ground observations in the small area is very limited.

The efficiency of small area estimators is strongly influenced by aggregation and disaggregation. Consider that georeferenced data are subject to aggregation and disaggregation when they are combined with other layers in a GIS. Pratesi and Petrucci (2014) assessed the sensitivity of several small area estimators to the level of aggregation of the underlying spatial data through simulation. For each small area, they computed the Average Relative Root MSE (AvRRMSE) for the original and for the aggregated population. They evaluated the scale effect through the percentage of increase of AvRRMSE for each predictor from the original population to the aggregated population. Not surprisingly, the small area estimators that strongly depend on the level of the spatial autocorrelation (like the SEBLUP predictor) showed a considerable increase of AvRRMSE, due to the decrease of the value of the spatial autocorrelation parameter. Since the ranking of the estimators, according to their error, changes when the area units of the auxiliary variable are aggregated, the scale effect should be taken into consideration, when choosing a specific small area estimator.

7. Conclusions

The paper has discussed some statistical issues and methodological developments in the production of agricultural statistics, taking into consideration different typologies of countries. Attention has been devoted to the kind of information available in the country and the structural characteristics of the agricultural sector, in order to delineate the most appropriate approaches for collecting data for producing agricultural statistics.

The use of correlograms and other specific analyses have been suggested for identifying the kind of area sampling frame that better fits the characteristics of specific typologies of countries and ensures efficient data collection. The inefficiency of point sampling where the quality of the road system is low, the travel cost is high and the fields are very small has been highlighted.

Advantages, requirements and some issues of the use of GIS, PDA and GPS and remote sensing for sampling frame construction and stratification and for data collection have been analysed for different kinds of sampling frame, highlighting the risks of bias when some areas are not surveyed.

The use of remote sensing data as auxiliary variables at the estimator level has been discussed, focusing on the suitable spatial resolution and the impact of inappropriate choices, and on the risk of bias, in case model based estimators are adopted. Moreover, the scale dependency of small area estimators has been discussed, suggesting to take into consideration the scale effect when choosing a specific small area estimator.

References

Arbia, G. (1993), The use of GIS in spatial statistical surveys. International Statistical Review. vol 63, n. 2, pp 339-359.

Battese G. E., Harter R.M., Fuller W.A. (1988), An error-components model for prediction of county crop areas using survey and satellite data. Journal of the American Statistical Association, 83, 28-36.

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., 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

Carfagna E., Giuiani D., Carfagna A.(2008), Optimisation of Area Frame Sample Designs through the use of Spatial Autocorrelation Functions, Atti della XLIV Riunione Scientifica della Società Italiana di Statistica, Università della Calabria, Arcavacata, 25-27 June 2008, pp. 1-2

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

Chambers, R. and Tzavidis, N. (2006), M-quantile models for small area estimation. Biometrika, 93, pp. 255-268

Chandra, H. and Chambers, R.L. (2005), Comparing Eblup And C-Eblup For Small Area Estimation. Statistics In Transition, 7, pp. 637-648

Chandra, H., Salvati, N. and Chambers, R. (2007), Small Area Estimation For Spatially Correlated Populations - A Comparison of Direct And Indirect Model - Based Methods. Statistics In Transition, 8, pp. 331-350

Gallego F. J. (2004) Remote sensing and land cover area estimation. International Journal of Remote Sensing, 25(15), 3019-3047.

Gallego, F.J. Carfagna, E. Peedell, S. (1999), The use of CORINE Land Cover to improve area frame survey estimates in Spain. Research in Official Statistics, 2, 99-122

Hanuschak, G. A., Sigman, R., Craig, M. E., Ozga, M., Luebbe, R. C., Cook, P. W., Kleweno D. D., Miller C. E., (1980). Crop-area estimates from landsat; transition from research and development timely results. IEEE Transactions on Geoscience and Remote Sensing, GE-18(2), 160-166

Petrucci, A. and Salvati, N. (2006), Small area estimation for spatial correlation in watershed erosion assessment, Journal of Agricultural, Biological, and Environmental Statistics, 11, pp. 169-182

Pontius Jr. R. G., Cheuk M. L. (2006) A generalized cross-tabulation matrix to compare soft-classified maps at multiple resolutions. International Journal of Geographical Information Science, 20(1), 1-30.

Pratesi M., Petrucci A. (2014), Developing robust and statistically based methods for spatial disaggregation and for integration of various kinds of geographical information and geo-referenced survey data, Second meeting of the Scientific Advisory Committee of the Global strategy to Improve Agricultural and Rural Statistics, FAO Headquarters, 29 - 30 January (2014)

Pratesi, M. and Salvati, N.(2009), Small area estimation in the presence of correlated random area effects, Journal of Official Statistics, 25, pp. 37-53

Qi, Y. and and Wu, J. (1996), Effects of changing spatial resolution on the results of landscape pattern analysis using spatial autocorrelation indices. Landscape Ecol., 11, pp. 39-49

Rao, J.N.K. (2003), Small area estimation, John Wiley & Sons, New York