QdS, vol. 16, 2014

Spatial data for agricultural statistics, focus on spatial resolution, change of support and transformations of spatial data

Elisabetta Carfagna
Department of Statistical Sciences, University of Bologna
E-mail: elisabetta.carfagna@unibo.it

Simone Maffei
TERRA NOVA GIS
E-mail:simone.maffei@gmail.com

Andrea Carfagna
Independent consultant
E-mail:andreacarfagna@virgilio.it

Summary: Big amounts of spatial data have become accessible at lower prices, fostering their use in the production of agricultural statistics. In this paper, we focus on spatial resolution, change of support and transformations of spatial data when they are used for sampling frames construction, sample designs, stratification, small area estimation and yield forecasting.

Keywords: Spatial data; Remote sensing; GIS; Small area estimation; Yield forecasting.

1. Introduction

Due to the technological development, in the last decades, different kinds of spatial data have become easily accessible at decreasing prices and have started to be used for producing statistics.

In this paper, we focus on spatial resolution of data, on change of support (section 2) and on some kinds of transformations like aggregation and disaggregation of spatial data when remote sensing data, Global Positioning Systems (GPS) and Geographic Information Systems (GIS) are used for producing agricultural statistics.

Particular attention is devoted to the impact of above-mentioned characteristics and transformations of spatial data on sampling frame construction and sample design (sections 3, 4 and 5), stratification (section 6), use of remote sensing data for agricultural statistics (section 7), small area estimation (section 8) and yield forecasting (section 9).

2. The change of support problem

The traditional spatial resolution for which statistical data are available are administrative areas, data on farms are collected at census enumeration area level, whilst the geographical units of auxiliary variables are areas defined by land use, land cover, soil type, watershed boundaries, and a variety of other biophysical and geophysical features. Thus, the possible support of geo-referenced data are usually points, lines, areas or surfaces. Census enumeration areas and their concomitant data sets seldom correspond to these geographic areas; consequently, data have to be interpolated, disaggregated and aggregated.

When the support of the spatial process of interest is different from the one of the observed data, a change of support problem arises.

The Modifiable Areal Unit Problem was introduced by Openshaw and Taylor (1979). It is a specification of the change of support problem and presents two facets (Arbia 1986; Arbia and Petrarca, 2013):

- "scale problem", which refers to the indeterminacy of any statistical measure with respect to changes in the level of data aggregation;
- the "aggregation (or zoning) problem", which concerns the indeterminacy of any statistical measure with respect to changes in aggregation criterion at a given spatial scale (e.g., two alternative partitions of the same area at a given spatial scale).

3. Use of satellite imagery for building a sampling frame

Various kinds of sampling frames are currently adopted for producing agricultural statistics (see Carfagna, 2013):

1. Population census enumeration areas used as first stage sampling unit and the list of households in selected enumeration areas is created;
2. List frames based on the population census: the list of farms or agricultural households is identified on the basis of specific agricultural questions included in the population census questionnaire (FAO and UNFPA, 2012; Keita and Gennari, 2013 and Carfagna et at., 2013).
3. Agricultural census enumeration areas (some enumeration areas are randomly selected and screened for farms);

4. List frames based on the agricultural census;
5. List of farms based on the integration of various administrative sources through record matching;
6. Area frames.

The sampling frames from 1) to 4) 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 fields 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.

Remote sensing data add the geographical dimension to the sampling frames from 1) to 4) mentioned above, providing 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 (not to the farms).

Lists of farms are often based on administrative sources, such as business registers or tax collections (see Carfagna et al. 2013). 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.

Area frames, in the general meaning, are probability sample surveys in which, at least for one sampling stage, the sampling units are land areas. They have a geographic dimension by definition. 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.

Point frames are generally considered area frames because points are small circles on the ground.

In some cases, samples are selected from different list frames and combined at the estimator level (Carfagna, 2001). From the point of view of the link with the land, this approach faces the same difficulties highlighted previously with reference to the list frames.

4. Estimating spatial autocorrelation to optimise sampling frames and designs

Classified satellite images provide a proxy variable for the spatial structure of land cover, that 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).

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 can be studied through the intra-cluster correlation, which can be computed as a weighted average of correlogram values (Carfagna, 1998, Gallego *et al.*, 1999).

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).

Moreover, correlograms based on remote sensing data can be used for feeding some sequential selection techniques which 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).

5. Use of geo-referencing technology for building sampling frames

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. Also 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.

6. Stratification

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 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 changes rapidly, this classification does not need to be updated frequently (every 10 years in relatively stable conditions).

The spatial, spectral, and temporal, resolutions of the sensors are important factors to take in account for building, updating or stratifying a sampling frame.

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.

An efficient and low cost stratification for agricultural estimates is based on percentages of agriculture that can be approximately derived from photo-interpretation of remote sensing images. 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).

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.

7. Use of remote sensing data for producing agricultural statistics

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):

$$\hat{\lambda}_{dir}(g) = P_g \Lambda_c$$

$$\hat{\lambda}_{inv}(g) = P_g'^{-1} \Lambda_c$$

where Λ_c is the column vector with the number of pixels classified into each class *c* and P_g and P_c are the error matrices with the proportions and for the sample.

$$P_g(g,c) = \frac{\lambda_{gc}}{\lambda_{g+}}$$

and

$$P_c(g,c) = \frac{\lambda_{gc}}{\lambda_{+c}}$$

for the sample.

The regression estimator (Hansen *et al.*, 1953, Cochran, 1977) has been used for crop area estimation since the early times of satellite EO (Hanuschak *et al.,* 1980):

$$\bar{y}_{reg} = \bar{y} + b(\bar{X} - \bar{x})$$

where: $\bar{y}$ and $\bar{x}$ are the sample means of the ground observations and the image classification, $\bar{X}$ is the population mean for the image classification and *b* is the angular coefficient of the regression between *y* and *x*. Ratio estimators are also used (Lathrop, 2006) and can be seen as a particular case of regression estimators, as well as difference estimators.

Small area estimators (Battese *et al.*, 1988) are also adopted to improve 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.

Let us focus on the spatial resolution of remote sensing data for producing agricultural statistics. The suitable spatial resolution mostly depends on the size 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 relationship between ground observations and the image classification, which plays a crucial role in reducing the variance of the estimate produced using only ground observation, as can be easily noticed in the formula of the variance of the regression estimator (Cochran, 1977, sect. 7.6):

$$V(\overline{y}_{reg}) = \frac{N-n}{N \times n}(1 + \frac{1}{n-3} + \frac{2G_x^2}{n^2})\sigma_y^2(1-\rho^2)$$

where

$$G_x = \frac{k_{3x}}{\sigma_x^3}$$

is the relative skewness.

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 *x* is not coherent with the computation of *X,* 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 described above. 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*.

8. The impact of aggregation of spatial data on small area estimation

The need of statistics for small geographical domains has fostered the use of small area estimators based on spatial auxiliary variables. Several small area estimators have been proposed, including for spatially correlated populations (Chandra *et al.* 2007), but limited attention has been devoted to their performance in case of change of support. Pratesi and Petrucci (2014) have assessed the sensitivity to the level of aggregation of the underlying spatial data of several small area estimators, namely EBLUP, 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).

The sensitivity analysis performed relies on a model based simulation study designed by Chandra *et al.* (2012) for comparing the performances of small area estimators. In the mentioned study, the number of small areas was fixed at $A = 20$. The model used to generate the population corresponded to a nested error regression model with random area effects for neighbouring areas distributed according to a simultaneously autoregressive spatial correlation structure with spatial autoregressive coefficient sets equal to 0.75 (high spatial correlation). This was of the form $y_{ij} = 100 + 1.5x_{ij} + v_i + e_{ij}$, where $x_{ij} \sim Chi^2(20), j=1,...,N_i, i=1,...,A$, with the random area effects a_i generated as $N(0, 23.52)$; $\boldsymbol{v} = (v_i) = (\boldsymbol{I} - \rho\mathbf{W})^{-1}\boldsymbol{a}$; $\boldsymbol{W}$ is a proximity matrix of order A; $\mathbf{I}$ is a diagonal matrix of order A, and ρ is the spatial autoregressive coefficient and it is set equal to 0.75 (high spatial correlation). The element W_{kl} of a contiguity matrix $\mathbf{W}$ takes the value 1 if area k shares an edge with area l and 0 otherwise.

The experiment carried out by Pratesi and Petrucci (2014) is based on about 10,000 points located randomly within 20 small areas, each representing an individual. The small area population sizes N_i are randomly drawn from a uniform distribution on $[450, 500]$ and kept fixed over the simulations. The location coordinates for each unit of the population are independently generated as $U[0,50]$.

In addition, it is assumed that the only spatial information available is the spatial coordinates of the sampled units and the spatial coordinates of the centroids of the small areas.

To examine the scale effect, the points are aggregated into *101* (in mean) areal units in each small area. The spatial aggregation is performed by aggregating a number of contiguous spatial units into one unit. A sample of size $n = 80$ is selected from each simulated population, with small area sample sizes proportional to the fixed small area population sizes, resulting in an average area sample size of $n_i = 4$. These area specific sample sizes n_i are kept fixed in the simulations, and the small areas are treated as strata, with the final sample selection carried out by randomly sampling within each small area. A total of $T = 500$ simulations is carried out.

For each small area, the Average Relative Bias (AvRBias) and the Average Relative Root MSE (AvRRMSE) are computed for the original and for the aggregated population.

The scale effect is evaluated through the percentage of increase of AvRRMSE for each predictor from the Original Population to the Aggregated Population. The SEBLUP predictor shows the highest increase in terms of AvRRMSE (+21.5%), due to the decrease of the value of the spatial autocorrelation parameter. The MQ-type estimator has the lowest increase of AvRRMSE (+0.4%) because the changes in geography do not affect the M-quantile coefficients at area level; although the lowest level of AvRRMSE for the aggregated population is showed by EBLUP_GC: 1.814% (1.534% for the original population).

9. The impact of change of support on yield forecasting

Several kinds of yield forecasting models have been developed in the last years, most of them include the use of remote sensing data (see for example Shi et al, 2007, Salazar *et al.*, 2008), sometimes combined with ground observations (Genovese *et al.,* 2001). Wall, *et al.* (2008) reviewed the most common approaches to the use of vegetation indices in crop yield forecasting models developed in the past two decades and concluded that the most accurate yield estimates from remote sensing data have been reported using regression analysis and extensive multi temporal datasets.

In this paper, we focus only on the implications of the use of geo-referenced data in statistical models. Main explanatory variables of the statistical yield forecasting models are the trend of historical yields and some agro-meteorological models (generally deterministic models like Penman's), which account for water and temperature stress.

Statistical forecasting models require variables with a common support (polygons or, more frequently raster) obtained through a series of GIS operations, like interpolation of point data (e.g. meteorological data, soil data, etc.), disaggregation and aggregation of database layers (land cover, crop masks, remote sensing data etc.). These operations generate a series of errors in the explanatory variables, due to interpolation, location, change of support, and so on.

More complex implications arise when these data are combined with other kinds of data, in the statistical model. In this case, the uncertainty or the bias of certain input data produces a propagated impact on the output of the statistical model.

We can identify four branches related to this issue:

1. methodologies to measure the uncertainty induced by some input in the final results;

2. operative indications at design stage of the procedure to reduce the impact of those disturbs on the final result;

3. methods to correct the disturb caused by some specific processing procedure on the input data;

4. specific statistical and mathematical procedures adjusted to take into account these disturbs.

Some authors have attempted to address the problem of error propagation with an analytical approach, through an extremely detailed "error" modelling and have derived analytically the consequences of the impact of the error on the final model. However, Crosetto and Tarantola (2001) state that the major limitation of the analytical approaches is that it applies only to particular kinds of GIS operations, or to particular types of data. Given the complexity of the relations among the different information layers, the analytical approach has not been followed extensively.

More frequently, a-posteriori analyses are carried out with the aim to optimize the procedures of data processing, to obtain the best result in terms of accuracy of the final estimates. The a-posteriori approach does not try to include the source of disturb/error into the final model (thus, taking into account the distortion on the final model/formulae), but tries to simulate/highlight the problems to measure the impact and minimize the effects. Veregin (1994) states that simulation modelling is an attractive alternative when little is known about error propagation mechanisms. In such cases, simulation modelling can be applied whether or not a formal error model has been developed, with the following aims:

- simulate the effects of GIS operations on the data, or, more generally, to simulate the presence of errors on the data

- quantitatively assess the impact on the final model

Where possible, above results are used to optimize the GIS procedure.

In early nineties, Cancellieri *et al.* (1993) adopted a sensitivity analysis approach for measuring the influence of the different variables and the effects of GIS operations on a statistical yield-forecasting model.

Saltelli, *et al.* (2012) distinguish between uncertainty and sensitivity analysis:

- uncertainty analysis is responsible for analysing the propagation, into results of models, of the uncertainty embedded in some variables. The uncertainty analysis answers the question "how reliable / uncertain this model is?"

- sensitivity analysis is concerned with measuring the strength of the impact / relationship between variables and model. That is: "what is the impact of "each" factor on the variability of the final outputs?".

The literature on sensitivity analysis is wide. However, this approach is seldom used for assessing the impact of GIS operations in yield forecasting models; thus, we believe that research is still needed in order to take into account the specificities of this field of application.

10. Conclusions

The paper has analysed the impact of spatial resolution, change of support, transformations and co-registration of spatial data, when remote sensing data, Global Positioning Systems (GPS) and Geographic Information Systems (GIS) are used for producing agricultural statistics. Particular attention has been devoted to the impact on sampling frame construction and sample design, stratification, use of remote sensing data for agricultural statistics, small area estimation and yield forecasting, highlighting advantages and warnings.

The analysis of the influence of the use of remote sensing data, GIS and GPS on the building process of most common sampling frames for agricultural statistics, has highlighted how remote sensing data add the geographical dimension to most commonly used sampling frames, providing 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; however, this geographic information cannot be associated to the farms, unless their borders are digitized.

From the point of view of the link with the land, the multiple frame approach faces the same difficulties showed for list frames.

Correlograms computed on remote sensing data are useful for identifying the most appropriate area frame sample design; namely the kind of area frame, segment size, number of stages, sample selection procedure involving the spatial autocorrelation. However, the spatial autocorrelation estimated through remote sensing data can be used for feeding the mentioned procedures only if the spatial resolution of remote sensing data is not too far from ground data, particularly where the field size is small.

The development of sampling frames has changed with the use of Geographic Information Systems (GIS), which allow overlapping and integrating different geographic information layers, such as borders of administrative areas, enumeration areas, fields, land cover databases, coordinates of headquarters of farms and households. In addition, the use of Global Positioning Systems (GPS) has influenced the development of sampling frames; in fact, GPS allows geo-referencing data collected on the ground, which can then be overlaid to the other geographic information layers through a GIS.

Remote sensing data have a very important role in the stratification of area frames. The trade-off between the relative efficiency and the long lasting of the stratification has to be taken into consideration, when evaluating the cost efficiency of remote sensing data for stratification. Some approaches can be very efficient but can generate high biases, particularly when the geographic information is used for eliminating parts of the area from the sampling domain.

The suitable spatial resolution of remote sensing data mostly depends on the size of parcels, when remote sensing data are used as auxiliary variables for producing agricultural statistics. Only images for which most pixels are fully inside a plot and a minority of pixels is shared by several plots should be used. In fact, mixed pixels reduce the linear relationship between ground observations and the image classification and inflate the skewness of the distribution of the image classification and the variance of the estimator and biases the estimates.

Sub-pixel analysis, like probabilistic, fuzzy or area-share classification have not proved yet to be operational.

When a geographic auxiliary variable is adopted for small area estimators, the ranking of the estimators, according to their error, changes when the area units of the auxiliary variable are aggregated.

Finally, evaluating the uncertainty of yield forecasting models is very relevant, when geo-referenced data are used in statistical models, which require variables with a common support, obtained through a series of GIS operations. Simulation models for sensitivity analysis can contribute to this evaluation.

References

Arbia, G. (1986). The Modifiable Areal Unit Problem and the Spatial Autocorrelation Problem: Towards a Joint Approach, *Metron*, **44**, 391–407.

Arbia, G. (1993). The use of GIS in spatial statistical surveys. *International Statistical Review*, **63:2**, 339–359.

Arbia, G., Petrarca, F. (2013). Effects of Scale in Spatial Interaction Models. *Journal of Geographical Systems*, **15**, 249–264.

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.

Cancellieri M.C., Carfagna E., Narciso G., Ragni P. (1993). Aspects of Sensitivity Analysis of a Spectro-Agro-Meteorological Yield Forecasting Model, EARSeL (European Association of Remote Sensing Laboratories), *Advances in Remote Sensing*, **2:2-IV**, 124–132.

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

Carfagna E. (2001). Multiple Frame Sample Surveys: Advantages, Disadvantages and Requirements, *Proceedings of the 53th ISI Conference*, International Statistical Institute, Seoul, 253–270.

Carfagna E. (2013). Using satellite imagery and geo-referencing technology for building a master sampling frame, *Proceedings of the 59th World Congress*, International Statistical Institute, Hong Kong, 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, pp.

45–61.

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, 1–2.

Carfagna, E., Pratesi M., Carfagna, A. (2013). Methodological developments for improving the reliability and cost-effectiveness of agricultural statistics in developing countries, *Proceedings of the 59th World Congress*, International Statistical Institute, Hong Kong, 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**, 255–268.

Chandra, H. and Chambers, R.L. (2005). Comparing Eblup And C-Eblup For Small Area Estimation, *Statistics In Transition*, **7**, 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**, 331–350.

Chandra, H., Salvati, N., Chambers, R. and Tzavidis, N. (2012). Small area estimation under spatial nonstationarity, *Computational Statistics and Data Analysis*, **56**, 2875–2888.

Cochran W., (1977). *Sampling Techniques*, John Wiley & Sons, New York.

Crosetto, M., Tarantola, S. (2001). Uncertainty and sensitivity analysis: tools for GIS-based model implementation, *International Journal of Geographical Information Science*, **15**, 415–437.

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

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.

Gallego F.J., DelincÃľ J. and Carfagna E. (1994). Two Stage Area Frame on Squared Segments for Farm Surveys, *Survey Methodology*,**20:2**, 107–115.

Gallego, F.J. Feunette, I. Carfagna, E. (1999). Optimising the size of sampling units in an area frame, in: J. Gómez -Hernández *et al.* (eds.): *GeoENV II - Geostatistics for Environmental applications*, Series: Quantitative geology and geostatistics, vol.10, Kluwer, pp. 393-404.

Genovese, G., Vignolles, C., Nègre, T., Passera, G., (2001). A methodology for a combined use of normalised difference vegetation index and CORINE land cover data for crop yield monitoring and forecasting. A case study on Spain. *Agronomie*, **21:1**, 91–111.

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

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.

Keita N. (2013). Assessing the effect of slope and weather conditions on area measurement using GPS, *Proceedings of the 59th World Congress*, International Statistical Institute, Hong Kong, http://2013.isiproceedings.org/ Files/IPS007-P2-A.pdf

Keita N., Gennari P. (2013). Building a Master Sampling Frame by Linking the Population and Housing Census with the Agricultural Census, *Proceedings of the 59th World Congress*, International Statistical Institute, Hong Kong, http://2013.isiproceedings.org/ Files/STS063-P1-S.pdf.

Kilic T., Zezza A., Carletto C., Savastano S. (2013). Missing(ness) in Action: Selectivity Bias in GPS-Based Land Area Measurements, *Proceedings of the 59th World Congress*, International Statistical Institute, Hong Kong, http://2013.isiproceedings.org/ Files/IPS007-P3-S.pdf.

Lathrop R. (2006). The application of a ratio estimator to determine confidence limits on land use change mapping, *International Journal of Remote Sensing*, **27:10**, 2033–2038.

Openshaw, S. and Taylor, P.G. (1979). A million or so correlation coefficients: three experiments on the modifiable areal unit problem. in: N. Wrigley (ed.): *Statistical Application in the Spatial Sciences*, Pion London, pp. 127–144.

Petrucci, A. and Salvati, N. (2006). Small area estimation for spatial correlation in watershed erosion assessment, *Journal of Agricultural, Biological, and Environmental Statistics*, **1**, 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.

Pratesi, M. and Salvati, N.(2009). Small area estimation in the presence of correlated random area effects, *Journal of Official Statistics*, **25**, 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**, 39–49

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

Salazar, L., Kogan, F., Roytman, L. (2008). Using vegetation health indices and partial least squares method for estimation of corn yield, *International Journal of Remote Sensing*, **29:1**, 175–189.

Saltelli, A., Ratto, M., Tarantola, S., Campolongo, F. (2012). Update 1 of Sensitivity Analysis for Chemical Models, *Chemical Reviews*, **105:7**, 2811-2828.

Shi, Z., Ruecker, G. R., Mueller, M., Conrad, C., Ibragimov, N., Lamers, J. P. A., Martius C., Strunz, G., Dech S., Vlek, P.L.G. (2007). Modeling of cotton yields in the Amu Darya river floodplains of Uzbekistan integrating multitemporal remote sensing and minimum field data. *Agronomy Journal*, **99:5**, 1317–1326.

Veregin, H. (1994). Integration of Simulation Modeling and Error Propagation for the Buffer Operation in GIS , *Photogrammetric Engineering and Remote Sensing*, **60**, 427–435.

Verbeiren, S., Eerens, H., Piccard, I., Bauwens, I., Van Orshoven, J. (2008). Subpixel classification of SPOT-VEGETATION time series for the assessment of regional crop areas in Belgium, *International Journal of Applied Earth Observation and Geoinformation*, **10:4**, 486–497.

Wall, L., Larocque, D., Léger, P. (2008). The early explanatory power of NDVI in crop yield modelling. *International Journal of Remote Sensing*, **29:8**, 2211–2225.

Woodcock, C. E., Gopal, S. (2000). Fuzzy set theory and thematic maps: Accuracy assessment and area estimation. *International Journal of Geographical Information Science*, **14:2**, 153–172.