

Geo-referenced information for agricultural statistics

L'Informazione Geo-referenziata per le Statistiche Agricole

Elisabetta Carfagna, Simone Maffei and Andrea Carfagna¹

Abstract In this paper, methods for evaluating the impact of interpolation, aggregation and disaggregation of geo-referenced information on models for producing agricultural statistics are discussed. The focus is on small area estimators and yield forecasting models.

Abstract In questo articolo, si discutono metodi per la valutazione dell'impatto di interpolazione, aggregazione e disaggregazione di informazioni geo-referenziate sui modelli per la produzione di statistiche agricole. Si pone l'attenzione su stimatori per piccole aree e modelli di previsione delle rese.

Key words: geo-referenced data, GIS operations, change of support, agricultural statistics

1 Introduction

In several countries, the quality of agricultural statistics is low; in other countries, accurate statistics are produced for main agricultural variables at country level, but the accuracy of estimates for smaller geographic domains, like lower level administrative units, is very poor.

Particularly for developing countries, allocating high shares of public resources to data collection for producing agricultural statistics is a difficult choice. Thus, there

¹

Elisabetta Carfagna, University of Bologna; email: elisabetta.carfagna@unibo.it

Simone Maffei, TERRA NOVA GIS; email: simone.maffei@gmail.com

Andrea Carfagna, Independent consultant; email: andreacarfagna@virgilio.it

is strong need to use efficient methods and some kinds of low cost information for improving the accuracy of estimates.

In some cases, administrative data play an important role, although some warnings should be kept in mind, see for example Carfagna and Carfagna (2010, 2013) and Carfagna *et al.* (2013). In other cases, some questions concerning agriculture are added to the population questionnaire, in order to create the list of rural households to be used for agricultural annual surveys (Keita and Gennari 2013). This approach allows saving the funds generally devoted to agricultural censuses, although the risk of under and over coverage can be relevant and very few auxiliary variables can be used for designing efficient annual sample surveys.

Due to the technological development, in the last decades, different kinds of geographic information have become easily accessible at decreasing prices and have started to be used for frame construction, sample design, estimation and yield forecasting. The role of Geographic Information Systems (GISs), Global Positioning Systems (GPSs) and remote sensing for creating different kinds of master sampling frame is analysed in Carfagna (2013).

A wide literature has been developed on the use of geo-referenced data for sample designs from various viewpoint, like optimal stratification, spatial sampling, and multiple probability proportional to size; Benedetti (2014) has made a very recent review and some methodological proposals on these topics.

In this paper, we focus on some statistical issues related to the use of geo-referenced information for agricultural statistics, namely the impact of interpolation, aggregation and disaggregation on the estimation of agricultural parameters and on yield forecasting models.

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 and 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. However, census enumeration areas and its concomitant data set seldom correspond to these geographic areas; consequently, data have to be interpolated, disaggregated and aggregated.

The possible support of geo-referenced data are usually points, lines, areas or surfaces. When the support of the spatial process of interest is different from the one of the observed data, a change of support problem arise. Multiple frame surveys for agricultural statistics are a typical case in which different kinds of data referring to the same phenomenon are taken into consideration. Generally, an area sample is combined with a sample selected from a list of farms, in order to improve the efficiency of estimates and reduce 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 forms 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.

The Modifiable Areal Unit Problem is a specification of the change of support problem and presents two facets (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.

Qi and Wu (1996) noted that Moran Coefficient, 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.

3 The impact of change of support on small area estimation

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 (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).

The sensitivity analysis performed relies on a model based simulation study designed for comparing the performances of small area estimators (Chandra *et al.*, 2012). 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 $x_{ij} \sim \text{Chi}^2(20)$, $j=1, \dots, N_i$, $i=1, \dots, A$, with the random area effects a_i generated as $N(0, 23.52)$; $\mathbf{v} = (v_i) = (\mathbf{I} - \rho\mathbf{W})^{-1}\mathbf{a}$; $\mathbf{W}$ 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 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. The small area sizes N_i^A of the aggregated population vary between 89 and 108. 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).

4 The impact of change of support on yield forecasting

Several kinds of yield forecasting models have been developed in the last years. 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 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 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 produce a propagated impact on the output of the statistical model.

We can identify four branches related to this issue:

- methodologies to measure the uncertainty induced by some input in the final results;
- operative indications at design stage of the procedure to reduce the impact of those disturbs on the final result;
- methods to correct the disturb caused by some specific processing procedure on the input data;

- 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 try to simulate/highlight the problems to measure the impact and minimize the effects. Veregin (1994) states that simulation modeling is an attractive alternative when little is known about error propagation mechanisms. In such cases, simulation modeling 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.

References

1. Arbia, G. & Petrarca, F.: Effects of Scale in Spatial Interaction Models. Journal of Geographical Systems (2013)
2. Bendetti R.: Developing more efficient and accurate methods for using remote sensing, Second meeting of the Scientific Advisory Committee of the Global strategy to Improve Agricultural and Rural Statistics, FAO Headquarters, 29 - 30 January (2014)

- 6 Carfagna, Maffei, Carfagna
3. Cancellieri M.C.: Carfagna E., Narciso G., Ragni P.: Aspects of Sensitivity Analysis of a Spectro-Agro-Meteorological Yield Forecasting Model, EARSeL (European Association of Remote Sensing Laboratories) Advances in Remote Sensing, vol. 2, n. 2 - VI, , pp. 124-132 (1993)
4. Carfagna E.: Multiple Frame Sample Surveys: Advantages, Disadvantages and Requirements, in International Statistical Institute, Proceedings, Invited papers, International Association of Survey Statisticians (IASS) Topics, Seoul August 22-29, pp. 253-270 (2001)
5. Carfagna E.: Using satellite imagery and geo-referencing technology for building a master sampling frame, the 59th World Statistical Congress, Invited Paper Session (IPS110) Developing a master sampling frame for integrated survey, Hong Kong, 25-30 August (2013) <http://www.statistics.gov.hk/wsc/IPS110-P1-S.pdf>
6. Carfagna, E., Carfagna, A.: 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 (2010)
7. Carfagna, E., Carfagna, A.: Appropriate Survey Methods for Different Country Profiles - Key Challenges, Gaps and Remaining Methodological Issues, ICAS VI, technical session IDCB 7 - Estimates, forecasts, expert opinion and assessment - their role in the official statistics, Rio de Janeiro, 23-25 October (2013) <http://eventos.ibge.gov.br/en/icas2013/programa-da-conferencia/programa-detalhado>
8. Carfagna, E. , Pratesi M., Carfagna, A.: Methodological developments for improving the reliability and cost-effectiveness of agricultural statistics in developing countries, the 59th World Statistical Congress, Special Topic Session (STS043) Using geospatial information in area sampling and estimation for agricultural and environmental surveys, Hong Kong, 25-30 August (2013) <http://www.statistics.gov.hk/wsc/STS043-P1-S.pdf>
9. Chambers, R. and Tzavidis, N.: M-quantile models for small area estimation. Biometrika, 93, pp. 255-268 (2006)
10. Chandra, H. and Chambers, R.L.: Comparing Eblup And C-Eblup For Small Area Estimation. Statistics In Transition, 7, pp. 637-648 (2005)
11. Chandra, H., Salvati, N. and Chambers, R.: Small Area Estimation For Spatially Correlated Populations - A Comparison of Direct And Indirect Model - Based Methods. Statistics In Transition, 8, pp. 331-350 2007.
12. Chandra, H., Salvati, N., Chambers, R. and Tzavidis, N.: Small area estimation under spatial nonstationarity. Computational Statistics and Data Analysis, 56, pp. 2875-2888 (2012)
13. Crosetto, M. & Tarantola, S.: Uncertainty and sensitivity analysis: tools for GIS-based model implementation. International Journal of Geographical Information Science (2001)
14. Keita N., Gennari P.: Building a Master Sampling Frame by Linking the Population and Housing Census with the Agricultural Census, the 59th World Statistical Congress, Special Topic Session (STS063) "Role of population and housing and agricultural censuses in the national statistical systems", Hong Kong, 25-30 August (2013) <http://2013.isiproceedings.org/Files/STS063-P1-S.pdf>
15. Petrucci, A. and Salvati, N.: Small area estimation for spatial correlation in watershed erosion assessment, Journal of Agricultural, Biological, and Environmental Statistics, 11, pp. 169-182 (2006)
16. Pratesi M., Petrucci A.: 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)
17. Pratesi, M. and Salvati, N.: Small area estimation in the presence of correlated random area effects, Journal of Official Statistics, 25, pp. 37-53 (2009)
18. Qi, Y. and Wu, J. (1996). 6. Effects of changing spatial resolution on the results of landscape pattern analysis using spatial autocorrelation indices. Landscape Ecol., 11, pp. 39-49 (1996)
19. Rao, J.N.K.: Small area estimation, John Wiley & Sons, New York (2003)
20. Saltelli, A., Ratto, M., Tarantola, S. & Campolongo, F.: Update 1 of Sensitivity Analysis for Chemical Models. Chemical Reviews (2012)
21. Veregin, H.: Integration of Simulation Modeling and Error Propagation for the Buffer Operation in GIS. Photogrammetric Engineering and Remote Sensing (1994)