

Multiple imputation in land cover EBLUP estimates

Roberto Benedetti, Danila Filipponi, Federica Piersimoni

Abstract AGRIT is an Italian point frame sample survey realized to produce land cover estimates over small areas (districts/provinces). To improve the precision of the estimates basic area level models have been used, where the small area direct estimator has been related to remote sensing satellite data. Spatial autocorrelation amongst neighbouring areas has been considered. Outliers and missing data are often present in the satellite information, mainly due to cloudy weather that does not allow an identification or a correct recognition of the cultures from the acquired digital images. The problem of outliers has been addressed by using a multiple imputation (MI) approach. Markov Chain Monte Carlo method has been used to generate repeated imputation of the missing covariate. The use of MI to fill in the missing information allow valid statistical inferences, even if determine a loss of the precision in order to reflect the uncertainty about the true value to impute.

Key words: mixed model, multiple imputation, small area, spatial sampling

1 Introduction

Land cover information receive a growing importance in order to implement and evaluate environmental policy and therefore, an high precision is usually expected. AGRIT is an Italian point frame sample survey realized to produce area estimates of the main crops and designed in a way to produce estimates with a prefixed sample error for the different crops. In the project are available remote sensing satellite data

Roberto Benedetti
University of Chieti-Pescara, Viale Pindaro 42, 65127 Pescara, Italy, e-mail: benedett@unich.it

Danila Filipponi
Italian Statistical Institute, Via Ravá 150, 00142 Roma, Italy e-mail: danila.filipponi@istat.it

Federica Piersimoni
Italian Statistical Institute, Via Ravá 150, 00142 Roma, Italy e-mail: federica.piersimoni@istat.it

covering the national agricultural area for the spring and summer periods.

When auxiliary information are available, the design-based regression estimator (Cochran, 1977) is a classical technique used to improve the precision of a sample estimation. This technique has been widely applied to improve the efficiency of crop area estimations since the early times of availability of satellite images data (Allen, 1990; Flores and Martnez, 2000). However, the regression estimator is not sufficiently precise to produce land cover estimate over small area. To increase the precision of the small area estimates a model-based approach can be used (Rao, 2003) defining statistical models to "borrow strength" from related small areas in order to increase the effective sample size. A good auxiliary information, linked to variables of interest, is very important for the success of any model-based methods. The location accuracy between the ground survey and satellite images and the difficulties to improve this accuracy through geometrical correction have been considered one of the main problem in relating remote sensing satellite data to crop areas, mainly in point frame sample surveys where the sampled point represents a very small portion of the territory. Moreover, outliers and missing data are often present in the satellite information; they are mainly due to cloudy weather that does not allow an identification or a correct recognition of the cultures from the acquired digital images. Here, the first problem has been overtaken by using the basic area level model to improve the land cover estimate at a small area level (Rao, 2003). Thus, the small area direct estimator is related to area specific auxiliary variables, i.e. number of pixels classified in each crop type according to the satellite data in each small area. The missing data problem has been addressed by using a multiple imputation (MI) procedure (Rubin, 1996, 2004; Schafer, 1997, 1999).

The following section gives a background to the AGRIT survey, describes the sampling design and the used direct estimator. The section 3 gives some details of the area specific models used. Spatial autocorrelation amongst the small area units has been also considered for improving the small area estimators. In section 4 is addressed the problem of missing values in the auxiliary variable. Finally, the concluding sections summarize and discuss some results.

2 The AGRIT Survey

The AGRIT is a survey projected to obtain areas and yields estimates of the main crops according to the LUCAS nomenclature of land cover valid at the European level. The survey covers the entire area of the Italian territory, a surface of about $301,300 \text{ km}^2$ and a high precision for the main categories of the nomenclature is expected. The adopted sampling design is a stratified two-phase sampling. In the first phase the reference frame is defined. The Italian territory is partitioned into N quadrants of 1000 meters size. In each quadrant n' units are selected using an aligned spatial systematic sample: one point is randomly selected within each quadrant and the remaining $n' - 1$ are placed in the same position within the corresponding quadrant. Using aerial photos each point is then classified according to a land use hierar-

chical nomenclature. The sampling frame obtained in the first phase is then divided into 103 provincial domains and stratified, using the nomenclature of land use. In each Italian province is then selected a stratified sample of about 150,000 points.

In order to define the problem formally, let us denote with A the area of the entire territory, A_d the domain d ($d = 1, \dots, D$) in a partition of the territory and c ($c = 1, \dots, C$) the crops for which are provided the estimate of land use. The parameters of interest are $\bar{Y}_A^c$, and $Y_A^c = A\bar{Y}_A^c$ that respectively denote the percentage and the total of land cover for the crop c on the entire territory, and $\bar{Y}_d^c$, and $Y_d^c = A_d\bar{Y}_d^c$ that denote the percentage and the total of land cover for the crop c in the domain d . The sample design of the first phase is assumed to be a simple random sampling, since it behaves like an aligned spatial systematic sampling, while the second phase sample is a stratified sampling. Thus, in the first phase sampling, n' represents the total number of sampled point, n'_h the number of sampled point in the strata h ($h = 1, \dots, H$) and $w_h = n'_h/n'$ is the proportion of sampled point belonging to the strata h , while in the second phase sampling, n represents the total number of point and n_h the number of sampled point in the strata h . Finally, y_{ih} represents the proportion of land cover for the crop c observed on the point i in the strata h .

In accordance with the formula of the two stage sampling when a simple random sampling is carried out in the first stage and stratified sampling in the second stage, the Horwitz-Thompson (HT) estimator of $\bar{Y}_A^c$ is given by:

$$\bar{y}_A^c = \sum_{h=1}^H (w_h \sum_{i=1}^{n_h} \frac{y_{ih}^c}{n_h}) = \sum_{h=1}^H w_h \bar{y}_h^c \quad (1)$$

while the unbiased estimator of the corresponding variance is

$$v(\bar{y}_A^c) = \sum_{h=1}^H \frac{w_h^2 s_h^2}{n_h} + \frac{1}{n'} \sum_{h=1}^H w_h (\bar{y}_h^c - \bar{y}_A^c)^2 \quad (2)$$

where $s_h^2 = \sum_{i=1}^{n_h} \frac{(y_{ih}^c - \bar{y}_h^c)^2}{n_h - 1}$.

Once the percentage of land cover for the crop c is estimated, the estimation of the total $Y_A^c = A\bar{Y}_A^c$ is straightforward and is given by $\hat{y}_A^c = A\bar{y}_A^c$, while the unbiased estimator of the corresponding variance is $v(\hat{y}_A^c) = A^2 v(\bar{y}_A^c)$ and $CV(\hat{y}_A^c) = \sqrt{v(\hat{y}_A^c)}/\hat{y}_A^c$. If the parameters of interest are $\bar{Y}_d^c$, and $Y_d^c = A_d\bar{Y}_d^c$, i.e. the percentage and the total of land cover for the crop c in the domain d , the previous theory is applicable by defining a domain membership variable $y_{d,ih}^c$, where $y_{d,ih}^c = y_{ih}^c$ if the unit $i \in A_d$ and $y_{d,ih}^c = 0$ otherwise.

3 EBLUP estimator of land cover in small area

Remote sensing satellite data provide a complete spectral resolution of an area that can be use to classify the area by crop types. The availability of such information

does not eliminate the need of ground data, since satellite data do not always have the accuracy required to estimate the different cultures, but can be used as auxiliary information to improve the precision of the direct estimates. The model suggested to estimate main crops at a provinces level is the basic area level model. Let's denote with $z_d(1 \times p) = (z_{d1}, z_{d2}, \dots, z_{dp})$ the vector containing the number of pixels classified in each crop type according to the satellite data in the small area $d, d = 1, \dots, D$, where $D \leq 103$ indicates the number of Italian provinces. It seems natural to assume that the area covered by a culture Y_d^c , in the small area d is in some way linked to z_d . A model based estimate of $Y^c(D \times 1)$, eventually transformed as $\theta^c = g(Y^c)$, is the Empirical Best Linear Unbiased Predictor (EBLUP) based on the linear mixed model:

$$\hat{\theta}^c = \beta^c Z + v + e \quad (3)$$

where $\hat{\theta}^c = g(\hat{Y}^c)$ is a D -component vector of the direct survey estimators of θ^c , $\beta(D \times 1)$ is the vector of regression parameters, $Z(D \times p)$ is the covariates matrix, $e(D \times 1)$ are the sampling errors assumed to be independent across areas with mean 0 and variance matrix equal to $R = \text{diag}(\psi_1, \psi_2, \dots, \psi_m)$ where ψ_i 's are the known sampling variances corresponding to the d^{th} small area and $v(D \times 1)$ are the model errors assumed to be independent and identically distributed with mean 0 and variance matrix equal to $G = \sigma_v^2 I$. Independence between v and e is also assumed. The EBLUP estimator of θ_d^c and estimation of the mean square error (MSE) of the EBLUP are described by Rao (2003).

The small area characteristics usually have spatial dependence. The spatial autocorrelation amongst neighbouring areas can be introduced to improve the small area estimation. An area level model with conditional spatial dependence among random effects (Cressie, 1991) can be considered an extension of the model 3 where all the parameters have the same meaning as previously explained, and the model errors v are assumed with mean 0 and variance-covariance matrix $G = \sigma_v^2(I - \rho W)^{-1}$. $W(D \times D)$ is a known proximity matrix that indicates the interaction between any pair of small areas. The elements of $W \equiv [W_{ij}]$ with $W_{ij} = 0 \forall i$ are binary values, $W_{ij} = 1$ if the j^{th} small area is physically contiguous to i^{th} small area and $W_{ij} = 0$ otherwise. The constant ρ is called spatial autoregressive coefficient and it is a measure of the overall level of spatial autocorrelation. The spatial model "borrows strength" from related small area by using two parameters: the regression parameters and the spatial autoregressive coefficient. By setting $\rho = 0$ we obtain the model 3. Considering the spatial model as a special case of generalized mixed models the BLUP of θ^c and the MSE of the BLUP can be easily obtained as:

$$\theta^{*c}(\rho, \sigma_v^2) = Z\hat{\beta}(\rho, \sigma_v^2) + \lambda(\rho, \sigma_v^2)[\hat{\theta} - Z\hat{\beta}(\rho, \sigma_v^2)] \quad (4)$$

$$MSE[\theta^{*c}(\rho, \sigma_v^2)] = g_1(\rho, \sigma_v^2) + g_2(\rho, \sigma_v^2) \quad (5)$$

where

$$\begin{aligned} g_1(\rho, \sigma_v^2) &= \lambda(\rho, \sigma_v^2)R, \\ g_2(\rho, \sigma_v^2) &= RV^{-1}(\rho, \sigma_v^2)Z(Z^T V^{-1}(\rho, \sigma_v^2)Z)^{-1}Z^T V^{-1}(\rho, \sigma_v^2)R \end{aligned}$$

and

$$\hat{\beta}(\rho, \sigma_v^2) = [Z^T V^{-1}(\rho, \sigma_v^2) Z]^{-1} Z^T V^{-1}(\rho, \sigma_v^2) \hat{\theta}, V(\rho, \sigma_v^2) = \sigma_v^2 A^{-1}(\rho, \sigma_v^2) + R, \\ \lambda(\rho, \sigma_v^2) = \sigma_v^2 A^{-1}(\rho, \sigma_v^2) V^{-1}(\rho, \sigma_v^2), A(\rho, \sigma_v^2) = (I - \rho W).$$

The BLUP of θ^c and the MSE of the BLUP are both functions of the parameter vector (ρ, σ_v^2) which is unknown and need to be estimated. Assuming normality, the parameters (ρ, σ_v^2) can be estimated both by maximum likelihood (ML) or restricted maximum likelihood (REML) procedures. Therefore, the empirical best linear unbiased predictor (EBLUP), $\theta^{*c}(\rho, \sigma_v^2)$, and the naive estimator of the MSE are obtained by replacing the parameter vector (ρ, σ_v^2) , with its estimator $(\hat{\rho}, \hat{\sigma}_v^2)$. The naive estimator of the MSE of the EBLUP, underestimates the MSE, since it doesn't take into account the additional variability due to the estimation of the parameters. If $(\hat{\rho}, \hat{\sigma}_v^2)$ is a REML estimator the an approximation of the $MSE[\theta^{*c}(\hat{\rho}, \hat{\sigma}_v^2)]$ is given by Prasad and Rao (1990).

4 Multiple imputation for missing data in satellite images

Let us denote with $\hat{Y}(D \times C)$ the estimations matrix of the areas covered by crop types and with $Z(D \times C)$ the matrix containing the number of pixels classified by crop types according to the satellite data in each small area. The matrix $\hat{Y}$ is considered fully observed while missing data are often present in satellite images. Outliers and missing data in satellite information are mainly due to cloudy weather that does not allow an identification or a correct recognition of the cultures from the acquired digital images.

An approach to incomplete data problem is the multiple imputation (MI). Rubin (2004) introduce the idea of MI and he described the MI approach as a three steps process. First, instead of imputing a single value for each missing data, $m > 1$ likely values are drawn from the predictive distribution $P(Y_{mis}/Y_{obs})$, where Y_{obs} is the observed part of Y and Y_{mis} is the missing one, in order to reflect the uncertainty about the true value to impute. Second, m possible alternative version of the complete data are produced substituting the i^{th} simulated value, $i = 1, \dots, m$, in the corresponding missing data. The m imputed data sets are analyzed using standard procedures for complete data. Finally, the results are combined in a way to produce statistical inferences that properly reflect the uncertainty due to missing values; a procedure that combines the results of the analysis and generates valid statistical inferences is presented in Rubin (2004). Reviews of multiple imputation have been published by Rubin (1996), Schafer (1997, 1999).

The choice of the imputation model has been carried on considering that the imputation model should preserve the relationships among variables measured on a subject, that is the joint distribution in the imputed values. For this reason, it not needed to make distinctions between dependent and independent variables, but the variables can be treated as a multivariate response. A model that preserves the multivariate distributions of the variable will preserve also the relations of any variables on the others. Here, it is seems reasonable to suppose that the area covered by a culture c , $c = 1, \dots, C$, $\hat{Y}^c(D \times 1)$ is somehow related to $Z^c(D \times 1)$ and therefore to impute

missing values in the explanatory variables Z^c , it is assumed that the random vectors $(\hat{Y}^c, Z^c)$ have a multivariate normal distribution, that is $(\hat{Y}^c, Z^c) \sim N(\mu^c, \Sigma^c)$, where (μ^c, Σ^c) are unknown parameters. Considering that it is not available information about (μ^c, Σ^c) an improper prior is applied. MCMC method is then used to obtain m independent drawn from the predictive distribution.

5 Results

The aim of the AGRIT survey is to produce area estimates of the main crops at a national level and for the geographical domains: regions and provinces. The direct estimates of the area covered by the crop type c , $c = 1, \dots, C$, (LCT_c), their estimated standard errors (SE) and the coefficients of variation (CV) have been obtained as described in section 2 for the 103 Italian provinces. The available explanatory variable is given by the number of pixels classified in the crop type c according to the satellite data in the small area d , $d = 1, \dots, D$, with $D \leq 103$.

When the explanatory variable Z^c has missing values we generate $M = 10$ complete data set $(\hat{y}^c, Z_1^c), \dots, (\hat{y}^c, Z_M^c)$ by imputing one set of the plausible values according to the procedure describe in section 4. The choice of $M = 10$ is justify by Rubin (2004). Let define with Model 1 the direct estimator. Two different mixed models

Table 1: Estimates of Parameters for Small Area Models of LCT

Crops	R^2	% missing	β_0	β_1	σ_v^2	β_0	β_1	ρ	σ_v^2
			Model -2			Model -3			
Durum wheat	0.94	26.1	113.6	0.09	1.064.6	98.2	0.09	13.5	1.069.9
Soft wheat	0.95	6.5	118.2	0.10	193.4	182.8	0.09	70.4	240.5
Barley	0.93	11.4	280.5	0.09	21.7	291.8	0.09	30.4	22.7
Maize	0.97	0.5	87.7	0.08	358.5	87.7	0.08	0.0	358.5
Sunflower	0.97	7.1	50.4	0.09	8.1	50.4	0.09	0.2	8.1
Soya	0.96	16.0	-139.3	0.09	29.6	-170.9	0.09	59.4	33.5
Sugar beet	0.94	13.0	23.2	0.09	5.6	23.2	0.09	0.9	5.6
Tomatoes	1.00	20.7	-3.8	0.08	0.9	-3.8	0.08	1.0	0.9

have been used to improve the estimates of LCT_c : an area level model (Model 2) and an area level model with correlation (Model 3). In the model 3 has been considered a spatially continuous covariance function $G = \sigma^2[\exp(-W/\theta)]$, where the elements of W are the distance between the small area i^{th} and the small area j^{th} . This choice can be motivated by a greater stability of the correlation parameter θ , with respect to the spatial autoregressive coefficient ρ , for the M imputed datasets. Table 1 presents the estimate parameters for the simple mixed effect model and for the mixed effect model with spatial correlation. The R^2 coefficients between the LCT estimates and

the auxiliary variables, and the percentage of missing data for each crop type are also given in the table. A summary of the estimated standard errors (based on the D

Table 2: Average SE of the Estimates under the Models 1-3

Crops	Model -1	Model -2	Model -3
Average SE			
Durum wheat	848,37	799,04	799,04
Soft wheat	561,24	481,14	466,33
Barley	529,22	341,10	340,43
Maize	782,02	656,14	656,14
Sunflower	400,07	222,06	222,48
Soya	536,34	367,78	353,50
Sugar beet	370,68	204,67	204,67
Tomatoes	258,12	140,47	138,28

small areas) under each models and for all crops are shown in table 2.

A measure of the gain in precision due to the small area models is given by the relative efficiency defined as:

$$RE_d = \left(\frac{MSE(\theta_d^{Mod_a})}{MSE(\theta_d^{Mod_b})} \right) \times 100 \quad (6)$$

that is the MSE of the estimator obtained using a model against the MSE of the estimator under a different model. Figures 1 shows the distribution of the provinces by

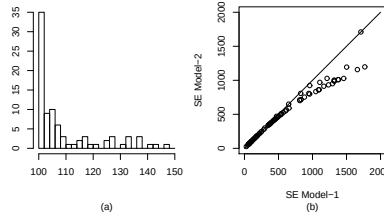


Fig. 1: (a) Number of the provinces by class of RE Model-1 vs Model-2; (b) SE Model-1 vs SE Model-2

RE for the soft wheat, where are compared model 1 vs model 2. The scatter plots of the SE for the different comparisons of models are also given. Figures 1 indicate a large gain of efficiency by using remote sensing data in the estimator (RE of model 1 vs model 2 and RE of model 1 vs model 3), whereas there is a small improvement by introducing a spatial effects. Remarkable is the distribution of the provinces by

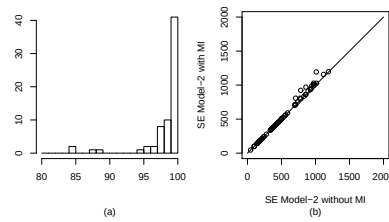


Fig. 2: (a) Number of the provinces by class of RE:Model-2 without MI vs Model-2 with MI; (b) SE Model-2 without MI vs SE Model-2 with MI

RE for the soft wheat, where are compared the model 2 with a single imputation and the model 2 with $M = 10$ different imputation to fill the missing information (Figure 2). The loss of efficiency underlined in the figure 2 reflects the uncertainty about the true value to impute (Rubin, 2004) as discussed in section 4.

The results of this study confirm the superiority of the small area models in comparison to the direct surveys estimator, that is the direct surveys estimates can be definitely improved by defining basic area models. The introduction of a spatial autocorrelations amongst the small area can be also used for increasing the efficiency, but the model must be apply after an evaluation of the significance of the spatial correlation among the small areas. Moreover, the choice of an appropriate proximity matrix W , that indicates the interaction between any pair of small areas, can be considered to strengthen the small area estimates.

References

1. Allen, J. D.: A look at the Remote Sensing Applications Program of the National Agricultural Statistics Service. *Journal of Official Statistics* **6**, 393–409 (1990)
2. Cochran, W.G.: *Sampling Techniques*. Wiley (1977)
3. Cressie, N.: Small-Area Prediction of Undercount Using the General Linear Model. *Proceedings of Statistic Symposium 90: Measurement and Improvement of Data Quality*, Ottawa: Statistics Canada. 93–105 (1991)
4. Flores L.A. and Martinez L.I.: Land Cover Estimation in Small Areas Using Ground Survey and Remote Sensing. *Remote Sensing of Environment* **74**, 240–248 (2000)
5. Meng, X.L.: Multiple-imputation inferences with uncongenial sources of input (with discussion). *Statistical Science* **10**, 538–573 (1995)
6. Prasad, N. and Rao, J.N.K.: The Estimation of the Mean Squared Error of Small-Area Estimators. *Journal of the American Statistical Association* **85**, 163–171 (1990)
7. Rao, J.N.K.: *Small Area Estimation*. Wiley (2003)
8. Rubin, D. B.: Inference and missing data. *Biometrika* **63**, 581–592 (1976)
9. Rubin, D. B.: Multiple imputation after 18+ years. *Journal of the American Statistical Association* **91**, 473–489 (1996)
10. Rubin, D. B.: *Multiple Imputation for Nonresponse in Surveys*. Wiley (2004)
11. Schafer, J. L.: *Analysis of incomplete multivariate data*. Chapman Hall (1997)
12. Schafer, J. L.: Multiple imputation: A primer. *Statistical Methods in Medical Research* **8**, 3–15 (1999)