EBLUP ESTIMATES BY USING REMOTELY SENSED DATA

R. Benedetti, University "G. d'Annunzio" Chieti-Pescara, Italy
M. Ciavattella, University "La Sapienza"Rome, Italy
D. Filipponi, Istat, Rome, Italy

Commission VIII, WG VIII/10

KEYWORDS: Small area models, Spatial autocorrelation, EBLUP, Multiple imputation.

ABSTRACT:

The knowledge of auxiliary variables for entire population is a necessary task in many agricultural applications; for example a method widely applied to improve the efficiency of crop area estimations is the regression estimator (Cochran, 1977), that exploit the correlation between classified satellite images (auxiliary information) and ground surveys' data. The regression analysis offer a lot of advantages, even if it's highly affected by two main problems: the presence of outliers which usually implies instability in the regression parameters estimates and the so called scale problem (MAUP) that is the use of aggregated data increase artificially the amount of correlation between variables (Openshaw e Taylor, 1979). According to the aforesaid problems and considering that generally the final goal is to obtain crop area estimations at small area level (usually districts/provinces), the idea is to improve the direct estimator using small area models that relate the small area surface to area specific auxiliary variables, instead of using regression analysis on the sampled point (which generally represent a very small portion of the territory). Generally the linking models based on random area-effects that account for between area variations other than the variation explained by auxiliary variables are called small area models and the indirect estimator based on small area models will be called "model-based estimators".
The aim of this work is to produce area estimates of the main crops at a provincial, regional and national level, using as sampling design a stratified two-phase sampling. We improve the direct Horwitz-Thompson estimator whit the two stage model of Fay and Herriot (1979) and the EBLUP (Empirical Best Linear Unbiased Predictors) estimators (Rao, 2003).
A spatial autocorrelation amongst the small area units has been also considered to improve the small area efficiency. Spatial models are a special case of a mixed linear model and therefore EBLUP estimator con be easily obtained.
Moreover, we have valuated the importance of good auxiliary data for the success of model-based methods. In fact, cloudy whether is usually the main source of missing data in satellite images. All the missing data and outliers (considered as missing) have been imputed, whit a technique called Multiple Imputation (MI - Rubin, 1987). In the MI instead of imputing a single value for each missing value, m value are drawn from the predictive distribution $P(Y_{miss} / Y_{obs})$ and then complete data analysis is repeated a given number of times, say h.
We confirmed that the combination of small area model whit the technique of spatial autocorrelation and Multiple Imputation improve the crop area estimations compared with direct estimations. To illustrate the proposed approach, we applied the spatial EBLUP estimation and the MI method for missing covariates to the AGRIT 2005 data.

1. INTRODUCTION

The regression estimator is a classical technique used to improve the precision of a sample estimation when an auxiliary variable is available for entire the population. This technique has been widely applied to improve the efficiency of crop area estimations when classified satellite images are available as auxiliary information. The collocation inaccuracy between the ground survey and satellite images, and the difficulties to improve the positioning through geometrical correction have been considered the main problem of this technique. However, the lack of gain in efficiency when remote sensing information are linked to the ground observations is mainly due to the statistical methodology used. In particular, the regression analysis is highly affected (i) by the presence of outliers, and (ii) by the scale problem (MAUP), that is the use of aggregated data increase the amount of correlation between variables. According to (i) and (ii) and considering that usually the final goal is to obtain crop area estimations at small area level (usually districts/provinces), the idea is to improve the direct estimator using small area models that relate the small area surface to area specific auxiliary variables, instead of using regression analysis on the sampled point (which generally represent a very small portion of the territory). The scope of the present work is to produce conjectural area estimates of the main crops at a provincial, regional and national level. The direct estimator is obtained using the Horwitz-Thompson estimator, where the sampling design is a stratified two-phase stratified sampling (i.e. the phase 1 is a stratification of the national frame using digital photo interpretation and the phase 2 is a field data collection on physical points sampled from the strata). In section 2 is described the area specific model used which is the two stage model of Fay Herriot (1979) and the EBLUP (Empirical Best Linear Unbiased Predictors) estimators. Spatial autocorrelation amongst the small area units has been also considered for improving the small area estimators. In section 3 is considered the problem of missing values and outliers in the auxiliary variable and a multiple imputation (Rubin 1976, 1987) has been used to fill-in the missing information. Finally, in the last section of the paper are discussed some results.

2. SMALL AREA ESTIMATION

Linking models based on random area-effects that account for between area variation other than the variation explained by auxiliary variables are called small area models. Indirect estimator based on small area models will be called "model-based estimators". Rao (2003) classifies the small area models in two types: (i) Basic area level models, that relates the small area direct estimators to area-specific covariates; (ii) basic units level models, that relate the unit values of a study variables to units specific covariates. Here, according to what has been

discussed earlier, a basic area level model is used. The two stage model of Fay and Herriot (1979) can be expressed as:

$$\theta_i = z_i^T \beta + b_i v_i , \qquad (1)$$

$$\hat{\theta}_i = \theta_i + \varepsilon_i , \; i = 1,...,m \qquad (2)$$

where θ_i is the characteristic under study in the small area i, related to a specific auxiliary data $z_i = (z_{1i},...,z_{pi})$, $\hat{\theta}_i$ is the direct survey estimator of θ_i, b_i 's are known positive constants , β is a $(p \times 1)$ vector of regression parameters; the v_i are the area random effect assumed to be independent and identically distributed with $E(v_i) = 0$, $Var(v_i) = \sigma_v^2 \quad i = 1,...,m$ (3) and the ε_i are the independent sampling errors with $E(\varepsilon_i / \theta_i) = \psi_i$, $i = 1,...,m$ (4), where ψ_i are the sampling variance usually assumed as known. The equation (1) is the design model while the equation (2) is the linking model; combining (1) e (2) we obtain the model

$$\theta_i = z_i^T \beta + b_i v_i + \varepsilon_i , \; i = 1,...,m \qquad (5)$$

which involves the design error v_i and the model error ε_i, assumed to be independent.

The model (5) is a special case of a linear mixed model; from the linear mixed model theory the best linear unbiased predictor (BLUP) estimator can be easily obtain to θ_i and the MSE of the BLUP estimator as measure of variability. Then, if the matrix of variance and covariance of v_i and ε_i are given by (3) and (4), the BLUP estimator of θ_i is a weighted average of the direct survey estimator $\hat{\theta}_i$ and the regression estimator $z_i^T \beta$, that is $\tilde{\theta}_i = \gamma_i \hat{\theta}_i + (1 - \gamma_i) z_i^T \tilde{\beta}$ where $\gamma_i = \sigma_v^2 b_i / (\psi_i + \sigma_v^2 b_i^2)$ The MSE of the BLUP estimator is given by: $MSE(\tilde{\theta}_i) = g_{1i}(\sigma_v^2) + g_{2i}(\sigma_v^2)$ where $g_{1i}(\sigma_v^2) = \gamma_i \psi_i$ and $g_{2i}(\sigma_v^2) = (1 - \gamma_i) z_i^T \left[\sum_i z_i z_i^T / (\psi_i + \sigma_v^2 b_i^2) \right]^{-1} z_i$. Comparing the leading term $\gamma_i \psi_i$ with ψ_i, the MSE of the direct estimator, it is clear that $\tilde{\theta}_i$ gains in efficiency when γ_i is small, that is when the variability of the model error is small relative to the total variability. The BLUP estimator and the relative MSE depend on the variance component σ_v^2 which is unknown. Various methods have been proposed to estimate the variance component in a linear mixed model. By replacing σ_v^2 with an asymptotically consistent estimator $\hat{\sigma}_v^2$, it is possible to obtain the empirical best linear unbiased predictor (EBLUP) and the MSE estimation. The model (5) assume iid small area effects v_i. In this paper, to improve the efficiency of small area estimators, a spatial autocorrelation amongst the small area v_i has been considered. The covariance matrix of the v_i, earlier set equal to $G = \sigma_v^2 I_m$, is now given by $G = \sigma_v^2 (I - \rho W)$. The elements of W have been defined as binary values that is $W_{ij} = 1$ if the small area i is physically continuous to the small area j and $W_{ij} = 0$ otherwise and the constant ρ is a measure of the level of spatial autocorrelation. Spatial models are a spatial case of a general linear model and therefore EBLUP estimator con be easily obtained.

3. IMPUTATION OF THE MISSING AUXILIARY VARIABLE

Availability of good auxiliary data are crucial for the success of any model-based methods and therefore attention have been given to their control. Missing data or outliers in satellite images are mainly due to cloudy whether and their detection have been carried out using graphical tools. All the missing data and outliers (considered as missing) have been imputed. In order to take into account, in the complete data standard errors, the error generated by the imputation, a multiple imputation (MI) has been carried on. In the MI instead of imputing a single value for each missing value, m value are drawn from the predictive distribution $P(Y_{miss} / Y_{obs})$ and then complete data analysis are repeated m times, once for each imputation. Important is the choice of the imputation model. If Y_1 is a fully observed variable, Y_2 given the observed Y_2 is high then better imputation can be obtained by using Y_1 and the observed Y_2 for the imputation; this means that an imputation model should preserve the relationships among variables measured on a subject and therefore a full parametric model for $Y = (Y_1, Y_2)$ to impute for non-response should be used.

Here, to impute missing values in the explanatory variables z_i, it is assumed that $y_i = (z_i, \hat{\theta}_i)$, $i = 1,...,m$ are independent realizations of a random vector $Y = (Z, \hat{\theta})$ with a multivariate normal distribution; that is $(z_i, \hat{\theta}_i)$, $i = 1,...,n / (\mu, \Sigma) \sim$ iid $N(\mu, \Sigma)$, where (μ, Σ) are the unknown parameters. Considering that are not available information about (μ, Σ) an improper prior is applied. MCMC method are then used to obtain m independent drawn from the predictive distribution $P(Y_{miss} / Y_{obs})$.

4. RESULTS

To apply the EBLUP estimators it is necessary to know the direct estimators and the relative sampling errors of the main crops at a province level. Moreover, σ_v^2 or (σ_v^2, ρ). if a spatial autocorrelation among the small areas is considered, need to be estimate. Here, σ_v^2 and (σ_v^2, ρ) have been estimate using REML; the EBUP estimators and RMSE have been calculated using the linear mixed models formulas. For all the crops both models (with/without spatial autocorrelation) have been fitted and the one with a larger variance reduction has been chosen.

Before fitting the linear mixed model, provinces with missing value or outliers (mainly due to cloudy whether) in the auxiliary variable, have been detected. All the outliers have been considered as missing and the unknown missing data have been replaced with m=100 simulate values.

Crop	A	B	R^2	Spatial Correlation	% missing	% missing outlier	% Variance reduction
Hard wheat	1934	0,099	0,9851	No	6,63	6,63	28,16
Wheat	407	0,097	0,93051	No	3,70	3,70	15,79
Barley	124	0,090	0,98269	No	30,81	30,81	34,47
Maize	439	0,083	0,98669	Yes	1,57	18,16	32,69
Sun-flawer	69	0,109	0,96738	Yes	16,56	37,79	30,83
Soya	-69	0,114	0,95965	Yes	0,32	17,43	38,99
Sugar beet	405	0,121	0,87175	No	7,92	16,81	18,37
Tomato	315	0,081	0,72850	No	40,15	60,64	14,97

Table 1: Estimates of Parameters for Small Area Estimation.

Some of the results have been summarized in the table 1. Here for each crop are shown the estimates of the parameters for small area estimation, the value of the correlation coefficient between the direct estimator and the auxiliary variable, the % of missing values and the reduction of the variance. The variance reduction in table 1 have been calculated as a national average, of course bigger variance reduction have been obtained for some province. However, the models for all the crops improve the direct survey. The spatial model further improves the estimates for three crop (maize, sunflower, soya).

References

Fay R.E., Herriott R.A. (1979) Estimates of income for small places: an application of James-Stein procedures to census data, *Journal of the American Statistical Association,* 78, 879-884.

Lahiri P.A., Rao J.N.K. (1995,) Robust estimation of mean squared error of small area estimators, *Journal of the American Statistical Association*, 82, 758-766.

Rao J.N.K., (2003) *Small Area Estimation*, Wiley, New York.

Rao J.N.K., Yu M. (1994) Small area estimation by combining time series and cross-sectional data, *Canadian Journal of Statistics*, 22, 511-528.

Rubin, D.B. (1987), *Multiple Imputation for Nonresponse in Survey*, Wiley, N.Y.

Schafer, J.L. (1997) *Analysis of Incomplete Multivariate Data*. Book number 72 in the Chapman & Hall series Monographs on Statistics and Applied Probability. London: Chapman & Hall.