The use of geographic information under the area-level approach to small area estimation

L'utilizzo dell'informazione spaziale nei modelli a livello di area nella stima per piccole aree

Stefano Marchetti, Caterina Giusti and Nicola Salvati

Abstract The Fay & Herriot (1979) area-level model, widely used in statistics to obtain reliable small area estimates, assumes that the area-level direct estimates are spatially uncorrelated, an hypothesis that can be unrealistic with agricultural, environmental, economic and epidemiological data. The aim of this paper is to compare, using some simulation studies, the classical Fay-Herriot model with three different extensions of it, that are able to take into account the spatial proximity effects between the small areas: the Semiparametric Fay-Herriot model proposed by Giusti et al. (2012), the Spatial Fay-Herriot model proposed by Salvati (2004), Singh et al. (2005) and Petrucci & Salvati (2006), and, finally, the Spatial Nonstationary Fay-Herriot model recently proposed by Chandra et al. (2014).

Abstract *Il modello proposto da Fay e Herriot (1979) é comunemente utilizzato per ottenere stime per domini non previsti quando i dati sono disponibili a livello di area. Tale modello ipotizza che gli effetti di area siano incorrelati, ipotesi che puó essere irrealistica nel caso di dati agricoli, ambientali, economici e epidemiologici. In questo lavoro proponiamo un confronto, attraverso opportuni studi di simulazione, tra il modello Fay-Herriot e tre sue estensioni che consentono di tenere in considerazione gli effetti di prossimitá tra le aree: il modello Semiparametric Fay-Herriot proposto da Giusti et al. (2012, il modello Spatial Fay-Herriot proposto da Salvati (2004), Singh et al. (2005) e Petrucci e Salvati (2006), e, infine, il modello Spatial Nonstationary Fay-Herriot recentemente proposto da Chandra et al. (2014).*

Key words: Fay-Herriot model, semiparametric models, nonstationary models, simulation studies

---

Stefano Marchetti
University of Pisa, Via Ridolfi, 10 56124 Pisa (PI), e-mail: stefano.marchetti@ec.unipi.it

Caterina Giusti
University of Pisa, Via Ridolfi, 10 56124 Pisa (PI), e-mail: caterina.giusti@ec.unipi.it

Nicola Salvati
University of Pisa, Via Ridolfi, 10 56124 Pisa (PI), e-mail: salvati@ec.unipi.it

1 Introduction

Often, sample surveys provide a cost-effective way to obtain direct estimates for population parameters of interest for planned domains or areas. However, on many occasions the interest is in estimating these parameters for unplanned domains/areas where the sample size can be small, resulting in unreliable direct estimates. These areas/domains are known in the literature as small areas (Rao, 2003). Model-based small area estimation techniques are commonly used to produce reliable estimates for these areas. Small area estimation methods can be based on unit-level or area-level models. Unit-level models are defined for individual survey data and includes small area effects. Area-level models allow to smooth out the variability for unreliable area-level direct estimates and they can be used when auxiliary variables are only available in aggregate form.

The Fay & Herriot (1979) model produces reliable small area estimates by combining small areas direct survey estimates with area-level auxiliary variables. The underlying assumption is that the direct survey estimators are linear function of the covariates. Let Y_i be the unobserved population parameter of interest for area i, with $i = 1, \ldots, m$ (for example the population mean of the variable Y), and y_i its unbiased survey direct estimator. Let $\mathbf{x}_i$ be a vector of known auxiliary variables for area i (usually coming from administrative and census records). The Fay-Herriot model can be expressed as:

$$y_i = \mathbf{x}_i^T \beta + u_i + e_i \tag{1}$$

where β is the $p \cdot 1$ vector of unknown parameters, u_i are independent and identically distributed regression errors, usually assumed to be normally distributed with $E(u_i) = 0$, $var(u_i) = \sigma_u^2$, and e_i represent the sampling errors, again usually assumed to be independent and identically normally distributed with $E(e_i|Y_i) = 0$, $var(e_i|Y_i) = \sigma_{ei}^2$. Defining by $\mathbf{y}$ the $m \cdot 1$ vector of the direct estimates y_i and by $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_q)$ the corresponding $m \times p$ matrix of the area level auxiliary variables, the Fay-Herriot model can also be expressed by

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{D}\mathbf{u} + \mathbf{e}. \tag{2}$$

Here $\mathbf{u}$ is $m \times 1$ vector of independent and identically distributed normal random variables with mean $\mathbf{0}$ and $m \times m$ variance matrix $\Sigma_u = \sigma_u^2 \mathbf{I}_m$, $\mathbf{D}$ is $m \times m$ matrix of known positive constants, $\mathbf{e}$ is the $m \times 1$ vector of normal independent sampling errors with mean $\mathbf{0}$ and known diagonal variance matrix $\Sigma_e = diag(\sigma_{e1}^2, \sigma_{e2}^2, \ldots, \sigma_{em}^2)$, and β is the $p \times 1$ vector of regression parameters. Thus, the Fay-Herriot model is a general linear mixed model with diagonal covariance structure $var(\mathbf{y}) = \mathbf{V} = \mathbf{D}\Sigma_u \mathbf{D}^T + \Sigma_e$. Estimating Σ_u by ML or REML and deriving the corresponding plug-in estimator $\hat{\mathbf{V}} = \mathbf{D}\hat{\Sigma}_u \mathbf{D}^T + \Sigma_e$, leads, under model (2), to the EBLUE (empirical best linear unbiased estimator) $\hat{\beta} = (\mathbf{X}^T \hat{\mathbf{V}}^{-1} \mathbf{X})^{-1} \hat{\mathbf{V}}^{-1} \mathbf{y}$ and to the EBLUP (empirical best linear unbiased predictor) $\hat{\mathbf{u}} = \hat{\Sigma}_u \mathbf{D}^T \hat{\mathbf{V}}^{-1} (\mathbf{y} - \mathbf{X}\hat{\beta})$. The corresponding EBLUP estimate of Y_i under model (1) is

$$\hat{Y}_i^{EBLUP} = \mathbf{x}_i^T \hat{\beta} + \mathbf{d}_i^T \hat{\Sigma}_u \mathbf{D}^T \hat{\mathbf{V}}^{-1} (\mathbf{y} - \mathbf{X}\hat{\beta}), \tag{3}$$

where $\mathbf{d}_i^T$ denotes the ith row of the $\mathbf{D}$ matrix.

The Fay-Herriot model assumes that the direct survey estimators are linear function of the covariates. When this assumption fails down, the Fay-Herriot model can lead to biased estimators of the small area parameters. A semiparametric specification of the Fay-Herriot model, which allows non linearities in the relationship between Y_i and the auxiliary variables $\mathbf{x}_i$, can be obtained by penalized-splines (Giusti et al., 2012). Bivariate smoothing under this approach can also be used when geographically referenced responses play a central role in the analysis and need to be converted to maps. The estimator proposed by Giusti et al. (2012) is

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{Z}\gamma + \mathbf{D}\mathbf{u} + \mathbf{e}, \tag{4}$$

where $\mathbf{X}$ is a matrix obtained stacking two matrixes $\mathbf{X}_a$ and $\mathbf{X}_b$, and $\beta = [\alpha, \eta]$. Here, $\mathbf{X}_a$ is the matrix of the auxiliary variables and α is the vector of the fixed effects, $\mathbf{X}_b$ is the matrix of the coordinates of the centroid of area i and η is the vector of the coefficients used in the polynomial part of a p-spline model of the type $m(\mathbf{x}_b; \eta, \gamma) = \eta_0 + \eta_1 \mathbf{x}_b^1 + \ldots + \eta_p \mathbf{x}_b^p + \sum_{k=1}^K \gamma_k (\mathbf{x}_b - \kappa_k)_+^p$. The $\sum_{k=1}^K \gamma_k (\mathbf{x}_b - \kappa_k)_+^p$ is expressed in (4) as $\mathbf{Z}\gamma$, where $\mathbf{Z}$ is a $2 \times K$ matrix of $(x_{bj} - \kappa_k)_+^p$ terms ($j = 1, 2$ and $k = 1, \ldots, K$) - with some technicalities explained in Giusti et al. (2012) - and γ is the coefficients vector for this part of the p-spline model which allows for handling departures from the p-polynomial in the structure of the relationship. The γ component can be treated as a vector of independent and identically distributed random variables with mean 0 and variance matrix $\Sigma_\gamma = \sigma_\gamma^2 \mathbf{I}_m$. Moreover, $\mathbf{Du}$ and $\mathbf{e}$ have the same meaning and variance structure as in model (2). So, now $\mathbf{V} = \mathbf{Z}\Sigma_\gamma \mathbf{Z}^T + \mathbf{D}\Sigma_u \mathbf{D}^T + \Sigma_e$. The unknown variance parameters Σ_γ, Σ_u and Σ_e can be estimated by REML or ML, obtaining $\hat{\mathbf{V}} = \mathbf{Z}\hat{\Sigma}_\gamma \mathbf{Z}^T + \mathbf{D}\hat{\Sigma}_u \mathbf{D}^T + \hat{\Sigma}_e$.

Model-based estimation of the small area parameters can be obtained by using the empirical best linear unbiased prediction (Henderson, 1975):

$$\hat{Y}_i^{NPEBLUP} = \mathbf{x}_i^T \hat{\beta} + (\mathbf{z}_i^T \hat{\Sigma}_\gamma \mathbf{Z}^T + \mathbf{d}_i^T \hat{\Sigma}_u \mathbf{D}^T) \hat{\mathbf{V}}^{-1} (\hat{\mathbf{Y}} - \mathbf{X}\hat{\beta}), \tag{5}$$

where $\hat{\beta} = (\mathbf{X}^T \hat{V}^{-1} \mathbf{X})^{-1} \mathbf{X}^T \hat{\mathbf{V}}^{-1} \hat{\mathbf{Y}}$, and $\mathbf{d}_i^T$ and $\mathbf{z}_i^T$ denote the ith row of the $\mathbf{D}$ and $\mathbf{Z}$ matrixes.

An alternative approach to incorporating spatial information in small area estimation is by assuming that the effects of neighboring area in the Fay-Herriot model are correlated. The Fay-Herriot model (1) assumes that the area effects are uncorrelated. However, it is often the case, especially with agricultural, environmental, economic and epidemiological data, that the effects of neighboring area are correlated, even if geographic covariates are included in the model. For these situations, Salvati (2004), Singh et al. (2005), Petrucci & Salvati (2006) proposed the introduction of spatial autocorrelation in small area estimation under the Fay-Herriot model. The specification of this model is of the form

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{D}\mathbf{u} + \mathbf{e}, \quad \mathbf{D} = (\mathbf{I} - \rho\mathbf{W})^{-1}, \tag{6}$$

where $\mathbf{X}$ and β are the same as in (2) and $\mathbf{D}$ represents the coefficients of random effects $\mathbf{u}$ that incorporate the spatial structure. Here, $\mathbf{W}$ is a known spatial weight matrix which shows the amount of interaction between any pair of small areas. The elements of $\mathbf{W}$ may depend on the distance between the centers of small areas or on the length of common boundary between them. The simplest form of $\mathbf{W}$ is a symmetric matrix with a diagonal of zeroes and with element $\mathbf{w}_{ij} = 1$ if area i and j are adjacent and $\mathbf{w}_{ij} = 0$ otherwise. The constant ρ is a measure of the overall level of spatial autocorrelation and its magnitude reflects the suitability of $\mathbf{W}$ for given $\mathbf{y}$ and $\mathbf{X}$. Further, $\mathbf{Du}$ and $\mathbf{e}$ have the same meaning and variance structure as in model (2). By ML or REML it is possible to estimate the variance matrix, obtaining $\hat{\mathbf{V}} = \mathbf{D}\hat{\Sigma}_u\mathbf{D}^T + \hat{\Sigma}_e$ where $\hat{\Sigma}_u = \hat{\sigma_u^2}[(\mathbf{I}_m - \hat{\rho}\mathbf{W})^T(\mathbf{I}_m - \hat{\rho}\mathbf{W})]^{-1}$. Estimates for small area target can be obtained by the empirical best linear unbiased prediction (Henderson, 1975)

$$\hat{Y}_i^{SEBLUP} = \mathbf{x}_i^T\hat{\beta} - \mathbf{d}_i^T\hat{\Sigma}_u\mathbf{D}^T\hat{\mathbf{V}}^{-1}(\hat{\mathbf{Y}} - \mathbf{X}\hat{\beta}), \tag{7}$$

where the terms of the equation are the same as in (2) with the exception of $\hat{\Sigma}_u$.

Recently, Chandra et al. (2014) proposed a spatially nonstationary Fay-Herriot model defined as

$$\mathbf{y}(\mathbf{l}) = \mathbf{X}\beta(\mathbf{l}) + \mathbf{D}\mathbf{u} + \mathbf{e} = \mathbf{X}\beta + \mathbf{Z}\gamma(\mathbf{l}) + \mathbf{D}\mathbf{u} + \mathbf{e}, \tag{8}$$

where $\mathbf{X}$, β, $\mathbf{D}$, $\mathbf{u}$ and $\mathbf{e}$ have the same meaning as before, $\mathbf{Z} = \{diag(\mathbf{x}_1), \ldots, diag(\mathbf{x}_p)\}$ is a $m \times pm$ matrix (here p is the number of auxiliary variables), $\gamma(\mathbf{l})$ is a vector of random effects that characterizes the spatial nonstationarity in the data and $\mathbf{l}$ is the coordinates matrix of the centroids of the areas. The variance structure of the model is $\mathbf{V} = \mathbf{Z}\Sigma_\gamma\mathbf{Z}^T + \mathbf{D}\Sigma_u\mathbf{D}^T + \Sigma_e$, where Σ_u and Σ_e are the same as in the Fay-Herriot model. Here, $\Sigma_\gamma = (\mathbf{c}\mathbf{c}^T) \odot \Omega$ with $\Omega = [d(\mathbf{l}_i, \mathbf{l}_j)]$ the matrix of the distances between areas and $\mathbf{c}$ a vector of positive unknown constants such that $(\mathbf{c}\mathbf{c}^T) \odot \Omega$ is a variance-covariance matrix ($\odot$ denotes the element-wise multiplication). Estimating the variance component Σ_u and Σ_γ, Chandra et al. (2014) obtained the best empirical unbiased predictor of Y_i, that is

$$\hat{Y}_i^{NSEBLUP} = \mathbf{x}_i^T\hat{\beta} + (\mathbf{z}_i^T\hat{\Sigma}_\gamma\mathbf{Z}^T\hat{\mathbf{V}}^{-1} + \mathbf{d}_i^T\hat{\Sigma}_u\mathbf{D}^T\hat{\mathbf{V}}^{-1})(\hat{\mathbf{Y}} - \mathbf{X}\hat{\beta}). \tag{9}$$

Here, $\hat{\mathbf{V}} = \mathbf{Z}\hat{\Sigma}_\gamma\mathbf{Z}^T + \mathbf{D}\hat{\Sigma}_u\mathbf{D}^T + \Sigma_e$ and $\hat{\beta} = (\mathbf{X}^T\hat{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\hat{\mathbf{V}}^{-1}\hat{\mathbf{Y}}$.

An analytic mean squared error estimator has been proposed for all the four presented small area estimators (EBLUP, NPEBLUP, SEBLUP, NSEBLUP), but is not discussed in this work.

2 Simulation experiment

In this section we propose a model-based Monte-Carlo simulations to empirically evaluate the performance of the EBLUP, NPEBLUP, SEBLUP and NSEBLUP.

The experiment setting is inspired by the work of Crainiceanu et al. (2005), Marley & Wand (2010) and Bocci & Rocco (2014). The populations will be generated using the following model:

$$Y_i = \alpha + \beta x_i + f(\mathbf{l}_i) + u_i$$
$$y_i = Y_i + e_i$$

Here Y_i is the mean of the target variable in small area i, while y_i is its direct estimate. The parameters α and β are set to 10 and 0.4, $x_i \sim Bin(m, 0.5)$, $f(\mathbf{l}_i)$ is the density function of a mixture of three bivariate normal variables with $\mathbf{l}_i$ the vector of latitude and longitude of the centroid of area i. As concerns the errors, $u_i \sim N(0, \sigma_u^2 = 0.025)$ and $e_i \sim N(0, \sigma_e^2 = 0.04)$. The bivariate normal mixture density is obtained as weighted mean of the following bivariate normal density

$$W_a = N\left(\begin{bmatrix} 0.25 \\ 0.5 \end{bmatrix}, \begin{bmatrix} 0.0225 & 0.1050 \\ 0.1050 & 0.0306 \end{bmatrix}\right)$$
$$W_b = N\left(\begin{bmatrix} 0.75 \\ 0.8 \end{bmatrix}, \begin{bmatrix} 0.0225 & 0 \\ 0 & 0.0306 \end{bmatrix}\right)$$
$$W_c = N\left(\begin{bmatrix} 0.75 \\ 0.2 \end{bmatrix}, \begin{bmatrix} 0.0225 & 0 \\ 0 & 0.0306 \end{bmatrix}\right),$$

with weights $3/7$ for W_a and W_b and $1/7$ for W_c.

We consider a squared region divided into $m = 100$ equal squared areas. Fig. 1 represents the bivariate mixture density function $f(\mathbf{l}_i)$, held constant for each Monte-Carlo run. The population is generated so that $f(\mathbf{l}_i)$ leads the magnitude of the Y_i values. The aim is to compare the performance of the estimators computing the relative bias (RB) and the relative root mean squared error (RRMSE):

$$RB_i = \frac{H^{-1}\sum_{h=1}^{H}(\hat{Y}_{ih} - Y_{ih})}{H^{-1}\sum_{h=1}^{H} Y_{ih}} \times 100, \quad RRMSE_i = \frac{\sqrt{H^{-1}\sum_{h=1}^{H}(\hat{Y}_{ih} - Y_{ih})^2}}{H^{-1}\sum_{h=1}^{H} Y_{ih}} \times 100,$$

Y_{ih} and $\hat{Y}_{ih}$ being the true and estimated values of the area i mean at simulation h.

Results of the simulation experiment are shown in table 1. The results seem indicating that all the estimators are unbiased, while in terms of variability the EBLUP is the worst, as expected given that it does not take into account the spatial information. Among the other estimators the NPEBLUP shows the best performance both in terms of relative bias and relative root mean squared error.

Further studies will be focused on comparing the performance of these estimators under different simulation scenarios.

Table 1 Mean and median over areas of the Relative Bias (RB) and Relative Root Mean Squared Error (RRMSE) of the target estimators (EBLUP, SEBLUP, NSEBLUP, NPEBLUP).

		EBLUP	SEBLUP	NSEBLUP	NPEBLUP
RB %	Mean	0.032	0.030	0.023	0.015
	Median	0.133	0.159	0.177	-0.022
RRMSE %	Mean	1.734	1.671	1.457	1.188
	Median	1.714	1.666	1.433	1.137

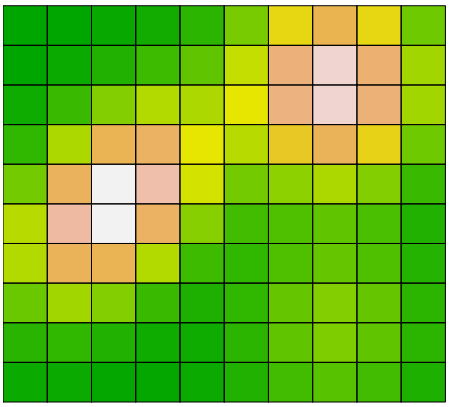

Fig. 1 Graphical representation of the $f(\mathbf{l}_i)$ values, where $\mathbf{l}_i$ is the vector of latitude and longitude of the centroid of area i.

References

1. Bocci, C. & Rocco, E.: Estimates for geographical domains through geoadditive models in presence of incomplete geographical information. Statistical Methods & Applications (2014) doi: 10.1007/s10260-014-0256-9
2. Chandra, H., Salvati, N., Chambers, R.: A Spatially Nonstationary Fay-Herriot Model for Small Area Estimation. Submitted (2014)
3. Crainiceanu, C. and Ruppert, D. and Wand, M.P.: Bayesian analysis for penalized spline regression using WinBUGS. Journal of Statistical Software **14**, 1–24 (2005)
4. Fay, R.E. & Herriot, R.A.: Estimation of income from small places: An application of James-Stein procedures to census data. Journal of the American Statistical Association, **74**, 269–277 (1979)
5. Giusti, C., Marchetti, S., Pratesi, M., Salvati, N.: Semiparametric Fay-Herriot Model using Penalized Splines. Journal of the Indian Society of Statistics **66**, 1–14 (2012)
6. Henderson, C.R.: Best linear unbiased estimation and prediction under a selection model. Biometrics **31**, 423–447 (1975)
7. Marley, J. & Wand, M.P.: Non-Standard Semiparametric Regression via BRugs. Journal of Statistical Software **37**, 1–30 (2010)
8. Petrucci, A. & Salvati N.: Small area estimation for spatial correlation in watershed erosion assessment. Journal of Agricultural, Biological, and Environmental Statistics **11**, 169–182 (2006)
9. Rao, J.N.K.: Small Area Estimation. New York, Wiley (2003)
10. Salvati N.: La correlazione spaziale nella stima per piccole aree: metodi proposti e casi di studio. PhD Thesis, University of Florence (2004)
11. Singh, B.B., Shukla, G.K., Kundu, D.: Spatio-Temporal Models in Small Area Estimation. Survey Methodology **31**, 183–195 (2005)