Available online at www.isas.org.in/jisas
JOURNAL OF THE INDIAN SOCIETY OF AGRICULTURAL STATISTICS 66(1) 2012 1-14

Semiparametric Fay-Herriot Model using Penalized Splines

C. Giusti, S. Marchetti, M. Pratesi* and N. Salvati
Department of Statistics and Mathematics Applied to Economics, University of Pisa, Via Ridolfi 10, 56124-Pisa, Italy

Received 24 December 2010; Revised 06 September 2011; Accepted 06 September 2011

SUMMARY

In this paper we propose a semiparametric Fay and Herriot area level model based on P-splines, which can handle situations where the functional form of the relationship between the variable of interest and the covariates is unknown. This is often the case when the data are supposed to be affected by spatial proximity effects. In these cases P-spline bivariate smoothing can easily introduce spatial effects in the area level model. By this spatial effect we can obtain estimates for out of sample areas and also for those areas where auxiliary information is unavailable. We focus here on the small area mean estimator and on an analytic and a bootstrap based mean squared error estimators. The proposed estimators of the small area means and mean squared errors are contrasted to the traditional ones by means of two simulations studies. We finally present results of the application of our semiparametric model to estimate the mean of the Acid Binding Capacity (ANC) and Calcium (CA) concentration in streams for each 8-digit Hydrologic Unit Code (HUC) within the Mid-Atlantic region of the US. ANC and CA concentration represent two of the key indicators to keep under control for environmental protection and preservation of natural resources. These results present evidence that the proposed estimators can be used to obtain accurate estimates in those areas where direct estimates are unreliable or even unavailable.

Keywords : Small area methods, Semiparametric models, Bootstrap methods, Environmental data.

1. INTRODUCTION

Traditional Fay-Herriot area level models (Fay and Herriot 1979) are based on linear mixed models, characterized by random area effects which allow for between area heterogeneity apart from that explained by the auxiliary variables (Rao 2003). These models are based on the hypothesis of a linear relationship between the variable of interest and the covariates, an hypothesis that can represent a serious restriction in many real data applications. Furthermore, traditional linear mixed models do not handle spatial proximity effects between the areas, an important feature in environmental studies where detailed geo-referenced information for the units of analysis is usually available. Indeed, in recent years extensions to random effects models have been proposed to allow for spatially correlated random area effects taking into account the information provided by neighboring areas (Petrucci and Salvati 2006; Pratesi and Salvati 2009), but these models still rely on the linearity assumption.

Here we propose a semiparametric version of the basic Fay-Herriot model that is based on P-splines, so that we can also handle situations where the functional form of the relationship between the variable of interest and the covariates cannot be specified a priori. This is often the case when the data are supposed to be affected by spatial proximity effects. In these cases P-spline bivariate smoothing can easily introduce spatial effects in the area level model. Opsomer *et al.* (2008) proposed a similar small area model based on

*Corresponding author : M. Pratesi
E-mail address : m.pratesi@ec.unipi.it

P-splines but under the assumption that all the data are available at the unit level, and this can be a restriction in some situations.

The model proposed here is applied to two case studies with focus on the ecological conditions of the waters in the Mid-Atlantic States of the US. Between years 1993 and 1998 the Environmental Monitoring and Assessment Program of the US Environmental Protection Agency conducted two surveys of streams to determine the ecological condition of these waters. Our target on these data is to estimate the mean of the Acid Binding Capacity (ANC) and Calcium (CA) concentration for each 8-digit Hydrologic Unit Code (HUC) within the region of interest. ANC and CA concentration represent two of the key indicators to keep under control for environmental protection and preservation of natural resources.

In the problem of estimating the ANC and CA concentration at Hidrologic Unit Code level the auxiliary variables can be known only at area level, making it necessary to specify area level models (Stoddard *et al.* 2006; Jones *et al.* 1997). In this situation the use of area level models also allows to obtain the estimates of interest for out of sample areas. Note in addition that area level specifications take into account the sampling design as they model direct estimates at area level.

Information like ANC and CA concentration are usually computed for macro geographical units like the Ecoregions (Stoddard *et al.* 2006; Whittier *et al.* 2008): the relevance would be higher if these values could be known for the areas identified by the HUCs. The HUCs identify parcels of lands drained by a given stream: thus, HUCs represent a meaningful subdivision to delineate areas of analysis in surveys on hydrological features. However, 8-digit HUCs are unplanned domains in EPA surveys, so that not all the HUCs are surveyed and even for surveyed HUCs the number of measurements do not suffice to compute reliable direct estimates. To compute the indicators of interest (e.g. the mean value of ANC) for each 8-digit HUC there is the need to resort to small area estimation techniques.

The paper is organized as follows. In Section 2 we extend the Fay-Herriot model to a semiparametric specification by introducing a P-spline component. In Section 3 we derive the analytical approximation for the mean squared error of the semiparametric estimator, while in Section 4 we propose two alternative estimators of the mean squared error based on a semiparametric bootstrap procedure. The performance of the proposed estimators is evaluated and contrasted to that of the traditional ones by means of two simulation studies presented in Section 5. In Section 6 we apply the semiparametric model to estimate the mean ANC and CA concentration in 126 8-digit HUCs in the Mid-Atlantic States of the US. Finally, in Section 7 we summarize the theoretical and applied advantages of the proposed methodology.

2. SEMIPARAMETRIC FAY-HERRIOT MODEL

The Fay-Herriot model produces reliable small area estimates by combining the design model and the regression model and then borrowing strength from other domains. It assumes that the direct survey estimators are linear function of the covariates. When this assumption falls 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 the response variable and the auxiliary variables, can be obtained by P-splines.

A semiparametric additive model (referred by semiparametric model hereafter) with one covariate x can be written as $\tilde{m}(x)$, where the function $\tilde{m}(\cdot)$ is unknown, but assumed to be sufficiently well approximated by the function

$$m(x; \beta, \gamma) = \beta_0 + \beta_1 x + \ldots + \beta_p x^p + \sum_{k=1}^{K} \gamma_k (x - \kappa_k)_+^p \qquad (2.1)$$

where $\beta = (\beta_0, \beta_1, \ldots, \beta_p)^T$ is the $(p + 1)$ vector of the coefficients of the polynomial function, $\gamma = (\gamma_1, \gamma_2, \ldots, \gamma_K)^T$ is the coefficient vector of the truncated polynomial spline basis (P-spline) and p is the degree of the spline $(t)_+^p = t^p$ if $t > 0$ and 0 otherwise. The latter portion of the model allows for handling departures from a p-polynomial t in the structure of the relationship. In this portion κ_k for $k = 1, \ldots, K$ is a set of fixed knots and if K is sufficiently large, the class of functions in (2.1) is very large and can approximate most smooth functions. Details on bases and knots choice can be found in Ruppert *et al.* (2003, Chapters 3 and 5). Since a P-spline model can be viewed as a random-effects model (Ruppert *et al.* 2003; Opsomer

et al. 2008), it can be combined with the Fay-Herriot model for obtaining a semiparametric small area estimation framework based on linear mixed model regression.

Let θ be the $m \times 1$ vector of the parameter of inferential interest (small area total y_i, small area mean $\bar{y}_i$ with $i = 1 \ldots m$) and assume that the $m \times 1$ vector of the direct estimator $\hat{\theta}$ is available and design unbiased. Given the β and γ vectors, define

$$\mathbf{X} = \begin{bmatrix} 1 & x_1 & \cdots & x_1^p \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_m & \cdots & x_m^p \end{bmatrix},$$

and

$$\mathbf{Z} = \begin{bmatrix} (x_1 - \kappa_1)_+^p & \cdots & (x_1 - \kappa_K)_+^p \\ \vdots & \ddots & \vdots \\ (x_m - \kappa_1)_+^p & \cdots & (x_m - \kappa_K)_+^p \end{bmatrix}.$$

If other variables are available that need to be included in the model as parametric terms, they can be added to the **X** fixed effect matrix. The semiparametric Fay-Herriot model can be written as

$$\hat{\theta} = \mathbf{X}\beta + \mathbf{Z}\gamma + \mathbf{D}\mathbf{u} + \varepsilon, \qquad (2.2)$$

where β is a vector of regression coefficients, the γ component can be treated as a $K \times 1$ vector of independent and identically distributed random variables with mean $\mathbf{0}$ and $K \times K$ variance matrix $\mathbf{\Sigma}_\gamma = \sigma_\gamma^2 \mathbf{I}_K$. Moreover, $\mathbf{u}$ is $m \times 1$ vector of independent and identically distributed random variables with mean $\mathbf{0}$ and $m \times m$ variance matrix $\Sigma_u = \sigma_u^2 \mathbf{I}_m$, $\mathbf{D}$ is a $m \times m$ matrix of known positive constants and ε is the $m \times 1$ vector of independent sampling errors with mean $\mathbf{0}$ and known diagonal variance matrix $\mathbf{R} = diag(\sigma_1^2, \sigma_2^2, \ldots, \sigma_m^2)$. The semiparametric Fay-Herriot model is a general linear mixed model with variance-covariance matrix $\mathbf{\Sigma}(\psi) = \mathbf{Z}\mathbf{\Sigma}_\gamma\mathbf{Z}^T + \mathbf{D}\mathbf{\Sigma}_u\mathbf{D}^T + \mathbf{R}$ where $\psi = (\sigma_\gamma^2, \sigma_u^2)^T$.

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

$$\tilde{\theta}^B(\psi) = \mathbf{X}\tilde{\beta}(\psi) + \mathbf{\Lambda}(\psi)[\hat{\theta} - \mathbf{X}\tilde{\beta}(\psi)] \qquad (2.3)$$

with $\mathbf{\Lambda}(\psi) = (\mathbf{Z}\mathbf{\Sigma}_\gamma\mathbf{Z}^T + \mathbf{D}\mathbf{\Sigma}_u\mathbf{D}^T)\mathbf{\Sigma}^{-1}(\psi)$ and $\tilde{\beta}(\psi) = (\mathbf{X}^T\mathbf{\Sigma}^{-1}(\psi)\mathbf{X})^{-1}\mathbf{X}^T\mathbf{\Sigma}^{-1}(\psi)\hat{\theta}$.

When geographically referenced responses play a central role in the analysis and need to be converted to maps, we can deal with utilise bivariate smoothing: $\tilde{m}(x_1 x_2) = m(x_1, x_2; \beta, \gamma)$. This is the case of environment, agricultural, public health and poverty mapping application fields. P-splines rely on a set of basis functions to handle non-linear structures in the data. So bivariate basis functions are required for bivariate smoothing. We will assume the following model (see details in Opsomer *et al.* 2008)

$$m(x_1, x_2; \beta, \gamma) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \mathbf{z}_i\gamma, \qquad (2.4)$$

where $\mathbf{z}_i$ is the i-th row of the following $n \times K$ matrix

$$\mathbf{Z} = \left[C(\tilde{\mathbf{x}}_i - \kappa_k)\right]_{\substack{1 \le i \le n \\ 1 \le k \le K}} \left[C(\kappa_k - \kappa_{k'})\right]_{1 \le k \le K}^{-1/2}, \qquad (2.5)$$

where $C(\mathbf{t}) = \| \mathbf{t} \|^2 \log \| \mathbf{t} \|$, $\tilde{\mathbf{x}}_i = (x_{1i}, x_{2i})$ *and* κ_k, $k = 1, \ldots, K$ are knots. $C(\mathbf{t})$ function is applied so that when knots correspond to all observations (that is the full rank case) the model for bivariate smoothing leads to thin plate spline (Green and Silverman 1994). More details on the **Z** matrix can be found in Ruppert *et al.* (2003, Chapter 13), Kammann and Wand (2003) and French *et al.* (2001). The use of matrix **Z** allows for simplification in the estimation procedure.

3. ANALYTICAL APPROXIMATION OF THE MSE

The Mean Squared Error (MSE) of $\tilde{\theta}^B(\psi)$, depending on the variance components $\psi = (\sigma_\gamma^2, \sigma_u^2)^T$, can be expressed as (Rao 2003):

$$MSE[\tilde{\theta}^B(\psi)] = g_1(\psi) + g_2(\psi) \qquad (3.1)$$

where the first term

$$g_1(\psi) = \mathbf{\Lambda}(\psi)\mathbf{R} = \mathbf{R} - \mathbf{R}\mathbf{\Sigma}^{-1}(\psi)\mathbf{R} \qquad (3.2)$$

is due to the estimation of random effects and it is of order $O(1)$, while the second term

$$g_2(\psi) = \mathbf{R}\mathbf{\Sigma}^{-1}(\psi)\mathbf{X}(\mathbf{X}^T\mathbf{\Sigma}^{-1}(\psi)\mathbf{X})^{-1}\mathbf{X}^T\mathbf{\Sigma}^{-1}(\psi)\mathbf{R} \qquad (3.3)$$

is due to the estimation of β and it is of order $O(m^{-1})$ for large m.

The estimator $\tilde{\theta}^B(\psi)$ depends on the unknown variance components σ_γ^2 and σ_u^2. Replacing the

parameters with estimators $\hat{\sigma}_\gamma^2$ and $\hat{\sigma}_u^2$, an empirical best linear unbiased predictor (EBLUP) $\tilde{\theta}^E(\hat{\psi})$ is

$$\tilde{\theta}^E(\hat{\psi}) = \mathbf{X}\hat{\beta}(\hat{\psi}) + \hat{\mathbf{\Lambda}}(\hat{\psi})[\hat{\theta} - \mathbf{X}\hat{\beta}(\hat{\psi})] \quad (3.4)$$

where $\hat{\beta}(\hat{\psi}) = (\mathbf{X}^T\hat{\mathbf{\Sigma}}^{-1}(\hat{\psi})\mathbf{X})^{-1}\mathbf{X}^T\hat{\mathbf{\Sigma}}^{-1}(\hat{\psi})\hat{\theta}$. Assuming normality of the random effects, σ_γ^2 and σ_u^2 can be estimated both by Maximum Likelihood (ML) and Restricted Maximum Likelihood (REML) procedures (Prasad and Rao 1990).

The ML and REML estimators possess the following properties (Datta *et al.* 2005): (i) they are $m^{1/2}$-consistent; (ii) they are even functions of $\hat{\theta}$, so that $\hat{\psi}(-\hat{\theta}) = \hat{\psi}(\hat{\theta})$; (iii) they are translation invariant functions, so that $\hat{\psi}(\hat{\theta} + \mathbf{G}c) = \hat{\psi}(\hat{\theta})$, for any $m \times (g + 1)$ matrix, $c \in \mathbb{R}^{g+1}$ and for all $\hat{\theta}$.

For any $\hat{\psi}$ satisfying (ii) and (iii), the MSE of $\tilde{\theta}^E(\hat{\psi})$ can be decomposed as

$$\begin{aligned} MSE[\tilde{\theta}^E(\hat{\psi})] &= g_1(\psi) + g_2(\psi) + E\{[\tilde{\theta}^E(\hat{\psi}) - \tilde{\theta}^B(\psi)]^2\} \\ &= g_1(\psi) + g_2(\psi) + g_3(\psi). \end{aligned} \quad (3.5)$$

Under the traditional Fay-Herriot model with diagonal covariance matrix $\mathbf{\Sigma}(\psi)$, Prasad and Rao (1990) obtained an approximation up to $o(m^{-1})$ terms of $g_3(\psi)$ through Taylor linearization. In case of the semi-parametric Fay-Herriot model the structure of the covariance matrix is not diagonal due to the introduction of the spline random component, then the results of Prasad and Rao (1990) can not be applied directly. The results of Opsomer *et al.* (2008) can be used for deriving a second order approximation to the $g_3(\psi)$ term. It can be given by

$$g_3(\psi) = \mathbf{L}^T(\psi)\,[\mathcal{I}^{-1}(\psi) \otimes \mathbf{\Sigma}(\psi)]\,\mathbf{L}(\psi) + o(\delta_m/m) \quad (3.6)$$

where

$$\mathbf{L}(\psi) = [\mathbf{L}_{\sigma_\gamma^2}(\psi), \mathbf{L}_{\sigma_u^2}(\psi)]^T,\ \mathbf{L}_i(\psi) = \frac{\partial \Lambda(\psi)}{\partial \psi_i},\ i = 1, 2.$$

Here $\otimes$ represents Kronecker product, $\mathcal{I}^{-1}(\psi)$ is the inverse of the information matrix with $\mathcal{I}_{ij}^{-1}(\psi) = 0.5\mathrm{tr}[\mathbf{P}(\psi)\mathbf{B}_i\mathbf{P}(\psi)\mathbf{B}_j]$, $i, j = 1, 2$, $\mathbf{P}(\psi) = \mathbf{\Sigma}^{-1}(\psi) - \mathbf{\Sigma}^{-1}(\psi)\mathbf{X}(\mathbf{X}^T\mathbf{\Sigma}^{-1}(\psi)\mathbf{X})^{-1}\mathbf{X}^T\mathbf{\Sigma}^{-1}(\psi)$, $\mathbf{B}_1 = \mathbf{Z}\mathbf{Z}^T$ and $\mathbf{B}_2 = \mathbf{D}\mathbf{D}^T$ and $\delta_m = o(\sqrt{m})$.

In practical applications, the EBLUP $\tilde{\theta}^E(\hat{\psi})$ should be accompanied by an estimate of the MSE. Again, under Fay-Herriot models with diagonal covariance matrix, Prasad and Rao (1990) obtained an approximately unbiased estimator of the MSE (3.5). Following the results of Prasad and Rao (1990) and Das *et al.* (2004), Opsomer *et al.* (2008) extended the Prasad-Rao MSE estimator to models with more general covariance structure. An approximately unbiased estimator of the MSE is

$$mse^{ana}[\tilde{\theta}^E(\hat{\psi})] = g_1(\hat{\psi}) + g_2(\hat{\psi}) + 2g_3(\hat{\psi}) \quad (3.7)$$

which is the same estimator derived by Prasad and Rao (1990). In formula (3.7), the term $g_3(\hat{\psi})$ appears twice due to a bias correction of $g_1(\hat{\psi})$.

4. NONPARAMETRIC BOOTSTRAP FOR ESTIMATING THE MSE

This section describes an alternative procedure for estimating the MSE of the EBLUP $\tilde{\theta}^E(\hat{\psi})$ based on bootstrapping according to the bootstrap procedure proposed by González-Manteiga *et al.* (2007), Opsomer *et al.* (2008) and Molina *et al.* (2009). In this procedure, the bootstrap random effects $(\gamma_1^*, ..., \gamma_K^*)^T$, $(u_1^*, ..., u_m^*)^T$ and the random errors $(\varepsilon_1^*, ..., \varepsilon_m^*)^T$ are obtained by resampling respectively from the empirical distribution of the predicted random elements $\hat{\boldsymbol{\gamma}} = (\hat{\gamma}_1, ..., \hat{\gamma}_K)^T$, $\hat{\mathbf{u}} = (\hat{u}_1, ..., \hat{u}_m)^T$, and the residuals $\hat{\mathbf{r}} = \hat{\theta} - \mathbf{X}\hat{\beta} - \mathbf{Z}\hat{\boldsymbol{\gamma}} - \mathbf{D}\hat{\mathbf{u}}$ $= (\hat{r}_1, ..., \hat{r}_m)^T$, previously standardized. This method avoids the need of distributional assumptions; therefore, it is expected to be more robust to non-normality of any of the random components of the model. The procedure works as follows:

1. Fit model (2.2) to the initial direct estimates $\hat{\theta}$, obtaining estimates $(\hat{\sigma}_\gamma^2, \hat{\sigma}_u^2)$ and $\hat{\beta}$.
2. With estimates obtained in step 1, calculate predictors of $\hat{\boldsymbol{\gamma}} = (\hat{\gamma}_1, ..., \hat{\gamma}_K)^T$ and

$\hat{\mathbf{u}} = (\hat{u}_1, ..., \hat{u}_m)^T$. Then take $\hat{\boldsymbol{\gamma}}^S = \hat{\boldsymbol{\Sigma}}_\gamma^{-1/2}\hat{\boldsymbol{\gamma}}$ and $\hat{\mathbf{u}}^S = \hat{\boldsymbol{\Sigma}}_u^{-1/2}\hat{\mathbf{u}}$ where $\hat{\boldsymbol{\Sigma}}_\gamma^{-1/2}$ and $\hat{\boldsymbol{\Sigma}}_u^{-1/2}$ are the root square of the generalized inverse of $\hat{\boldsymbol{\Sigma}}_\gamma^{-1/2} = \mathbf{Z}\hat{\boldsymbol{\Sigma}}_\gamma\mathbf{Z}^T\mathbf{3}(\hat{\boldsymbol{\psi}})\mathbf{Z}^T\hat{\boldsymbol{\Sigma}}_\gamma\mathbf{Z}$ and $\hat{\boldsymbol{\Sigma}}_u^{-1/2} = \mathbf{D}\hat{\boldsymbol{\Sigma}}_u\mathbf{D}^T\mathbf{P}(\hat{\boldsymbol{\psi}})\mathbf{D}^T\hat{\boldsymbol{\Sigma}}_u\mathbf{D}$ respectively, obtained by the spectral decomposition. It is convenient to re-scale the elements $\hat{\boldsymbol{\gamma}}_k^S$ and $\hat{\mathbf{u}}_i^S$ so that they have sample means exactly equal to zero and sample variances $\hat{\sigma}_\gamma^2$, $\hat{\sigma}_u^2$. This is achieved by the transformations

$$\hat{\gamma}_k^{SS} = \frac{\hat{\sigma}_\gamma\left\{\hat{\gamma}_k^S - K^{-1}\Sigma_{j=1}^K\hat{\gamma}_j^S\right\}}{\sqrt{K^{-1}\Sigma_{d=1}^K\left\{\hat{\gamma}_d^S - K^{-1}\Sigma_{j=1}^K\hat{\gamma}_j^S\right\}^2}}, \quad k = 1, ..., K$$

$$\hat{u}_i^{SS} = \frac{\hat{\sigma}_u\left\{\hat{u}_k^S - m^{-1}\Sigma_{j=1}^m\hat{u}_j^S\right\}}{\sqrt{m^{-1}\Sigma_{d=1}^m\left\{\hat{u}_d^S - m^{-1}\Sigma_{j=1}^m\hat{u}_j^S\right\}^2}}, \quad i = 1, ..., m.$$

Construct the vectors $\boldsymbol{\gamma}^* = (\gamma_1^*, ..., \gamma_K^*)^T$ and $\mathbf{u}^* = (u_1^*, ..., u_m^*)^T$, whose elements are obtained by extracting a simple random sample with replacement of size K and m from the sets $\hat{\boldsymbol{\gamma}}^{SS} = (\hat{\gamma}_1^{SS}, ..., \hat{\gamma}_K^{SS})^T$ and $\hat{\mathbf{u}}^{SS} = (\hat{u}_1^{SS}, ..., \hat{u}_K^{SS})^T$, respectively. Then calculate the bootstrap quantity of interest $\theta^* = \mathbf{X}\hat{\boldsymbol{\beta}} + \mathbf{Z}\boldsymbol{\gamma}^* + \mathbf{D}\mathbf{u}^* = (\theta_1^*, ..., \theta_m^*)^T$.

3. Compute the vector of residuals $\hat{\mathbf{r}} = \hat{\boldsymbol{\theta}} - \mathbf{X}\hat{\boldsymbol{\beta}} - \mathbf{Z}\hat{\boldsymbol{\gamma}} - \mathbf{D}\hat{\mathbf{u}} = (\hat{r}_1, ..., \hat{r}_m)^T$. Standardize the residuals by $\hat{\mathbf{r}}^S = (\mathbf{RP}(\hat{\boldsymbol{\psi}})\mathbf{R})^{-1/2}\hat{\mathbf{r}}$. Re-standardized these values

$$\hat{r}_i^{SS} = \frac{\left\{\hat{r}_i^S - m^{-1}\Sigma_{j=1}^m\hat{r}_j^S\right\}}{\sqrt{m^{-1}\Sigma_{d=1}^m\left\{\hat{r}_d^S - m^{-1}\Sigma_{j=1}^m\hat{r}_j^S\right\}^2}}, \quad i = 1, ..., m.$$

Construct the vector $\mathbf{r}^* = (r_1^*, ..., r_m^*)^T$, whose elements are obtained by extracting a simple random sample with replacement of size m from the set $\hat{\mathbf{r}}^{SS} = (\hat{r}_1^{SS}, ..., \hat{r}_m^{SS})^T$. Then take $\boldsymbol{\varepsilon}^* = (\varepsilon_1^*, ..., \varepsilon_m^*)^T$ where $\varepsilon_i^* = \sigma_i r_i^*$.

4. Construct bootstrap data from the model,

$$\hat{\boldsymbol{\theta}}^* = \boldsymbol{\theta}^* + \boldsymbol{\varepsilon}^* = \mathbf{X}\hat{\boldsymbol{\beta}} + \mathbf{Z}\boldsymbol{\gamma}^* + \mathbf{D}\mathbf{u}^* + \boldsymbol{\varepsilon}^* = (\hat{\theta}_1^*, ..., \hat{\theta}_m^*)^T.$$

5. Regarding $\hat{\boldsymbol{\beta}}, \hat{\sigma}_\gamma^2$ and $\hat{\sigma}_u^2$ as the true values of $\boldsymbol{\beta}$, σ_γ^2 and σ_u^2, fit the model (2.2) to the bootstrap data $\hat{\boldsymbol{\theta}}^*$. The obtained estimates $\boldsymbol{\beta}^*, \sigma_\gamma^{2*}$ and σ_u^{2*} will be called bootstrap estimators.

6. Calculate the bootstrap small area estimator using $\boldsymbol{\beta}^*$, σ_γ^{2*} and σ_u^{2*} in place of the 'true' $\hat{\boldsymbol{\beta}}, \hat{\sigma}_\gamma^2$ and $\hat{\sigma}_u^2$,

$$\tilde{\boldsymbol{\theta}}^{E*}(\hat{\boldsymbol{\psi}}^*) = \mathbf{X}\hat{\boldsymbol{\beta}}^*(\hat{\boldsymbol{\psi}}^*) + \hat{\boldsymbol{\Lambda}}^*(\hat{\boldsymbol{\psi}}^*)[\hat{\boldsymbol{\theta}}^* - \mathbf{X}\boldsymbol{\beta}^*(\hat{\boldsymbol{\psi}}^*)]$$

7. Repeat steps 2-6 B times. In the b-th bootstrap replication, let $\theta_i^{*(b)}$ be the quantity of interest in area i, $\tilde{\boldsymbol{\theta}}_i^{E*}(\hat{\boldsymbol{\psi}}^{*(b)})$ be the bootstrap estimator for area i.

A naïve bootstrap estimator for the MSE for area i is

$$mse_i^{naNPB}[\tilde{\boldsymbol{\theta}}_i^E(\hat{\boldsymbol{\psi}})] = B^{-1}\sum_{b=1}^B\left\{\tilde{\boldsymbol{\theta}}_i^{E*(b)}(\hat{\boldsymbol{\psi}}^{*(b)}) - \boldsymbol{\theta}_i^{*(b)}\right\}^2. \quad (4.1)$$

Another MSE estimate can be obtained by adding the bootstrap estimate $g_{3i}^{NPB}(\hat{\boldsymbol{\psi}})$ to the analytical estimates $g_{1i}(\hat{\boldsymbol{\psi}})$ and $g_{2i}(\hat{\boldsymbol{\psi}})$, and then including a bootstrap bias correction of $g_{1i}(\hat{\boldsymbol{\psi}}) + g_{2i}(\hat{\boldsymbol{\psi}})$ (Pfeffermann and Tiller 2006), as

$$mse_i^{bcNPB}[\tilde{\boldsymbol{\theta}}_i^E(\hat{\boldsymbol{\psi}})] = 2[g_{1i}(\hat{\boldsymbol{\psi}}) + g_{2i}(\hat{\boldsymbol{\psi}})] - B^{-1}\sum_{b=1}^B\left[g_{1i}(\hat{\boldsymbol{\psi}}^{*(b)}) + g_{2i}(\hat{\boldsymbol{\psi}}^{*(b)})\right] + g_{3i}^{NPB}(\hat{\boldsymbol{\psi}}). \quad (4.2)$$

where

$$g_{3i}^{NPB}(\hat{\boldsymbol{\psi}}) = B^{-1}\Sigma_{b=1}^B\left\{\tilde{\boldsymbol{\theta}}_i^{E*(b)}(\hat{\boldsymbol{\psi}}^{*(b)}) - \tilde{\boldsymbol{\theta}}^{BLUP*(b)}(\hat{\boldsymbol{\psi}})\right\}^2$$

with $\tilde{\boldsymbol{\theta}}^{BLUP*(b)} = \mathbf{X}\hat{\boldsymbol{\beta}}^* + \mathbf{Z}\boldsymbol{\gamma}^* + \mathbf{D}\mathbf{u}^*$.

5. SIMULATION STUDIES

5.1 Estimation of the Small Area Means

In this section we develop a simulation study to compare the performance of the $\tilde{\boldsymbol{\theta}}^E(\hat{\boldsymbol{\psi}})$ estimator of the small area mean under the proposed semiparametric specification (denoted by NPEBLUP hereafter) to that under the traditional Fay-Herriot specification (denoted by EBLUP).

We consider five models for creating the true underlying relationship between the covariate x and the expected value of the response variable θ, $E(\theta \mid x) = m(x)$:

Linear. $m(x) = 10 + 2(x)$;

Jump. $m(x) = 1 + 2(x - 1.5)I(x \leq 1.5) + 2I(x > 1.5)$.

Exponential. $m(x) = 2 + exp(3x)/400$.

Bump. $m(x) = 10 + 2(x - 1.5) + 5exp(-200(x - 1.5)^2)$.

Cycle. $m(x) = 10 + 10 \sin(2\pi x)$;

Under the random intercepts model $\hat{\bar{y}}_i = m(x) + u_i + \varepsilon_i$, where $i = 1, ..., 200$, we generate $T = 500$ data sets for each model with x drawn from a Uniform distribution [0, 1], the area effects u_i drawn from N(0, 0.04) and the error effects ε_i independently generated from N(0, σ_i^2). Here, the sampling errors σ_i^2 vary in the pattern: 0.08, 0.10, 0.12, 0.14, 0.16. There are five groups of small areas and 40 small areas in each group; the sampling variances σ_i^2 are the same for areas within the same group. The chosen pattern corresponds to an intra-area correlation that varies between 0.2 to 0.33.

The linear case represents a situation in which the EBLUP is based on a good representation of the true model, while the NPEBLUP may be too complex and overparametrized. The jump model is a discontinuous function for which EBLUP and NPEBLUP are based on a misspecified model; the Exponential, Bump and Cycle models define increasingly more complicated structures of the relationship between y and x.

For each data set the EBLUP and the NPEBLUP estimators have been used to estimate the small area means $\bar{y}_i$.

Then, for each estimator and for each small area i we averaged over Monte Carlo replications $t = 1, ..., T$ to estimate the Bias

$$B_{iMC} = \frac{1}{T}\sum_{t=1}^{T}(\hat{\bar{y}}_{it} - \bar{y}_{it}) \tag{5.1}$$

and with it the percentage relative bias

$$RB_i\% = \frac{B_{MC}}{\frac{1}{T}\sum_{t=1}^{T}\bar{y}_{it}}\ 100; \tag{5.2}$$

the Root Mean Squared Error

$$RMSE_{iMC} = \sqrt{\frac{1}{T}\sum_{t=1}^{T}(\hat{\bar{y}}_{it} - \bar{y}_{it})^2}\ , \tag{5.3}$$

and the corresponding percentage Relative Root Mean Squared Error

$$RRMSE_i\% = \frac{RMSE_{MC}}{\frac{1}{T}\sum_{t=1}^{T}\bar{y}_{it}}100. \tag{5.4}$$

To evaluate the RB% and the RRMSE% across the 200 small areas we consider these summary statistics: the minimum value, the first quartile, the mean and the median value, the third quartile and the maximum value.

Tables 1 and 2 report respectively the summary statistics for the RB% and the RRMSE% values obtained for the estimation of the small area means under the Linear, Jump, Exponential, Bump and Cycle signals.

The results are promising. First note that the performance of the two estimators is essentially equivalent under the Linear signal, both in terms of bias and variability. Then, from Table 1 we see that the mean and median biases of the NPEBLUP estimator are always lower with respect to the EBLUP estimator, with the only exception of the mean value under the Jump signal. Moreover, in many cases there is a high gain also in terms of minimum and maximum values of the RB%, that is, the bias of the NPEBLUP estimator in estimating the 200 small area means varies in a range of smaller size than the EBLUP. In terms of variability (Table 2) the results show a similar behavior: NPEBLUP is always a good competitor of the EBLUP.

Table 1. Percentage Relative Bias (RB%) of the estimators of the small area means.

RB% Point Estimation						
Estimator	Min	1st quartile	Mean	Median	3rd quartile	Max
Linear Signal						
EBLUP	−0.74	−0.12	−0.01	0.00	0.13	0.57
NPEBLUP	−0.44	−0.13	0.00	−0.01	0.11	0.47
Jump Signal						
EBLUP	−204.21	−12.84	0.64	−4.70	10.25	399.83
NPEBLUP	−108.79	−0.63	4.36	0.52	03.45	79.01
Exponential Signal						
EBLUP	−15.92	−4.54	1.29	1.78	8.24	13.42
NPEBLUP	−2.74	−0.82	−0.06	−0.14	0.67	2.80
Bump Signal						
EBLUP	−12.08	0.59	0.19	0.82	1.01	1.79
NPEBLUP	−10.46	−0.11	0.12	0.13	0.60	3.49
Cycle Signal						
EBLUP	−9.23	−2.24	33.14	−0.49	6.13	769.18
NPEBLUP	−46.69	−0.16	−0.62	−0.01	0.22	11.68

Table 2. Percentage Relative Root Mean Squared Error (RRMSE%) of the estimators of the small area means.

RRMSE% Point Estimation						
Estimator	Min	1st quartile	Mean	Median	3rd quartile	Max
Linear Signal						
EBLUP	4.37	5.03	5.66	5.54	6.28	7.25
NPEBLUP	4.45	5.00	5.67	5.60	6.24	7.36
Jump Signal						
EBLUP	23.29	26.81	108.66	42.74	106.62	2114.17
NPEBLUP	22.61	24.34	113.22	42.40	109.19	2487.47
Exponential Signal						
EBLUP	08.41	27.16	34.90	41.02	43.89	47.84
NPEBLUP	04.58	21.05	27.99	33.32	35.70	38.55
Bump Signal						
EBLUP	5.51	6.61	7.87	7.71	8.91	13.10
NPEBLUP	5.63	6.51	7.80	7.89	8.74	11.62
Cycle Signal						
EBLUP	4.85	5.74	86.00	8.13	23.03	1884.93
NPEBLUP	3.45	4.10	55.80	6.22	18.35	1317.73

5.2 Estimation of the Mean Squared Error

In this Section we present a simulation experiment carried out to contrast the three alternative estimators of the Mean Squared Error of the NPEBLUP estimator $\tilde{\boldsymbol{\theta}}^{E}(\hat{\boldsymbol{\psi}})$ described in Sections 3 and 4. Namely, the estimators we consider are the analytical estimator mse^{ana} (3.7), the naïve nonparametric bootstrap estimator mse^{naNPB} (4.1) and the combined analytical and bootstrap estimator mse^{bcNPB} (4.2).

The simulation study is carried out using real data coming from the Italian Agricultural Census of year 2000 for the Tuscany region, as in Molina *et al.* (2009), under two different settings. The small areas of interest are the 287 municipalities of the region, with N_i, $i = 1, \ldots, m$, given by the census and the n_i randomly generated from a Binomial distribution with parameters N_i and $p = 0.05$. These sampling data are used to compute, for each municipality i, the direct estimator of the mean agrarian surface area used for production of grape in hectares (θ_i) and its sampling variance (σ_i^2). Information on the agrarian surface area used for production in hectares (x_{1i}) and on the average number of working days in the reference year (x_{2i}) for each municipality i is also available from the census data.

Thus, in the simulation study the goal is the estimation of the mean agrarian surface area used for production of grape in hectares ($\bar{y}_i$) for all the municipalities of the region, using as explanatory variables x_{1i} and x_{2i}, which have a linear relation with $\bar{y}_i$, and an intercept term. The centroids of the small areas are also available as spatial reference points (latitude and longitude) and are used in the **Z** matrix when fitting the semiparametric model under both settings. Since the true sampling variances σ_i^2 were equal to 0 for nine areas, in the simulation experiment we consider $m = 278$. Note that the true sampling variances σ_i^2 have a highly right-skewed distribution with a range of 102745; this skewness is caused by few municipalities with atypically large sampling variances.

More in detail, in the first simulation setting the Monte Carlo samples are generated at each step as follows: first, the random errors e_i are generated from a normal distribution with mean 0 and variance σ_i^2; second, the random effects u_i are generated from a normal distribution with mean 0 and variance σ_u^2 taken equal to the estimated value obtained fitting a linear model with random area effects to the census data, that is $\sigma_u^2 = 56.23$ for all the iterations, under the hypothesis of non–informative design; then, using the values of the covariates $\mathbf{x}_i = (1, x_{1i}, x_{2i})$ obtained from the census together with the true vector of coefficients $\boldsymbol{\beta} = (-3.72, -0.0095, 0.51)$, the vector of responses is generated under a Fay-Herriot model. In a second alternative setting the steps of Monte Carlo experiment are the same as in the first setting but the vector **y** of responses is generated under the model (2.2), with γ random errors generated under a normal distribution with mean 0 and variance $\sigma_\gamma^2 = 15$.

Under both settings we consider $T = 500$ Monte Carlo samples and we compute the three MSE estimators of interest, setting the replicates of the two bootstrap procedures to $B = 250$; the final estimates are computed taking the mean over the replicates. The empirical values of the MSEs, that is the reference values, were computed previously under both settings with 5000 Monte Carlo replicates to ensure better accuracy. Figs. 1 and 2 display for each of the $m = 278$ small areas the ratios of the three estimated RMSE (analytical root mse^{ana}, naïve nonparametric bootstrap root mse^{naNPB} and combined analytical and bootstrap root mse^{bcNPB}) over the empirical values (represented by the straight lines), under the first and the second setting respectively. Note that to allow a better comparison of the results, the scale used in the two figures has been zoomed out to the interval 0.9-1.25.

The main result standing from the simulation results is that the two proposed bootstrap estimators of the MSE outperform the analytical one, under both settings. As regards the comparison between the estimator mse^{naNPB} and the estimator mse^{bcNPB}, the first seems to better follow the empirical values (see Fig. 1). This behavior is the same even considering the second setting, where the model used to generate the $\bar{y}_i$ values has a spline component: in this case we can observe a slightly higher variability of the estimates, while the estimators are more correct, as expected. Thus, the estimation of the g_3 term of the MSE seems to play an important role in this estimation context.

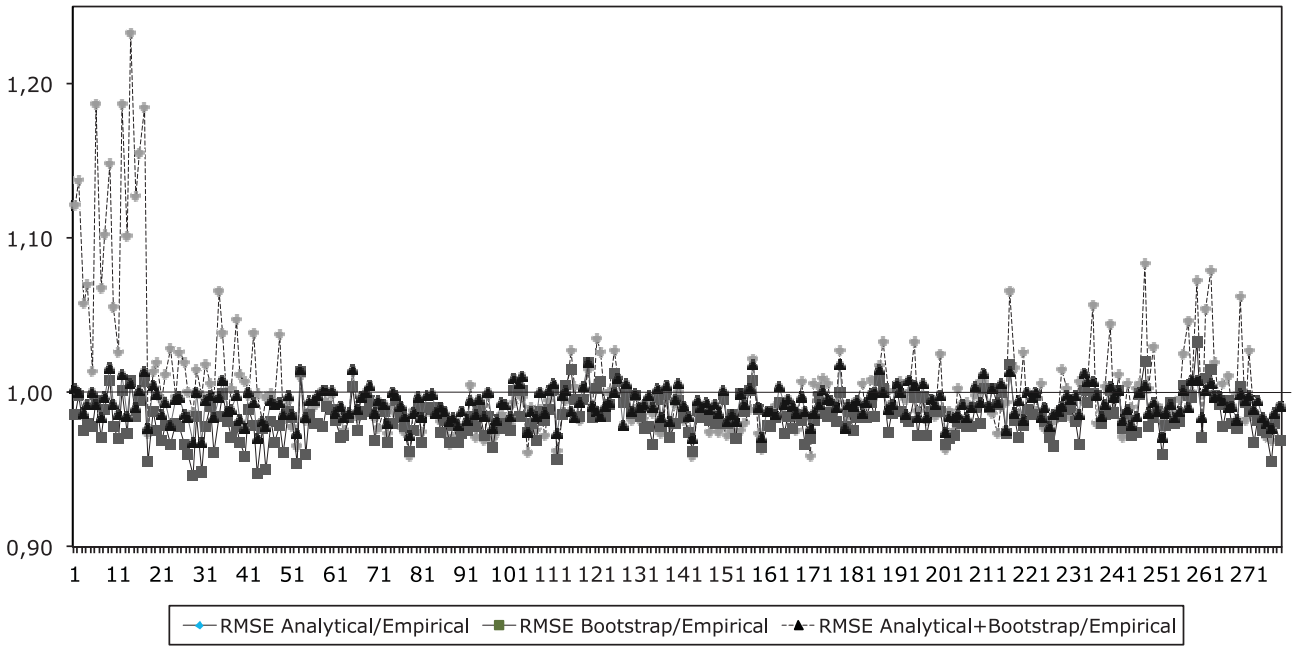

Fig. 1. Ratios of analytical Root Mean Squared Error (RMSE), naïve nonparametric bootstrap RMSE and combined analytical and bootstrap RMSE over empirical values for the m = 278 small areas, model with $\sigma_\gamma^2 = 0$.

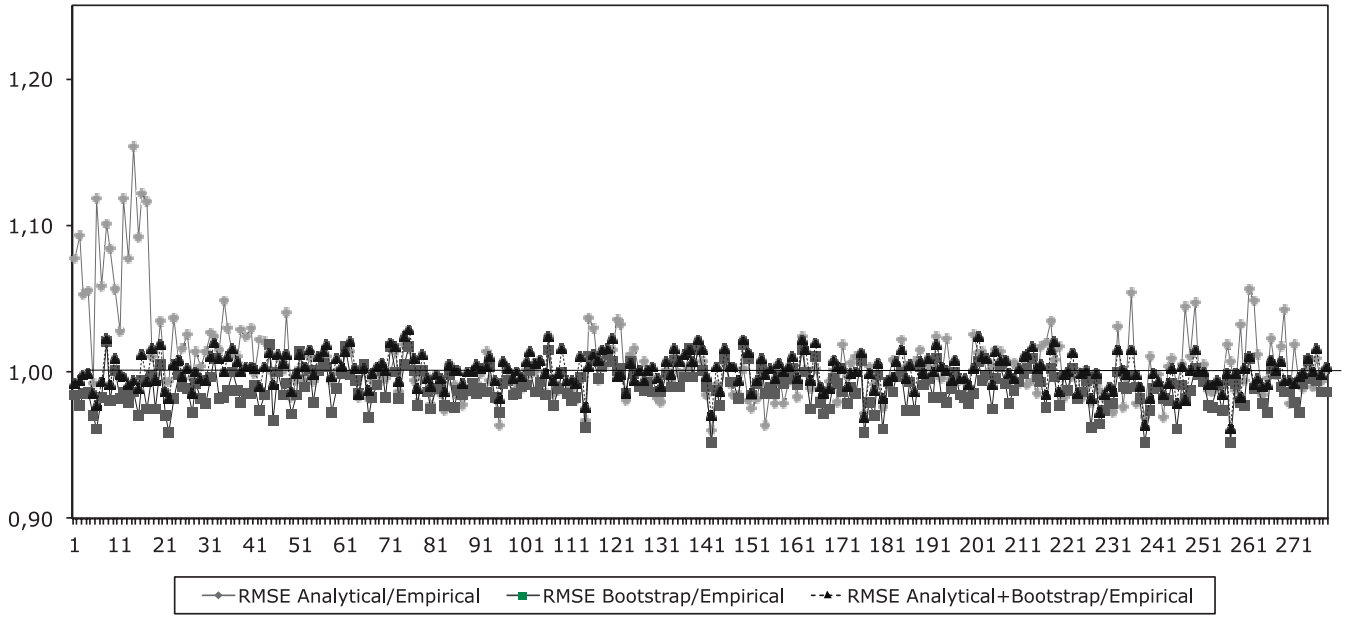

Fig. 2. Ratios of analytical Root Mean Squared Error (RMSE), naïve nonparametric bootstrap RMSE and combined analytical and bootstrap RMSE over empirical values for the m = 278 small areas, model with $\sigma_\gamma^2 = 15$.

6. APPLICATION

In this Section we estimate the mean Acid Binding Capacity (ANC) and Calcium (CA) concentration in 126 8-digit HUCs in the Mid-Atlantic area of the US using the proposed semiparametric Fay-Herriot model.

The ANC measures the buffering capacity of water against negative changes in pH-values and is often used as an indicator of the risk of acidification of water bodies in water resource surveys. An ANC level smaller than zero (measured in μeq/L, microequivalents per liter) means that the water is acidic; the higher the ANC, the larger the amount of acid a lake can neutralize before pH drops. Thus, low values of ANC identify critical situations for the streams water. The CA concentration is another relevant indicator: for example it has been argued that the distribution of some invasive and alien species like zebra mussels is associated with calcium concentration in surface waters. Thus, the development of a map showing CA concentrations in stream and waters could indicate the areas at risk of invasion (Whittier *et al.* 2008).

Previous studies (Opsomer *et al.* 2008; Pratesi *et al.* 2008) have shown the usefulness of a semiparametric specification to link the responses with available auxiliary information in this area of study. Direct estimates of ANC and CA concentrations were computed using data from two surveys of streams in the Mid-Atlantic States of the US (Stoddard *et al.* 2006). Covariate information at the area level was available from Jones *et al.* (1997), the Atlas of Environmental Assessment of the Mid-Atlantic region of the United States, done using measurements derived from satellite imagery and spatial data bases. Both data collections were available on the website www.epa.gov.

A semiparametric model to estimate the ANC mean in the areas of interest can be specified as:

$$s(\mathbf{x}) = m(x_1, x_2) + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5, \quad (6.1)$$

where x_1 and x_2 are the coordinates (latitude and longitude) of the centroid of the HUC, x_3 is the HUC proportion of total stream length that has roads within 30 meters, x_4 the proportion of watershed with potential soil loss greater than one ton per acre per year, x_5 the proportion of watershed area with suitable forest edge habitat (65 hectare scale), and $m(\cdot)$ is an unknown smooth bivariate function.

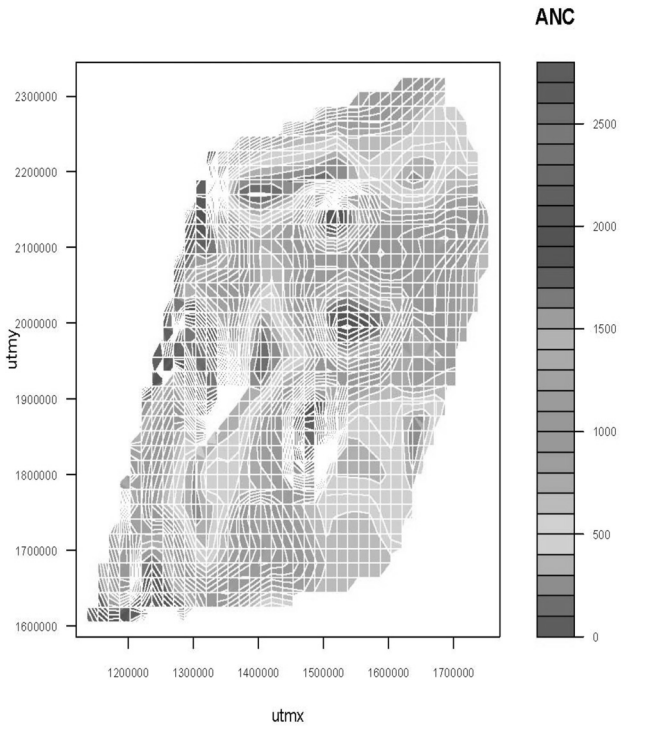

Fig. 3. Image of the bivariate spline effect in the region of interest for the ANC mean values using Universal Transverse Mercator coordinates.

To estimate the smooth bivariate function we use P-splines, assuming that $m(\cdot)$ can be approximated sufficiently well by model (2.4), where the last part of the model allows to handle nonlinearities in the structure of the relationship between the study variable and the covariates. Note that the choice of model (6.1) was driven by preliminary analyses specifying different combinations of parametric and non-parametric specifications for the available variables. Fig. 3 confirms the need of using a P-spline when specifying the relation between the direct mean ANC values and the coordinates. The same model specification was used for the estimation of the CA mean in the areas; in this case the variables that enter linearly in model (6.1) are: the human use index, that is the proportion of watershed area with agriculture or urban land cover (x_3 variable), the proportion of total stream length that has roads within 30 meters (x_4 variable) and the proportion of watershed area with suitable forest edge habitat (7 hectare scale, x_5 variable).

The choice of knots in two dimensions is more challenging than in one. To choose the number of knots K, Ruppert *et al.* (2003) suggest using K = max[20;

min(n/4; 150)] in two dimensions. Following these suggestions and performing the estimation with different number of knots, we finally chose K = 40, since we found that the approximation ability of the spline stabilizes after this number of knots for the estimation of ANC and CA means. As concerns the choice of the knots, two solutions suggested in the literature that provide a subset of observations nicely scattered to cover the domain are space filling designs (Nychka and Saltzman 1998) and the clara algorithm (Kaufman and Rousseeuw 1990, Chapter 3). The first one is based on the maximal separation principle of K points among the unique $\tilde{\mathbf{x}}_i$ and is implemented in the fields package of the R language (R Development Core Team 2005). As in Opsomer *et al.* (2008) we used in our application the second one, that is based on clustering and it selects K representative objects out of n; it is implemented in the package cluster of R.

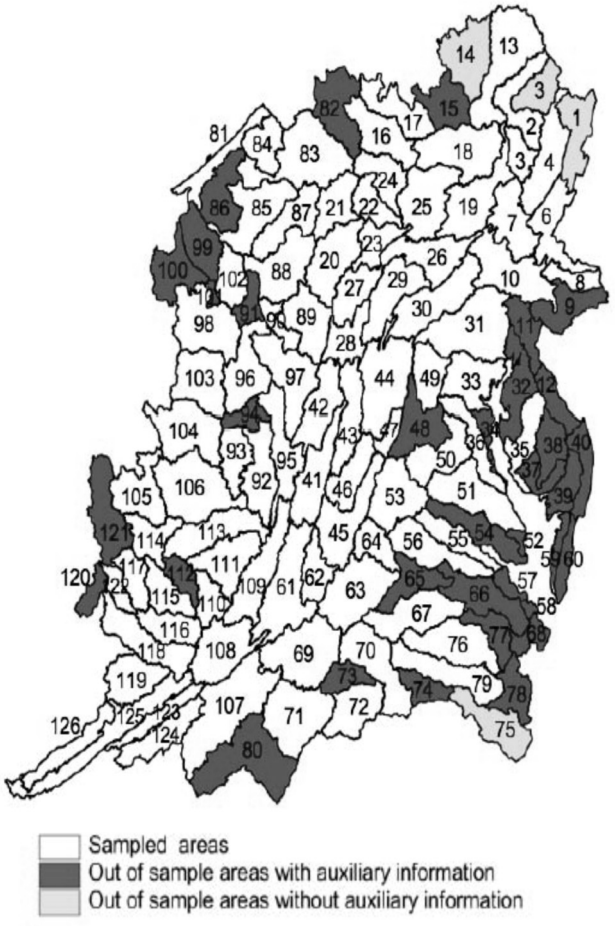

Fig. 4. Map of the small areas. The codes are used in Fig. 7 and Fig. 8 to report the confidence intervals.

It should be noted, then, that this estimating framework can be used to handle univariate smoothing and bivariate smoothing by suitably changing the parametric and the spline part of the model, i.e. once the **X** and the **Z** matrices are set up. Finally, other continuous or categorical variables can be easily inserted parametrically in the model by adding columns to the **X** matrix.

As concerns our estimation problem, note that between the HUCs of interest some are out of sample areas with available covariate information while for other HUCs also the covariate information is not available (see Fig. 4). This is because the target geographical areas of the EMAP surveys and of the landscape atlas of the Mid-Atlantic region do not coincide. For both types of out of sample HUCs we can predict the mean ANC and CA using the proposed semiparametric model.

Figs. 5 and 6 show the maps of the estimated mean ANC and CA concentration respectively. The areas with lower ANC mean values are spread in the

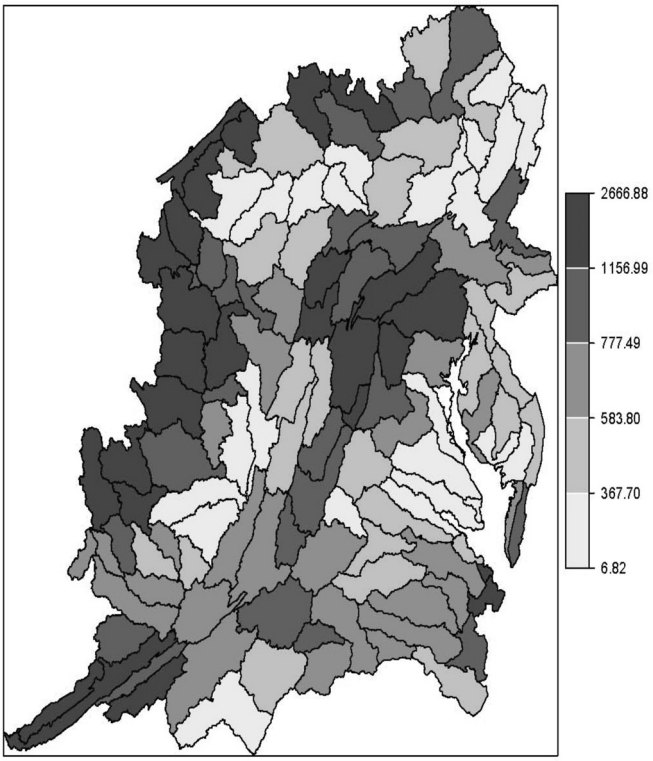

Fig. 5. Map of Mid-Atlantic Area with semiparametric estimates for average ANC.

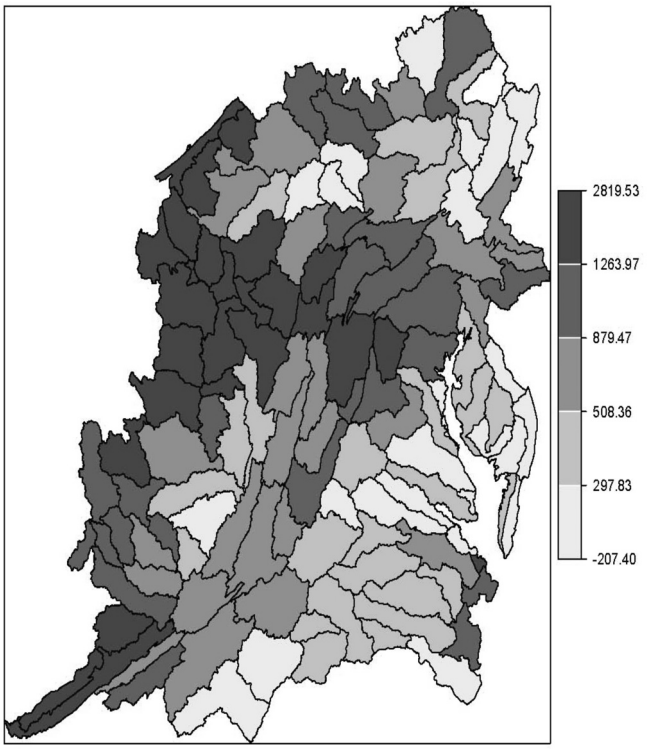

Fig. 6. Map of Mid-Atlantic Area with semiparametric estimates for average CA.

Mid-Atlantic region: some of them are on the coast, others in the internal regions corresponding to the North and Central Appalachians ecoregions. The higher CA concentration levels are estimated for the HUCs in the Western Allegheny Plateau, Central Appalachian and Ridge and Valley ecoregions. These results are coherent with previous studies on CA concentration in the Mid-Atlantic region, where however the availability of the data at a wider detail did not suffice to classify all the areas of interest (Whittier *et al.* 2008).

Using the nonparametric bootstrap estimator (4.1) we computed the confidence intervals for the mean ANC and CA estimates in each area of interest. Figs. 7 and 8 represent the mean values with corresponding bootstrapped 95% confidence intervals, with the HUCs ordered by increasing estimated values of the variable of interest. Looking at these results in connection with Fig. 4 we can see that, as expected, the wider confidence intervals are usually obtained for the out of sample areas or for areas at the boundary of the region of interest, where the calculated indicators are probably not as reliable as the indicators calculated

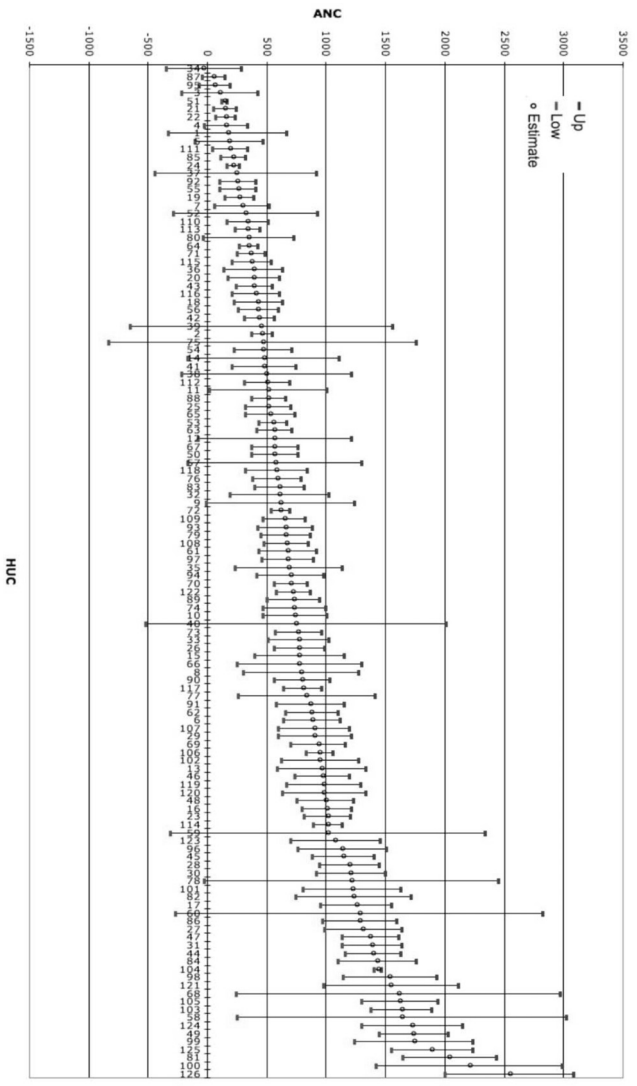

Fig. 7. Estimate of the average ANC and bootstrap based 95% confidence interval for the 126 HUCs. Numbers denoting HUCs can be paired with the map in Fig. 4.

for watersheds that had complete data coverage (Jones *et al.* 1997).

7. CONCLUSIONS

In this paper we have proposed a semiparametric version based on P-splines of the basic area level Fay-Herriot model. This model can be used in situations where the information is available only at the area level and the functional form of the relationship between the variable of interest and the covariates cannot be specified a priori. Furthermore, spatial effects can easily be introduced in the proposed

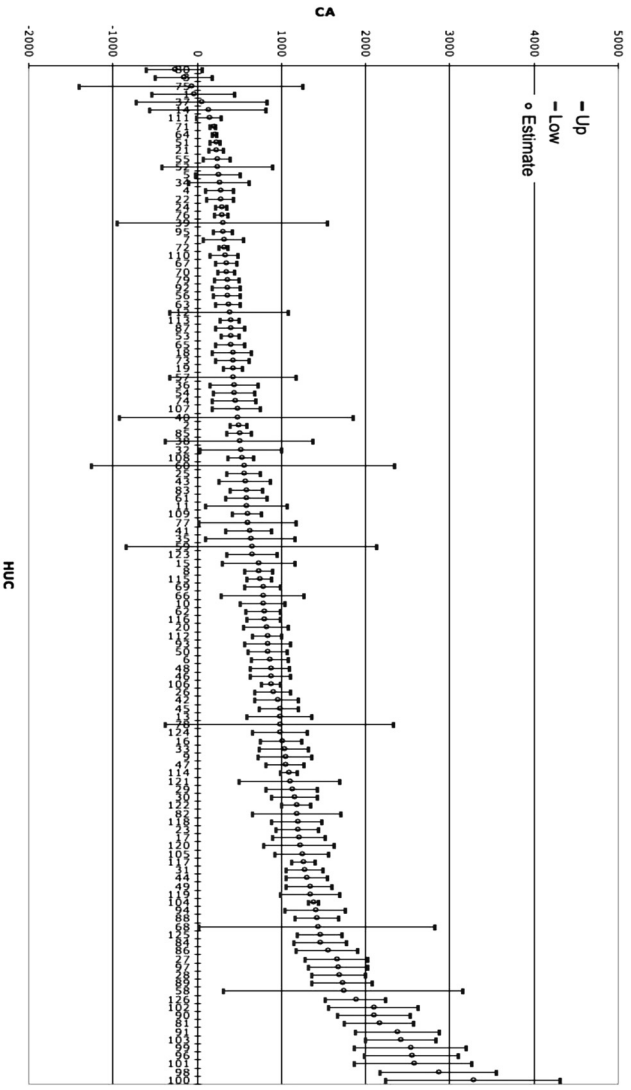

Fig. 8. Estimate of the average CA and bootstrap based 95% confidence interval for the 126 HUCs. Numbers denoting HUCs can be paired with the map in Fig. 4.

semiparametric area level model, thus making this model very useful for environmental studies where geo-referenced information is usually available.

The results of the model based simulation study suggest that the proposed small area mean estimator allows us to obtain an appreciable improvement of the estimates under many different hypothesis for the relation of the study variable with the covariates; moreover, the semiparametric estimator is still competitive also when the linear assumption of the traditional Fay-Herriot model holds. In addition, the design based simulation study has shown the good performance of the two proposed estimators of the MSE, both based on resampling techniques.

Using the proposed semiparametric model we have considered the problem of estimating the mean Acid Binding Capacity (ANC) and Calcium (CA) concentration in 126 8-digit HUCs in the Mid-Atlantic area of the US using EPA surveys data. Note that in this case the auxiliary variables are available at the area level and that geographic information (latitude and longitude of the HUCs) has been included in the model to take into account spatial proximity effects. The results we obtained are concordant with those obtained in previous studies were however the use of direct estimators did not suffice to classify all the areas of interest.

Until now there is no contribution on the relationship between the expected results obtained modeling the same data set at unit level and at area level. Besides, it is well known that the spatial relation active at unit level is not likely to be the same when the analysis is done at a more aggregate level. However, we did not approach this problem here. In future analysis we will address these relevant issues on the same case study comparing the results obtained using our proposed semiparametric area level model with those obtained using the nonparametric unit level proposed by Opsomer *et al.* (2008).

ACKNOWLEDGEMENTS

This work is financially supported by the European Project SAMPLE “Small Area Methods for Poverty and Living Condition Estimates”, funded by the European Commission, 7th FP (www.sample-project.eu).

REFERENCES

Das, K., Jiang, J. and Rao, J. (2004). Mean squared error of empirical predictor. *Ann. Stat.*, **32**, 818-840.

Datta, G., Rao, J. and Smith, D. (2005). On measuring the variability of small area estimator under a basic area level model. *Biometrika*, **92**, 183-196.

Fay, R. and Herriot, R. (1979). Estimates of income for small places: An application of James-Stein procedures to census data. *J. Amer. Statist. Assoc.*, **74**, 269-277.

French, J., Kammann, E. and Wand, M. (2001). Comment on paper by Ke and Wang. *J. Amer. Statist. Assoc.*, **96**, 1285-1288.

González-Manteiga, W., Lombarda, M., Molina, I., Morales, D. and Santamara, L. (2007). Estimation of the mean squared error of predictors of small area linear parameters under logistic mixed model. *Comp. Statist. Data Anal.*, **51**, 2720–2733.

Green, P.J. and Silverman, B.W. (1994). *Nonparametric Regression and Generalized Linear Models: A Roughness Penalty Approach*. Chapman & Hall Ltd.

Henderson, C. (1975). Best linear unbiased estimation and prediction under a selection model. *Biometrics*, **31**, 423-447.

Jones, K., Riitters, K., Wickham, J., Tankersley Jr., R., O'Neill, R., Chaloud, D., Smith, E. and Neale, A. (1997). An Ecological Assessment of the United States Mid-Atlantic Region: A Landscape Atlas. EPA-United States Environmental Protection Agency, Washington, DC.

Kammann, E.E. and Wand, M.P. (2003). Geoadditive models. *J. Roy. Statist. Soc.*, **C52**, 1-18.

Kaufman, L. and Rousseeuw, P. (1990). *Finding Groups in Data: An Introduction to Cluster Analysis*. Wiley, New York.

Molina, I., Salvati, N. and Pratesi, M. (2009). Bootstrap for estimating the MSE of the Spatial EBLUP. *Comp. Statist*, **24**, 441-458.

Nychka, D. and Saltzman, N. (1998). Design of air quality monitoring networks. In: *Nychka, Douglas, Piegorsch, Walter W. and Cox, Lawrence H. (eds), Case studies in environmental statistics*, pages 51-76.

Opsomer, J.D., Claeskens, G., Ranalli, M.G., Kauermann, G., and Breidt, F.J. (2008). Nonparametric small area estimation using penalized spline regression. *J. Roy. Statist. Soc.*, **B70**, 265-286.

Petrucci, A. and Salvati, N. (2006). Small area estimation for spatial correlation in watershed erosion assessment. *J. Agri., Bio. Envir. Statist.*, **11**, 169-182.

Pfeffermann, D. and Tiller, R. (2006). Bootstrap approximation to prediction mse for state-space models with esti-mated parameters. *J. Time Series Anal.*, **26**, 893-916.

Prasad, N. and Rao, J. (1990). The estimation of mean squared error of small-area estimators. *J. Amer. Statist. Assoc.*, **85**, 163-171.

Pratesi, M., Ranalli, M.G., and Salvati, N. (2008). Semiparametric M-quantile regression for estimating the proportion of acidic lakes in 8-digit HUCs of the Northeastern US. *Environmetrics*, **19**, 659-764.

Pratesi, M. and Salvati, N. (2009). Small area estimation in the presence of correlated random area effects. *J. Official Statist.*, **25**, 37-53.

R Development Core Team (2005). R: *A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

Rao, J.N.K. (2003). *Small Area Estimation*. John Wiley & Sons.

Ruppert, D., Wand, M.P. and Carroll, R. (2003). *Semiparametric Regression*. Cambridge University Press, Cambridge, New York.

Stoddard, J., Herlihy, A., Hill, B.H., Hughes, R., Kaufmann, P., Klemm, D., Lazorchak, J., McCormick, F., Peck, D., Paulsen, S., Olsen, A., Larsen, D., Van Sickle, J. and Whittier, T. (2006). *Mid-Atlantic Integrated Assessment (MAIA). State of the Flowing Waters Report*. EPA -United States Environmental Protection Agency, Washington, DC.

Whittier, T.M., Ringold, P.L., Herlihy, A.T. and Pierson, S.M. (2008). A calcium-based invasion risk assessment for zebra and quagga mussels (*Dreissena* spp). *Frontiers Ecol. Environ.*, **6(4)**, 180-184.