Computational Statistics and Data Analysis 53 (2009) 3616–3629

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

Spline smoothing in small area trend estimation and forecasting

M.D. Ugarte [a,*], T. Goicoa [a], A.F. Militino [a], M. Durbán [b]

[a] Departamento de Estadística e Investigación Operativa, Universidad Pública de Navarra. Campus de Arrosadía, 31006 Pamplona, Spain

[b] Departamento de Estadística, Universidad Carlos III de Madrid, Escuela Politécnica Superior, Campus de Leganés, Madrid, Spain

ARTICLE INFO

Article history:
Received 30 July 2008
Received in revised form 25 February 2009
Accepted 26 February 2009
Available online 9 March 2009

ABSTRACT

Semiparametric models combining both non-parametric trends and small area random effects are now currently being investigated in small area estimation (SAE). These models can prevent bias when the functional form of the relationship between the response and the covariates is unknown. Furthermore, penalized spline regression can be a good tool to incorporate non-parametric regression models into the SAE techniques, as it can be represented as a mixed effects model. A penalized spline model is considered to analyze trends in small areas and to forecast future values of the response. The prediction mean squared error (MSE) for the fitted and the predicted values, together with estimators for those quantities, are derived. The procedure is illustrated with real data consisting of average prices per squared meter of used dwellings in nine neighborhoods of the city of Vitoria, Spain, during the period 1993–2007. Dwelling prices for the next five years are also forecast. A simulation study is conducted to assess the performance of both the small area trend estimator and the prediction MSE estimators. The results confirm a good behavior of the proposed estimators in terms of bias and variability.

1. Introduction

Small area estimation techniques have experienced a quick evolution in the last few years, motivated by the necessity of precise information for small domains. As the sample size for these regions is very scarce, traditional design-based methods are not appropriate and model-based techniques have been widely applied. The mixed model approach to small area estimation and the use of the empirical best linear unbiased predictor (EBLUP) have been thoroughly studied (Rao, 2003), particularly the calculation of the mean squared error (MSE) of the EBLUP, one of the most challenging problems in this field.

Different models have been investigated in small area estimation and researchers have come across new problems related to the estimation of the MSE. One such model consists of a combination of small area random effects and a non-parametric specified trend. Using penalized splines (P-splines) as a representation of the non-parametric trend can be very advantageous because P-splines are very flexible and they have some useful properties. Firstly, they do not show boundary effects as other types of smoothers, that is, they do not bend toward zero outside the domain of the data. This is very useful for forecasting purposes. Secondly, they are low rank smoothers, that is, the size of the basis is smaller than the dimension of the data, and then they present computational advantages. Finally, the choice of knots is not a very important matter due to the introduction of the penalty. In addition, one remarkable feature of P-splines is that they can be represented as mixed effects models, and then they benefit from the existing theory and the well-developed software for mixed model analysis. For a connection between smoothing splines and mixed models, see for example Brumback et al. (1999), Coull et al. (2001), Parise et al. (2001), Aerts et al. (2002) or Wand (2003) and the references therein. However, a particular problem in this context that

* Corresponding author. Tel.: +34 948169202; fax: +34 948169204.
E-mail address: lola@unavarra.es (M.D. Ugarte).

0167-9473/$ – see front matter © 2009 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2009.02.027

has not been treated in depth relates to the estimation of the prediction mean squared error (MSE). Expressions commonly used to estimate the prediction MSE do not usually take into account the uncertainty associated to the estimation of the variance components. This matter is particularly interesting in this setting where the variance–covariance matrix of the data is not block-diagonal. This means that the results provided by Prasad and Rao (1990) and Das et al. (2004) concerning prediction of the MSE in the small area literature do not apply directly. Very recently, Opsomer et al. (2008) proposed a second order correct approximation to the MSE for the EBLUP of the small area mean under a semiparametric model. They also derive an appropriate estimator of that quantity. Pratesi et al. (2008) consider a semiparametric M-quantile regression model for small area estimation and bootstrap techniques for the MSE estimation.

In this paper, a semiparametric longitudinal model is proposed with the objective of forecasting future values of the response within the small areas. The model combines both a non-parametric time trend and a specific random effect for each area. Penalized splines and their representation as mixed effects models will be used for obtaining the EBLUP for particular observations within the small areas. The prediction MSE of both fitted and forecast values, as well as estimators of those quantities, are derived. The estimators take account of the uncertainty due to the estimation of the variance components and the variability of the particular observations, and include a bias correction term to make them consistent for the prediction MSE.

The main contributions of this paper relative to the work by Opsomer et al. (2008) can be summarized in three main points. First, B-splines basis are used in the P-spline model. Then, the representation of the P-splines as a mixed effect model is not as straightforward as in the case of truncated polynomial or radial basis. However, B-spline basis have more stable numerical properties and their use is advisable (see page 70 of Ruppert et al. (2003)). Second, a longitudinal model is considered to deal with the temporal trend, and an expression for the forecast values is provided from the P-spline mixed model representation. This requires extending the B-spline basis and to consider an augmented mixed model. Finally, the MSE of the predicted and forecast values is provided. As we are not predicting the small area mean or total (which is typical in small area estimation) but particular values, an extra term related to the variability of the particular values is obtained. A comparison with the variability measure widely used in the P-spline literature is provided. The procedure is illustrated with real data consisting of average prices per squared meter of used dwellings in nine neighborhoods of the city of Vitoria, Spain, during the period 1993–2007. Dwelling prices for the next five years are also forecast. The housing problem in Spain has become very important in the last ten years. The increased demand for dwelling in combination with low interest rate mortgages has created an environment conducive to property speculation which in turn has driven real estate prices to new record highs. This has also been strengthened by the traditional Spanish preference for buying rather than renting a house. Several hedonic models analyzing Spanish dwelling prices have been considered in the literature (see for example, Militino et al. (2004) and Ugarte et al. (2004)). The procedures developed in this paper are used to study dwelling price evolution in small areas. Concerning the real application, the local Government is interested in monitoring housing prices in the different neighborhoods of the city to analyze the extent of the crisis and the adjustment of the housing market. This is necessary because the recent global economical crisis and the reluctance of the banks to lend money are leading to a downturn in the Spanish housing market. Unfortunately, this is becoming a global problem nowadays.

The plan of the paper is as follows. Section 2 briefly reviews the theory of P-splines with B-spline basis, the use of different basis, and the mixed model representation of penalized splines. Section 3 presents a semiparametric model for small area trend estimation and prediction. In Section 4, the prediction MSE for the fitted and the predicted values, together with analytical and bootstrap estimators for those quantities, are derived. Section 5 discusses hypothesis testing procedures. In Section 6, the procedures developed in the paper are illustrated with the real data set. In Section 7, a simulation study is conducted to assess the performance of the small area trend estimator and the MSE estimators. The article ends with a discussion.

2. Penalized splines

Penalized splines or P-splines were popularized by Eilers and Marx (1996) and their use has increased during the last few years (see Ruppert et al. (2003) for a complete account of the use of P-splines). In this section the theory of P-splines with B-splines basis is briefly reviewed. Let us assume paired data (y_i, x_i) and let us consider the model

$$y_i = f(x_i) + \epsilon_i = a_0 B_0(x_i) + a_1 B_1(x_i) + \cdots + a_K B_K(x_i) + \epsilon_i, \quad i = 1, \ldots n,$$

or in matrix form

$$\mathbf{Y} = \mathbf{f} + \boldsymbol{\epsilon} = \mathbf{Ba} + \boldsymbol{\epsilon}, \tag{1}$$

where $\mathbf{f} = \mathbf{Ba}$ is a smooth function to be estimated using P-splines, $\mathbf{B} = (\mathbf{B}_0, \ldots, \mathbf{B}_K)$ is the matrix of the spline basis obtained from the covariate $\mathbf{x}$, $\mathbf{a}' = (a_0, \ldots, a_K)$ is the vector of the basis coefficients, $\boldsymbol{\epsilon}' = (\epsilon_1, \ldots, \epsilon_n)' \sim N(\mathbf{0}, \sigma^2 \mathbf{I}_n)$, and $\mathbf{I}_n$ is the identity matrix. The P-spline approach minimizes the penalized sum of squares

$$\mathbf{S}(\mathbf{a}; \mathbf{Y}, \lambda) = \sum_{i=1}^{n} (y_i - f(x_i))^2 + \lambda \mathbf{a}'\mathbf{D}'\mathbf{Da} = (\mathbf{Y} - \mathbf{Ba})'(\mathbf{Y} - \mathbf{Ba}) + \lambda \mathbf{a}'\mathbf{D}'\mathbf{Da}, \tag{2}$$

where $\mathbf{Y}$ is the vector of sampled observations and $\mathbf{D}'\mathbf{D}$ is a penalty matrix imposing smoothness on the adjacent coefficients. The matrix $\mathbf{D}$ is a difference matrix of order d. An example for $d = 2$ and a vector of five coefficients is

$$\mathbf{D} = \begin{pmatrix} 1 & -2 & 1 & 0 & 0 \\ 0 & 1 & -2 & 1 & 0 \\ 0 & 0 & 1 & -2 & 1 \end{pmatrix}.$$

Minimizing the penalized equation (2) for a fixed λ leads to the solution

$$\hat{\mathbf{a}} = (\mathbf{B}'\mathbf{B} + \lambda\mathbf{D}'\mathbf{D})^{-1}\mathbf{B}'\mathbf{Y}, \tag{3}$$

and the fitted values are calculated as

$$\hat{\mathbf{Y}} = \mathbf{B}\hat{\mathbf{a}} = \mathbf{B}(\mathbf{B}'\mathbf{B} + \lambda\mathbf{D}'\mathbf{D})^{-1}\mathbf{B}'\mathbf{Y} = \mathbf{H}\mathbf{Y},$$

where $\mathbf{H} = \mathbf{B}(\mathbf{B}'\mathbf{B} + \lambda\mathbf{D}'\mathbf{D})^{-1}\mathbf{B}'$. The matrix $\mathbf{H}$ is commonly called a hat matrix, and it is very important because its trace gives a measure of the effective degrees of freedom, df, of the model (Hastie and Tibshirani, 1990, p. 52).

The parameter λ determines the influence of the penalty matrix. Note that it is not estimated from the data. A common way to optimize the smoothing parameter is using cross-validation techniques (see Hastie and Tibshirani (1990), p. 43). Standard errors for $\hat{\mathbf{a}}$ might be calculated from the covariance matrix

$$\widehat{\mathrm{var}}(\hat{\mathbf{a}}) = \hat{\sigma}^2(\mathbf{B}'\mathbf{B} + \lambda\mathbf{D}'\mathbf{D})^{-1}\mathbf{B}'\mathbf{B}(\mathbf{B}'\mathbf{B} + \lambda\mathbf{D}'\mathbf{D})^{-1},$$

where $\hat{\sigma}^2 = \sum_{i=1}^{n}\hat{\epsilon}_i^2/(n - df)$. However, inferences based on this covariance matrix generally give poor results (Wood, 2006). In the smoothing spline context, Wahba (1983) and Silverman (1985) solve this problem by using a Bayesian approach. The posterior covariance for the vector of parameters $\mathbf{a}$ is estimated by

$$\widehat{\mathrm{var}}(\hat{\mathbf{a}}) = \hat{\sigma}^2(\mathbf{B}'\mathbf{B} + \lambda\mathbf{D}'\mathbf{D})^{-1}.$$

Standard errors for the fitted values are obtained as the squared root of the diagonal elements of

$$\widehat{\mathrm{var}}(\hat{\mathbf{Y}}) = \widehat{\mathrm{var}}(\mathbf{B}\hat{\mathbf{a}}) = \hat{\sigma}^2\mathbf{B}(\mathbf{B}'\mathbf{B} + \lambda\mathbf{D}'\mathbf{D})^{-1}\mathbf{B}'. \tag{4}$$

2.1. Basis, knots, and mixed model representation

P-spline regression depends on a regression basis and a penalty matrix. There is a considerable amount of literature on the construction of bases (see for example Eilers et al. (2006), Wand (2003) and Wood (2003) among others). Each of them has advantages and disadvantages, and their performance is different depending on the data. Truncated polynomial bases have been widely used in the literature, and the representation of P-splines with truncated polynomial bases as a mixed model is well established (see for example Wand (2003)). However, in this paper B-splines bases are considered because they have better numerical properties, and penalized splines with B-spline bases can also be represented as mixed effects models (Eilers, 1999; Currie and Durbán, 2002). B-spline bases are constructed from polynomial pieces joined at certain values of x, called knots. Once the knots are given, it is easy to compute B-splines using a recursive algorithm (de Boor, 1978). The number and position of knots is not fixed. Usually the knots are located at equally spaced sample quantiles of x and the number of knots is chosen according to the rule: min{number of unique values of $x/4$, 40} (Ruppert, 2002).

The mixed model representation for penalized splines with a B-spline basis $\mathbf{B}$ is based on a reparametrization of Model (1). A transformation $\mathbf{T}$ is needed, such that $\mathbf{BT} = [\mathbf{X} : \mathbf{Z}]$, so one can reparameterize $\mathbf{Ba} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Zu}$, where $\mathbf{a} = \mathbf{T}\begin{pmatrix}\boldsymbol{\beta}\\ \mathbf{u}\end{pmatrix}$ (or $\begin{pmatrix}\boldsymbol{\beta}\\ \mathbf{u}\end{pmatrix} = \mathbf{T}^{-1}\mathbf{a}$). Here, $\boldsymbol{\beta}$ is a set of unpenalized parameters (fixed effects) and $\mathbf{u}$ are penalized coefficients (random effects). This transformation is not unique, and the one commonly used is based on the singular value decomposition of the penalty matrix

$$\mathbf{D}'\mathbf{D} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{U}'.$$

Here $\mathbf{U} = [\mathbf{U}_1 : \mathbf{U}_2]$ is a matrix whose columns are the singular vectors of $\mathbf{D}'\mathbf{D}$, $\mathbf{U}_1$ and $\mathbf{U}_2$ are the matrices of singular vectors corresponding to the null and non-zero eigenvalues respectively, and $\boldsymbol{\Sigma} = \mathrm{diag}(0, \ldots, 0, \tau_1, \ldots, \tau_c)$ is a diagonal matrix whose elements are the eigenvalues of $\mathbf{D}'\mathbf{D}$ (note that a penalty of order d, has d eigenvalues equal to zero). Then,

$$\mathbf{T} = \mathbf{U}\begin{pmatrix}\mathbf{I}_d & \\ & \widetilde{\boldsymbol{\Sigma}}^{-1/2}\end{pmatrix} = [\mathbf{U}_1 : \mathbf{U}_2\widetilde{\boldsymbol{\Sigma}}^{-1/2}],$$

where $\widetilde{\boldsymbol{\Sigma}}$ is a matrix whose diagonal elements are the non-zero eigenvalues of the penalty. Then,

$$\mathbf{Ba} = \mathbf{BT}\begin{pmatrix}\boldsymbol{\beta}\\ \mathbf{u}\end{pmatrix} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Zu},$$

with

$$\mathbf{X} = \mathbf{BU}_1, \qquad \mathbf{Z} = \mathbf{BU}_2\widetilde{\boldsymbol{\Sigma}}^{-1/2}.$$

Using this transformation, Eq. (2) can be written as

$$\begin{aligned}\mathbf{S}(\boldsymbol{\beta}, \mathbf{u}; \mathbf{Y}, \lambda) &= (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\mathbf{u})'(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\mathbf{u}) + \lambda \begin{pmatrix}\boldsymbol{\beta}' & \mathbf{u}'\end{pmatrix} \mathbf{T}'\mathbf{D}'\mathbf{D}\mathbf{T} \begin{pmatrix}\boldsymbol{\beta} \\ \mathbf{u}\end{pmatrix}, \\ &= (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\mathbf{u})'(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\mathbf{u}) + \lambda \mathbf{u}'\mathbf{u}. \end{aligned} \tag{5}$$

Dividing Eq. (5) by σ_e^2, setting $\lambda = \sigma_e^2/\sigma_u^2$, and differentiating with respect to $\boldsymbol{\beta}$ and $\mathbf{u}$, the mixed model equations corresponding to the mixed model

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \boldsymbol{\epsilon}, \quad \mathbf{u} \sim N(\mathbf{0}, \sigma_u^2\mathbf{I}), \boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma_e^2\mathbf{I})$$

are attained.

Note that, although $\mathbf{X} = \mathbf{B}\mathbf{U}_1$, any matrix of rank d whose columns are linearly independent from those of $\mathbf{Z}$ can be used. Here, for simplicity, the following matrix is considered

$$\mathbf{X} = \begin{pmatrix} 1 & x_1 & x_1^2 & \cdots & x_1^{d-1} \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ 1 & x_n & x_n^2 & \cdots & x_n^{d-1} \end{pmatrix}, \tag{6}$$

where d is the order of the differences in the penalty matrix. Observe that this is the common matrix used in ordinary regression, and so the linear model is a special case of the spline mixed model when $\sigma_u^2 = 0$.

The representation of the penalized spline regression as a mixed model allows the estimation of the smoothing parameter as $\hat{\lambda} = \hat{\sigma}_e^2/\hat{\sigma}_u^2$ using restricted maximum likelihood (REML). Then, the uncertainty arising from the estimation of the variance components has to be taken into account, unlike Expression (4), which considers λ as a known parameter.

3. A semiparametric small area model

In this section, a semiparametric model combining both a non-parametric trend and a small area random effect is considered. Namely

$$y_{ij} = f(x_{ij}) + v_i + \epsilon_{ij}, \quad v_i \sim N(0, \sigma_v^2), \epsilon_{ij} \sim N(0, \sigma_e^2), i = 1, \ldots, I, j = 1, \ldots, n_i, \tag{7}$$

where in the ith area, y_{ij} and x_{ij} are the j-values of the response and the explanatory variables respectively, v_i is the small area effect, and ϵ_{ij} is the error term. Note that the total sample size is $n = \sum_i n_i$. Using P-splines, Model (7) can be equivalently written as the following mixed effects model

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \mathbf{W}\mathbf{v} + \boldsymbol{\epsilon}, \tag{8}$$

where $\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u}$ represents the spline function, and $\mathbf{W}\mathbf{v}$ corresponds to the small area random effect. Here, $\mathbf{Y} = (\mathbf{y}_1', \ldots, \mathbf{y}_I')'$, $\mathbf{y}_i' = (y_{i1}, \ldots, y_{in_i})$, $\mathbf{X}$ is the fixed effect matrix coming from the mixed model representation of the spline, and $\mathbf{Z}$ is the design matrix of the spline random effect. Note that matrices $\mathbf{X}$ and $\mathbf{Z}$ take different forms depending on the spline basis. The matrix $\mathbf{W} = \text{diag}(\mathbf{1}_{n_1}, \ldots, \mathbf{1}_{n_I})$ is the matrix corresponding to the small area random effect $\mathbf{v}$, and $\mathbf{1}_{n_i}$'s are columns of ones. Finally, $\mathbf{u} \sim N(\mathbf{0}, \boldsymbol{\Gamma}_u = \sigma_u^2\mathbf{I}_L)$, where L is the number of columns of $\mathbf{Z}$, $\mathbf{v} \sim N(\mathbf{0}, \boldsymbol{\Gamma}_v = \sigma_v^2\mathbf{I}_I)$, and $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \boldsymbol{\Gamma}_e = \sigma_e^2\mathbf{I}_n)$. Then, the covariance matrix of the variable $\mathbf{Y}$ is given by $\mathbf{V} = \mathbf{Z}\boldsymbol{\Gamma}_u\mathbf{Z}' + \mathbf{W}\boldsymbol{\Gamma}_v\mathbf{W}' + \boldsymbol{\Gamma}_e$. If the variance components are known, the best linear unbiased estimator (BLUE) of $\boldsymbol{\beta}$ is obtained as (Henderson et al., 1959)

$$\tilde{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{Y}, \tag{9}$$

and the best linear unbiased predictors (BLUP) of $\mathbf{u}$ and $\mathbf{v}$ are given by (see for example, Henderson (1963)and McCulloch and Searle (2001))

$$\begin{aligned}\tilde{\mathbf{u}} &= \boldsymbol{\Gamma}_u\mathbf{Z}'\mathbf{V}^{-1}(\mathbf{Y} - \mathbf{X}\tilde{\boldsymbol{\beta}}), \\ \tilde{\mathbf{v}} &= \boldsymbol{\Gamma}_v\mathbf{W}'\mathbf{V}^{-1}(\mathbf{Y} - \mathbf{X}\tilde{\boldsymbol{\beta}}). \end{aligned} \tag{10}$$

Usually, in small area estimation, the quantity of interest is the small area mean (or small area total)

$$\bar{y}_i = \bar{\mathbf{x}}_i\boldsymbol{\beta} + \bar{\mathbf{z}}_i\mathbf{u} + v_i, \tag{11}$$

where $\bar{\mathbf{x}}_i$ and $\bar{\mathbf{z}}_i$ are the true means of the rows of $\mathbf{X}$ and $\mathbf{Z}$ corresponding to the ith small area respectively. Note that Eq. (11) ignores the mean of the errors ϵ_{ij}, something which is usually done in small area estimation, since $E[\epsilon_{ij}] = 0$. Then, an estimator is given by

$$\tilde{\bar{y}}_i = \bar{\mathbf{x}}_i\tilde{\boldsymbol{\beta}} + \bar{\mathbf{z}}_i\tilde{\mathbf{u}} + \tilde{v}_i.$$

However, in this paper, the interest relies on predicting the value of particular observations in the small areas

$$y_{ij} = f(x_{ij}) + v_i + \epsilon_{ij} = \mathbf{x}_{ij}\boldsymbol{\beta} + \mathbf{z}_{ij}\mathbf{u} + \mathbf{w}_{ij}\mathbf{v} + \epsilon_{ij},$$

where $\mathbf{x}_{ij}$ and $\mathbf{z}_{ij}$ are the jth rows of the $\mathbf{X}$ and $\mathbf{Z}$ matrices corresponding to the ith small area, and $\mathbf{w}_{ij}$ is a vector with 1 in the ith position and zeros elsewhere. Note that in this case, the error term is considered because we are not predicting a mean, but a particular observation. The BLUP predictor is given by

$$\tilde{y}_{ij} = \mathbf{x}_{ij}\tilde{\boldsymbol{\beta}} + \mathbf{z}_{ij}\tilde{\mathbf{u}} + \mathbf{w}_{ij}\tilde{\mathbf{v}}. \tag{12}$$

The derivation of predictor (12) assumes that the variance components σ_u^2, σ_v^2, and σ_e^2 are known. However, this rarely happens in practice and they have to be estimated from the data. Well-known methods, such as REML, can be used to estimate the variance components (see for example Searle et al. (1992)). Then, the estimator $\hat{\boldsymbol{\beta}}$ and the predictors $\hat{\mathbf{u}}$ and $\hat{\mathbf{v}}$ are obtained from Eqs. (9)–(10) replacing the true variances by their sample estimates. Similarly, an empirical best linear unbiased predictor (EBLUP) for a particular observation is obtained from Eq. (12) as

$$\hat{y}_{ij} = \mathbf{x}_{ij}\hat{\boldsymbol{\beta}} + \mathbf{z}_{ij}\hat{\mathbf{u}} + \mathbf{w}_{ij}\hat{\mathbf{v}}. \tag{13}$$

3.1. Predicting new observations

One of the main interests of this paper is the prediction of new observations, in particular future observations, using the semiparametric model (8). Currie et al. (2004) propose a method for fitting and forecasting simultaneously with P-splines models when the coefficients are estimated, as in Eq. (3). However, the small area model (7) includes a small area random effect and formula (3) cannot be applied. Here, the mixed model expression (8) will be used for predicting new observations. The prediction process is as follows. Suppose we wish to predict n_1 new values. Then, the set of knots used to compute the basis $\mathbf{B}$ has to be extended to include n_1 new values, and the predicted value is

$$\hat{y}_{i0} = \mathbf{x}_{i0}\hat{\boldsymbol{\beta}} + \mathbf{z}_{i0}\hat{\mathbf{u}} + \mathbf{w}_{i0}\hat{\mathbf{v}},$$

where $\mathbf{x}_{i0}$ and $\mathbf{z}_{i0}$ are the rows of the extended $\mathbf{X}$ and $\mathbf{Z}$ matrices corresponding to the new observation in the ith small area, and $\mathbf{w}_{i0}$ is a vector with 1 in the ith position if the new observation is in the small area i and 0 elsewhere. When using a B-spline basis, the extended $\mathbf{Z}$ matrix is not straightforward and some algebra is needed. Let us consider the B-spline extended basis $\mathbf{B}^*$

$$\mathbf{B}^* = \begin{pmatrix} \mathbf{B} & \mathbf{0} \\ \mathbf{B}_1 & \mathbf{B}_2 \end{pmatrix},$$

where $\mathbf{B}_1$ and $\mathbf{B}_2$ are the matrices corresponding to the extended basis. Since it is necessary to increase the number of knots to cover the range of the extended covariate, there are more parameters, and therefore, the difference matrix used in the penalty has also to be extended. Let us consider the difference matrix $\mathbf{D}^*$ with the same number of columns as the number of parameters in the extended B-spline basis $\mathbf{B}^*$. Then, this matrix can be expressed as

$$\mathbf{D}^* = \begin{pmatrix} \mathbf{D} & \mathbf{0} \\ \mathbf{E} & \mathbf{F} \end{pmatrix}.$$

Now, a transformation is needed to reparameterize the extended model as was done in Section 2. There are different possibilities of choosing this transformation. The only restriction is that it needs to preserve the reparameterization used to fit the data. We take

$$\mathbf{T}^* = \begin{pmatrix} \mathbf{T} & \\ & \mathbf{F}^{-1} \end{pmatrix},$$

where $\mathbf{T}$ is the transformation defined in Section 2, and $\mathbf{F}$ is given in $\mathbf{D}^*$. Then,

$$\mathbf{B}^*\mathbf{T}^* = \begin{pmatrix} \mathbf{X} & \mathbf{Z} & 0 \\ \mathbf{X}_0 & \mathbf{Z}_1 & \mathbf{Z}_2 \end{pmatrix}, \quad \text{where } \mathbf{Z}_1 = \mathbf{B}_1\mathbf{U}_2\tilde{\boldsymbol{\Sigma}}^{-1/2}, \text{ and } \mathbf{Z}_2 = \mathbf{B}_2\mathbf{F}^{-1}.$$

The mixed model representation with this parametrization is

$$\begin{pmatrix} \mathbf{y} \\ \mathbf{y}_0 \end{pmatrix} = \begin{pmatrix} \mathbf{X} \\ \mathbf{X}_0 \end{pmatrix}\boldsymbol{\beta} + \begin{pmatrix} \mathbf{Z} & 0 \\ \mathbf{Z}_1 & \mathbf{Z}_2 \end{pmatrix}\begin{pmatrix} \mathbf{u} \\ \mathbf{u}_0 \end{pmatrix} + \begin{pmatrix} \mathbf{W} \\ \mathbf{W}_0 \end{pmatrix}\mathbf{v} + \begin{pmatrix} \boldsymbol{\epsilon} \\ \boldsymbol{\epsilon}_0 \end{pmatrix},$$

where $\mathbf{y}_0$ are the observations to be predicted, $\mathbf{X}_0$ is the extended $\mathbf{X}$ matrix (note that one only has to add the rows corresponding to the new values in the matrix given by (6)), and $\mathbf{W}_0$ is a matrix whose rows are vectors with 1 in the ith position and 0 elsewhere if the observation to be predicted belongs to the ith area. The vectors $\mathbf{u}_0$ and $\boldsymbol{\epsilon}_0$ denote the random effects and random errors not present in the data set, but drawn from the same population as $\mathbf{u}$ and $\boldsymbol{\epsilon}$. By including

the values to be predicted in Eq. (2) and using the transformation $\mathbf{T}^*$, the covariance matrix between the estimated and predicted random effects is given by

$$\mathrm{Cov}\begin{pmatrix}\mathbf{u}\\ \mathbf{u}_0\end{pmatrix} = \begin{pmatrix}\mathbf{I} & -\widetilde{\boldsymbol{\Sigma}}^{-1/2}\mathbf{U}_2'\mathbf{E}'\\ -\mathbf{E}\mathbf{U}_2\widetilde{\boldsymbol{\Sigma}}^{-1/2} & \mathbf{I}+\mathbf{E}\mathbf{U}_2\widetilde{\boldsymbol{\Sigma}}^{-1}\mathbf{U}_2'\mathbf{E}'\end{pmatrix}.$$

Using results in Gilmour et al. (2004), we find that

$$\hat{\mathbf{u}}_0 = -\mathbf{E}\mathbf{U}_2\widetilde{\boldsymbol{\Sigma}}^{-1/2}\hat{\mathbf{u}},$$

and

$$\hat{\mathbf{y}}_0 = \mathbf{X}_0\hat{\boldsymbol{\beta}} + \mathbf{Z}_1\hat{\mathbf{u}} + \mathbf{Z}_2\hat{\mathbf{u}}_0 + \mathbf{W}_0\hat{\mathbf{v}} = \mathbf{X}_0\hat{\boldsymbol{\beta}} + \mathbf{Z}_0\hat{\mathbf{u}} + \mathbf{W}_0\hat{\mathbf{v}},$$

where

$$\mathbf{Z}_0 = \mathbf{Z}_1 - \mathbf{Z}_2\mathbf{E}\mathbf{U}_2\boldsymbol{\Sigma}^{-1/2}. \tag{14}$$

4. Prediction mean squared error

4.1. Analytical MSE estimators

In this section, both the prediction MSE of the EBLUP given by (13) and the corresponding expression for the EBLUP predictor of a new observation y_{i0} in the ith small area are derived. Two alternative estimators of those quantities are also provided.

Firstly, the prediction error $\tilde{y}_{ij} - y_{ij}$ is derived following results in Opsomer et al. (2008), although extra terms arising from considering the errors ϵ_{ij} are added. To make the algebra easier, let us denote $\mathbf{M} = [\mathbf{Z}, \mathbf{W}]$, $\boldsymbol{\xi} = (\mathbf{u}', \mathbf{v}')'$ and $\boldsymbol{\Gamma}_m = \begin{pmatrix}\boldsymbol{\Gamma}_u & \mathbf{0}\\ \mathbf{0} & \boldsymbol{\Gamma}_v\end{pmatrix}$, and let us considered $\theta_{ij} = \mathbf{x}_{ij}\boldsymbol{\beta} + \mathbf{z}_{ij}\mathbf{u} + \mathbf{w}_{ij}\mathbf{v}$ and $\boldsymbol{\sigma}^2 = (\sigma_u^2, \sigma_v^2, \sigma_e^2)'$. Then, the mean squared error of the BLUP (12) is given by

$$\begin{aligned}\mathrm{MSE}[\tilde{y}_{ij}] &= E[(\tilde{y}_{ij} - y_{ij})^2] = E[(\tilde{y}_{ij} - \theta_{ij} - \epsilon_{ij})^2]\\ &= E[(\tilde{y}_{ij} - \theta_{ij})^2] + \sigma_e^2 - 2E[\tilde{y}_{ij}\epsilon_{ij}].\end{aligned} \tag{15}$$

To derive the first term in the right hand side of Eq. (15), we first note that

$$\tilde{y}_{ij} - \theta_{ij} = \mathbf{q}_{ij}(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}) + \mathbf{m}_{ij}(\boldsymbol{\Gamma}_m\mathbf{M}'\mathbf{V}^{-1}(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) - \boldsymbol{\xi}), \tag{16}$$

where $\mathbf{q}_{ij} = \mathbf{x}_{ij} - \mathbf{m}_{ij}\boldsymbol{\Gamma}_m\mathbf{M}'\mathbf{V}^{-1}\mathbf{X}$, and $\mathbf{m}_{ij}$ is the row of the $\mathbf{M}$ matrix corresponding to the jth observation in the ith small area. It is easy to show that the two terms in Eq. (16) are uncorrelated. Using that $E[\tilde{y}_{ij}] = \mathbf{x}_{ij}\boldsymbol{\beta}$ and $E[\theta_{ij}] = \mathbf{x}_{ij}\boldsymbol{\beta}$, it follows that

$$\begin{aligned}E[(\tilde{y}_{ij} - \theta_{ij})^2] &= \mathrm{var}[(\tilde{y}_{ij} - \theta_{ij})]\\ &= \mathrm{var}[\mathbf{q}_{ij}(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta})] + \mathrm{var}[\mathbf{m}_{ij}(\boldsymbol{\Gamma}_m\mathbf{M}'\mathbf{V}^{-1}(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) - \boldsymbol{\xi})],\end{aligned}$$

where (see Henderson (1975))

$$\mathrm{var}[\mathbf{m}_{ij}(\boldsymbol{\Gamma}_m\mathbf{M}'\mathbf{V}^{-1}(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) - \boldsymbol{\xi})] = \mathbf{m}_{ij}(\boldsymbol{\Gamma}_m - \boldsymbol{\Gamma}_m\mathbf{M}'\mathbf{V}^{-1}\mathbf{M}\boldsymbol{\Gamma}_m)\mathbf{m}_{ij}' := g_1(\boldsymbol{\sigma}^2),$$

$$\mathrm{var}[\mathbf{q}_{ij}(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta})] = \mathbf{q}_{ij}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{q}_{ij}' := g_2(\boldsymbol{\sigma}^2).$$

To calculate the last term in Eq. (15), we first note that

$$\tilde{y}_{ij}\epsilon_{ij} = [\mathbf{x}_{ij}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{Y} + \mathbf{m}_{ij}\boldsymbol{\Gamma}_m\mathbf{M}'\mathbf{P}\mathbf{Y}]\epsilon_{ij},$$

where $\mathbf{P} = \mathbf{V}^{-1} - \mathbf{V}^{-1}\mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}$, and $E[y_{ij}\epsilon_{rs}] = \sigma_e^2$ if $i = r$ and $j = s$, and 0 otherwise. Then,

$$\begin{aligned}E[\tilde{y}_{ij}\epsilon_{ij}] &= [\mathbf{x}_{ij}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1} + \mathbf{m}_{ij}\boldsymbol{\Gamma}_m\mathbf{M}'\mathbf{P}]E[\mathbf{Y}\epsilon_{ij}],\\ &= [\mathbf{x}_{ij}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1} + \mathbf{m}_{ij}\boldsymbol{\Gamma}_m\mathbf{M}'\mathbf{P}]\sigma_e^2\mathbf{h}_{ij},\\ &:= g_4(\boldsymbol{\sigma}^2),\end{aligned}$$

where $\mathbf{h}_{ij}$ is a vector of length $n = \sum_{i=1}^{I} n_i$ with 1 in the jth position corresponding to the ith small area, and zeros elsewhere. Then, the MSE of the BLUP (12) is given by

$$\mathrm{MSE}[\tilde{y}_{ij}] = E[(\tilde{y}_{ij} - y_{ij})^2] = g_1(\boldsymbol{\sigma}^2) + g_2(\boldsymbol{\sigma}^2) + \sigma_e^2 - 2g_4(\boldsymbol{\sigma}^2). \tag{17}$$

Note that notation g_4 is used because g_3 is kept to denote the well-known term in the small area literature. The corresponding expression for the predictor of a new observation in the ith small area y_{i0} is

$$\mathrm{MSE}[\tilde{y}_{i0}] = E[(\tilde{y}_{i0} - y_{i0})^2] = g_1(\boldsymbol{\sigma}^2) + g_2(\boldsymbol{\sigma}^2) + \sigma_e^2. \tag{18}$$

In this case the term g_4 vanishes because the observed values $\mathbf{Y}$ and the error term ϵ_{i0} are uncorrelated.

The variance components are rarely known in practice and they must be estimated from the data. It is known that plugging the variance components estimators in Eqs. (17) and (18) might lead to an underestimation of the MSE of the EBLUP (13). The literature in small area estimation studying different approximations to the MSE is rich. Prasad and Rao (1990) derive a second-order approximation for the prediction MSE and a correct estimator up to second order when the variance components are estimated by the method of moments (see for example (Searle et al., 1992)). Das et al. (2004) extend these results to more general models and different methods of estimating the variance components. These results are applicable when the covariance matrix of the data is block-diagonal. Although this is not the case here, it is still possible to obtain similar results following the work by Opsomer et al. (2008).

The MSE of the EBLUP (13) is given by

$$\mathrm{MSE}[\hat{y}_{ij}] = E[(\tilde{y}_{ij} - \theta_{ij})^2] + E[(\hat{y}_{ij} - \tilde{y}_{ij})^2] + E[\epsilon_{ij}^2] - 2E[\hat{y}_{ij}\epsilon_{ij}]. \tag{19}$$

The second term in the right hand side of Eq. (19) can be approximated by

$$E[(\hat{y}_{ij} - \tilde{y}_{ij})^2] \approx tr[\mathbf{SVS}'\mathcal{I}^{-1}] := g_3(\boldsymbol{\sigma}^2),$$

where $\mathbf{S}$ is a matrix with rows $\mathbf{S}_j = \mathbf{m}_{ij}\left(\frac{\partial \boldsymbol{\Gamma}_m}{\partial \sigma_j^2}\mathbf{M}'\mathbf{V}^{-1} + \boldsymbol{\Gamma}_m\mathbf{M}'\frac{\partial \mathbf{V}^{-1}}{\partial \sigma_j^2}\right)$, $j = 1, 2, 3$ ($\sigma_1^2 = \sigma_u^2$, $\sigma_2^2 = \sigma_v^2$ and $\sigma_3^2 = \sigma_e^2$), $\frac{\partial \boldsymbol{\Gamma}_m}{\partial \sigma_u^2} = \mathrm{diag}[\mathbf{I}_K, \mathbf{0}]$, $\frac{\partial \boldsymbol{\Gamma}_m}{\partial \sigma_v^2} = \mathrm{diag}[\mathbf{0}, \mathbf{I}_I]$, $\frac{\partial \boldsymbol{\Gamma}_m}{\partial \sigma_e^2} = \mathbf{0}$, $\frac{\partial \mathbf{V}^{-1}}{\partial \sigma_u^2} = \mathbf{ZZ}'$, $\frac{\partial \mathbf{V}^{-1}}{\partial \sigma_v^2} = \mathbf{WW}'$, $\frac{\partial \mathbf{V}^{-1}}{\partial \sigma_e^2} = \mathbf{I}_n$, and $\mathcal{I}^{-1}$ is the asymptotic covariance matrix of the variance component estimators derived from the REML equations (see Harville (1977)). An appropriate estimator for the prediction MSE given by (19) is

$$\widehat{\mathrm{MSE}}[\hat{y}_{ij}] = g_1(\hat{\boldsymbol{\sigma}}^2) + g_2(\hat{\boldsymbol{\sigma}}^2) + 2g_3(\hat{\boldsymbol{\sigma}}^2) + \hat{\sigma}_e^2 - 2\widehat{E[\tilde{y}_{ij}\epsilon_{ij}]}.$$

Note that the extra term $g_3(\hat{\boldsymbol{\sigma}}^2)$ is introduced because $g_1(\hat{\boldsymbol{\sigma}}^2)$ is not an unbiased estimator of $g_1(\boldsymbol{\sigma}^2)$ (see Prasad and Rao (1990)). Finally, the estimator of the MSE for the EBLUP of a new observation y_{i0} is given by

$$\widehat{\mathrm{MSE}}[\hat{y}_{i0}] = g_1(\hat{\boldsymbol{\sigma}}^2) + g_2(\hat{\boldsymbol{\sigma}}^2) + 2g_3(\hat{\boldsymbol{\sigma}}^2) + \hat{\sigma}_e^2.$$

In the P-spline literature (see for example Ruppert et al. (2003)), it is common to use the following estimator of the prediction error to build prediction intervals

$$\mathrm{var}_p := g_1(\hat{\boldsymbol{\sigma}}^2) + g_2(\hat{\boldsymbol{\sigma}}^2) + \hat{\sigma}_e^2. \tag{20}$$

Summing up, two alternative mean squared error estimators of the prediction MSE of the EBLUP are considered. Namely,

$$mse_1 := g_1(\hat{\boldsymbol{\sigma}}^2) + g_2(\hat{\boldsymbol{\sigma}}^2) + 2g_{3*}(\hat{\boldsymbol{\sigma}}^2) + \hat{\sigma}_e^2 - 2\widehat{E[\tilde{y}_{ij}\epsilon_{ij}]}, \tag{21}$$

where $g_{3*}(\hat{\boldsymbol{\sigma}}^2) = (\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}})'\hat{\mathbf{S}}'\hat{\mathcal{I}}^{-1}\hat{\mathbf{S}}(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}})$, and

$$mse_2 := g_1(\hat{\boldsymbol{\sigma}}^2) + g_2(\hat{\boldsymbol{\sigma}}^2) + 2g_{3**}(\hat{\boldsymbol{\sigma}}^2) + \hat{\sigma}_e^2 - 2\widehat{E[\tilde{y}_{ij}\epsilon_{ij}]}, \tag{22}$$

where $g_{3**}(\hat{\boldsymbol{\sigma}}^2) = tr[\hat{\mathbf{S}}\hat{\mathbf{V}}\hat{\mathbf{S}}'\hat{\mathcal{I}}^{-1}]$.

The corresponding estimators of the MSE for the EBLUP of a new observation are the same above quantities with the last term omitted.

4.2. Bootstrap MSE estimators

Bootstrap MSE estimators have been proposed in the literature for their simplicity. See for example, Lahiri (2003) and González-Manteiga et al. (2008) for different bootstrap proposals. Opsomer et al. (2008) consider a simple bootstrap estimator which is not second-order correct as it does not take into account the estimation of the variance components. However, they show that the error in estimating the variance components is small relative to the prediction error. Ugarte et al. (2009) also show in a simulation study that a simple bootstrap estimator can be appropriate. In this paper we consider the approach followed by Opsomer et al. (2008) and Ugarte et al. (2009) which consists in generating bootstrap observations

$$\mathbf{Y}^* = \mathbf{X}\hat{\boldsymbol{\beta}} + \mathbf{Zu}^* + \mathbf{Wv}^* + \boldsymbol{\epsilon}^*,$$

where $\mathbf{u}^*$, $\mathbf{v}^*$ and $\boldsymbol{\epsilon}^*$ are bootstrap replicates of the random effects and the random error of the model. The final bootstrap MSE estimator is calculated as

$$mse^{boot} = \frac{1}{B}\sum_{b=1}^{B}(\hat{y}_{ij}^{*b} - y_{ij}^{*b}), \tag{23}$$

where B is the number of bootstrap replicates, and $\hat{y}_{ij}^{*b}$ is the prediction for the jth observation of the ith small area in the bth bootstrap sample, and y_{ij}^{*b} is the corresponding simulated value. The corresponding bootstrap MSE estimator for the forecast

values is similarly defined. If the random elements are generated from Model (8), the bootstrap is known as parametric bootstrap, but if the random elements are sampled from the standardized predicted random effects and the standardized estimated residuals, the bootstrap is known as non-parametric bootstrap. In this case, the standardized random effects and the standardized residuals take the form

$$\begin{aligned}\hat{\mathbf{u}}_{std} &= (\mathbf{Z}'\mathbf{P}\mathbf{Z})^{-1/2}\hat{\mathbf{u}}/\sigma_u,\\ \hat{\mathbf{v}}_{std} &= (\mathbf{W}'\mathbf{P}\mathbf{W})^{-1/2}\hat{\mathbf{v}}/\sigma_v,\\ \hat{\boldsymbol{\epsilon}}_{std} &= (\mathbf{P})^{-1/2}\hat{\boldsymbol{\epsilon}}/\sigma_e.\end{aligned} \tag{24}$$

Note that $\hat{\mathbf{u}}_{std} \sim N(\mathbf{0}, \sigma_u^2\mathbf{I}_L)$, $\hat{\mathbf{v}}_{std} \sim N(\mathbf{0}, \sigma_v^2\mathbf{I}_l)$ and $\hat{\boldsymbol{\epsilon}}_{std} \sim N(\mathbf{0}, \sigma_e^2\mathbf{I}_n)$, since $\text{var}(\hat{\mathbf{u}}) = \sigma_u^4\mathbf{Z}'\mathbf{P}\mathbf{Z}$, $\text{var}(\hat{\mathbf{v}}) = \sigma_v^4\mathbf{W}'\mathbf{P}\mathbf{W}$ and $\text{var}(\hat{\boldsymbol{\epsilon}}) = \sigma_e^4\mathbf{P}$.

It is interesting to remark that a non-parametric bootstrap for the forecast values cannot be defined because the true values are unknown and it is not possible to resample from the corresponding residuals. In this work, the non parametric bootstrap is used for the observed values. In the following, we will denote the parametric bootstrap as mse^{pboot} and the non-parametric bootstrap as mse^{npboot}.

5. Hypothesis testing

To test whether or not the spline and the area effect components are significant is reduced to test the following hypothesis

$$H_0 : \sigma_u^2 = 0, \text{ versus } H_1 : \sigma_u^2 > 0, \quad \text{and}$$
$$H_0 : \sigma_v^2 = 0, \text{ versus } H_1 : \sigma_v^2 > 0.$$

Likelihood ratio tests (or restricted likelihood ratio tests) can be used for this purpose, but standard theory does not apply as we are testing at the boundary of the parameter space. Self and Liang (1987) show that the asymptotic distribution of the likelihood ratio test in the case of independent and identically distributed data is an equal mixture of a point mass at zero and a chi-square distribution with one degree of freedom, $\frac{1}{2}\chi_0^2 + \frac{1}{2}\chi_1^2$. However, Crainiceanu and Ruppert (2004) show that this approximation is not very good for spline components. Crainiceanu et al. (2005) propose to obtain the exact distribution of the restricted likelihood ratio test using spectral decompositions. Here, the bootstrap-based alternative proposed by Opsomer et al. (2008) is used. A similar technique has been proposed by Militino et al. (2006). The procedure to test the spline component (analogously the area effect) is as follows.

(1) Fit the mixed model under the null $H_0 : \sigma_u^2 = 0$ and under the alternative $H_1 : \sigma_u^2 > 0$, and compute the restricted likelihood ratio statistic $L_{obs} = 2(L_1 - L_0)$, where L_0 and L_1 are the restricted log-likelihood under the null and the alternative respectively.
(2) Generate R bootstrap replicates as $\mathbf{Y}^* = \mathbf{X}\hat{\boldsymbol{\beta}} + \mathbf{W}\mathbf{v}^* + \boldsymbol{\epsilon}^*$, where $\hat{\boldsymbol{\beta}}$ is the estimated fixed effects under the null, $\mathbf{v}^*$ and $\boldsymbol{\epsilon}^*$ are the bootstrap random effects and random errors respectively. They are sampled with replacement from the estimated standardized area random effects and residuals respectively (see Eq. (24)).
(3) For each bootstrap sample r, compute $L^{*r} = 2(L_1^{*r} - L_0^{*r})$. The bootstrap p-value is obtained as

$$P_{boot} = \frac{1 + \#\{L^{*r} \geq L_{obs}\}}{R + 1},$$

where $\#\{L^{*r} \geq L_{obs}\}$ denotes how many times the bootstrap likelihood ratio statistics L^{*r} is greater than or equal to the observed value L_{obs}.

6. Illustration

The Spanish real estate market has experienced tremendous development during the last ten years. The increased demand for dwelling in combination with low interest rate mortgages has created an environment conducive to property speculation which, in turn, has driven real estate prices to new record highs. However, the recent global economical crisis is leading to a downturn in the Spanish real estate market. This is a very serious problem because Spanish economy is very dependent on the construction industry, and thousands of people are losing their jobs. On the other hand, experts think that dwellings in Spain are overvalued and a price adjustment is necessary. In this paper, the price evolution of used dwellings is analyzed in the city of Vitoria, Spain. The local Government is interested in monitoring housing prices in the different neighborhoods of the city to analyze the extent of the crisis and the adjustment of the housing market. The data set consists of average prices per squared meter (m^2) of used dwellings in nine neighborhoods of the city during the period 1993–2007 (15 observations in each neighborhood). Prices are given in thousands of euros. The neighborhoods are codified as $Z1$, $Z2$, $Z3A$, $Z3B$, $Z4A$, $Z4B$, $Z4C$, $Z5$, and $Z6$. Dwelling prices for the next five years (2008–2012) are also forecast.

Model (8) is fitted to the data using P-splines with a cubic B-spline basis and a penalty of order 2. A neighborhood random effect is also considered. Note that although there are 135 observations, the explanatory variable takes the values 1993–2007. Then the B-spline basis is defined using these values. In this application, 6 knots have been considered. The general rule given

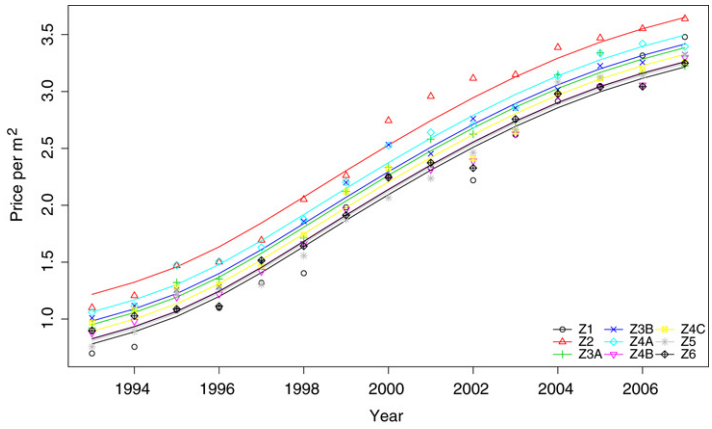

Fig. 1. Temporal trend for each neighborhood.

Table 1
Parameter estimates and their significance.

Estimator	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\sigma}_u^2$	$\hat{\sigma}_v^2$	$\hat{\sigma}_e^2$
Estimate	−311.902	0.157	0.100	0.021	0.012
p-value	0.000	0.000	0.001	0.001	–

by Ruppert (2002) points towards 4 inner knots, however the total number of knots for the construction of the B-spline basis is the number of interior knots plus twice the degree of the polynomials (see Eilers and Marx (1996)). As a result, with 4 inner knots the B-spline basis covers the years to be predicted, making the prediction unfeasible. Consequently, 6 knots are considered. The smoothing parameter is the same for all neighborhoods.

Fig. 1 shows the temporal trend for each neighborhood. From this picture, it is clear that the average price per m^2 is different from one district to another, but the price evolution is the same. Model (8) is used to fit parallel curves in each neighborhood. A model allowing for different curves in each neighborhood has been also fitted. However, the AIC criterion, defined as $AIC = -2 * \text{loglikelihood} + 2 * df$, selects Model (8) as the best candidate. The AIC values are −136.979 and −83.603 for models with parallel curves and non-parallel curves respectively. Note that df is the trace of the matrix $\mathbf{H}$, where $\hat{\mathbf{Y}} = \mathbf{HY}$.

Table 1 displays the parameter estimates, and their corresponding p-values for model (8). To assess the significance of the neighborhood effect a bootstrap test (see Section 5) has been conducted using a bootstrap sample of size $R = 1000$. The significance of the variance component coming from the spline has also been tested. The small p-values reveal the significance of both components. Note that the significance of the spline means that a curve is more appropriate than a straight line for the price evolution.

Fig. 2 displays the fitted housing price trend together with the price predictions for the next five years for each neighborhood. Pointwise prediction bands, based on the different prediction errors, are also shown. For the bootstrap MSE, 1000 replicates have been considered. The bands are practically equal for the observed prices, but they present some differences for the forecast values. The band based on the variance estimator tends to be slightly narrower than the bands based on the MSE estimators. In particular the parametric bootstrap MSE estimator produces the wider confidence bands for the forecast values. Bands based on the non-parametric bootstrap are not shown here because it is not applicable to the forecast values. For the observed values, the results are similar to the parametric bootstrap. A simple idea to build a non-parametric bootstrap for the forecast values consists in using the residuals from the model fitting instead of the unknown forecasting residuals. However, this does not produce good results (at least in our case).

Fig. 3 exhibits, in detail, the price trend and the forecast values for one neighborhood randomly chosen. The moderate increase in average prices for the next five years is clear, something which agrees with the current economic perspective, because nowadays the prices of used dwellings are increasing at a rate below the consumer price index (CPI).

Table 2 displays the predicted prices (in thousands of euros) for the coming years (2008–2012) in the different districts, showing the slow increasing trend in average prices in future years.

7. Simulation study

In this section, a simulation study is conducted to assess the performance of the small area predictor given by (13) and the different prediction error estimators given by (20)–(23) and their corresponding versions for the forecast values. A total of $B = 10\,000$ data sets have been generated from Model (8) using the parameter estimates obtained in the analysis of the real data (see Table 1). New observations corresponding to years 2008–2012 have also been generated using the corresponding

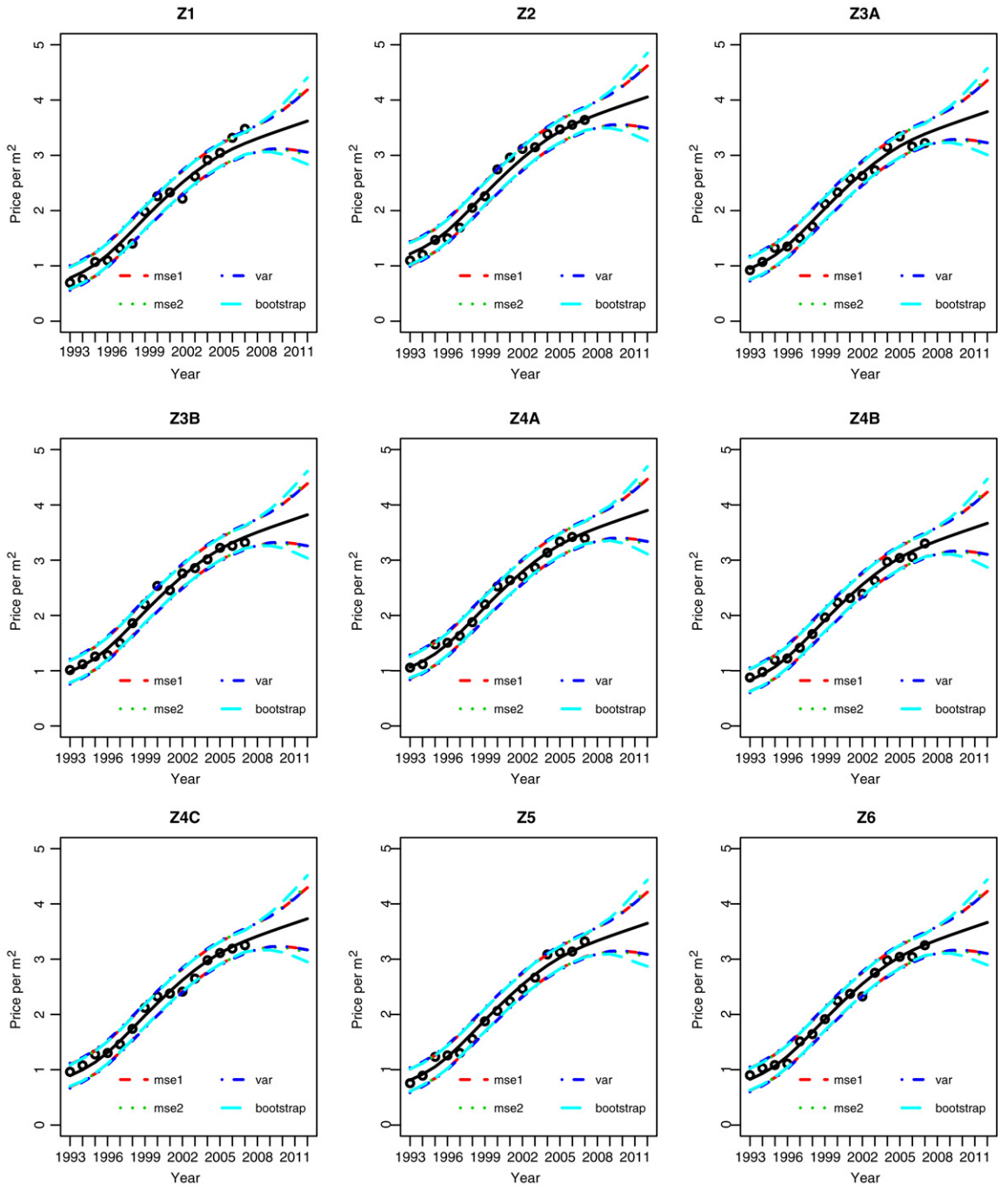

Fig. 2. Temporal trend and predicted values for each neighborhood.

Table 2
Predicted prices for the five nearest years in the different districts.

Year	*Z1*	*Z2*	*Z3A*	*Z3B*	*Z4A*	*Z4B*	*Z4C*	*Z5*	Z6
2008	3.304	3.740	3.474	3.506	3.585	3.351	3.416	3.334	3.349
2009	3.387	3.822	3.556	3.588	3.668	3.434	3.498	3.417	3.432
2010	3.466	3.901	3.635	3.667	3.747	3.513	3.577	3.496	3.511
2011	3.543	3.978	3.712	3.744	3.824	3.590	3.654	3.573	3.588
2012	3.619	4.055	3.789	3.821	3.900	3.667	3.731	3.649	3.664

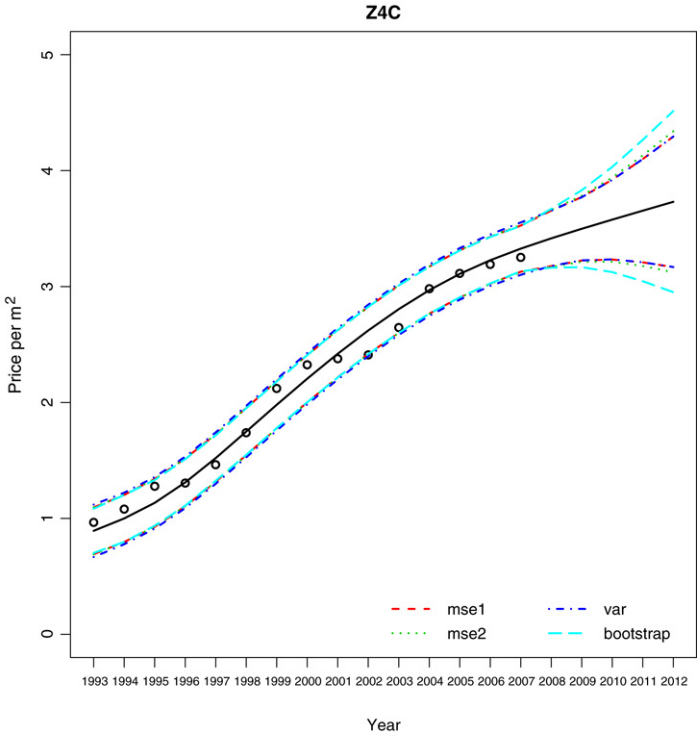

Fig. 3. Temporal trend and predicted values for neighborhood Z4C.

extended matrix $\mathbf{Z}_0$ given by (14). For each data set, the model is fitted using the `lme` function in the R package `nlme` written by Pinheiro et al. (2008). The prediction error estimates var_p, mse_1, and mse_2 are calculated for each fitted value. In addition, predictions and their corresponding errors are computed for years 2008–2012. Empirical measures of relative bias (*RB*) and relative root mean squared error (*RRMSE*) of the small area predictor (13) and the corresponding predictor for future values are used to assess their performance. They are given by

$$RB_{ij} = \frac{1}{B}\frac{\sum_{b=1}^{B}(\hat{y}_{ij}^b - y_{ij})}{y_{ij}},$$
$$RRMSE_{ij} = \frac{\sqrt{\frac{1}{B}\sum_{b=1}^{B}(\hat{y}_{ij}^b - y_{ij})^2}}{y_{ij}}. \tag{25}$$

As global measures in every small area, we consider

$$MARB_i = \frac{1}{J}\sum_{j=1}^{J}|RB_{ij}|, \qquad MRRMSE_i = \frac{1}{J}\sum_{j=1}^{J} RRMSE_{ij}. \tag{26}$$

Fig. 4 displays the MARB and the MRRMSE of the predictor (13) for both the observed and the forecast values in every small area. The behavior of the estimator is similar in all the areas. This is expected because our setting is very regular, with the same number of observations in every small area. The relative bias is practically negligible (less than 0.5%) for both, observed and forecast values. The lower values of the MRRMSE (less than 7%) indicate that the predictor is very precise. We have also plotted (not shown here, to conserve space) these quantities for the five forecast values in every neighborhood. The results point to an increase of RB and MRRMSE with time, indicating that forecasting should be done with caution.

To assess the performance of the prediction MSE estimators we also consider empirical measures of relative bias and relative root mean squared error, similarly defined as in Eq. (25) but replacing $\hat{y}_{ij}^b$ by $\widehat{MSE}_{ij}^b$ and y_{ij} by $EMSE_{ij}$, where $\widehat{MSE}_{ij}^b$ is the prediction error (that is var_p, mse_1, mse_2, mse^{pboot} and mse^{npboot}) evaluated in the bth sample and $EMSE_{ij} = \sum_{b=1}^{B}(\hat{y}_{ij}^b - y_{ij}^b)^2/B$ is the true MSE for the predictor $\hat{y}_{ij}$. As global measures, the mean absolute relative bias ($MARB^{mse}$) and the mean relative root mean squared error ($MRRMSE^{mse}$) over the I small areas are considered. Table 3 displays the results for the fitted values and the predictions of future observations. Prediction MSE estimators, mse_1, mse_2, and mse^{boot} perform similarly for the observed

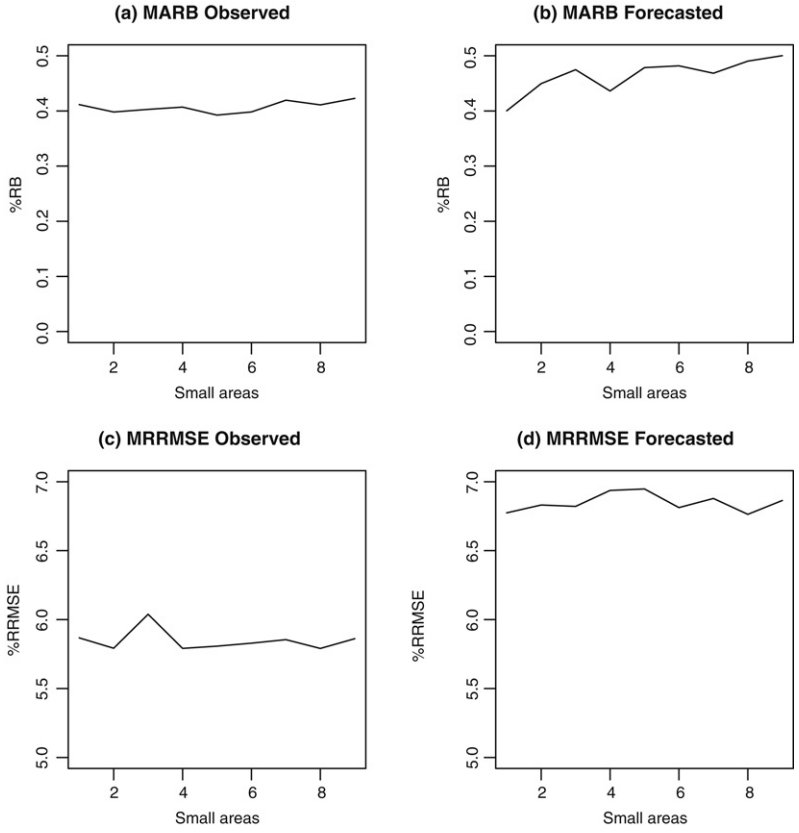

Fig. 4. MARB and MRRMSE of the small area estimator for both the observed and the forecast values. Graph (a) exhibits the MARB for the observed values, graph (b) corresponds to the MARB for the forecast values. Graph (c) and (d) display the MRRMSE for observed and forecast values respectively. Results are displayed for each neighborhood.

Table 3
MARB and MRRMSE for the different estimators of the prediction error.

	var_p	mse_1	mse_2	mse^{pboot}
	Fitted values			
$MARB^{mse}$	0.205	0.022	0.027	0.011
$MRRMSE^{mse}$	0.257	0.137	0.138	0.140
	Forecasted values			
$MARB^{mse}$	0.351	0.328	0.284	0.033
$MRRMSE^{mse}$	0.411	0.393	0.362	0.406

values, with low mean relative bias and error. However, the estimator mse^{boot} exhibits a much smaller bias for the forecast values than the rest of estimators, indicating that it is more appropriate to calculate standard errors of these predictions. The var_p estimator is clearly inferior to mse_1, mse_2, and mse^{boot} for both observed and forecast values. In general, all the estimators lead to larger $MARB^{mse}$ and $MRRMSE^{mse}$ for the forecast values. This is somehow expected because forecasting new values is always a difficult task. Nevertheless, the parametric bootstrap estimator performs very well even for the forecast values. Finally, it is interesting to remark that the $MARB^{mse}$ and $MRRMSE^{mse}$ of the non-parametric bootstrap estimator for the observed values are 0.035 and 0.075 respectively, indicating that it is a good estimator for the MSE of the observed values. As it has been mentioned in the application, we have tried to build a non-parametric bootstrap for the forecast values using the residuals from the fitting, instead of the predicted residuals, which are unknown. The $MARB^{mse}$ and $MRRMSE^{mse}$ are unacceptably large, indicating that this simple idea is not appropriate. Further research on how to perform non-parametric bootstrap for forecast values is needed.

8. Discussion

Semiparametric models combining both a non-parametric trend and a small area random effect offer a flexible approach to small area estimation problems. The use of P-spline regression is very appealing because P-splines can be represented

as mixed effects models and then they benefit from the existing theory and the well-developed software for mixed models analysis. For instance, the estimation of the smoothing parameter is reduced to the estimation of the variance components using standard methods such as REML. In this paper, we show how to predict future values using the mixed model representation of a semiparametric model including a small area random effect. In addition, two alternative analytical estimators of the prediction MSE of a particular observation are derived. These estimators take into account the uncertainty arising from the estimation of the variance components. Traditionally, the P-spline literature does not consider this uncertainty when estimating the variability of the fitted non-parametric curve. Bootstrap MSE estimators are also provided because of their simplicity. In the real data analysis presented in this paper, the performance of the different variability measures is rather similar, but the simulation study reveals that the proposed analytical estimators and the parametric bootstrap estimator outperform the traditional variance estimator in terms of bias and variability. All the analytical estimators considered in this paper exhibit large biases when the aim is to forecast future values, but the parametric bootstrap seems to be a good alternative. In addition, we have observed that the spline predictor used in this paper is a very good tool for trend estimation and for forecasting, as the relative bias and error are practically negligible. However, the forecasting should be done with caution and only for a short period of time. The procedures developed in this paper have been used to analyze the housing price evolution in the districts of Vitoria, and a moderate price increase is observed for the coming years.

Acknowledgments

The work of Ugarte, Goicoa, and Militino has been supported by the Spanish Ministry of Science and Innovation (project MTM 2008-03085). The work of Durbán has been supported by the Spanish Ministry of Science and Innovation (project MTM 2008-02901). The authors would like to thank the company LKS Tasaciones for providing the data.

References

Aerts, M., Claeskens, G., Wand, M.P., 2002. Some theory for penalized spline generalized additive models. Journal of Statistical Planning and Inference 103, 455–470.
Brumback, B.A., Ruppert, D., Wand, M.P., 1999. Comment on "Variable selection and function estimation in additive nonparametric regression using a data-based prior", by Shively, T.S., Kohn, R., Wood, S. Journal of the American Statistical Association 94, 794–797.
Coull, B.A., Ruppert, D., Wand, M.P., 2001. Simple incorporation of interactions into additive models. Biometrics 57, 539–545.
Crainiceanu, C.M., Ruppert, D., 2004. Likelihood ratio test in linear mixed models with one variance component. Journal of the Royal Statistical Society. Series B 66, 165–185.
Crainiceanu, C., Ruppert, D., Claeskens, G., Wand, M.P., 2005. Exact likelihood ratio tests for penalized splines. Biometrika 92, 91–103.
Currie, I.D., Durbán, M., 2002. Flexible smoothing with P-splines: A unified approach. Statistical Modelling 4, 333–349.
Currie, I.D., Durbán, M., Eilers, P.H., 2004. Smoothing and forecasting mortality rates. Statistical Modelling 4, 279–298.
Das, K., Jiang, J., Rao, J.N.K., 2004. Mean squared error of empirical predictor. Annals of Statistics 32, 818–840.
de Boor, C., 1978. A Practical Guide to Splines. Springer, Berlin.
Eilers, P.H.C., 1999. Comment on "The analysis of designed experiments and longitudinal data by using smoothing splines", by Verbyla, A.P., Cullis, B.R., Kenward, M.G., Welham, J.S. Applied Statistics 48, 269–311.
Eilers, P.H.C., Currie, I.D., Durbán, M., 2006. Fast and compact smoothing on large multidimensional grids. Computational Statistics and Data Analysis 50, 61–76.
Eilers, P.H.C., Marx, B.D., 1996. Flexible smoothing with B-splines and penalties. Statistical Science 11, 89–1216.
Gilmour, A., Cullis, B., Welham, S., Gogel, B., Thompson, R., 2004. An efficient computing strategy for prediction in mixed linear models. Computational Statistics and Data Analysis 44, 571–586.
González-Manteiga, W., Lombardía, M.J., Molina, I., Morales, D., Santamaría, L., 2008. Analytic and bootstrap approximations of prediction errors under a multivariate Fay–Herriot model. Computational Statistics and Data Analysis 52, 5242–5252.
Harville, D.A., 1977. Maximum likelihood approaches to variance component estimation and to related problems. Journal of the American Statistical Association 72, 320–340.
Hastie, T., Tibshirani, R., 1990. Generalized Additive Models. Chapman and Hall, London.
Henderson, C.R., 1963. Selection index and expected genetic advance. In: Statistical Genetics and Plant Breeding, National Academy of Science, National Research Council, Publication 982, Washington, DC, pp. 141–163.
Henderson, C.R., 1975. Best linear unbiased estimation and prediction under a selection model. Biometrics 31, 423–447.
Henderson, C.R., Kempthorne, O., Searle, S.R., von Krosigk, C.N., 1959. Estimation of environmental and genetic trends from records subject to culling. Biometrics 13, 192–218.
Lahiri, P., 2003. On the impact of bootstrap in survey sampling and small-area estimation. Statistical Science 18, 199–210.
McCulloch, C.E., Searle, S.R., 2001. Generalized, Linear, and Mixed Models. John Wiley and Sons, New York.
Militino, A.F., Ugarte, M.D., García-Reinaldos, L., 2004. Alternative models for describing spatial dependence among dwelling selling prices. Journal of Real Estate Finance and Economics 29, 193–209.
Militino, A.F., Ugarte, M.D., Goicoa, T., González-Audícana, M., 2006. Using small area models to estimate the total area occupied by olive trees. Journal of Agricultural, Biological, and Environmental Statistics 11, 450–461.
Opsomer, J.D., Claeskens, G., Ranalli, M.G., Kauermann, G., Breidt, F.J., 2008. Non-parametric small area estimation using penalized spline regression. Journal of the Royal Statistical Society. Series B 70, 265–286.
Parise, H., Wand, M.P., Ruppert, D., Ryan, L., 2001. Incorporation of historical controls using semi-parametric mixed models. Journal of the Royal Statistical Society. Series C 50, 31–42.
Pinheiro, J., Bates, D., DebRoy, S., Sarkar, D., the R Core team. 2008. nlme: Linear and Nonlinear Mixed Effects Models. R package version 3.1-89.
Pratesi, M., Ranalli, M.G., Salvati, N., 2008. Semiparametric M-quantile regression for estimating the proportion of acidic lakes in 8-digit HUCs of the Northeastern US. Environmetrics 19, 687–701.
Prasad, N.G.N., Rao, J.N.K., 1990. The estimation of mean squared error of small area estimators. Journal of the American Statistical Association 85, 163–171.
Rao, J.N.K., 2003. Small Area Estimation. In: Wiley Series in Survey Methodology, Hoboken, New Jersey.
Ruppert, D., 2002. Selecting the number of knots for penalized splines. Journal of Computational and Graphical Statistics 11, 735–757.
Ruppert, D., Wand, M.P., Carroll, R.J., 2003. Semiparametric Regression. Cambridge University Press, New York.
Searle, S.R., Casella, G., McCullogh, C.E., 1992. Variance Components. John Wiley and Sons, New York.

Self, S.G., Liang, K.Y., 1987. Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. Journal of the American Statistical Association 82, 605–610.
Silverman, B.W., 1985. Some aspects of the spline smoothing approach to non-parametric regression curve fitting (with discussion). Journal of the Royal Statistical Society. Series B 47, 1–52.
Ugarte, M.D., Goicoa, T., Militino, A.F., 2004. Searching for housing submarkets using mixtures of linear models. Advances in Econometrics 18, 259–276.
Ugarte, M.D., Militino, A.F., Goicoa, T., 2009. Benchmarked Estimates in Small Areas Using Linear Mixed Models with Restrictions. Test. doi:10.1007/s11749-008-0094-x.
Wahba, G., 1983. Bayesian confidence intervals for cross-validated smoothing splines. Journal of the Royal Statistical Society. Series B 45, 133–150.
Wand, M.P., 2003. Smoothing and mixed models. Computational Statistics 18, 223–249.
Wood, S., 2003. Thin plate splines regression. Journal of the Royal Statistical Society. Series B 65, 95–114.
Wood, S., 2006. On confidence intervals for generalized additive models based on penalized regression splines. Australian and New Zealand Journal of Statistics 48, 445–464.