Università degli Studi di Firenze
Dipartimento di Statistica "G. Parenti"

***Dottorato di Ricerca in Statistica Applicata***
***XXII ciclo - SECS-S/01***

# Geoadditive Models for Data with Spatial Information

**Chiara Bocci**

Tutor: **Prof. Alessandra Petrucci**

Coordinatore: **Prof. Fabio Corradi**

# *Acknowledgments*

*I would like to thank Prof. Matt Wand and Prof. Raymond Chambers for the opportunity to spend part of my Ph.D. research program at the Centre for Statistical and Survey Methodology of the University of Wollongong (NSW - Australia). In particular I express my gratitude for the time they dedicated to me, for their knowledge and assistance. I would also like to thank Anica Damcevski, for her priceless support in all the aspects of my Australian life.*

*I must also acknowledge Dr. Gero Carletto, Senior Economist in the Development Research Group of the World Bank and member of the Living Standards Measurement Study team, for providing the datasets of the 2002 Living Standard Measurement Study and the 2001 Population and Housing Census of Albania.*

*I would like to thank my tutor, Prof. Alessandra Petrucci, for her guide. A special thanks goes to Prof. Silvana Salvini, Dr. Emilia Rocco and my mother for their support and editing assistance throughout this work of thesis.*

*Finally, I must thank my entire family for their comprehension and specifically my siblings for their support in all the impossible matters. Last, but surely not least, I greatly thank Cristian for his immense patience.*

*Chiara Bocci*
*31st December 2009*

# Contents

# List of Figures

# List of Tables

# *Introduction*

*Geostatistics* is concerned with the problem of producing a map of a quantity of interest over a particular geographical region based on, usually noisy, measurement taken at a set of locations in the region. The aim of such a map is to describe and analyze the geographical pattern of the phenomenon of interest.

Geostatistical methodologies are born and apply in areas such as environmental studies and epidemiology, where the spatial information is traditionally recorded and available. However, in the last years the diffusion of spatially detailed statistical data is considerably increased and these kind of procedures - possibly with appropriate modifications - can be used as well in other fields of application, for example to study demographic and socio-economic characteristics of a population living in a certain region.

Basically, to obtain a surface estimate we can exploit the exact knowledge of the spatial coordinates (latitude and longitude) of the studied phenomenon by using *bivariate smoothing* techniques, such as kernel estimate or *kriging* (Cressie, 1993; Ruppert et al., 2003). However, usually the spatial information alone does not properly explain the pattern of the response variable and we need to introduce some covariates in a more complex model.

*Geoadditive models*, introduced by Kammann and Wand (2003), answer this problem as they analyze the spatial distribution of the study variable while accounting for possible non-linear covariate effects. They represent such effects by merging an additive model (Hastie and Tibshirani, 1990) - that accounts for the non-linear relationship between the variables - and a kriging model - that accounts for the spatial correlation - and by expressing both as a *linear mixed model*. The linear mixed model representation is a useful instrument because it allows estimation using mixed model methodology and software. Moreover, we can extend geoadditive model to include generalized responses, small area estimation, longitudinal data, missing data and so on (Ruppert et al., 2009).

A first aim of this work was to present the application of geoadditive models in fields that differ from environmental and epidemiological studies.

In particular, a geoadditive small area estimation model is applied in order to estimate the mean of household log per-capita consumption expenditure for the Albanian Republic at district level.

As we said, the geographical information is now more available in socio-economic data. However sometimes we don't know the exact location of all the population units, just the areas to which they belong - like census districts, blocks, municipalities, etc - while we know the coordinates for sampled units. How can we continue to use the geoadditive model under these circumstances? The classic approach is to locate all the units belonging to the same area by the coordinates (latitude and longitude) of the area center. This is obviously an approximation, induced by nothing but a geometrical property, and its effect on the estimates can be strong and increases with the area dimension.

We decided to proceed differently, treating the lack of geographical information as a particular problem of *measurement error*: instead of use the same coordinates for all the units, we impose a distribution for the locations inside each area. To analyze the performance of this approach, various MCMC experiments are implemented with different scenarios: missing variable (univariate and bivariate), distribution (uniform and beta) and data (simulated and real). The results show that, with the right hypothesis, the estimates under the measurement error assumption are better than that under the classic approach.

The thesis is organized in four chapters, followed by some concluding remarks.

In the first chapter we present a particular class of semiparametric models - the *additive models* - that maintain the simple additive structure of the classic regression model without imposing any assumption on the functional relation between covariates and response variable. The fitting of such models relays on nonparametric methods and we focus on penalized splines smoothers and their mixed model representation, that allows estimation and inference using mixed model methodology. Then we focused on the flexible smoothing of point clouds to obtain surface estimates, like the kriging algorithm and the radial smoothers family. Finally we present the geoadditive models.

In the second chapter we introduce the concept of statistical analysis of spatial data, presenting both potentialities and problems that arise from this spatial approach. In addition. we present a general review on the use of the spatial information in the main areas of statistical research: official statistics, epidemiology, environmental statistics, demography and social statistics, and econometrics. A particular emphasis is posed on the methods of small area estimation in presence of spatially referenced data.

In the third chapter we present the concept of measurement errors in spatial data analysis. We define the problem of uncertainty and errors in GIS from the point of view of the geographical information science, and we illustrate the statistical approach to measurement error analysis. Then, we deal with the matter of applying a geaodditive model to produce estimates for some geographical domains in the absence of point referenced auxiliary data. The performance of our measurement error approach is evaluated through various Markov Chain Monte Carlo experiments implemented under different scenarios.

The last chapter is devoted to the application of a geoadditive model in the field of poverty mapping at small area level. In particular, we apply a geoadditive small area estimation model in order to estimate the district level mean of the household log per-capita consumption expenditure for the Republic of Albania. We combine the model parameters estimated using the dataset of the 2002 Living Standard Measurement Study with the 2001 Population and Housing Census covariate information. After the definition of the geoadditive SAE model, we illustrates the results of the application. In addition. we discuss the use of two possible MSE estimators through a desing-based simulation study.

# Chapter 1

# Semiparametric Regression Models

## 1.1 Introduction[1]

In this chapter we present a particular class of semiparametric models - the *additive models* - that maintain the simple additive structure of the classic regression model without imposing any assumption on the functional relation between covariates and response variable. The fitting of such models relays on nonparametric methods and we focus on spline-base smoothers. In particular, in Section 1.3 we introduce smoothing splines in a generic structure, while we present penalized splines more in detail in Sections 1.4 and 1.5. In Section 1.6 we derive the mixed model representation of penalized splines, that allows estimation and inference using mixed model methodology. Section 1.7 is focused on the flexible smoothing of point clouds to obtain surface estimates: the kriging algorithm (subsection 1.7.2) and the radial smoothers family (subsection 1.7.3) are presented. Finally, geoadditive models, that merge additive models and kriging under a common mixed model framework, are discussed in Section 1.8.

## 1.2 Additive Models

The additive model, firstly introduced in the early 1980s (Friedman and Stuetzle, 1981) and described in detail in the monograph of Hastie and Tibshirani (1990), is a generalization of the usual linear regression model. It

[1] For the writing of this chapter, we mainly followed the structure of Ruppert, Wand and Carroll (2003).

gained popularity in applied research as a flexible and interpretable regression technique because it maintains the assumption of additivity of the covariates effects, allowing nonetheless the presence of nonlinear relationships with the response variable.

In general, considering $k$ continuous covariates $x_i$, $i = 1, ..., k$, the model has a structure similar to

$$y = f(x_1) + g(x_2, x_3) + ... + h(x_k) + \varepsilon, \tag{1.1}$$

where the $f$,$g$,...,$h$ can be both parametric or smooth functions of one or more covariates. There are several available methods to represent these smooth functions, in the following we choose to focus on spline-base smoothers and in particular on *penalized splines.*

## 1.3 Spline-base Smoothers

To introduce how splines work, let us consider a simpler version of (1.1), containing one smooth function $f$ of one covariate $x_i$

$$y_i = f(x_i) + \varepsilon_i, \tag{1.2}$$

where $y_i$ is the response variable and the $\varepsilon_i$ are i.i.d. $N(0, \sigma_\varepsilon^2)$ random variables. The function $f$ in (1.2) needs to be estimated from the observation $(x_i, y_i)$, so usually this operation is known as *scatterplot smoothing.*

To estimate $f$, we require that it is represented in such a way that (1.2) becomes a linear model. This can be done by choosing a *basis*, that is by defining the space of functions of which $f$ (or a close approximation to it) is an element. To choose the basis means to choose some *basis functions* $B_j(x)$ that span the defined space so that we obtain

$$f(x) = \sum_{j=1}^{q} \beta_j B_j(x), \tag{1.3}$$

for some values of the unknown parameters $\beta_j$. Substituting (1.3) into (1.2) clearly yields to the linear model

$$y_i = \sum_{j=1}^{q} \beta_j B_j(x_i) + \varepsilon_i, \qquad \varepsilon_i \sim N(0, \sigma_\varepsilon^2) \tag{1.4}$$

and permits to use classic estimation methods.

For example, if we decide to use a *truncated polynomial spline basis of degree p*, the basis functions are

$$1, x, ..., x^p, (x - \kappa_1)_+^p, ..., (x - \kappa_K)_+^p,$$

where $(x - \kappa)_+$ indicates the positive part of the function $(x - \kappa)$ and the values $\kappa_1, ..., \kappa_K$, called *knots*, are the points at which the sections joint. The spline function (1.3) with this basis becomes

$$f(x) = \beta_0 + \beta_1 x + ... + \beta_p x^p + \sum_{k=1}^{K} \beta_{pk}(x - \kappa_k)_+^p \qquad (1.5)$$

and the unknown parameters $\beta_0, ..., \beta_p, \beta_{p1}, ..., \beta_{pK}$ can be estimated with the ordinary least square method.

The truncated polynomial basis is just one of the possible bases available in literature, some others are the *cubic spline* basis, the *B-spline* basis or the *radial* basis. For a more complete and detailed description of bases functions we refer to Ruppert, Wand and Carroll (2003, 2009) and Wood (2006).

In principle, a change of basis does not change the fit, though some bases are more numerically stable and allow computation of a fit with greater accuracy. Usually, reason for selecting one basis over another are ease of implementation or interpretability (generally not so important since we are usually interested in the fit, not the estimated coefficients of the spline). More effective on the fit is the degree of the spline model.

After the selection of the basis, the locations and the number of the knots must be chosen as well. Typically the knots would be evenly spaced through the range of observed $x$ values, or placed at quantiles of the distribution of unique $x$ values (see Ruppert et al., 2003). Finally, the choice of the number $K$ of knots (and consequently the number of basis functions) influences the smoothness of the spline model: for $K = 0$ the equation (1.5) corresponds to a polynomial regression, on the other hand for $K$ equal to the number $n$ of observations the spline model corresponds to an exact interpolation of the data. Our interest is to estimate the underlying trend between $x$ and $y$ in regression (1.4) so an intermediate value $0 < K < n$ is required. However, we need to choose carefully: with a value of $K$ too small the resulting fitting could be too smooth and ignore the real pattern of the data, but if we use too many knots the fit could be too flexible and overfit the data.

Following the definitions in Hastie (1996), we call *low-rank* the smoothers that use considerably less than $n$ basis functions, while we call *full-rank* those with a number of basis functions approximately the same as the sample size. Hastie (1996) shows that both the smoothers produce approximatively the same fits, as the low-rank splines tend to discard components of the full-rank splines that aren't significant to the final smooth.

The use of low-rank scatterplot smoothers go back at least to Parker and Rice (1985), O'Sullivan (1986, 1988) and Kelly and Rice (1990), but they have reach a high diffusion in the last years after the articles of Eilers and Marx (1996) and Hastie (1996). When we have large sample sizes or several smoothers, the use of $K << n$ knots produces more parsimonious models and reduce significantly the computational costs.

## 1.4 Penalized Splines

By choosing the basis dimension of the spline we can control the degree of smoothing. An alternative way to approach this problem is to maintain fixed the number of knots at a size a little larger than what we believe necessary but to constrain their influence adding a penalty to the least squares fitting objective.

Consider (1.4) in its matrix form

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{1.6}$$

with $\mathbf{y} = [y_i]$, $\mathbf{X} = [B_j(x_i)]$, $\boldsymbol{\beta} = [\beta_j]$ and $\boldsymbol{\varepsilon} = [\varepsilon_i]$. The *penalized least squares* fit of (1.6) can be written as

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}}, \quad \text{where } \hat{\boldsymbol{\beta}} \text{ minimizes } \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2 + \lambda\boldsymbol{\beta}^T\mathbf{D}\boldsymbol{\beta}, \tag{1.7}$$

for some number $\lambda \geq 0$ and a symmetric positive semidefinite matrix $\mathbf{D}$, and its solution is

$$\hat{\boldsymbol{\beta}} = \left(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{D}\right)^{-1}\mathbf{X}^T\mathbf{y}. \tag{1.8}$$

The fitted values for a penalized spline regression are then given by

$$\hat{\mathbf{y}} = \mathbf{X}\left(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{D}\right)^{-1}\mathbf{X}^T\mathbf{y}$$

and, analogously to the usual linear regression model, we can define the *hat matrix* $\mathbf{S}_\lambda$ (also known as *smoother matrix*) as

$$\mathbf{S}_\lambda = \mathbf{X}\left(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{D}\right)^{-1}\mathbf{X}^T. \tag{1.9}$$

The *penalty* $\boldsymbol{\beta}^T\mathbf{D}\boldsymbol{\beta}$ is such that induces a constrain only on the parameters $\boldsymbol{\beta}$ corresponding to the knots, while leaves unconstrained the others. The *smoothing parameter* $\lambda$ controls the trade off between model fit and model smoothness: $\lambda = 0$ corresponds to the unconstrained estimation, while $\lambda \to \infty$ leads to the $p$th degree polynomial fit.

We can choose various type of penalty, depending on the selected basis and on the model assumptions. If we consider the truncated polynomial spline basis of (1.5), a possible simple penalty (Wand, 1999) is to constrain the sum of squares of the knots coefficients $\beta_{pk}$ so that

$$\sum_{k=1}^{K} \beta_{pk}^2 < C, \tag{1.10}$$

with $C$ constant. The constrain (1.10) is analogous to define

$$\mathbf{D} = \left[ \begin{array}{cc} \mathbf{0}_{(p+1)\times(p+1)} & \mathbf{0}_{(p+1)\times K} \\ \mathbf{0}_{K\times(p+1)} & \mathbf{I}_{K\times K} \end{array} \right]. \tag{1.11}$$

As we said, the penalty imposed by this matrix D is one of many possible penalties: we can constrain other functions of the knots coefficients or of the smooth function $f$. Some common alternative to (1.11) are the *P-splines* suggested by Eilers and Marx (1996), which use a B-spline basis with a difference penalty applied to the knots parameters, or the *smoothing spline* penalty, which is related to the integrated squared derivative measure of roughness (see Green and Silverman, 1994, ch.2).

Differently from the unpenalized splines case, once we provided enough knots to cover the range of value of $x_i$ reasonably well, their number and positioning does not make much difference to the fit result. However, as there are computational advantages of keeping the number of knots relatively low, studies have been done to evaluate the knots influence. Ruppert and Carroll (2000) and Ruppert (2002) present two automatic algorithms for determining the value $K$, the *myopic* algorithm and the *full-search* algorithm. However, Ruppert et al. (2003) suggest to use these algorithms only after a preliminary inspection of the complexity of the data and propose a default *rule of thumb* for general cases:

- knots locations: $\kappa_k = \big(\frac{k+1}{K+2}\big)$th sample quantile of unique $x_i$ for $k = 1, ..., K$,
- knots number: $K = \min\big(\frac{1}{4} \times \text{number of unique } x_i,\ 35\big)$.

### 1.4.1 Selection of the Smoothing Parameter

The selection of the smoothing parameter $\lambda$ is much more important as it has a profound influence on the fit. At this aim, model selection criteria can be used to choose the appropriate value of $\lambda$. Several methods are proposed in literature, the most common are the *cross validation* and the *generalized* *cross validation* criteria. Both these methods are based on the ideal purpose of choose $\lambda$ so that $\hat{f}$ is as close as possible to the real $f$, but they differ on the way to measure that closeness.

The *cross validation* criterion selects the value $\lambda$ that minimizes

$$\mathrm{CV}(\lambda) = \sum_{i=1}^{n} \left[ y_i - \hat{f}_{-i}(x_i; \lambda) \right]^2 = \sum_{i=1}^{n} \frac{(y_i - \hat{y_i})^2}{(1 - \mathbf{S}_{\lambda,ii})^2},$$

where $\hat{f}_{-i}$ indicates the penalized spline regression estimator applied to all the data but the $(x_i, y_i)$ unit and $\mathbf{S}_{\lambda,ii}$ is the $i$th diagonal element of (1.9).

The *generalized cross validation* criterion selects the value $\lambda$ that minimizes

$$\mathrm{GCV}(\lambda) = \sum_{i=1}^{n} \frac{(y_i - \hat{y_i})^2}{\left(1 - \dfrac{\mathrm{tr}(\mathbf{S}_\lambda)}{n}\right)^2}.$$

The CV and GCV criteria are quite similar, but the latter has the nice property of invariance (see Wood, 2006, ch.4). Other model selection criteria, like the *Mallows $C_p$* criterion or the *Akaike's information criterion* (AIC), can be used as well.

### 1.4.2 Degrees of Freedom of a Smoother

The choice of the smoothing parameter $\lambda$ has a great influence on the fitting result, however the value of $\lambda$ does not have a direct interpretation as the amount of "structure" that is being imposed on the fit.

Generalizing the concept of degrees of freedom for a linear model, we can define the *degrees of freedom* of the fit corresponding to the smoothing parameter $\lambda$ as

$$df_{\mathrm{fit}} = \mathrm{tr}(\mathbf{S}_\lambda). \tag{1.12}$$

This quantity can be interpreted as the *equivalent number of parameters* that we need to obtain the same fit with a parametric model.

Considering 1.9, we have

$$df_{\mathrm{fit}} = \mathrm{tr}(\mathbf{X}\left(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{D}\right)^{-1}\mathbf{X}^T) = \mathrm{tr}(\left(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{D}\right)^{-1}\mathbf{X}^T\mathbf{X}).$$

For a penalized spline smoother with $K$ knots and degree $p$, it is easily shown that

$$\begin{aligned} &\mathrm{tr}(\mathbf{S}_0) = p + 1 + K, \\ &\mathrm{tr}(\mathbf{S}_\lambda) \to p + 1 \quad \text{as} \quad \lambda \to \infty, \end{aligned}$$

so positive values of $\lambda$ correspond to

$$p + 1 < df_{\mathrm{fit}} < p + 1 + K.$$

## 1.5 Models with Multiple Explanatory Variables

Now suppose that we have two continuous covariates $x$ and $z$ for the response variable $y$. The appropriate additive model is

$$y_i = f(x_i) + g(z_i) + \varepsilon_i, \tag{1.13}$$

where $\varepsilon_i \sim N(0, \sigma^2_\varepsilon)$ and $f$, $g$ are unknown smooth functions.

Each smooth function can be represented using penalized spline regression in the same way as for the simple univariate model. Using the spline basis seen in section 1.3, we have

$$\begin{aligned} f(x) &= \beta_0 + \beta_1 x + \sum_{k=1}^{K} \beta_{k+1}(x - \kappa^x_k)_+, \\ g(z) &= \gamma_0 + \gamma_1 z + \sum_{h=1}^{H} \gamma_{h+1}(z - \kappa^z_h)_+, \end{aligned}$$

where $\beta_0, ..., \beta_{K+1}$ and $\gamma_0, ..., \gamma_{H+1}$ are the unknown parameters for $f$ and $g$ respectively, while $\kappa^x_1, ..., \kappa^x_K$ and $\kappa^z_1, ..., \kappa^z_H$ are the knot locations for the two functions. To simplify notation we use truncated linear splines to represent both $f$ and $g$, however it is perfectly possible to use any others degree of polynomial or any other alternative basis.

By substitution, (1.13) becomes

$$y_i = \beta_0 + \beta_1 x_i + \sum_{k=1}^{K} \beta_{k+1}(x_i - \kappa^x_k)_+ + \gamma_1 z_i + \sum_{h=1}^{H} \gamma_{h+1}(z_i - \kappa^z_h)_+ + \varepsilon_i, \tag{1.14}$$

with $\gamma_0$ constrained equal to zero to avoid identifiability problems, and can be written in matrix form $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ by defining

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_0 & \beta_1 & \gamma_1 & \beta_2 & ... & \beta_{K+1} & \gamma_2 & ... & \gamma_{H+1} \end{bmatrix}^T$$

and

$$\mathbf{X} = \begin{bmatrix} 1 & x_1 & z_1 & (x_1 - \kappa^x_1)_+ & ... & (x_1 - \kappa^x_K)_+ & (z_1 - \kappa^z_1)_+ & ... & (z_1 - \kappa^z_H)_+ \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_n & z_n & (x_n - \kappa^x_k)_+ & ... & (x_n - \kappa^x_K)_+ & (z_n - \kappa^z_1)_+ & ... & (z_n - \kappa^z_H)_+ \end{bmatrix}$$

The parameters $\boldsymbol{\beta}$ of the model (1.14) can be obtained by minimization of the penalized least squares objective

$$\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2 + \lambda_x \boldsymbol{\beta}^T \mathbf{D}_x \boldsymbol{\beta} + \lambda_z \boldsymbol{\beta}^T \mathbf{D}_z \boldsymbol{\beta},$$

where $\lambda_x$ and $\lambda_z$ are the smoothing parameters and $\mathbf{D_x} = \text{diag}(0, 0, \mathbf{1}_K, \mathbf{0}_H)$ and $\mathbf{D_z} = \text{diag}(0, 0, \mathbf{0}_K, \mathbf{1}_H)$ are the penalty matrices.

Defining $\mathbf{\Lambda} \equiv \lambda_x \mathbf{D}_x + \lambda_z \mathbf{D}_z$, the estimated parameters are obtained as

$$\hat{\boldsymbol{\beta}} = \left(\mathbf{X}^T\mathbf{X} + \mathbf{\Lambda}\right)^{-1} \mathbf{X}^T\mathbf{y}.$$

The total degrees of freedom of the fit are

$$df_{\text{fit}} = \text{tr}(\mathbf{X}\left(\mathbf{X}^T\mathbf{X} + \mathbf{\Lambda}\right)^{-1} \mathbf{X}^T) = \text{tr}(\left(\mathbf{X}^T\mathbf{X} + \mathbf{\Lambda}\right)^{-1} \mathbf{X}^T\mathbf{X}).$$

However, we can also compute the degrees of freedom for each component. Let $q = 1 + (K + 1) + (H + 1)$ denote the number of columns in $\mathbf{X}$ and let

$$\{I_0, I_1, I_2\}$$

be a partition of the columns indices $\{1, ..., q\}$ such that $I_0$ corresponds to the intercept $\beta_0$, $I_1$ corresponds to $f$ and $I_2$ corresponds to $g$. That is:

$$I_0 = \{1\}, \quad I_1 = \{2, 4, ..., (K + 3)\}, \quad I_2 = \{3, (K + 4), ..., (H + K + 3)\}.$$

Define $\mathbf{E}_j$, $j = 0, 1, 2$, to be the $q \times q$ diagonal matrix with ones in the diagonal elements with indices $I_j$ and zeros elsewhere, then corresponding degrees of freedom for the $j$th smoother can be computes as

$$df_j = \text{tr}(\mathbf{E}_j \left(\mathbf{X}^T\mathbf{X} + \mathbf{\Lambda}\right)^{-1} \mathbf{X}^T\mathbf{X}),$$

which is the sum over the indices $I_j$ of the diagonal elements of the matrix $\left(\mathbf{X}^T\mathbf{X} + \mathbf{\Lambda}\right)^{-1} \mathbf{X}^T\mathbf{X}$. Thus, we have that $df_{\text{fit}} = df_0 + df_1 + df_2$.

The selection of the numbers $K$ and $H$ of knots and of the smoothing parameters $\lambda_x$ and $\lambda_z$ follows the same rules and criteria presented for the simple univariate model. Moreover, the extension to the additive model with a higher numbers of smooth functions is straightforward.

## 1.6 Linear Mixed Model Representation of Penalized Splines

A convenient way to work with penalized spline functions is to consider their mixed model representation (Wand, 2003). In fact, as shown in Brumback, Ruppert and Wand (1999), smoothing methods that use penalized basis functions can be formulated as maximum likelihood estimators and best predictors in a mixed model framework. This formulation is a useful instrument because it allows estimation and inference using mixed model methodology and software.

Consider once again the penalized spline regression with the truncated polynomial spline basis. To simplify explanation, some changes in notation occur and we rewrite model (1.6) as

$$y_i = \beta_0 + \beta_1 x_i + ... + \beta_p x_i^p + \sum_{k=1}^{K} u_k (x_i - \kappa_k)_+^p + \varepsilon_i, \tag{1.15}$$

with $\varepsilon_i \sim N(0, \sigma_\varepsilon^2)$. Let

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \vdots \\ \beta_p \end{bmatrix} \quad \text{and} \quad \mathbf{u} = \begin{bmatrix} u_1 \\ \vdots \\ u_K \end{bmatrix}$$

be the coefficients of the polynomial functions and the truncated functions, respectively. Corresponding to these vectors, define

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

and

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

Then the penalized fitting criterion (1.7), divided by $\sigma_\varepsilon^2$, can be written as

$$\frac{1}{\sigma_\varepsilon^2} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\mathbf{u}\|^2 + \frac{\lambda}{\sigma_\varepsilon^2} \|\mathbf{u}\|^2. \tag{1.16}$$

Instead of treat all the coefficients of (1.15) as unknown but fixed elements, as in the previous sections, now we define the $u_k$ as i.i.d. random variables with distribution $N(0, \sigma_u^2)$ uncorrelated with the error component. Then the formula (1.16) corresponds to Henderson's criterion to obtain the best linear unbiased predictor (BLUP) for linear mixed models (Robinson, 1991), with $\sigma_u^2 = \frac{\sigma_\varepsilon^2}{\lambda}$.

In summary, the mixed model representation of the penalized spline regression is

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \boldsymbol{\varepsilon}, \qquad \text{Cov}\begin{bmatrix} \mathbf{u} \\ \boldsymbol{\varepsilon} \end{bmatrix} = \begin{bmatrix} \boldsymbol{\Sigma}_u & 0 \\ 0 & \boldsymbol{\Sigma}_\varepsilon \end{bmatrix} \tag{1.17}$$

with $\mathbf{\Sigma}_u = \sigma_u^2 \mathbf{I}_K$, $\mathbf{\Sigma}_\varepsilon = \sigma_\varepsilon^2 \mathbf{I}_n$ and the smoothing parameter $\lambda$ is automatically selected as $\dfrac{\sigma_\varepsilon^2}{\sigma_u^2}$.

Let

$$\mathrm{Var}(\mathbf{y}) \equiv \mathbf{V} = \mathbf{Z}\mathbf{\Sigma}_u\mathbf{Z}^T + \mathbf{\Sigma}_\varepsilon \tag{1.18}$$

be the covariance matrix of $\mathbf{y}$. Following from the linear mixed model theory (Henderson, 1975), if the variance components $\sigma_u^2$ and $\sigma_\varepsilon^2$ are known, we derive that the BLUPs of the model coefficients are

$$\tilde{\boldsymbol{\beta}} = \left(\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{V}^{-1}\mathbf{y}, \tag{1.19}$$

$$\tilde{\mathbf{u}} = \mathbf{\Sigma}_u\mathbf{Z}^T\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\tilde{\boldsymbol{\beta}}). \tag{1.20}$$

However, in practical applications the variance components are usually unknown and must be estimated from the observed data.

Replacing the unknown parameters with their estimated values $\hat{\sigma}_u^2$ and $\hat{\sigma}_\varepsilon^2$ in (1.19) and (1.20), we obtain the empirical best linear unbiased predictors (EBLUPs) of the model coefficients as

$$\hat{\boldsymbol{\beta}} = \left(\mathbf{X}^T\hat{\mathbf{V}}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}^T\hat{\mathbf{V}}^{-1}\mathbf{y}, \tag{1.21}$$

$$\hat{\mathbf{u}} = \hat{\mathbf{\Sigma}}_u\mathbf{Z}^T\hat{\mathbf{V}}^{-1}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}), \tag{1.22}$$

where the *hat* symbol indicates that the covariance matrices $\mathbf{V}$, $\mathbf{\Sigma}_u$ and $\mathbf{\Sigma}_\varepsilon$ contain the estimated values $\hat{\sigma}_u^2$ and $\hat{\sigma}_\varepsilon^2$. Similarly, the smoothing parameter can be selected as $\lambda = \dfrac{\hat{\sigma}_\varepsilon^2}{\hat{\sigma}_u^2}$.

### 1.6.1 Estimation of the Variance Components

There is a large and varied literature on estimation of the variance components in mixed models, however *maximum likelihood* (ML) and *restricted maximum likelihood* (REML) are today the most common methods for estimating the parameters in covariance matrices (McCulloch and Searle, 2001).

If we consider again the covariance matrix $\mathbf{V}$ defined in (1.18), we can rewrite it to show directly the connection with the unknown parameters $\sigma_u^2$ and $\sigma_\varepsilon^2$:

$$\mathbf{V} = \sigma_u^2\mathbf{Z}\mathbf{Z}^T + \sigma_\varepsilon^2\mathbf{I}_n.$$

The ML estimate of the matrix $\mathbf{V}$ is based on the model

$$\mathbf{y} \sim N(\mathbf{X}\boldsymbol{\beta}, \mathbf{V}).$$

The log-likelihood of $\mathbf{y}$ under this model is

$$\ell(\boldsymbol{\beta}, \mathbf{V}) = -\frac{1}{2}\left[n\log(2\pi) + \log|\mathbf{V}| + (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})\right] \tag{1.23}$$

and the ML estimates of $(\boldsymbol{\beta}, \mathbf{V})$ are the ones that maximize $\ell(\boldsymbol{\beta}, \mathbf{V})$.

We first optimize over $\boldsymbol{\beta}$ and we obtain, for any fixed $\mathbf{V}$,

$$\tilde{\boldsymbol{\beta}}_{ML} = \left(\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{V}^{-1}\mathbf{y},$$

which corresponds to the BLUP shown in (1.19). On substitution into (1.23) we obtain the *profile log-likelihood* for $\mathbf{V}$:

$$\ell_p(\mathbf{V}) = -\frac{1}{2}\left\{n\log(2\pi) + \log|\mathbf{V}| + \mathbf{y}^T\mathbf{V}^{-1}\left[\mathbf{I} - \mathbf{X}\left(\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{V}^{-1}\right]\mathbf{y}\right\}$$

and the ML estimates of the parameters $\sigma_u^2$ and $\sigma_\varepsilon^2$ in $\mathbf{V}$ can be found by numerical maximization of $\ell_p(\mathbf{V})$ over those parameters.

To obtain the REML estimate of $\mathbf{V}$ we proceed similarly to the ML estimation, but the criterion function to be maximize over the parameters $\sigma_u^2$ and $\sigma_\varepsilon^2$ is the *restricted log-likelihood*

$$\ell_r(\mathbf{V}) = \ell_p(\mathbf{V}) - \frac{1}{2}\log|\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X}|.$$

The main advantage of REML over ML is that REML takes into account the degrees of freedom for the fixed effects in the model. For small sample sizes REML is expected to be more accurate than ML, but for large samples there should be little difference between the two approaches.

### 1.6.2 Multiple Explanatory Variables

If we use a penalized spline regression model with two or more explanatory variables, like in model (1.14), the linear mixed model representation is easily straightforward. We can rewrite regression (1.14) as

$$y_i = \beta_0 + \beta_x x_i + \beta_z z_i + \sum_{k=1}^{K} u_k^x(x_i - \kappa_k^x)_+ + \sum_{h=1}^{H} u_h^z(z_i - \kappa_h^z)_+ + \varepsilon_i,$$

with $\varepsilon_i \sim N(0, \sigma_\varepsilon^2)$, $u_k^x \sim N(0, \sigma_x^2)$ and $u_h^z \sim N(0, \sigma_z^2)$, all uncorrelated.

Define the coefficients vectors

$$\boldsymbol{\beta} = \begin{bmatrix}\beta_0\\ \beta_x\\ \beta_z\end{bmatrix}, \quad \mathbf{u}_x = \begin{bmatrix}u_1^x\\ \vdots\\ u_K^x\end{bmatrix} \quad \text{and} \quad \mathbf{u}_z = \begin{bmatrix}u_1^z\\ \vdots\\ u_H^z\end{bmatrix}$$

and the corresponding design matrices

$$\mathbf{X} = \begin{bmatrix} 1 & x_1 & z_1 \\ \vdots & \vdots & \vdots \\ 1 & x_n & z_n \end{bmatrix}$$

and

$$\mathbf{Z}_x = \begin{bmatrix} (x_1 - \kappa_1^x)_+ & ... & (x_1 - \kappa_K^x)_+ \\ \vdots & \vdots & \vdots \\ (x_n - \kappa_1^x)_+ & ... & (x_n - \kappa_K^x)_+ \end{bmatrix}, \mathbf{Z}_z = \begin{bmatrix} (z_1 - \kappa_1^z)_+ & ... & (z_1 - \kappa_H^z)_+ \\ \vdots & \vdots & \vdots \\ (z_n - \kappa_1^z)_+ & ... & (z_n - \kappa_H^z)_+ \end{bmatrix}.$$

The penalized fitting criterion (1.7) divided by $\sigma_\varepsilon^2$, now becomes

$$\frac{1}{\sigma_\varepsilon^2} \left\| \mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}_x\mathbf{u}_x - \mathbf{Z}_z\mathbf{u}_z \right\|^2 + \frac{\lambda_x}{\sigma_\varepsilon^2} \left\| \mathbf{u}_x \right\|^2 + \frac{\lambda_z}{\sigma_\varepsilon^2} \left\| \mathbf{u}_z \right\|^2$$

and the smoothing parameters are selected as $\lambda_x = \dfrac{\sigma_\varepsilon^2}{\sigma_x^2}$ and $\lambda_z = \dfrac{\sigma_\varepsilon^2}{\sigma_z^2}$.

The linear mixed model representation is then

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}_x\mathbf{u}_x + \mathbf{Z}_z\mathbf{u}_z + \boldsymbol{\varepsilon}, \qquad \text{Cov} \begin{bmatrix} \mathbf{u_x} \\ \mathbf{u_z} \\ \boldsymbol{\varepsilon} \end{bmatrix} = \begin{bmatrix} \boldsymbol{\Sigma}_x & 0 & 0 \\ 0 & \boldsymbol{\Sigma}_z & 0 \\ 0 & 0 & \boldsymbol{\Sigma}_\varepsilon \end{bmatrix} \tag{1.24}$$

with $\boldsymbol{\Sigma}_x = \sigma_x^2\mathbf{I}_K$, $\boldsymbol{\Sigma}_z = \sigma_z^2\mathbf{I}_H$, $\boldsymbol{\Sigma}_\varepsilon = \sigma_\varepsilon^2\mathbf{I}_n$,

$$\text{Var}(\mathbf{y}) \equiv \mathbf{V} = \mathbf{Z}_x\boldsymbol{\Sigma}_x\mathbf{Z}_x^T + \mathbf{Z}_z\boldsymbol{\Sigma}_z\mathbf{Z}_z^T + \boldsymbol{\Sigma}_\varepsilon$$

and the EBLUPs of the model coefficients are

$$\hat{\boldsymbol{\beta}} = \left(\mathbf{X}^T\hat{\mathbf{V}}^{-1}\mathbf{X}\right)^{-1} \mathbf{X}^T\hat{\mathbf{V}}^{-1}\mathbf{y}, \tag{1.25}$$

$$\hat{\mathbf{u}}_x = \hat{\boldsymbol{\Sigma}}_x\mathbf{Z}_x^T\hat{\mathbf{V}}^{-1}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}), \tag{1.26}$$

$$\hat{\mathbf{u}}_z = \hat{\boldsymbol{\Sigma}}_z\mathbf{Z}_z^T\hat{\mathbf{V}}^{-1}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}), \tag{1.27}$$

with $\hat{\sigma}_x^2$, $\hat{\sigma}_z^2$ and $\hat{\sigma}_\varepsilon^2$ obtained by ML or REML estimation.

If we have more than two explicative variables, the mixed model representation doesn't change: smoothing components are added as a new random effects term $\mathbf{Zu}$, while linear components can be incorporated as fixed effects in the $\mathbf{X}\boldsymbol{\beta}$ term. Moreover, the mixed model structure provides a unified and modular framework that allows to easily extend the model to include generalized responses, small area estimation, longitudinal data, hazard regression models, missing data and so on (Ruppert, Wand and Carroll, 2009).

## 1.7 Bivariate Smoothing

In the previous sections we described how to handle smoothing functions of one continuous variable. Analogously to scatterplot smoothing, bivariate smoothing deals with the flexible smoothing of point clouds to obtain surface estimates.

Bivariate smoothing is of central interest in application areas such as environmental study, mining, hydrology and epidemiology, where is common the use of *geostatistics* methods to analyze geographically referenced responses. The main tool of geostatistics is *kriging* and it has a close connection with penalized spline smoothing.

The geographical application, however, is not the only use of bivariate smoothing as the method can be applied to handle the non-linear relation between any two continuous predictors and a response variable.

### 1.7.1 Bivariate Basis Functions

Bivariate smoothing extends the penalized spline structure in two dimensions using bivariate basis functions.

If we consider two continuous predictors $s$ and $t$ of the response variable $y$, the general bivariate smoothing model is

$$y_i = f(s_i, t_i) + \varepsilon_i, \tag{1.28}$$

where $f$ is an unknown real-valued bivariate function.

The natural extension for truncated polynomial splines is to form all the pairwise products of the univariate bases functions

$$\begin{gathered} 1, s, ..., s^p, (s-\kappa_1^s)_+^p, ..., (s-\kappa_K^s)_+^p, \\ 1, t, ..., t^q, (t-\kappa_1^t)_+^q, ..., (t-\kappa_H^t)_+^q. \end{gathered}$$

The resulting basis is known as a *tensor product* basis and the relative regression spline model, for $p = q = 1$, is

$$\begin{split} y_i = \beta_0 + \beta_1 s_i + \beta_2 t_i + \beta_3 s_i t_i + \sum_{k=1}^{K} u_k^s (s_i - \kappa_k^s)_+ + \\ \sum_{h=1}^{H} u_h^t (t_i - \kappa_h^t)_+ + \sum_{k=1}^{K}\sum_{h=1}^{H} u_{kh}^{st} (s_i - \kappa_k^s)_+ (t_i - \kappa_h^t)_+ + \varepsilon_i. \end{split} \tag{1.29}$$

An inconvenience of tensor product splines is that the number of coefficients in model (1.29) increases really fast with the knots numbers $K$, $H$ and the polynomial degrees $p$, $q$. Moreover, the basis depends on the orientation of the coordinates axes and it is not *rotational invariant.* This property is not so relevant in a non-geographical application, but for geographical smoothing is a desirable characteristic for the result to be independent of axis orientation.

Rotational invariance can be achieved through the use of *radial basis functions*, that are of the form

$$C\left(\left\|(s,t)-(\kappa^s,\kappa^t)\right\|\right)$$

for some univariate function $C$. Since the value of the function at $(s,t)$ depends only on the distance from the knot $(\kappa^s,\kappa^t)$, the function is radially symmetric about this point.

### 1.7.2 Kriging

The term *kriging* refers to a widely used method for interpolating or smoothing spatial data.

Given a set of data $y_i$, $i=1,...,n$, at spatial location $\mathbf{x}_i$, $\mathbf{x}\in\Re^2$, the simple kriging model for interpolating the underlying spatial surface is

$$y_i=\mu+S(\mathbf{x}_i)+\varepsilon_i, \tag{1.30}$$

where $S(\mathbf{x})$ is a zero-mean stationary stochastic process in $\Re^2$ and the $\varepsilon_i$ are assumed to be independent zero-mean random variables with common variance $\sigma^2_\varepsilon$ and distributed independently of $S$ (Cressie, 1993). Interpolation at an arbitrary location $\mathbf{x}_0\in\Re^2$ is done through

$$\hat{y}_0=\bar{y}+\hat{S}(\mathbf{x}_0), \tag{1.31}$$

where $\hat{S}(\mathbf{x}_0)$ is the best linear predictor of $S(\mathbf{x}_0)$ based on the data in $\mathbf{y}$.

For a known covariance structure of $S$, the resulting predictor is

$$\hat{S}(\mathbf{x}_0)=\mathbf{c}_0^T(\mathbf{C}+\sigma^2_\varepsilon\mathbf{I}_n)(\mathbf{y}-\mu\mathbf{1}), \tag{1.32}$$

where

$$\mathbf{C}\equiv\mathrm{Cov}\begin{bmatrix}S(\mathbf{x}_1)\\ \vdots\\ S(\mathbf{x}_n)\end{bmatrix}\quad\text{and}\quad\mathbf{c}_0=\begin{bmatrix}\mathrm{Cov}\{S(\mathbf{x}_0),S(\mathbf{x}_1)\}\\ \vdots\\ \mathrm{Cov}\{S(\mathbf{x}_0),S(\mathbf{x}_n)\}\end{bmatrix}$$

The practical implementation of equation (1.32 requires the definition of the covariance structure of $S(\mathbf{x})$. The usual approach is to define a parsimonious model for $\mathrm{Cov}\{S(\mathbf{x}),S(\mathbf{x}+\mathbf{h})\}$, estimate the required parameters to derive the estimates of $\hat{\mathbf{C}}$ and $\hat{\mathbf{c}}_0$ and then substitute in (1.31) to obtain:

$$\hat{y}_0=\bar{y}+\hat{\mathbf{c}}_0^T(\hat{\mathbf{C}}+\hat{\sigma}^2_\varepsilon\mathbf{I}_n)(\mathbf{y}-\bar{y}\mathbf{1}). \tag{1.33}$$

Usually, a common assumption to simplify the covariance structure of $S$ is the assumption of *isotropy*, that is

$$\mathrm{Cov}\{S(\mathbf{x}), S(\mathbf{x}+\mathbf{h})\} \text{ depends only on } \|\mathbf{h}\|. \qquad (1.34)$$

This is a stronger assumption than stationary, because it says that the covariance is independent both of location and direction, and sometimes it couldn't be valid.

Condition (1.34) implies that

$$\mathbf{C} = \left[C\left(\|\mathbf{x}_i - \mathbf{x}_j\|\right)\right]_{1\leq i,j\leq n},$$

where

$$C(r) \equiv \sigma_S^2 C_0(r), \qquad \sigma_S^2 \equiv \mathrm{Var}\left[S(\mathbf{x})\right],$$

with $C_0(0) = 1$. The functions $C$ and $C_0$ are respectively the *covariance function* and the *correlation function* of the isotropic process $S(\mathbf{x})$ and they should be chosen to ensure that $\mathbf{C}$ is a valid covariance matrix.

A simple way to characterize the class of functions that can be chosen as correlation function is given by the *Bochner's theorem*. It states that $C_0$ is a valid correlation functions is and only if it is the characteristic function of a symmetric random variable (Ruppert et al., 2003).

The resulting class of candidate correlation functions is quite big. Quite diffuse are the *exponential* correlation function

$$C_0(r) = e^{-\left|\frac{r}{\rho}\right|}, \qquad (1.35)$$

for some $\rho > 0$, known as *range* parameter; and the *gaussian* correlation function

$$C_0(r) = e^{-r^2}. \qquad (1.36)$$

Moreover, Stein (1999) strongly suggests the use of the *Matérn family.*

The classic approach to selecting $C_0$ and its parameters, as well as $\sigma_S^2$ and $\sigma_\varepsilon^2$, is through the *variogram* analysis. For a detailed description of procedures and others possible correlation functions we refer to Cressie (1993), Ruppert et al. (2003) and Stein (1999).

### 1.7.3 Radial Smoothers

Kriging provides one method of radial smoothing, however it is not the only one. Moreover, we will show that it belongs to a bigger family of radial smoothers, known as *general radial smoothers.*

To simplify explanation, we firstly present this family of radial smoothers in one dimension. Because of the radial nature of the smoothing, the higher-dimensional extension is immediate.

Recall the problem of scatterplot smoothing (1.2) using full-rank penalized truncated linear splines and define

$$\mathbf{X} = \begin{bmatrix} 1 & x_i \end{bmatrix}_{1 \le i \le n} \quad \text{and} \quad \mathbf{Z} = \begin{bmatrix} (x_i - x_j)_+ \end{bmatrix}_{1 \le i,j \le n}. \tag{1.37}$$

We have seen in Section 1.4 that the fitted values are given by

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}} + \mathbf{Z}\hat{\mathbf{u}},$$

where $\hat{\boldsymbol{\beta}}$ and $\hat{\mathbf{u}}$ are obtained minimizing

$$\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\mathbf{u}\|^2 + \lambda \begin{bmatrix} \boldsymbol{\beta} \\ \mathbf{u} \end{bmatrix}^T \mathbf{D} \begin{bmatrix} \boldsymbol{\beta} \\ \mathbf{u} \end{bmatrix},$$

with $\mathbf{D} = \text{diag}(0, 0, 1, ..., 1)$.

Now consider a linear transformation of the truncated linear basis in such a way that $\mathbf{X}$ remain unchanged and $\mathbf{Z}$ becomes the radially symmetric matrix

$$\mathbf{Z}_R = \begin{bmatrix} |x_i - x_j| \end{bmatrix}_{1 \le i,j \le n}.$$

Such transformation can be expressed as

$$\begin{bmatrix} \mathbf{X} & \mathbf{Z}_R \end{bmatrix} = \begin{bmatrix} \mathbf{X} & \mathbf{Z} \end{bmatrix} \mathbf{L},$$

where $\mathbf{L}$ is an $(n+2) \times (n+2)$ matrix. The vector of fitted values is now obtained as

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}}_R + \mathbf{Z}_R\hat{\mathbf{u}}_R$$

and $\hat{\boldsymbol{\beta}}_R$ and $\hat{\mathbf{u}}_R$ are obtained minimizing

$$\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}_R\mathbf{u}\|^2 + \lambda \begin{bmatrix} \boldsymbol{\beta} \\ \mathbf{u} \end{bmatrix}^T \mathbf{L}^T\mathbf{D}\mathbf{L} \begin{bmatrix} \boldsymbol{\beta} \\ \mathbf{u} \end{bmatrix}.$$

This new penalty is not easy to extend to the multivariate case and, more important, it is still not radially symmetric. A simple way to answer both the requests is to replace it with $\lambda\mathbf{u}^T\mathbf{Z}_R\mathbf{u}$, so that the criterion becomes

$$\begin{bmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\mathbf{u}} \end{bmatrix} = \arg\min_{\boldsymbol{\beta},\mathbf{u}} \left( \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}_R\mathbf{u}\|^2 + \lambda\mathbf{u}^T\mathbf{Z}_R\mathbf{u} \right). \tag{1.38}$$

In addition, it can be shown (Green and Silverman, 1994) that the use of (1.38) corresponds to the *thin plate spline* family of smoothers, where we penalize the integral of a squared derivative of $f(x_i)$. Specifically, in this case we penalize the first derivative, which is appropriate for a linear spline, however, it is possible to penalize the $m$th derivative for any $m$ such that $2m - d > 0$, where $d$ is the dimension of $x$.

Analogously to Section 1.6, we want to rewrite the estimates of $\boldsymbol{\beta}$ and $\mathbf{u}$ as EBLUPs of a linear mixed model. The criterion (1.38) corresponds to fitting the model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}_R\mathbf{u} + \boldsymbol{\varepsilon}, \qquad \text{Cov}\begin{bmatrix}\mathbf{u} \\ \boldsymbol{\varepsilon}\end{bmatrix} = \begin{bmatrix}\sigma_u^2\mathbf{Z}_R^{-1} & 0 \\ 0 & \sigma_\varepsilon^2\mathbf{I}_n\end{bmatrix}.$$

However, this is not a valid linear mixed model because it implies that $\text{Cov}(\mathbf{Z}_R\mathbf{u}) = \sigma_u^2\mathbf{Z}_R$ even though $\mathbf{Z}_R$ is not a proper covariance matrix as it is not necessarily semi-positive definite.

Possible ways to obtain a valid mixed model are to replace $\mathbf{Z}_R$ using its *positive definitization* $\mathbf{Z}_P = (\mathbf{Z}_R^{1/2})^T\mathbf{Z}_R^{1/2}$ or to use *generalized covariance functions* (French et al., 2001). Another way is to use a proper covariance matrix and this approach corresponds to the kriging method.

### Low-Rank Radial Smoothers

Whatever is the final choice for the radial basis matrix, finally the radial smoother parameters are EBLUPs for a mixed model of the form

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}_C\mathbf{u} + \boldsymbol{\varepsilon},$$

where $\mathbf{X}$ is defined by (1.37),

$$\text{Cov}(\mathbf{u}) = \sigma_u^2(\mathbf{Z}_C^{-1/2})(\mathbf{Z}_C^{-1/2})^T$$

and

$$\mathbf{Z}_C \equiv \left[C\left(|x_i - x_j|\right)\right]_{1\leq i,j\leq n}$$

for some real-valued function $C$ possibly containing parameters.

Such a smoother is full-rank, however for practical implementation the *low-rank* version of radial smoothers is much more interesting.

Let $\kappa_1, ..., \kappa_K$ be a set of knots corresponding to the basis functions

$$C\left(|x - \kappa_k|\right), \qquad \text{with } 1 \leq k \leq K.$$

Then the low-rank penalized radial smoothing spline is equivalent to fitting the model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}_K\mathbf{u} + \boldsymbol{\varepsilon},$$

with

$$\mathrm{Cov}(\mathbf{u}) = \sigma_u^2(\boldsymbol{\Omega}_K^{-1/2})(\boldsymbol{\Omega}_K^{-1/2})^T,$$

$$\mathbf{Z}_K \equiv \left[C\left(|x_i - \kappa_k|\right)\right]_{1\leq i\leq n, 1\leq k\leq K}, \tag{1.39}$$

$$\boldsymbol{\Omega}_K \equiv \left[C\left(|\kappa_k - \kappa_h|\right)\right]_{1\leq k,h\leq K}. \tag{1.40}$$

Using the transformation $\mathbf{Z} = \mathbf{Z}_K\boldsymbol{\Omega}_K^{-1/2}$, the final model can be written as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \boldsymbol{\varepsilon}, \qquad \mathrm{Cov}\begin{bmatrix}\mathbf{u}\\ \boldsymbol{\varepsilon}\end{bmatrix} = \begin{bmatrix}\sigma_u^2\mathbf{I}_K & 0\\ 0 & \sigma_\varepsilon^2\mathbf{I}_n\end{bmatrix}$$

and can be estimated through mixed model software.

### Higher-Dimensional Radial Smoothers

Radial smoothers in one dimension present performances similar to the ordinary penalized splines presented in the previous sections. The real interest in this kind of smoothers arises from their multivariate application.

Since their dependence on the data is, by construction, only through the point-to-point distances

$$|x_i - \kappa_k|, \qquad 1 \leq i \leq n,\ 1 \leq k \leq K, \tag{1.41}$$

the extension to $\mathbf{x}_i \in \Re^d$ essentially involves replacing the distances (1.41) with

$$\|\mathbf{x}_i - \boldsymbol{\kappa}_k\|, \qquad 1 \leq i \leq n,\ 1 \leq k \leq K.$$

For $\mathbf{x}_i$ and $\boldsymbol{\kappa}_k \in \Re^d$, *low-rank thin plate splines* of higher dimensions can be obtained by taking the design matrices $\mathbf{X}$ to have columns spanning the space of all $d$-dimensional polynomials in the components of the $\mathbf{x}_i$ with degree less than $m$ and

$$\mathbf{Z} = \left[C\left(\|\mathbf{x}_i - \boldsymbol{\kappa}_k\|\right)\right]_{1\leq i\leq n, 1\leq k\leq K} \cdot \left[C\left(\|\boldsymbol{\kappa}_h - \boldsymbol{\kappa}_k\|\right)\right]_{1\leq k,h\leq K}^{-1/2},$$

where

$$C(\mathbf{r}) = \begin{cases}\|\mathbf{r}\|^{2m-d} & \text{for } d \text{ odd},\\ \|\mathbf{r}\|^{2m-d}\log\|\mathbf{r}\| & \text{for } d \text{ even},\end{cases} \tag{1.42}$$

and $m$ is an integer such as $2m-d>0$ that control the smoothness of $C(\cdot)$.

Alternatively, we could use *low-rank radial basis* functions corresponding to a proper covariance function as seen in Section 1.7.2. For example, the two simplest member of the *Matérn class* are

$$C(\mathbf{r}) = \begin{cases}\exp\left(-\|\mathbf{r}\|/\rho\right) & \nu = 1/2,\\ \exp\left(-\|\mathbf{r}\|/\rho\right)\left(1+\|\mathbf{r}\|/\rho\right) & \nu = 3/2.\end{cases} \tag{1.43}$$

### 1.7.4 Knots Selections

If we are working with full-rank smoothers, like in the classic kriging approach, the knots correspond to the predictors. With low-rank smoothers however a set of $K < n$ knots in $\Re^d$ needs to be chosen.

One possible approach is to put down a rectangular lattice of knots that covers the range of all the predictors. This method is the multivariate version of choosing equispaced knots on an interval in one dimension. If the predictors are regularly spaced on the surface, this approach results in a good selection of knots, otherwise it tend to waste a lot of knots by covering empty areas.

A reasonable alternative strategy is to have the knots follow the distribution of the predictor space. In one dimension, this corresponds to choose the knots on the quantiles of the predictor distribution, but the extension to higher dimensions is not straightforward as we lose the notion of quantile.

A way to handle the $d > 1$ case is to recall that select the sample quantiles of $x$ corresponds to *maximize the separation* of $K$ points among the unique values $x_i$. In higher dimensions, *space filling designs* are based on the same principle of maximal separation (Nychka and Saltzman, 1998). The use of space filling designs, like the `clara` algorithm of Kaufman and Rousseeuw (1990), usually supported by some software packages, ensure coverage of the covariate space as well as parsimony in the number of knots.

## 1.8 Geoadditive Models

As we presented in the previous section, we can obtain a map of the mean of a response variable exploiting the exact knowledge of the spatial coordinates (latitude and longitude) of the studied phenomenon by using bivariate smoothing techniques. However, usually the spatial information alone does not properly explain the pattern of the response variable and we need to introduce some covariates in a more complex model.

*Geoadditive models*, introduced by Kammann and Wand (2003), answer this problem as they analyze the spatial distribution of the study variable while accounting for possible linear or non-linear covariate effects. Under the additivity assumption they can handle such covariate effects by combining the ideas of additive models and kriging, both represented as linear mixed model.

Let $s_i$ and $t_i$, $1 \leq i \leq n$, be continuous predictors of $y_i$ at spatial location $\mathbf{x}_i$, $\mathbf{x} \in \Re^2$. A geoadditive model for such data can be formulated as

$$y_i = f(s_i) + g(t_i) + h(\mathbf{x}_i) + \varepsilon_i, \qquad \varepsilon_i \sim N(0, \sigma_\varepsilon^2), \tag{1.44}$$

where $f$ and $g$ are unspecified smooth functions of one variable and $h$ is an unspecified bivariate smooth functions. Following the representation presented in Sections 1.6 and 1.7.3, the model (1.44) can be written as a mixed model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \boldsymbol{\varepsilon}, \qquad \text{Cov}\begin{bmatrix}\mathbf{u}\\ \boldsymbol{\varepsilon}\end{bmatrix} = \begin{bmatrix}\sigma_s^2\mathbf{I}_{K_s} & 0 & 0 & 0\\ 0 & \sigma_t^2\mathbf{I}_{K_t} & 0 & 0\\ 0 & 0 & \sigma_x^2\mathbf{I}_{K_x} & 0\\ 0 & 0 & 0 & \sigma_\varepsilon^2\mathbf{I}_n\end{bmatrix} \tag{1.45}$$

where

$$\boldsymbol{\beta} = \left[\beta_0, \beta_s, \beta_t, \beta_x^T\right], \qquad \mathbf{u} = \left[u_1^s, ..., u_{K_s}^s, u_1^t, ..., u_{K_t}^t, u_1^x, ..., u_{K_x}^x\right],$$

$$\mathbf{X} = \left[1, s_i, t_i, \mathbf{x}_i^T\right]_{1\le i\le n}$$

and $\mathbf{Z}$ is obtained by concatenating the matrices containing spline basis functions to handle $f$, $g$, and $h$, respectively

$$\mathbf{Z} = \left[\mathbf{Z}_s | \mathbf{Z}_t | \mathbf{Z}_x\right],$$

$$\begin{aligned}
\mathbf{Z}_s &= \left[(\mathbf{s_i} - \kappa_1^s)_+, ..., (\mathbf{s_i} - \kappa_{K_s}^s)_+\right]_{1\le i\le n},\\
\mathbf{Z}_t &= \left[(\mathbf{t_i} - \kappa_1^t)_+, ..., (\mathbf{t_i} - \kappa_{K_t}^t)_+\right]_{1\le i\le n},\\
\mathbf{Z}_x &= \left[C\left(\mathbf{x}_i - \boldsymbol{\kappa}_k^x\right)\right]_{1\le i\le n, 1\le k\le K_x} \left[C\left(\boldsymbol{\kappa}_h^x - \boldsymbol{\kappa}_k^x\right)\right]_{1\le h,k\le K_x}^{-1/2}.
\end{aligned}$$

This linear mixed model representation permits to fit model (1.44) simultaneously using mixed model methodology and software, to obtain the estimates $\hat{\boldsymbol{\beta}}$ and $\hat{\mathbf{u}}$ by the EBLUPs (1.21) and (1.22) and $\hat{\sigma}_s^2$, $\hat{\sigma}_t^2$, $\hat{\sigma}_x^2$ and $\hat{\sigma}_\varepsilon^2$ by REML/ML estimation.

The addition of others explicative variables is straightforward: smoothing components are added in the random effects term $\mathbf{Zu}$, while linear components can be incorporated as fixed effects in the $\mathbf{X}\boldsymbol{\beta}$ term. Moreover, the mixed model structure provides a unified and modular framework that allows to easily extend the model to include various kind of generalization and evolution.

# Chapter 2

# Statistical Data and Spatial Information

## 2.1 Introduction[1]

In this second chapter we introduce the concept of statistical analysis of spatial data and we present a general overview on the use of the spatial information in main statistical research areas.

Specifically, in Section 2.2 we define the concept of spatial data analysis and we classify the types of spatial data, in relation to the specific objectives of analysis in which they are involved. Section 2.3 presents the Geographical Information System (GIS), a powerful instrument for statistical spatial analysis. The potentiality of the use of GIS in statistical analysis is illustrated with some generic examples, presenting also new issues and problems that arise from this spatial approach. Finally, in Section 2.4 we present a review on the use of spatial information in the main areas of statistical research: official statistics (subsection 2.4.1), epidemiology (subsection 2.4.2), environmental statistics (subsection 2.4.3), demography and social statistics (subsection 2.4.4) and econometrics (subsection 2.4.5). A particular emphasis is posed on the methods of small area estimation in presence of spatially referenced data (subsection 2.4.6).

[1] For the writing of this chapter, we mainly referred to Petrucci, Bocci, Borgoni, Civardi, Salvati, Salvini and Vignoli (2009), final report of the "Indagine sulla georeferenziazione dei dati nella statistica ufficiale" [Investigation on data georeferencing in official statistics] promoted by the *Commissione per la Garanzia dell'Informazione Statistica, Presidenza del Consiglio dei Ministri.*

## 2.2 Spatial Data Analysis

Over the last twenty years, spatial data analysis has become a relevant instrument in most areas of observational sciences, from epidemiology to environmental to social sciences, since the focus on geographical locations and on possible spatial patterns and relationships can help our understanding of the studied phenomena.

Bailey and Gatrell (1995, p.21) define spatial data analysis as an analysis that ≪involves the accurate description of data relating to a process operating in space, the exploration of patterns and relationship in such data, and the search for explanation of such patterns and relationships≫. The object of such analysis is to increase our knowledge of the process, evaluate the evidence in accord with some hypotheses concerning it, or predict values in areas where observations have not been collected. The data that we elaborate constitute a sample of observations on the process from which we attempt to infer its overall behaviour.

Obviously, not all data that can be located in space need to be subject to this kind of analysis. Spatial data analysis is involved when data are spatially located *and* explicit consideration is given to the possible importance of their spatial distribution in the analysis or in the interpretation of results.

Bailey and Gatrell (1995) define four classes of data involving spatial data analysis and for each one they outline specific objectives of analysis. The first class is composed by a set of point events, or a *point pattern*, and we want to investigate whether the proximity of the events, that is their spatial configuration, represents a significant pattern. Sometimes these points have some attributes associated with them distinguishing one kind of event from another, but it is the spatial arrangement of the events themselves that is of interest.

The second class of data comprise again a set of point locations, but the pattern of these locations is not itself the subject of analysis. This time, the locations are simply the sampled points at which a continuous variable is measured and the aim of the analysis is to understand the process generating these values and to use the information to model the variable of interest and to make predictions elsewhere on the map. This kind of data is common in the environmental sciences and we refer to it as *spatially continuous data*, while the analysis techniques are usually known collectively as *geostatistics*.

The third class is *area data*, that is data that have been aggregated to a set of areal units, such as districts, municipalities, census enumeration districts, and so on. One or more variables are measured over this set of zones and the analysis object is to understand the spatial arrangement of these values, to detect patterns and to examine relationships among the set of variables.

The final class of data is *spatial interaction data* and is composed by data on flows that link a set of locations, either areas or points. The analysis target is to understand and to model the arrangement of flows, and to use this information to predict how the flows may change under different scenarios.

## 2.3 Statistical Spatial Data Analysis

The set of computational tools that lets handle spatial data analysis is known as *Geographical Information System* (GIS). We can find a exact definition of GIS in Bailey and Gatrell (1995, p.52): ≪A Geographical Information System is a computer-based set of tools for capturing (collecting), editing, storing, integrating, analyzing and displaying spatially referenced data≫. In the last years, we observed a big increment in the use of GISs in every area of applied statistics. Starting from the merely use of GIS as a graphical tool, now it has been discovered as a complete and powerful instrument for statistical spatial analysis.

If we have two spatially referenced datasets, referring to the same region but coming from different sources, we can easily join them together with the use of GIS. This operation produces a new dataset that could be more informative than both the single datasets together, since relationships among the different sets of variables can now be evaluated. For example, in epidemiology we can relate a dataset that records the incidence of a particular illness with a dataset of environmental variables in order to evaluate the possible presence of clusters of risk levels. Obviously, the more precise and detailed is the spatial information, the more accurate will be the linkage.

When we want to join two datasets, but the point locations of the two sets of variables do not coincide, we are in presence of a *spatial misalignment* problem. A way to overcome it is to use some methods of spatial interpolation (Madsen et al., 2008; Gryparis et al., 2009) or other GIS tools. Otherwise, if one or both the datasets are composed by area data, usually we need to transform the spatial supports in order to obtain a common spatial reference for the join dataset. GIS methods and tools are available to solve this problem of *change of support*. The problems presented here are connected to the more general matter of analysis of measurement error in the GIS framework; an introduction of this subject will be presented in the next chapter.

As we said, the object of many statistical spatial data analysis is to evaluate the spatial pattern of the studied phenomenon. However, the behaviour of the single statistical units can be another target of analysis, especially in demography and social statistics. Even when we are doing an *individual level* analysis, the availability of spatial information can be very important.

Firstly, the relative position of each unit to the others and its proximity to specific points in the space (like schools, hospitals, etc...) can be relevant to explain the spatial variability. This analysis of the spatial arrangement of points involves the definition of spatial relations like *distance*, *direction* and *proximity* between points and areas.

Secondly, the spatial information can be a proxy for some useful background variable that have not been or cannot be measured. Haining (2003) define two sources of background influence: *compositional effect* and *contextual effect*. The compositional effect refers to the difference between areas in the composition of the population of statistical units. Such variability of the areas can produce a spatial variability on the economical, social and demographical phenomena that are influenced by the population structure. The contextual effect, on the other hand, is related to the difference between areas in term of exposure to factors that might have a direct or indirect influence on the studied phenomenon. These factors can be biological (like exposition to urban pollution), economical, cultural, and so on.

At this point, it is clear that the spatial influence on the studied process depends on the geographical scale at which the analysis is performed (*scale* or *aggregation effect*) and on the "shape" of the areas of analysis (*zoning effect*). When our data are the results of measurements aggregated on a set of zones, an issue related to the previous effects is the *modifiable area unit problem* (MAUP) (Holt et al., 1996). With this name, we refer to the important fact that any results obtained from the analyses of these area aggregations may be conditional upon the set of zones itself. If we have data with highly detailed spatial information, we can try to perform the analyses on alternative configurations of zones to evaluate the magnitude of the MAUP.

The need of spatial information in the analysis follows from the *first law of geography* (Tobler, 1970) that says: ≪everything is related to everything else, but near things are more related that distant things≫. This statement highlights the fact that spatial observations are not mutually independent and tend to be more "similar" to their neighbours. The presence of this similarity between observation is usually measured by a *spatial correlation* function and classical statistical methodologies need to be modified in order to account for such relation. The application of spatial prediction methods, like kriging or other geostatistical tools, relies on the preliminary study and definition of a conform spatial correlation structure (Cressie, 1993).

As we said in the previous section, these kind of techniques needs to work with point referred data. If we observe data that are strictly areal, or if we have only area aggregated measurements, a possible way to continue to use geostatistical techniques is to represent the areas by a set of points, one for each polygon. Typically, the geographical centre or *centroid* of each areal unit is used. The problems that arise from this kind of approximation and the proposal of a different approach are the objectives of Chapter 3.

## 2.4 Spatial Information in Research Areas of Statistics

Spatial data analysis applies in every area of statistical study. In all fields we have the common target of evaluate the spatial pattern of a studied phenomenon, but methods and specifications can differ because of the nature of the phenomenon itself. In the following subsections, we present a general overview on the use of the spatial information in main statistical research areas (Petrucci et al., 2009).

### 2.4.1 Official Statistics

The knowledge of the exact spatial location of statistical units is an important instrument for the production of official statistics.

First, it can increase the performance of surveys and censuses conducted by the official statistics producers, like Istat[2] and the other SISTAN offices[3], as it eases the procedure of detection and interview of the units, allows more control on the collection operations, and enables spatio-temporal comparisons and corrections between different definitions of spatial areas.

Second, the spatial referenced measurements allows the production of thematic maps and atlas to portray the spatial pattern of studied events on various spatial scales.

Last, the increasing availability of spatially referred microdata from administrative sources supports the use of GIS, which increases the data quality and the production of statistics referred to geographical areas specifically connected to the studied phenomenon, such as local labour systems, local economical system or agricultural areas (Calzaroni, 2008; Romei and Petrucci, 2003).

[2] The Italian National Statistical Institute.

[3] SISTAN is the Italian National Statistical System, that is a network of about 10.000 statistical operators belonging to the statistical offices of: Ministries, national agencies, Regions and autonomous Provinces, Provinces, Municipalities, Chambers of Commerce, local governmental offices, private agencies and subjects with specific characteristics stated by the law (http://www.sistan.it/english/index.htm).

### 2.4.2 Spatial Epidemiology

Spatial epidemiology is concerned with describing and understanding spatial variation in disease risk in relation to demographic, genetic, environmental and socio-economic factors . Considering the aims and use of spatial analyses in epidemiology, Elliott et al. (2000) distinguish four types of study: disease mapping; geographical correlation studies; assessment of risk in relation to a point or line-source; clusters detection and disease clustering.

*Disease mapping* is carried out to summarize spatial and spatio-temporal variation in risk. This information can be used for simple descriptive purposes of the spatial distribution of the studied phenomenon, to provide information on health needs of a population to establish context for further studies or to obtain clues on a disease aetiology by comparing the estimated risk map with an exposure map.

*Geographical correlation studies* are focused on aetiological aspects of a disease, evaluated by examine geographical variations in exposure to environmental variables (measured in air, water or soil) and lifestyle factors (such as smoking and diet behaviour) in relation to health outcomes measured on a geographical scale.

The studies of the *assessment of risk in relation to a point or line-source* are appropriate to evaluate *local* increments in a disease risk in relation to a potential source of environmental hazard. The source could be either a point (like a radio transmitter or a chimney stack) or a linear source (like a road or a power-line). These studies required an highly localized approach as any increased exposure due to the potential source is likely to extend only over a small region.

*Disease clustering* is the tendency of disease cases to occur in a non-random spatial pattern relative to the pattern of the non-cases. *Clusters detection* studies are carried out to provide an early detection of raised incidence if a disease when there is no specific aetiological hypothesis. Mainly, the aim of such studies is to support the activity of monitoring and surveillance of a geographical region, however they can be the preliminary stage of a more detailed study about the disease diffusion.

For these spatial epidemiology studies (Lawson and Cressie, 2000; Waller and Gotway, 2004), the ideal data would consist of precise information on the population of a study region, including individual characteristics, personal exposures in time and space and health records. Usually, such information is not available for the whole population. In case-control studies, for example, we can have detailed information collected for the cases, while it could be harder to have it for the controls. Moreover, as pointed out by Gryparis et al. (2009, p.1), ≪in many environmental epidemiology studies, the locations and/or times of exposure measurements and health assessments do not match. In such settings, health effects analyses often use the predictions from an exposure model as a covariate in a regression model≫.

Typically, population data are based on area aggregated counts. In such situation, we can exploit area data to obtain surveillance atlas through a range of methodologies, from simple choropleth maps (Cromley, 1996; Boffi, 2004) to more complex statistical models that account for possible sparsity of the spatial information and for the spatial heterogeneity of the areas. This kind of models has a big development, mainly as hierarchical Bayesian models (Banerjee et al., 2004; Lawson, 2009), and is strongly connected with the small area estimation problem, that will be presented in Subsection 2.4.6.

As seen in Section 2.3, we ought to remember that these procedure can be subject to measurement errors due to spatial misalignment of the sources or to data quality. Moreover, if the data are measured on some geographical scale, the analysis could be influenced by the modifiable area unit problem.

### 2.4.3 Environmental Statistics

Spatially referenced data are extremely valuable to analyze and describe environmental phenomena and to understand the connections and interactions between environment and human activities (Patil and Rao, 1994).

Similarly to spatial epidemiology, we can define four categories of environmental studies: mapping of environmental indicators; estimation of the spatial pattern of some environmental factor; clusters detection; planning of spatial survey samplings and environmental monitoring networks.

The use of thematic maps of environmental indicators is connected with the monitoring process of the environmental quality, like air or water quality. This study is usually the first step for a mitigation intervention or for a deeper analysis of the causes.

The study of the spatial pattern of a target environmental variable, such as the presence of a specific substance in the ground or the level of a pollutant in the air, or the estimate of the number of people living around some urban infrastructure, like a highway or an airport, are all example of typical environmental spatial analyses. Usually, such analyses are conducted with geostatistical methodologies and their results are useful in many research areas, such as epidemiology, economics and natural science, but are also exploited for legislative regulation.

Clusters detection studies are carried out to identify possible cluster of high risk of natural events, like earthquakes, avalanches or landslides, in order to predict and avoid possible disasters and protect the population.

Obviously, this kind of analyses require a highly detailed information about locations and sources of the studied phenomenon. To collect such information we usually need the implementation of specific survey sampling strategies and monitoring schemes. A detailed description of these sampling procedures can be found in de Gruijter et al. (2006).

### 2.4.4 Spatial Demography and Social Statistics

In the literature of last years, it has been noticed a re-emerging interest of social sciences in issues concerning social processes embedded within a spatial context (Goodchild and Janelle, 2004), that has introduced spatial analysis methodologies among the more usual social and demographical tools. The recent growing number of applications in spatial demography addresses space in several ways, ranging from visualization of one or more variables in a map, to sophisticated spatial statistical models that seek to explain why a particular spatial pattern is observed. These applications try to explain current demographic issues, and represent important information for the design and evaluation of realistic public policies (de Castro, 2007).

Voss (2007, p.458) defines *spatial demography* as ≪the formal demographic study of areal aggregates, i.e., of demographic attributes aggregated to some level within a geographic hierarchy.≫, while a somewhat similar definition (Woods, 1984, p.43) states that spatial demography is ≪demography viewed from the spatial perspective. [...] Spatial, together with temporal, variations in mortality, fertility and migration are studied as preliminaries to the investigation of population structure in its entirety.≫

Both definitions presented are very generic, and would include a significant list of studies, covering a wide range of topics, and applying a variety of methods and spatial tools. Among these are: mapping demographic variables; analyzing the spatial and temporal patterns of variables of interest; including variables describing location as covariates in regression models; multilevel models, where the hierarchical structure of the data refers also to spatial areas chosen a priori; and applications of geostatistical methods. In summary, any demographic analysis performed with a spatial perspective would fall under the definition of spatial demography (de Castro, 2007).

For a detailed discussion about spatial demography research on the three demographic components of fertility, mortality and migration we refer to de Castro (2007); while Goodchild and Janelle (2004) collect spatial application in various fields of social sciences.

### 2.4.5 Regional Economics and Spatial Econometrics

Richardson (1970, p.1) defined *regional economics* as the field concerned with the role of ≪space, distance and regional differentiation in economics≫.

The analysis of the regional spatial pattern of socio-economic processes has become a relevant area of economics for several reasons: first, the spatial clustering of economic activities is a product of the regional differences and could reflect individual inequalities that are object of policies; second, the geographical pattern can have great influence on the results of economic policies; and third, exploring spatial clustering of economic activities is a relevant input to model economic theories at a regional scale.

Research in this area focuses on the specification and estimation of spatial effects in a theoretical economic model, and on the use of such estimates to obtain spatial interpolations and predictions of the study variables. The set of methodologies concerned with this target belongs to the field of *spatial econometrics* (Anselin, 1988; Arbia, 2006), that is defined by Anselin (1988, p.7) as ≪the collection of techniques that deal with the peculiarities caused by space in the statistical analysis of regional science models≫.

In Paelinck and Klaassen (1979), we can find five fundamental characteristics of spatial econometric methodologies: the role of spatial interdependence in spatial models, the asymmetry of spatial relations, the importance of explanatory factors located in other spaces, the differentiation between ex-post and ex-ante interaction, and the explicit modeling of space (Arbia, 2006).

For a detailed review of spatial econometrics applications we refer to Arbia and Baltagi (2008), and Arbia (2006).

### 2.4.6 Small Area Estimation

Over the last decade there has been growing demand from both public and private sectors for producing estimates of population characteristics at disaggregated geographical levels, often referred to as *small areas* or *small domains* Rao (2003). This increasing request of small area statistics is due, among other things, to their growing use in formulating policies and programs, in the allocation of government funds and in regional planning. Demand from the private sector has also increased because business decisions, particularly those related to small businesses, rely heavily on the local socio-economic, environmental and other conditions.

Censuses provide “total” information, but only on a limited number of characteristics and once every ten years. Statistical surveys produce high quantities of data and estimates, but cost constraints in the design of sample surveys lead to small sample sizes within small areas. As a result, direct estimation using only the survey data is inappropriate as it yields estimates with unacceptable levels of precision. In such cases *small area estimation* (SAE) can be performed via models that "borrow strength" by using all the available data and not only the area specific data. The most popular class of models for small area estimation is linear mixed models that include independent random area effects to account for between area variation beyond that explained by auxiliary variables (Fay and Herriot, 1979; Battese et al., 1988). Following mixed models methodology (Jiang and Lahiri, 2006), a BLUP estimator is used to obtain the small area parameter of interest (usually the mean or total of the study variable). If, as usual, the variance components are unknown, the correspondent EBLUP estimator is used instead (see Rao (2003, Chapters 6-7) for a detailed description).

Under the classic SAE model we make the assumption of *independence* of the area-specific random effects. If the small domain of study are geographical areas, this assumption means that we don't take into account any possible spatial structure of the data. Remembering again the *first law of geography* however, it is reasonable to suppose that close areas are more likely to have similar values of the target parameter than areas which are far from each other, and that ≪an adequate use of geographic information and geographical modeling can help in producing more accurate estimates for small area parameters≫(Petrucci et al., 2005, p.610). In addition, Pratesi and Salvati (2008, p.114) noted that geographical ≪small area boundaries are generally defined according to administrative criteria without considering the eventual spatial interaction of the variable of interest≫. From all these considerations, it is reasonable to assume that the random effects between the neighbouring areas (defined, for example, by a contiguity criterion) are correlated and that the correlation decays to zero as distance increases.

The first studies that connect spatial relations and SAE methods are Cressie (1991) and Pfeffermann (2002). In the following years, many papers have been published showing how the use of geographical information improves the estimation of the small area parameter, both increasing efficiency and diminishing bias. We refer, among others, to Saei and Chambers (2005), Petrucci et al. (2005), Petrucci and Salvati (2006), Singh et al. (2005) and Pratesi and Salvati (2008). In all these studies, the classical hypothesis of independence of the random effects is overcome by considering correlated random area effects between neighbouring areas modeled through a *Simultaneously Autoregressive* (SAR) process with *spatial autocorrelation coefficient* $\rho$ and *proximity matrix* $\mathbf{W}$ (Anselin, 1988). The corresponding estimators of the small area parameters are usually known as Spatial EBLUP (SEBLUP).

In addition, the use of SAE models with spatially correlated random area effects gives a possible solution to the problem of estimating the parameter of interest for the areas in which no sample observations are available. With the traditional SAE model, the only prevision available for non-sampled areas is given by the “fixed term” of the mixed model, since the estimation of the random effect is not possible. On the contrary, the hypothesis of correlated random effects allows the estimation of the area-specific effects for all areas, both sampled and non-sampled. The addition of these estimated random effects to the fixed component of the model gives the prediction of the small area parameter in every area.

Spatial SAE models are applied in many area of statistical research: environmental statistics, economics, demography, epidemiology, and so on. Every study shows that the use spatially referred data produces estimates more reliable than that obtained by traditional methods.

Until now, we have considered the spatial structure of the data at the area level: the only information used to built the proximity matrix of the SAR process is about the small area locations. However, if the spatial location is available for every unit, we can try to use it directly as a covariate of the SAE model. As we have presented in Section 1.7, the application of bivariate smoothing methods, like kriging, produces a surface interpolation of the variable of interest. In particular, the geoadditive model defined in Section 1.8 analyzes the spatial distribution of the study variable while accounting for possible covariate effects through a linear mixed model representation. Exploiting the common linear mixed model framework of both small area estimation models and geoadditive models, we can define the *geoadditive SAE model*. This model will have two random effect components: the area-specific effects and the spatial effects.

The geoadditive SAE model belongs to a more general class of models introduced by Opsomer et al. (2008), called *non-parametric SAE model*, where the non-parametric component is a penalized spline model that accounts for a generic non-linear covariate.

The model will be discussed in the last chapter of this work, where an application on the estimation of the mean of household log per-capita consumption expenditure for the Albanian Republic at different geographical levels is presented.

# Chapter 3

# Geoadditive Models and Measurement Error

## 3.1 Introduction

In this chapter we present the concept of measurement errors in spatial data analysis. In particular, in Section 3.2 we define the problem of uncertainty and errors in GIS from the point of view of the geographical information science, while in Section 3.3 we illustrate the statistical approach to measurement error analysis. In Section 3.4 we deal with the matter of applying a geaodditive model to produce estimates for some geographical domains in the absence of point referenced auxiliary data. Instead of using the classic approach, that locate all the units belonging to the same area by the coordinates of the centroid of each area, we treat this lack of geographical information following a measurement error approach and imposing a distribution for the locations inside each area. The performance of our measurement error approach is evaluated through various Markov Chain Monte Carlo experiments implemented under different scenarios. Results are presented in Section 3.5.

## 3.2 Measurement Error in Spatial Analysis

In Chapter 2 we presented the concept of spatial data analysis in statistics and we introduced some peculiar issues that occur when we deal with spatially referenced data (like change of support or spatial misalignment) and that are connected with uncertainty and measurement error problems in spatial data.

Over the last years, a research area of geographical information science has been developed to investigate how the uncertainty in spatial data arises and distributes through GIS operations, and to assess the plausible effects on subsequent decision-making (Heuvelink, 1998; Zhang and Goodchild, 2002). As pointed out by Leung et al. (2004), ≪with the ever increasing volume of georeferenced data being generated, transferred, and utilized, the amount of uncertainty embedded in spatial databases has become a major issue of crucial theoretical importance and practical consideration≫.

Uncertainty in spatial databases, both in attribute values and in positions, generally involves accuracy, statistical precision, and bias in initial values or in estimated coefficients. Moreover, spatial uncertainty includes the estimation of errors in the final output that result from the propagation of external and internal uncertainty. It is thus important to be able to track the occurrence and propagation of uncertainties (Goodchild, 1991). Research on accuracy is strictly associated with the study of errors in GIS, and the literature on this subject has been extensive and diverse (Goodchild and Gopal, 1989; Heuvelink, 1998; Leung and Yan, 1998; Mowrer and Congalton, 2000; Stanislawski et al., 1996; Wolf and Ghilani, 1997; Zhang and Goodchild, 2002).

The error taxonomy of Veregin (1989) recognizes that different classes of spatial data exhibit different types of errors, and that errors may be introduced and propagated in various stages of data manipulation and spatial processing. Errors in spatial databases are generally divided in *inherent errors* and *operational errors*: inherent errors are the errors present in source documents, including the errors in the map used as input to a GIS; operational errors occur throughout data manipulation and spatial modeling and are introduced during the process of data entry or through the data capture and manipulation functions of a GIS (Leung et al., 2004). Moreover, from the modeling point of view, the errors can also be classified as either systematic or random. While the systematic component can usually be removed by model modification, it is impossible to avoid random errors in measurements entirely (Wolf and Ghilani, 1997). Dealing with such measurement error, is one of the most important problems in the use of georeferenced data.

In order to support the determination of error structures of location coordinates in GIS, the concept of a *measurement-based* GIS (MBGIS) has been proposed by Goodchild (1999). A MBGIS is ≪a system that provides access to measurements used to determine the locations of objects, to the geographical procedures (transformation functions) that link measurements to quantities to be measured, and to the rules used to determine interpolated positions≫. The basic idea is to retain details of measurements so that error analysis can be made possible. Moreover, Leung et al. (2004) propose a general framework within which the statistical approach to measurement error analysis and error propagation can be formulated.

## 3.3 Measurement Error in Statistics

The measurement error analysis approach presented in the previous section is a geographical science approach. It involves some statistical tools and concepts, but it is mainly a "technical" approach.

In statistics, the measurement error problem is concerned with the inference on regression models where some of the independent variables are contaminated with errors or otherwise not measured accurately on all subjects. In literature, it is well established that disregarding the measurement error in a predictor distorts its estimated relationship with the response variable and produces biased estimates of the regression coefficients, both in linear (Buonaccorsi, 1995; Fuller, 1987, Chapter 1) and in nonlinear models (Carroll et al., 2006, Chapter 3). Hence, the most part of measurement error analysis is about correcting for such effects.

Measurement error models are commonly composed of two components. First, we have an underlying model for the response variable **y** in terms of some predictors, distinguished between predictors measured without error, indicated with **z**, and predictors that cannot be observed exactly, indicated with **x**. Second, we can observe a variable **w**, which is related to the unobservable **x**. The parameters in the model relating **y** and (**z**,**x**) cannot be estimated directly, since **x** is not observed. The goal of measurement error modeling is to obtain nearly unbiased estimates of these parameters indirectly by fitting a model for **y** in terms of (**z**,**w**). In assessing measurement error, careful attention must be given to the type and nature of the error, and the sources of data that allow modeling of this error (Carroll et al., 2006).

A fundamental prerequisite for analyzing a measurement error problem is the specification of a model for the measurement error process. There are two general types: *classical error* model, where the conditional distribution of **w** given (**x**,**z**) is modeled; and *Berkson error* model, where the conditional distribution of **x** given (**w**,**z**) is modeled (Berkson, 1950). In their simplest formulation, the two models correspond to:

- Classical error model:

$$\mathbf{w}_i = \mathbf{x}_i + \mathbf{u}_i, \qquad \text{with} \quad \mathrm{E}(\mathbf{u}_i|\mathbf{x}_i) = 0$$

- Berkson error model:

$$\mathbf{x}_i = \mathbf{w}_i + \mathbf{u}_i, \qquad \text{with} \quad \mathrm{E}(\mathbf{u}_i|\mathbf{w}_i) = 0$$

where **u** can be distributed in various way.

For a detailed description of both the specifications we refer to Fuller (1987) and Carroll et al. (2006), however the basic difference between the two types of error models is that we choose the classical model if the error-prone variable is necessarily measured uniquely for every individual, while we choose the Berkson model if all individuals in a small group or strata are given the same value of the error-prone covariate.

The literature on statistical measurement error analysis is enormous, some example are Carroll et al. (1993); Bollinger (1998); Richardson et al. (2002); Chesher and Schluter (2002); Wang (2004); Carroll et al. (2004); Ganguli et al. (2005); Ybarra and Lohr (2008); Torabi et al. (2009). In particular, in the last years various applications on models with spatial measurement error has been published (Zhuly et al., 2003; Gryparis et al., 2007; Madsen et al., 2008; Goovaerts, 2009; Gryparis et al., 2009).

## 3.4 Lack of Geographical Information as Measurement Error

In Section 2.3 we observed that the implementation of geostatistical methods, like the geoadditive model, needs the statistical units to be referenced at point locations. If the aim of our study is to analyze the spatial pattern or to produce a spatial interpolation of a studied phenomenon, then we require such spatial information only for the sampled statistical units. If, however, we use a geoadditive model to produce estimates of a parameter of interest for some geographical domains, the spatial information is required for all the population units.

This information is not always easily available, especially when socio-economic data are involved. Typically, we know the coordinates for sampled units (which could be specifically collected for the analysis), but we don't know the exact location of all the non-sampled population units, just the areas to which they belong (like census districts, blocks, municipalities, etc).

In such situation, the classic approach that allows the use of geostatistical techniques is to locate all the units belonging to the same area by the coordinates (latitude and longitude) of the geographical centre or *centroid* of each area. This is obviously an approximation, induced by nothing but a geometrical property, and its effect on the estimates can be strong, depending on the level of nonlinearity in the spatial pattern and on the area dimension.

Instead of using the centroids, we decided to treat this lack of geographical information following a measurement error approach. In particular, we impose a distribution for the locations inside each area.

Let $\mathbf{x}_{ij}$ be the vector of the exact spatial coordinates for the unit $i$ belonging to the area $j$ and let $\mathbf{w}_j$ be the coordinates of the centroid of the area $j$, thus our hypothesis can be formulated as a Berkson-type error model:

$$\mathbf{x}_{ij} = \mathbf{w}_i + \mathbf{u}_{ij}, \tag{3.1}$$

where $\mathrm{E}(\mathbf{u}_{ij}|\mathbf{w}_i) = 0$ and $\mathbf{u}$ can assume distributions with different parameters in each area.

Our model is not a "complete" measurement error model as we assume that the measurement error doesn't influence the estimation of the geoadditive models parameters (as the spatial information is available for the sample), while it occurs when we predict the parameter of interest for the areas with the whole population covariates.

In order to evaluate the performance of our approach with respect to the centroids classic approach, various Markov Chain Monte Carlo (MCMC) experiments are implemented under various scenarios.

## 3.5 MCMC Experiments

Considering the hierarchical Bayesian models formulation of additive models (Ruppert et al., 2003, Chapter 16), we can exploit MCMC software for analyzing the performance of our measurement error approach. For the implementation of our experiments, we follow the settings and examples presented in Crainiceanu et al. (2005) and Marley and Wand (2010).

All the analyses are implemented using the `WinBUGS` Bayesian inference package (Lunn et al., 2000), a Windows interface to the `BUGS` inference engine (Spiegelhalter et al., 2003). We access `WinBUGS` using the package `BRugs` (Ligges et al., 2009) in the `R` computing environment (R Development Core Team, 2009). As pointed out in (Marley and Wand, 2010, p.2), ≪employment of `BRugs` has the advantage that an entire analysis can be managed using a single `R` script and accompanying `BUGS` script. Because `R` is used at the front-end and back-end of the analysis, one can take advantage of `R`'s functionality for data input and pre-processing, as well as summary and graphical display≫.

### 3.5.1 Model Specification

Consider the generic penalized spline regression with basis functions $b_k$

$$y_i = \beta_0 + \beta_1 x_i + \sum_{k=1}^{K} u_k b_k(x_i) + \varepsilon_i, \qquad \begin{aligned} \varepsilon_i &\sim N(0, \sigma_\varepsilon^2), \\ u_k &\sim N(0, \sigma_u^2). \end{aligned}$$

The *hierarchical Bayesian* formulation is

$$\begin{aligned} y_i|\beta_0, \beta_1, \sigma_u^2, \sigma_\varepsilon^2 &\overset{\text{ind}}{\sim} N\left(\beta_0 + \beta_1 x_i + \sum_{k=1}^{K} u_k b_k(x_i), \sigma_\varepsilon^2\right), \\ u_k|\sigma_u^2 &\overset{\text{ind}}{\sim} N(0, \sigma_u^2), \end{aligned} \tag{3.2}$$

and we need to define the *prior* distributions for the parameters $\beta_0, \beta_1, \sigma_u^2, \sigma_\varepsilon^2$. As suggested in Crainiceanu et al. (2005), we use the following non-informative priors:

$$\begin{cases} \beta_0, \beta_1 & \overset{\text{ind}}{\sim} N(0, 10^8) \\ \sigma_u^{-2}, \sigma_\varepsilon^{-2} & \overset{\text{ind}}{\sim} \text{Gamma}(10^{-8}, 10^{-8}). \end{cases} \tag{3.3}$$

The parametrization of the Gamma(a,b) distribution implies that the parameter has mean $a/b = 1$ and variance $a/b^2 = 10^8$. Moreover, it should be noticed that we parametrize the inverse of the variance, that is the *precision* parameter $\tau$, accordingly with `WinBUGS` specification.

In addition to (3.2) and (3.3), we need to specify our data structure, the measurement error hypothesis and the mean estimators. We present here the general case, that we will specify in detail for every experiment.

Suppose to have a population of $N$ units divided in $Q$ regions, and to be interested in estimate the regional mean of a study variable $y$. We take a sample of $n$ units from which we collect the response variable $y$, the location $s$ and, possibly, some other covariates (that are known without error for all the population units). To obtain the regional mean, we want to apply a model-based mean estimator based on (3.2):

$$\hat{\bar{y}}_q = \frac{1}{N_q}\left[\sum_{i \in S_q} y_i + \sum_{i \in R_q}\left(\hat{\beta}_0 + \hat{\beta}_1 s_{iq} + \sum_{k=1}^{K} \hat{u}_k b_k(s_{iq})\right)\right], \tag{3.4}$$

where $N_q$ is the total number of units in region $q$, $q = 1, ..., Q$, and $S_q$ and $R_q$ indicate respectively the indexes of the sampled units and of the non-sampled units belonging to region $q$.

We obtain the estimated parameters from the sampled units, but we cannot use directly (3.4) as we don't know $s$ for the not-sample units. Thus, the two working approaches are:

- **Naive approach**. Substitute $s_{iq}$ with the region centroid $c_q$, that is a constant for all the units in region $q$;

- **ME approach**. Define a distribution for $s_{iq}$ inside region $q$ and "sample" from it.

As we have noticed in the previous section, the ME approach is equivalent to define a similar Berkson-type measurement error model for $s_{iq}$

$$s_{iq} = c_q + \nu_{iq}, \qquad \nu_{iq} \overset{\text{ind}}{\sim} f_\nu(\boldsymbol{\theta}_q), \tag{3.5}$$

where $\boldsymbol{\theta}_q$ are the parameters of $\nu$ distribution and depend on the region $q$.

To better analyze the performance of the two approaches, we decided to work both with $s$ *univariate* (so that the regions are actually intervals) and $s$ *bivariate* and to insert some known covariates; moreover, $f_\nu$ is considered *uniform* or *beta* and, finally, we used datasets *completely* simulated and *partially* simulated. The list of scenarios is presented in table 3.1.

**Table 3.1:** Scenarios of the MCMC experiments.

| Scenario Name | Type of $s$ | $f_\nu$ | Type of data |
|---|---|---|---|
| Univariate Uniform | univariate | uniform | completely simulated |
| Univariate Beta | univariate | beta | completely simulated |
| California | univariate | uniform | partially simulated |
| Bivariate Uniform | bivariate | uniform | completely simulated |
| Albania | bivariate | uniform | real |

### 3.5.2 Univariate Uniform Model

The model for the first experiment is

$$\begin{aligned} y_{iq} &= \beta_t t_{iq} + f(s_{iq}) + \varepsilon_i = \\ &= \beta_0 + \beta_t t_{iq} + \beta_s s_{iq} + \sum_{k=1}^{K} u_k (s_{iq} - \kappa_k)_+ + \varepsilon_i, \end{aligned} \tag{3.6}$$

where $\varepsilon_i \sim N(0, \sigma_\varepsilon^2)$, $u_k \sim N(0, \sigma_u^2)$, $t$ is a dummy variable known for the whole population and $s$ is a univariate variable that we *hypothesize* have a uniform distribution in every region. The data are divided in $Q = 4$ intervals $[a_q; b_q]$. To model $f(s)$, we consider a penalized truncated linear spline function with $K = 30$ knots selected at the quantiles of $s$ (as seen in Section 1.4).

The appropriate hierarchical Bayesian model for this situation is

$$\begin{gathered} y_{iq} | \beta_0, \beta_t, \beta_s, \mathbf{u}, \sigma_\varepsilon^2 \overset{\text{ind}}{\sim} N\left(\beta_0 + \beta_t t_{iq} + \beta_s s_{iq} + \sum_{k=1}^{K} u_k (s_{iq} - \kappa_k)_+, \sigma_\varepsilon^2\right), \\ \mathbf{u} | \sigma_u^2 \sim N(\mathbf{0}, \sigma_u^2 \mathbf{I}_K), \qquad \beta_0, \beta_t, \beta_s \overset{\text{ind}}{\sim} N(0, 10^8), \\ \tau_u, \tau_\varepsilon \overset{\text{ind}}{\sim} \text{Gamma}(10^{-8}, 10^{-8}). \end{gathered} \tag{3.7}$$

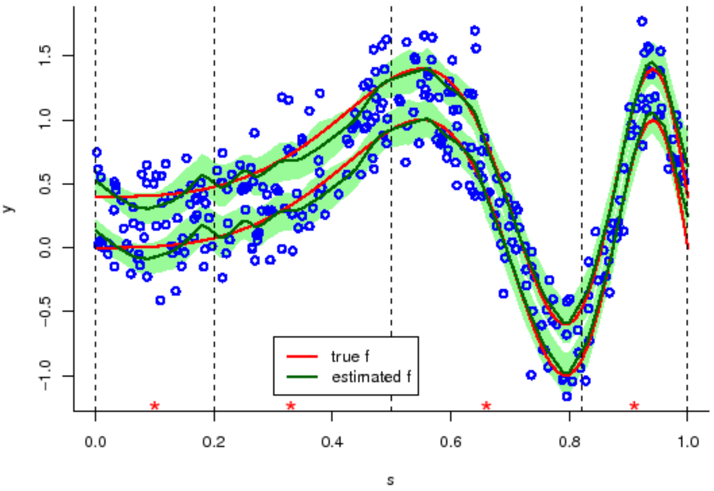

**Figure 3.1:** MCMC-base fitting of the univariate uniform model. The upper line corresponds to $t = 1$, the lower to $t = 0$ and the pale green shaded regions are pointwise 95% credible sets. The blue points are the sampled units. The vertical dashed lines delimit the regions and the red stars indicate the centroids.

The measurement error hypothesis is

$$\begin{aligned} c_q &= \frac{(a_q + b_q)}{2} \\ s_{iq} &= c_q + \nu_{iq}, \qquad \nu_{iq} \overset{\text{ind}}{\sim} \text{Unif}\left(\frac{(a_q - b_q)}{2}, \frac{(b_q - a_q)}{2}\right), \end{aligned} \tag{3.8}$$

where $a_q, b_q$ are the known boundaries of the interval $q$. It is immediate to derive that the parametrization (3.8) corresponds to the hypothesis

$$s_{iq} \overset{\text{ind}}{\sim} \text{Unif}(a_q; b_q).$$

We fitted model (3.7) to a set of simulated data with

$$\begin{gathered} N = 2000, \quad n = 300, \quad \beta_t = 0.4, \quad \sigma_\varepsilon = 0.2, \\ f(s) = sin(3\pi s^3), \quad t \sim \text{Ber}(0.5), \quad s \sim \text{Unif}(0; 1), \\ a = [0, 0.2, 0.5, 0.82], \quad b = [0.2, 0.5, 0.82, 1] \end{gathered} \tag{3.9}$$

We implement the MCMC analysis[1] with a *burn-in* period of 15000 iterations and then we retain 5000 iterations, that are thinned by a factor of 5,

[1] The WinBUGS model code is presented in appendix (Section A.1).

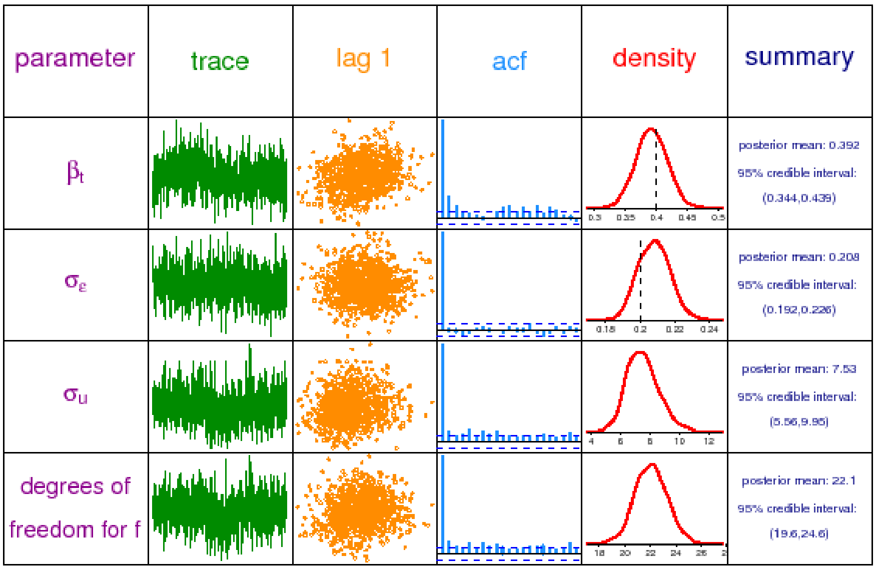

**Figure 3.2:** Graphical summary of MCMC-based inference for the parameters of the univariate uniform model. The columns are: parameter, trace plot of MCMC sample, plot of sample against 1-lagged sample, sample autocorrelation function, kernel estimates posterior density and basic numerical summaries. The vertical dashed lines in the density plots, where present, correspond to the true values of the parameters according to the simulation set-up.

resulting in a sample of size 1000 collected for inference. Figure 3.1 shows the fitted function for model (3.6) as well as the pointwise 95% credible intervals. In Figure 3.2 we summarize the MCMC output for the model parameters: the true values (3.9) from which the data were simulated, shown as vertical dashed lines in the posterior density plots, are inside the 95% credible sets; the credible interval for $\sigma_u$ is away from zero, which confirm the non-linearity of the effect of $s$; and all chains are seen to be well-behaved.

Once we have obtained a good estimate of the model (3.6), we apply the mean estimator (3.4) under the two approaches: the naive one and the ME one with hypothesis (3.8). The posterior density distributions of the region mean estimator[2] are presented in Figure 3.3: for each region, the red line corresponds to the ME approach, the purple line to the naive approach and the vertical green line is the true mean value (that is known since we are using a simulated dataset). As we can see, the ME estimator has a better performance, since the naive estimator underestimates the mean in region 1 and 2 and overestimates the mean in region 3 and 4.

[2] The summary of the MCMC output for the region mean estimators is presented in appendix (Figure A.1).

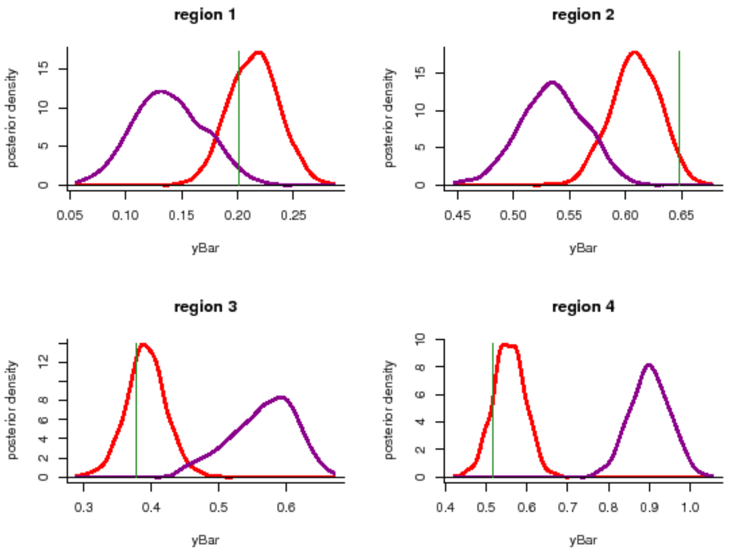

**Figure 3.3:** Posterior density of the region mean estimator for the univariate uniform model. The red lines correspond to the ME approach, the purple lines to the naive approach and the vertical green lines are the true mean values.

In addition, we implemented model (3.6) also without the dummy covariate $t$. Since the results are quite similar to the case in which the variable $t$ is included, we don't present the results here. However, the MCMC results are included in appendix (Section A.1).

### 3.5.3 Univariate Beta Model

In the second experiment we consider again the model defined in (3.6)-(3.7), but we change the hypothesis on the distribution of $s$: this time we suppose that $s$ has a beta distribution with different parameters in every region.

Thus, the measurement error hypothesis becomes

$$\begin{aligned} s_{iq} &= c_q + \nu_{iq}, \\ \nu_{iq} &\overset{\text{ind}}{\sim} \text{Beta}_{\text{gen}}\left(c_q; d_q\right) \quad \text{on } \left[\frac{a_q - b_q}{2}, \frac{b_q - a_q}{2}\right]. \end{aligned} \tag{3.10}$$

With $\text{Beta}_{\text{gen}}$ we indicate a beta distribution defined on a generic interval (instead of standard [0,1]). From (3.10) we derive the corresponding hypothesis

$$s_{iq} \overset{\text{ind}}{\sim} \text{Beta}_{\text{gen}}\left(c_q; d_q\right) \quad \text{on } \left[a_q, b_q\right].$$

The generic beta distribution can be derived from the standard beta distribution with the same parameters $c_q, d_q$ using a simple linear transformation:

$$\text{if} \quad ss_{iq} \sim \text{Beta}\,(c_q; d_q) \quad \text{then} \quad s_{iq} = a_q + (b_q - a_q)ss_{iq}.$$

The parameters $c_q, d_q$ are estimated directly in the MCMC process[3], by adding the following priors to the hierarchical Bayesian model (3.7)

$$ss_{iq}|c_q, d_q \overset{\text{ind}}{\sim} \text{Beta}(c_q; d_q), \qquad c_q, d_q \overset{\text{ind}}{\sim} \text{Unif}(0; 100).$$

Again, we fitted the model (3.7) to a set of simulated data with settings

$$\begin{gathered} N = 2000, \quad n = 400, \quad \beta_t = 0.4, \quad \sigma_\varepsilon = 0.2, \\ f(s) = sin(3\pi s^3), \quad t \sim \text{Ber}(0.5), \quad s_q \sim \text{Beta}_{\text{gen}}(c_q; d_q), \\ a = [0, 0.2, 0.5, 0.82], \quad b = [0.2, 0.5, 0.82, 1], \\ c = [2, 4, 1.5, 6], \quad d = [3, 2, 2, 5.2]. \end{gathered} \tag{3.11}$$

The simulated distribution of $s$ is presented in Figure 3.4.

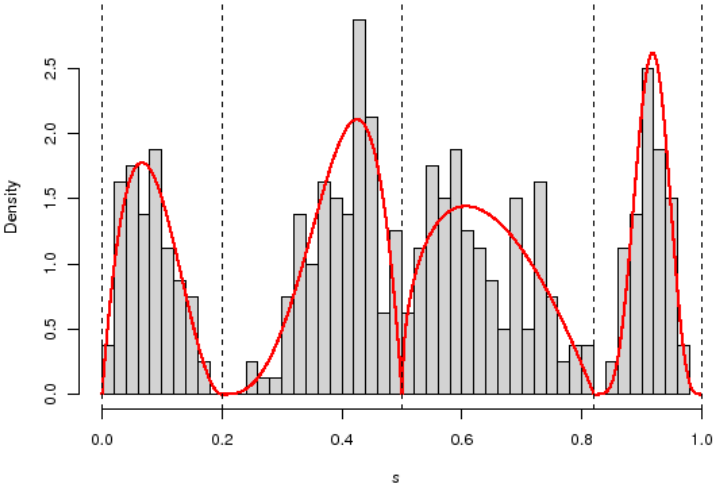

**Figure 3.4:** Distribution of the sampled covariate $s$ for the univariate beta model. The vertical dashed lines delimit the regions and the red lines correspond to the true beta distributions according to the simulation set-up.

[3] The summary of the MCMC output for the parameters of the beta distribution is presented in appendix (Figure A.8).

We implement the MCMC analysis[4] with a *burn-in* period of 15000 iterations and then we retain 5000 iterations, that are thinned by a factor of 5, resulting in a sample of size 1000 collected for inference.

Figure 3.5 shows the fitted function and the pointwise 95% credible intervals. It should be noticed the different distribution of the sampled units with respect to the previous experiment: the data are now "grouped" in every interval. In Figure 3.6 we summarize the MCMC output for the model parameters and again we notice that the true values are inside the 95% credible sets, the credible interval for $\sigma_u$ is away from zero and the chains are well-behaved.

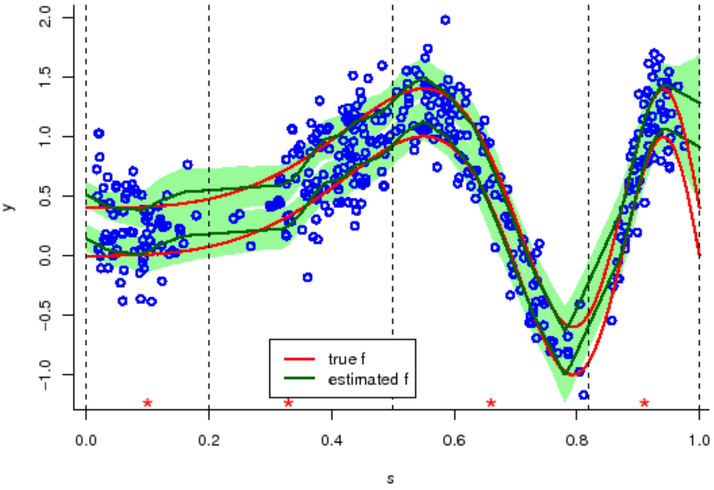

**Figure 3.5:** MCMC-base fitting of the univariate beta model. The upper lines correspond to $t = 1$, the lower ones to $t = 0$ and the pale green shaded regions correspond to pointwise 95% credible sets. The blue points are the sampled units. The vertical dashed lines delimit the regions and the red stars indicate the centroids of each region.

The posterior density distributions of the two region mean estimators (ME and naive)[5] are presented in Figure 3.7: the ME estimator has a good performance, especially in the regions where the function is more non-linear or where the distribution of $s$ is more asymmetric (as in region 2).

[4] The WinBUGS model code is presented in appendix (Section A.2).

[5] The summary of the MCMC output for the region mean estimators is presented in appendix (Figure A.9).

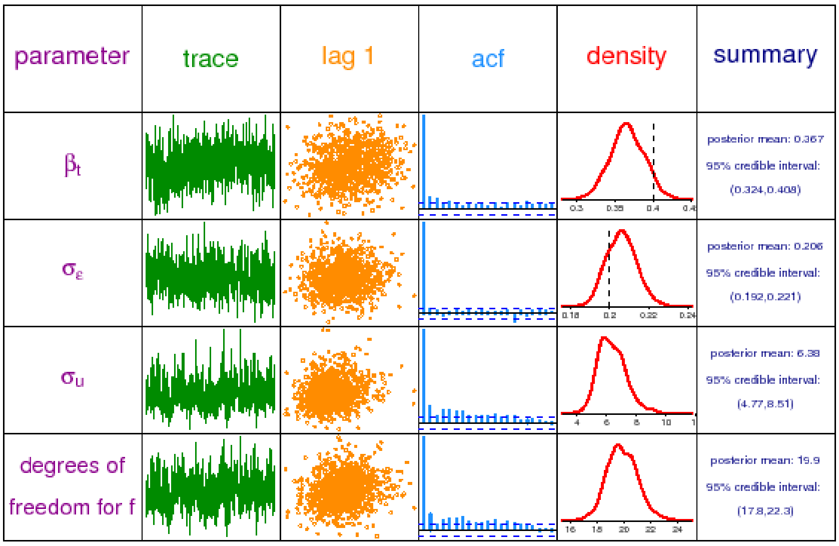

**Figure 3.6:** Graphical summary of MCMC-based inference for the parameters of the univariate beta model. The columns are: parameter, trace plot of MCMC sample, plot of sample against 1-lagged sample, sample autocorrelation function, kernel estimates posterior density and basic numerical summaries. The vertical dashed lines in the density plots, where present, correspond to the true values of the parameters according to the simulation set-up.

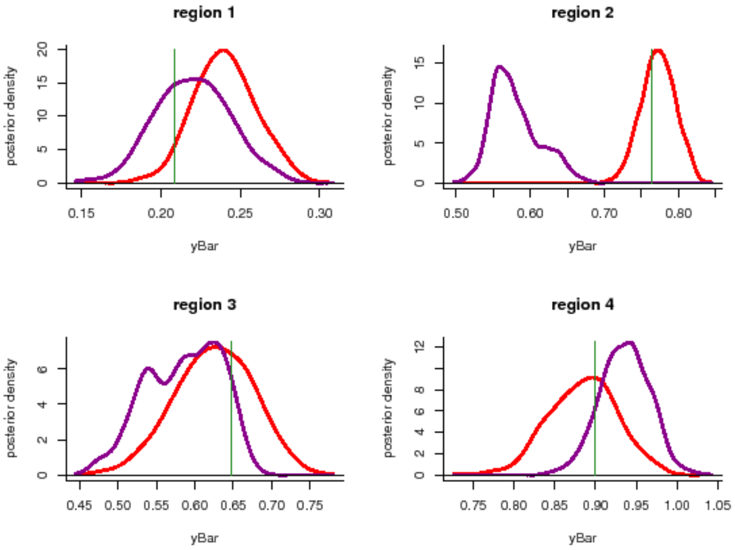

**Figure 3.7:** Posterior density of the region mean estimator for the univariate beta model. The red lines correspond to the ME approach, the purple lines to the naive approach and the vertical green lines are the true mean values.

### 3.5.4 California Model

In the third experiment we use a partially simulated dataset. We consider the *California air pollution* dataset presented in Breiman and Friedman (1985) that consists of 345 sets of daily measurements of ozone concentration and meteorology in the Los Angeles basin in 1976. The response is the ozone concentration (in ppm) at Sandburg Air Force Base (CA) and we selected three covariates: pressure gradient at Daggett (CA), in mmHg; inversion base height, in feet; and inversion base temperature, in Fahrenheit degrees.

As suggested in (Marley and Wand, 2010, p.4), we standardize all the variables before commencing the Bayesian analysis because ≪this makes the priors scale invariant and can also lead to better behaviour of the MCMC≫. Thus, we define the standardized variables:

- s = standardized *pressure gradient at Daggett*. We treat it as the unknown variable;
- x = standardized *inversion base height*;
- t = standardized *inversion base temperature.*

Instead of using the ozone variable, we decide to simulate a new response

$$y = sin(4\pi s - 0.2) + cos(6\pi x + 0.35) + sin(3\pi t^3 + 0.1) + \varepsilon,$$

with $\sigma_\varepsilon = 0.1$. Figure 3.8 shows the pairwise scatterplots for the "new" variables. The final dataset is composed by the variables $s, x, t, y$ with 320 observations (observations that present extremely high or low values of the variable $s$ were deleted).

We fit the additive model

$$\begin{aligned} y_{iq} &= f_x(x_{iq}) + f_s(s_{iq}) + f_t(t_{iq}) + \varepsilon_i = m(x_{iq}, s_{iq}, t_{iq}) + \varepsilon_i = \\ &= \beta_0 + \beta_x x_{iq} + \beta_s s_{iq} + \beta_t t_{iq} + \sum_{k=1}^{K_x} u_k^x (x_{iq} - \kappa_k^x)_+ + \\ &+ \sum_{k=1}^{K_s} u_k^s (s_{iq} - \kappa_k^s)_+ + \sum_{k=1}^{K_t} u_k^t (t_{iq} - \kappa_k^t)_+ + \varepsilon_i, \end{aligned} \tag{3.12}$$

where $\varepsilon_i \sim N(0, \sigma_\varepsilon^2)$, $u_x \sim N(0, \sigma_x^2)$, $u_s \sim N(0, \sigma_s^2)$, $u_t \sim N(0, \sigma_t^2)$. The data are divided in $Q = 4$ intervals $[a_q; b_q]$. To model $f_x$, $f_s$ and $f_t$, we consider three penalized truncated linear spline functions with $K_x = K_s = K_t = 15$ knots selected at the quantiles of $x, s$ and $t$ respectively.

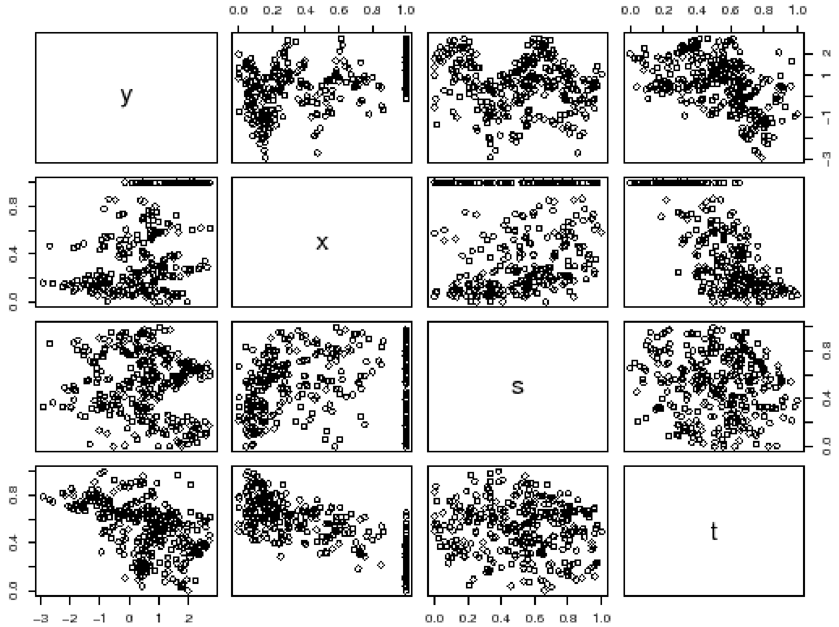

**Figure 3.8:** Pairwise scatterplots for the variables of the California model.

The correspondent hierarchical Bayesian model is

$$
\begin{gathered}
y_{iq}|\beta_0, \beta_x, \beta_s, \beta_t, \mathbf{u}_x, \mathbf{u}_s, \mathbf{u}_t, \sigma_\varepsilon^2 \overset{\text{ind}}{\sim} N\left(m(x_{iq}, s_{iq}, t_{iq}), \sigma_\varepsilon^2\right), \\
\mathbf{u}_x|\sigma_x^2 \sim N(\mathbf{0}, \sigma_x^2\mathbf{I}_{K_x}), \quad \mathbf{u}_s|\sigma_s^2 \sim N(\mathbf{0}, \sigma_s^2\mathbf{I}_{K_s}), \\
\mathbf{u}_t|\sigma_t^2 \sim N(\mathbf{0}, \sigma_t^2\mathbf{I}_{K_t}), \quad \beta_0, \beta_x, \beta_s, \beta_t \overset{\text{ind}}{\sim} N(0, 10^8), \\
\tau_x, \tau_s, \tau_t, \tau_\varepsilon \overset{\text{ind}}{\sim} \text{Gamma}(10^{-8}, 10^{-8}).
\end{gathered} \tag{3.13}
$$

Observing the empirical distribution of $s$, we decide to assume the measurement error uniform hypothesis:

$$s_{iq} \overset{\text{ind}}{\sim} \text{Unif}(a_q; b_q).$$

Finally, the experiment settings are

$$N = 320, \quad n = 96, \quad a = [0, 0.2, 0.5, 0.82], \quad b = [0.2, 0.5, 0.82, 1] \tag{3.14}$$

and the MCMC analysis[6] is implemented with a *burn-in* period of 15000 iterations and a retain of 10000 iterations with a thinning factor of 5, resulting in a sample of size 2000 collected for inference.

[6] The WinBUGS model code is presented in appendix (Section A.3).

Figure 3.9 shows the fitted functions for model (3.12). The fitting is quite good, but for the $x$ variable: this is due to the great quantity of units with value $x = 1$, and the complete absence of units with $0.9 < x < 1$. In Figure 3.10 we summarize the MCMC output for the model parameters. The credible intervals for $\sigma_x$, $\sigma_s$, and $\sigma_t$ are quite wide, this high variability should be due to the moderate number of sampled observation.

Notwithstanding the high variability of the parameters, the posterior density distributions of the two region mean estimators (ME and naive)[7] presented in Figure 3.11 show that the ME estimator has a really good performance, while the naive estimator is really poor. This result is influenced by the high non-linearity between $s$ and $y$. Moreover we want to highlight that the uniform distribution of the measurement error hypothesis produce good results, even if the data are not completely uniform.

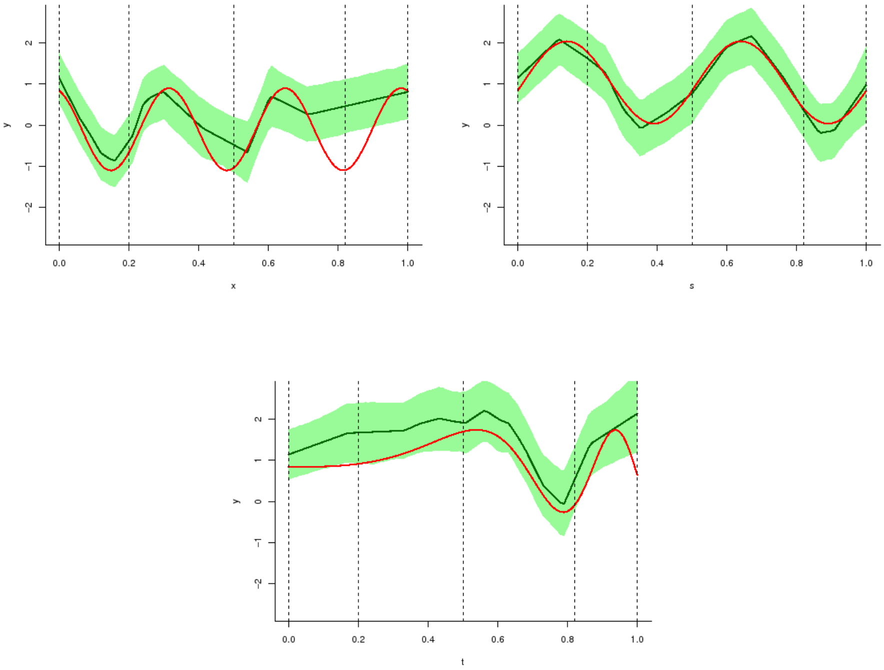

**Figure 3.9:** MCMC-base fitting of the California model. The green shaded regions correspond to pointwise 95% credible sets. The vertical dashed lines delimit the regions and the red stars indicate the centroids of each region.

[7] The summary of the MCMC output for the region mean estimators is presented in appendix (Figure A.11).

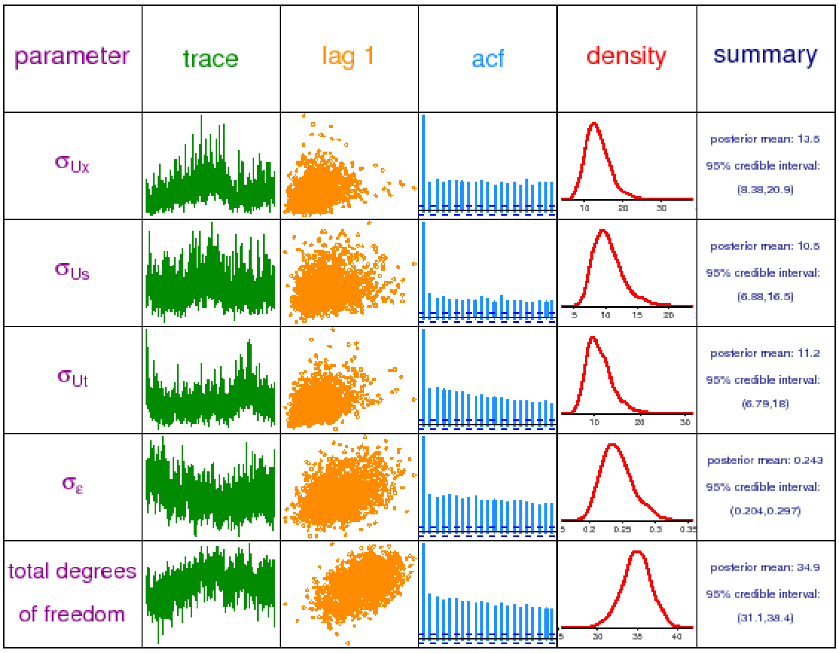

**Figure 3.10:** Graphical summary of MCMC-based inference for the parameters of the California model.

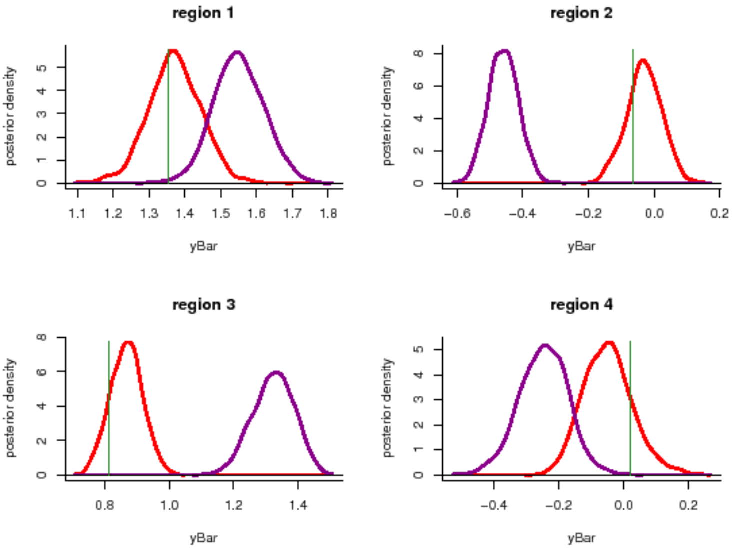

**Figure 3.11:** Posterior density of the region mean estimator for the California model. The red lines correspond to the ME approach, the purple lines to the naive approach and the vertical green lines are the true mean values.

### 3.5.5 Bivariate Uniform Model

The forth experiment implements a bivariate smoothing model, where the coordinates are unknown for the non-sampled units. This scenario is more related to the measurement error issue that we introduced in Section 3.4.

The generic model for this experiment is

$$y_{iq} = \beta_t t_{iq} + f(\mathbf{s}_{iq}) + \varepsilon_i$$

where $\varepsilon_i \sim N(0, \sigma_\varepsilon^2)$, $t$ is a dummy variable known for the whole population and $s$ is a bivariate variable that we *hypothesize* have a uniform distribution in every region. The data are divided in $Q = 9$ rectangular regions that can be represented by their vertices $\{(a_{1q}, b_{1q}), (a_{2q}, b_{1q}), (a_{2q}, b_{2q}), (a_{1q}, b_{2q})]$. As we presented in Section 1.7, there are at least two family of bivariate smoothers that we can use to model $f(s)$: the tensor product smoother and the radial smoother. Both the models have pros and cons: the tensor product smoother, using the truncated linear basis functions (1.29) is relatively easy to implement, but the number of random effects $u$ increases really fast; the thin plate splines (1.42) are more complex computationally (since the $\mathbf{Z}$ matrix for the non sampled units needs to be computed inside BUGS) but tend to have good numerical properties. In particular, as pointed out by Crainiceanu et al. (2005, p.2), ≪the posterior correlation of parameters of the thin-plate splines is much smaller than for other basis (e.g. truncated polynomials) which greatly improves mixing≫.

We decided to implement both the models, however we present here only the thin plate splines model as it produces better results. The model (3.5.5) becomes

$$y_{iq} = \beta_0 + \beta_t t_{iq} + \boldsymbol{\beta}_s \mathbf{s}_{iq} + \sum_{k=1}^{K} u_k z_k(\mathbf{s}_{iq}) + \varepsilon_i, \tag{3.15}$$

where $z_k(\mathbf{s}_{iq})$ are defined by (1.42), $\varepsilon_i \sim N(0, \sigma_\varepsilon^2)$, $u_k \sim N(0, \sigma_u^2)$, We consider $K = 64$ knots selected on a regular grid on the space.

The hierarchical Bayesian model for this situation is

$$\begin{gathered} y_{iq} | \beta_0, \beta_t, \boldsymbol{\beta}_s, \mathbf{u}, \sigma_\varepsilon^2 \overset{\text{ind}}{\sim} N\left(\beta_0 + \beta_t t_{iq} + \boldsymbol{\beta}_s \mathbf{s}_{iq} + \sum_{k=1}^{K} u_k z_k(\mathbf{s}_{iq}), \sigma_\varepsilon^2\right), \\ \mathbf{u}|\sigma_u^2 \sim N(\mathbf{0}, \sigma_u^2 \mathbf{I}_K), \qquad \beta_0, \beta_t, \boldsymbol{\beta}_s \overset{\text{ind}}{\sim} N(0, 10^8), \\ \tau_u, \tau_\varepsilon \overset{\text{ind}}{\sim} \text{Gamma}(10^{-8}, 10^{-8}). \end{gathered} \tag{3.16}$$

The measurement error hypothesis is

$$\mathbf{c}_q = \left[\frac{(a_{1q}+a_{2q})}{2}; \frac{(b_{1q}+b_{2q})}{2}\right], \qquad \mathbf{s}_{iq} = \mathbf{c}_q + \nu_{iq}, \tag{3.17}$$
$$\nu_{iq} \overset{\text{ind}}{\sim} \text{Unif}\left(\frac{(a_{1q}-a_{2q})}{2}; \frac{(a_{2q}-a_{1q})}{2}; \frac{(b_{1q}-b_{2q})}{2}; \frac{(b_{2q}-b_{1q})}{2}\right),$$

where Unif is a bivariate uniform distribution. The parametrization (3.17) corresponds to the hypothesis

$$\mathbf{s}_{iq} \overset{\text{ind}}{\sim} \text{Unif}(a_{1q}; a_{2q}; b_{1q}; b_{2q}). \tag{3.18}$$

We fitted model (3.16) to a set of simulated data with

$$N = 2000, \quad n = 300, \quad \beta_t = 0.4, \quad \sigma_\varepsilon = 0.2,$$
$$t \sim \text{Ber}(0.5), \quad s \sim \text{Unif}(0; 0; 1; 1), \tag{3.19}$$

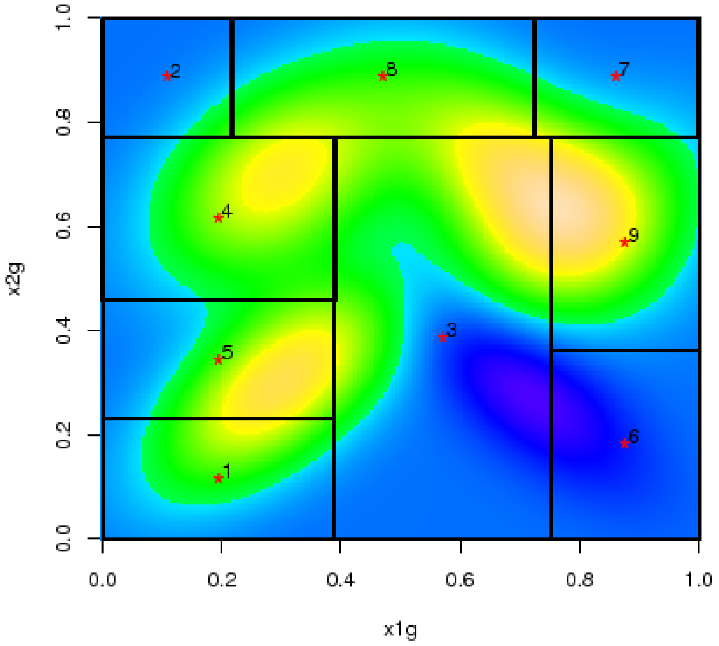

**Figure 3.12:** Bivariate normal mixture density for the bivariate uniform model. The rectangular are the regions and the red stars indicate the centroids.

The function $f(\mathbf{s})$ is generated as a bivariate normal mixture density (following Wand and Jones (1993)) and is represented in Figure 3.12. The regions are obtained using a *random binary splitting* procedure.

We implement the MCMC analysis[8] with a *burn-in* period of 15000 iterations and then we retain 5000 iterations, thinned by a factor of 5, resulting in a sample of size 1000 collected for inference.

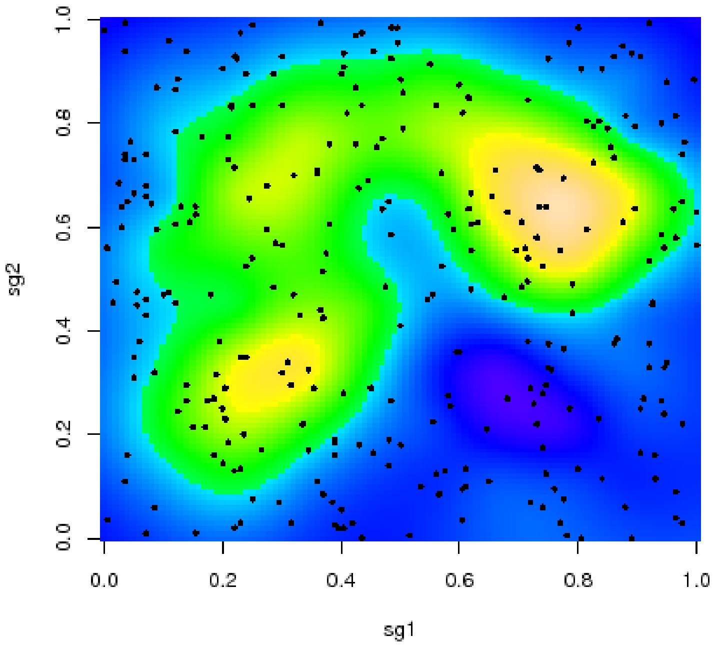

**Figure 3.13:** MCMC-base fitting of the bivariate uniform model. The points indicate the sampled units.

Figure 3.13 shows the interpolated spatial function for model (3.15). Compared with the true function showed in Figure 3.12, we see that the model produces a good fitting. In Figure 3.14 we summarize the MCMC output for the model parameters: the true values (3.19) are inside the 95% credible sets; the credible interval for $\sigma_u$ is away from zero; and all chains are well-behaved.

The posterior density distributions of the region mean estimators[9] are presented in Figure 3.15: the red line corresponds to the ME approach, the

[8] The WinBUGS model code is presented in appendix (Section A.4).

[9] The summary of the MCMC output for the region mean estimators is presented in appendix (Figure A.13).

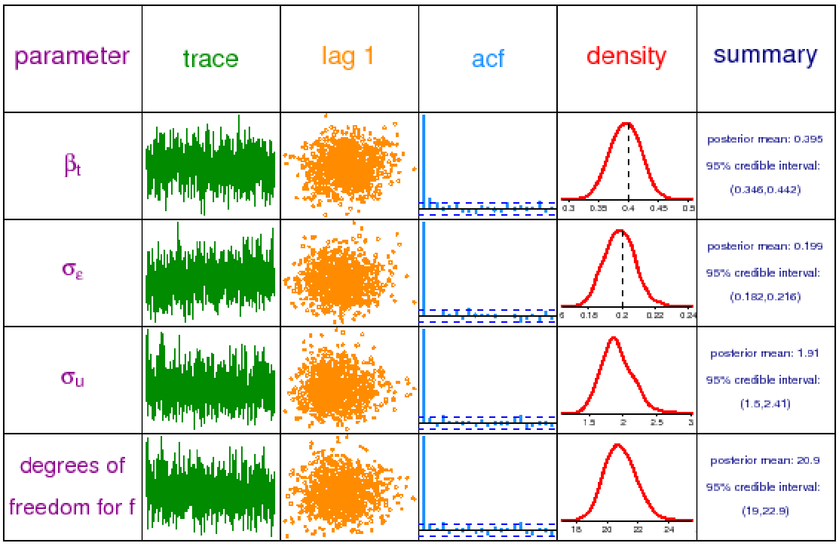

**Figure 3.14:** Graphical summary of MCMC-based inference for the parameters of the bivariate uniform model. The columns are: parameter, trace plot of MCMC sample, plot of sample against 1-lagged sample, sample autocorrelation function, kernel estimates posterior density and basic numerical summaries. The vertical dashed lines in the density plots, where present, correspond to the true values of the parameters according to the simulation set-up.

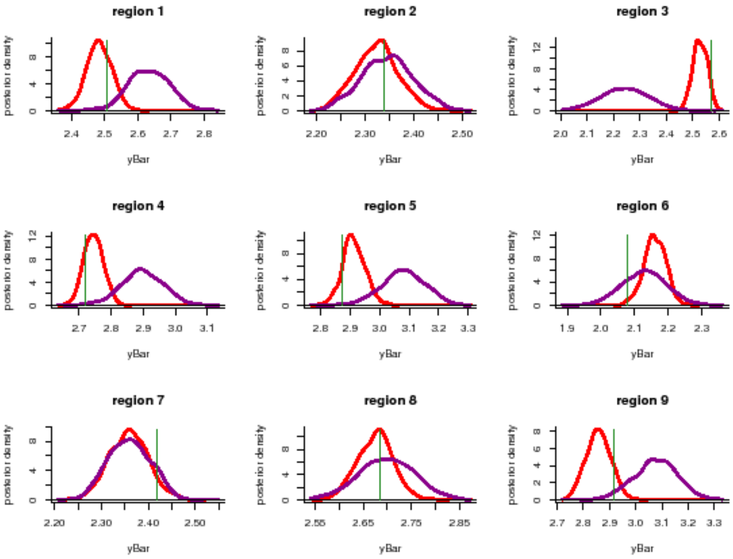

**Figure 3.15:** Posterior density of the region mean estimator for the bivariate uniform model. The red lines correspond to the ME approach, the purple lines to the naive approach and the vertical green lines are the true mean values.

purple line to the naive approach and the vertical line is the true mean value. As we can see, the ME estimator has generally a good performance: in particular, the two methods have quite similar results if the relationship between $s$ and $y$ is nearly linear (as in region 2 and 7); however, when the relationship is highly non-linear (like, for example,in region 3, 4 and 5) the ME estimator produces a really good result, while the naive estimator fails.

### 3.5.6 Albania Model

In the final experiment we apply model (3.15) (without covariate $t$) to some variables from the dataset of the 2002 Living Standard Measurement Study (LSMS) conducted in Albania[10]. In this survey, the data are referred to various geographical levels, and the spatial location of each household is collected. Thus, we decide to model the *household log per-capita consumption expenditure* (response variable $y$) against the household spatial location ($s$) and to estimate the mean of $y$ for the 36 districts of Albania (presented in Figure 3.16(a)).

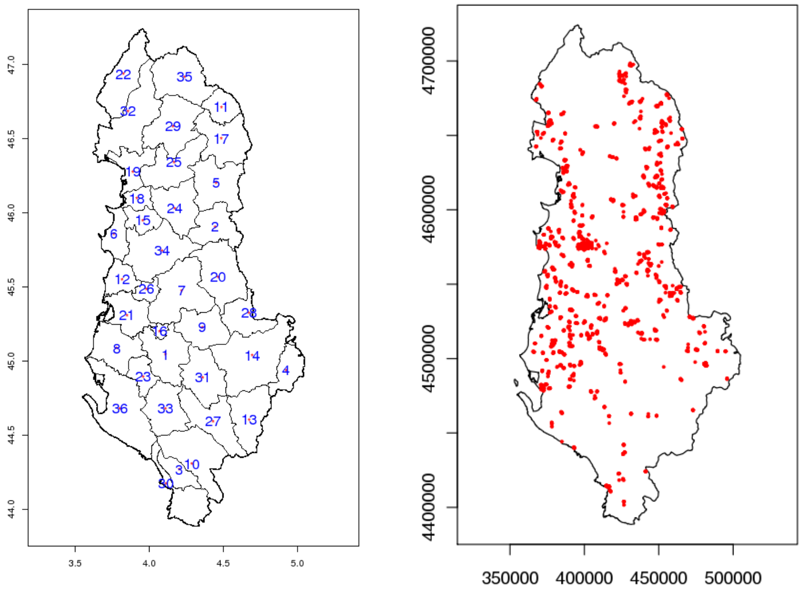

**Figure 3.16:** Map of the districts of the Republic of Albania with the corresponding codes and locations of the households in LSMS dataset.

Again, we make the measurement error hypothesis of uniform distribution

[10] For a detailed description of the dataset we refer to Chapter 4.

of $s$ inside each region. Observing Figure 3.16(b), we are aware that this assumption is not quite realistic, nonetheless we decided to proceed as a possible first step to other future assumptions.

Moreover, differently from the simulated data utilized in the previous experiment, the districts polygons are extremely irregular. This introduce the issue on how to define the uniform distribution on the areas. We proceeded in two ways (but only results of the second way are presented here):

- first, $s_{iq}$ has uniform distribution (3.18) on the *bounding box* of region $q$, that is the rectangular region with vertices $[(a_{1q}, b_{1q}), (a_{2q}, b_{1q}), (a_{2q}, b_{2q}), (a_{1q}, b_{2q})]$ that includes the polygon $q$;
- second, $s_{iq}$ has uniform distribution on the points that compose the polygon $q$.

The first approach is quite easy to implement, but it includes as plausible values of $s$ all the points that lay outside the polygon but inside the bounding box. On the other hand, the second approach considers only the points that are inside the polygon but is more computationally intensive, as it needs to define the list of points that lay inside the polygon and to associate an equal selection probability to each point.

In addition, we want to highlight that the use of this second approach can theoretically be generalized to other bivariate distributions of $s$ by changing the selection probabilities of the points.

The hierarchical Bayesian model is

$$\begin{gathered} y_{iq}|\beta_0, \boldsymbol{\beta}_s, \mathbf{u}, \sigma_\varepsilon^2 \overset{\text{ind}}{\sim} N\left(\beta_0 + \boldsymbol{\beta}_s \mathbf{s}_{iq} + \sum_{k=1}^{K} u_k z_k(\mathbf{s}_{iq}), \sigma_\varepsilon^2\right), \\ \mathbf{u}|\sigma_u^2 \sim N(\mathbf{0}, \sigma_u^2 \mathbf{I}_K), \qquad \beta_0, \boldsymbol{\beta}_s \overset{\text{ind}}{\sim} N(0, 10^8), \\ \tau_u, \tau_\varepsilon \overset{\text{ind}}{\sim} \text{Gamma}(10^{-8}, 10^{-8}), \end{gathered} \tag{3.20}$$

and the experiment settings are $N = 3591$ and $n = 718$ (corresponding to a sampling fraction of 20%).

The MCMC analysis[11] is implemented with a *burn-in* period of 30000 iterations and a retain of 10000 iterations with a thinning factor of 5, resulting in a sample of size 2000 collected for inference. Figure 3.17 summarizes the MCMC output for the model parameters.

The posterior density distributions for the mean estimators[12] (3.4) in the 36 districts are presented in Figure 3.18. There are little differences

[11] The WinBUGS model code is presented in appendix (Section A.5).

[12] The summary of the MCMC output for the region mean estimators is presented in appendix (Figure A.15).

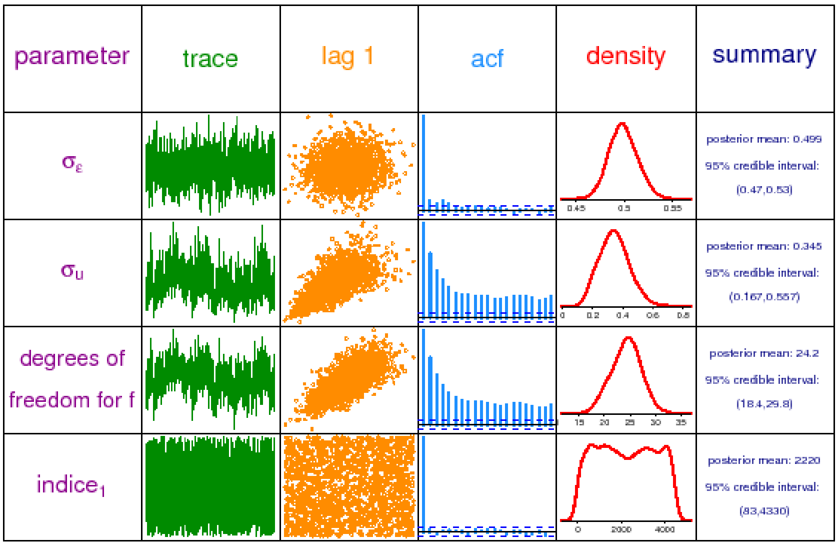

**Figure 3.17:** Graphical summary of MCMC-based inference for the parameters of the Albania model. The columns are: parameter, trace plot of MCMC sample, plot of sample against 1-lagged sample, sample autocorrelation function, kernel estimates posterior density and basic numerical summaries.

between the performances of the ME estimator and the naive estimator and the results do not show a definitive trend in all the regions[13]. These results are not unexpected, as we have already pointed out that the distribution of the households on the Albanian surface is not uniform (see Figure 3.16(b)). Moreover, in some regions the number of sampled units is really small and this influences the reliability of the estimates, especially for the provinces that are near the country boundaries. For example, if we consider the district 22, we observe that both the estimators underestimate the true parameter.

Notwithstanding this single result, we are confident - on the basis of the procedure's properties discussed in this chapter - that the measurement error approach considering a more realistic hypothesis on spatial distribution for the households can improve the estimates of district mean of the household log per-capita consumption expenditure, with respect to the centroids classic approach. Definitely, further investigations should be done in this direction.

[13] Similar results are obtained from the experiment with the hypothesis of $s$ uniformly distributed on the bounding box.

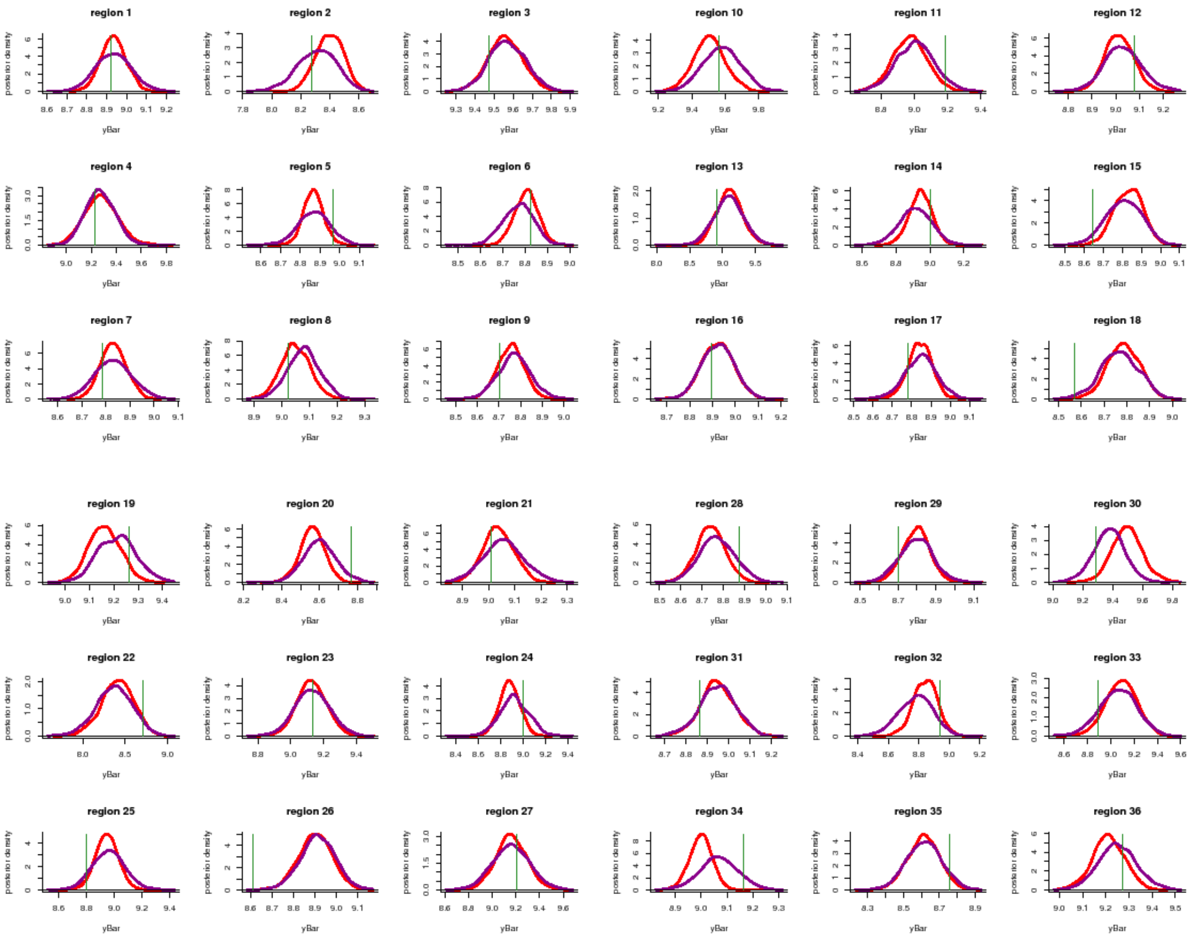

**Figure 3.18:** Posterior density of the region mean estimator for the Albania model. The red lines correspond to the ME approach, the purple lines to the naive approach and the vertical green lines are the true mean values.

# Chapter 4

# Geoadditive Model for the Estimation of Consumption Expenditure in Albania

## 4.1 Introduction

This last chapter is devoted to the application of a geoadditive model in the field of poverty mapping at small area level. In particular, we apply a geoadditive small area estimation model in order to estimate the district level mean of the household log per-capita consumption expenditure for the Republic of Albania. We combine the model parameters estimated using the dataset of the 2002 Living Standard Measurement Study with the 2001 Population and Housing Census covariate information. In Section 4.2 we present the general structure of the geoadditive SAE model. Section 4.3 illustrates the application presenting the data (subsection 4.3.1) and the results (subsection 4.3.2). Finally, in Section 4.4 we discuss the use of two possible MSE estimators through a desing-based simulation study.

## 4.2 Geoadditive SAE Model

In Section 2.4.6 we generally introduced small area estimation methods and their possible formulations for the analysis of spatial data. In this section we present more in detail the geoadditive SAE model, that is obtained by the union of the geoadditive model presented in Section 1.8 and the classic SAE model (Rao, 2003) under the linear mixed model framework.

Suppose that there are $T$ small areas for which we want to estimate a quantity of interest and let $y_{it}$ denote the value of the response variable for the $i$th unit, $i = 1, ..., n$, in small area $t$, $t = 1, ..., T$. Let $\mathbf{x}_{it}$ be a vector of $p$ linear covariates associated with the same unit, then the *classic SAE model* is given by

$$y_{it} = \mathbf{x}_{it}\boldsymbol{\beta} + u_t + \varepsilon_{it}, \qquad \varepsilon_{it} \sim N(0, \sigma_\varepsilon^2), \quad u_t \sim N(0, \sigma_u^2), \tag{4.1}$$

where $\boldsymbol{\beta}$ is a vector of $p$ unknown coefficients, $u_t$ is the random area effect associated with small area $t$ and $\varepsilon_{it}$ is the individual level random error. The two error terms are assumed to be mutually independent, both across individuals as well as across areas.

If we define the matrix $\mathbf{D} = [d_{it}]$ with

$$d_{it} = \begin{cases} 1 & \text{if observation } i \text{ is in small area } t, \\ 0 & \text{otherwise} \end{cases} \tag{4.2}$$

and $\mathbf{y} = [y_{it}]$, $\mathbf{X} = \left[\mathbf{x}_{it}^T\right]$, $\mathbf{u} = [u_t]$ and $\boldsymbol{\varepsilon} = [\varepsilon_{it}]$, then the matrix notation of (4.1) is

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{D}\mathbf{u} + \boldsymbol{\varepsilon}, \tag{4.3}$$

with

$$\mathrm{E}\begin{bmatrix} \mathbf{u} \\ \boldsymbol{\varepsilon} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \qquad \mathrm{Cov}\begin{bmatrix} \mathbf{u} \\ \boldsymbol{\varepsilon} \end{bmatrix} = \begin{bmatrix} \sigma_u^2\mathbf{I}_T & 0 \\ 0 & \sigma_\varepsilon^2\mathbf{I}_n \end{bmatrix}.$$

The covariance matrix of $\mathbf{y}$ is

$$\mathrm{Var}(\mathbf{y}) \equiv \mathbf{V} = \sigma_u^2\mathbf{D}\mathbf{D}^T + \sigma_\varepsilon^2\mathbf{I}_n$$

and the BLUPs of the model coefficients are

$$\begin{aligned} \boldsymbol{\beta} &= \left(\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{V}^{-1}\mathbf{y}, \\ \mathbf{u} &= \sigma_u^2\mathbf{D}^T\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}). \end{aligned}$$

If the variance components $\sigma_u^2$ and $\sigma_\varepsilon^2$ are unknown, they are estimated by REML or ML methods and the model coefficients are obtained with the EBLUPs.

The formulation (4.3) is a linear mixed model, analogous to the geoadditive model (1.45), thus it is straightforward to compose the geoadditive SAE model. Consider again the response $y_{it}$ and the vector of $p$ linear covariates $\mathbf{x}_{it}$, and suppose that both are measured at a spatial location $\mathbf{s}_{it}$, $\mathbf{s} \in \Re^2$. The *geoadditive SAE model*[1] for such data is a linear mixed model with two random effects components:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\gamma} + \mathbf{D}\mathbf{u} + \boldsymbol{\varepsilon}, \tag{4.4}$$

[1] The same model formulation is in Opsomer et al. (2008), where is presented a model, called by the authors *non-parametric SAE model*, that accounts for a generic non-linear covariate.

with

$$\mathrm{E}\begin{bmatrix}\boldsymbol{\gamma}\\ \mathbf{u}\\ \boldsymbol{\varepsilon}\end{bmatrix}=\begin{bmatrix}0\\0\\0\end{bmatrix},\qquad \mathrm{Cov}\begin{bmatrix}\boldsymbol{\gamma}\\ \mathbf{u}\\ \boldsymbol{\varepsilon}\end{bmatrix}=\begin{bmatrix}\sigma^2_\gamma\mathbf{I}_K & 0 & 0\\ 0 & \sigma^2_u\mathbf{I}_T & 0\\ 0 & 0 & \sigma^2_\varepsilon\mathbf{I}_n\end{bmatrix}.$$

Now $\mathbf{X}=\left[\mathbf{x}^T_{it},\mathbf{s}^T_{it}\right]_{1\le i\le n}$ has $p+2$ columns, $\boldsymbol{\beta}$ is a vector of $p+2$ unknown coefficients, $\mathbf{u}$ are the random small area effects, $\boldsymbol{\gamma}$ are the thin plate spline coefficients (seen as random effects) and $\boldsymbol{\varepsilon}$ are the individual level random errors. Matrix $\mathbf{D}$ is still defined by (4.2) and $\mathbf{Z}$ is the matrix of the thin plate spline basis functions

$$\mathbf{Z}=\left[C\left(\mathbf{s}_i-\boldsymbol{\kappa}_k\right)\right]_{1\le i\le n,1\le k\le K}\left[C\left(\boldsymbol{\kappa}_h-\boldsymbol{\kappa}_k\right)\right]^{-1/2}_{1\le h,k\le K},$$

with $K$ knots $\boldsymbol{\kappa}_k$ and $C(\mathbf{r})=\|\mathbf{r}\|^2\log\|\mathbf{r}\|$.

Again, the unknown variance components are estimated via REML or ML estimators and are indicated with $\hat{\sigma}^2_\gamma$, $\hat{\sigma}^2_u$ and $\hat{\sigma}^2_\varepsilon$. The estimated covariance matrix of $\mathbf{y}$ is

$$\hat{\mathbf{V}}=\hat{\sigma}^2_\gamma\mathbf{Z}\mathbf{Z}^T+\hat{\sigma}^2_u\mathbf{D}\mathbf{D}^T+\hat{\sigma}^2_\varepsilon\mathbf{I}_n \tag{4.5}$$

and the EBLUP estimators of the model coefficients are

$$\hat{\boldsymbol{\beta}}=\left(\mathbf{X}^T\hat{\mathbf{V}}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}^T\hat{\mathbf{V}}^{-1}\mathbf{y}, \tag{4.6}$$

$$\hat{\boldsymbol{\gamma}}=\hat{\sigma}^2_\gamma\mathbf{Z}^T\hat{\mathbf{V}}^{-1}(\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}}), \tag{4.7}$$

$$\hat{\mathbf{u}}=\hat{\sigma}^2_u\mathbf{D}^T\hat{\mathbf{V}}^{-1}(\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}}). \tag{4.8}$$

For a given small area $t$, we are interested in predicting the mean value of $y$

$$\bar{y}_t=\overline{\mathbf{x}}_t\boldsymbol{\beta}+\overline{\mathbf{z}}_t\boldsymbol{\gamma}+u_t$$

where $\overline{\mathbf{x}}_t$ and $\overline{\mathbf{z}}_t$ are the true means over the small area $t$ and are assumed to be known. The EBLUP for the quantity of interest is

$$\hat{\bar{y}}_t=\overline{\mathbf{x}}_t\hat{\boldsymbol{\beta}}+\overline{\mathbf{z}}_t\hat{\boldsymbol{\gamma}}+\mathbf{e}_t\hat{\mathbf{u}} \tag{4.9}$$

where $\mathbf{e}_t$ is a vector with 1 in the $t$-th position and zeros elsewhere.

## 4.3 Estimation of the Household Per-capita Consumption Expenditure in Albania

Poverty maps are useful tools to describe the spatial distribution of poverty in a country, especially when they represent small geographic units, such as municipalities or districts. This information is extremely useful to policymakers and researchers in order to formulate efficient policies and programs.

As pointed out in Neri et al. (2005), ≪in order to produce poverty maps, large data sets are required which include reasonable measures of income or consumption expenditure and which are representative and of sufficient size at low levels of aggregation to yield statistically reliable estimates. Household budget surveys or living standard surveys covering income and consumption usually used to calculate distributional measures are rarely of such a sufficient size; whereas census or other large sample surveys large enough to allow disaggregation have little or no information regarding monetary variables≫. Then, the required small area estimates are usually based on a combination of sample surveys and administrative data.

The Republic of Albania is divided in 3 geographical levels: prefectures, districts and communes. There are 12 prefectures, 36 districts and 374 communes, however the Living Standard Measurement Study survey, which provides valuable information on a variety of issues related to living conditions in Albania, is stratified in 4 big strata (Costal Area, Central Area, Mountain Area and Tirana) and these strata are the smaller domain of direct estimation. In order to map and estimate the mean of the household log per-capita consumption expenditure for the districts of Albania, we apply a geoadditive SAE model combining the model estimated using the survey data and the census covariate information.

### 4.3.1 Data

The two main sources of statistical information available in Albania are the 2001 Population and Housing Census (PHC) and the 2002 Living Standard Measurement Study (LSMS), both conducted in Albania by the INSTAT (Albanian Institute of Statistics).

The 2002 LSMS provides individual level and household level socio-economic data from 3,599 households drawn from urban and rural areas in Albania. The sample was designed to be representative of Albania as a whole, Tirana, other urban/rural locations, and the three main agro-ecological areas (Coastal, Central, and Mountain). The survey was carried out by the Albanian Institute of Statistics (INSTAT) with the technical and financial assistance of the World Bank.

Four survey instruments were used to collect information for the 2002 Albania LSMS: a household questionnaire, a diary for recording household food consumption, a community questionnaire, and a price questionnaire. The household questionnaire included all the core LSMS modules as defined in Grosh and Glewwe (2000), plus additional modules on migration, fertility, subjective poverty, agriculture, and nonfarm enterprises. Geographical referencing data on the longitude and latitude of each household were also recorded using portable GPS devices (World Bank and INSTAT, 2003).

The covariates selected to fit the geoadditive SAE model are chosen following prior studies on poverty assessment in Albania (Betti, G. and Ballini, F. and Neri, L., 2003; Neri et al., 2005). We have selected the following household level covariates:

- *size of the household* (in term of number of components)

- *information on the components of the household*:
  - age of the householder,
  - marital status of the householder,
  - age of the spouse or husband of the household,
  - number of children 0-5 years,
  - age of the first child,
  - number of components without work,
  - highest level of education in the household;

- *information on the house*:
  - building with 2-15 units,
  - built with brick or stone,
  - built before 1960,
  - number of rooms per person,
  - house surface $< 40$ m$^2$,
  - house surface $40 - 69^2$,
  - wc inside;

- *presence of facilities in the dwelling*:
  - TV,
  - parabolic,
  - refrigerator,
  - washing machine,
  - air conditioning,
  - computer,
  - car;

- *ownership of agricultural land*

All these variables are are available both in LSMS and PHC surveys (see Neri et al. (2005) for comparability between the two sources); in addition, the geographical location of each household is available for the LSMS data.

The response variable is the logarithm of the household per-capita consumption expenditure. The use of the logarithmic transformation is typical for this type of data as it produce a more suitable response for the regression model (see the distributions presented in Figure 4.1).

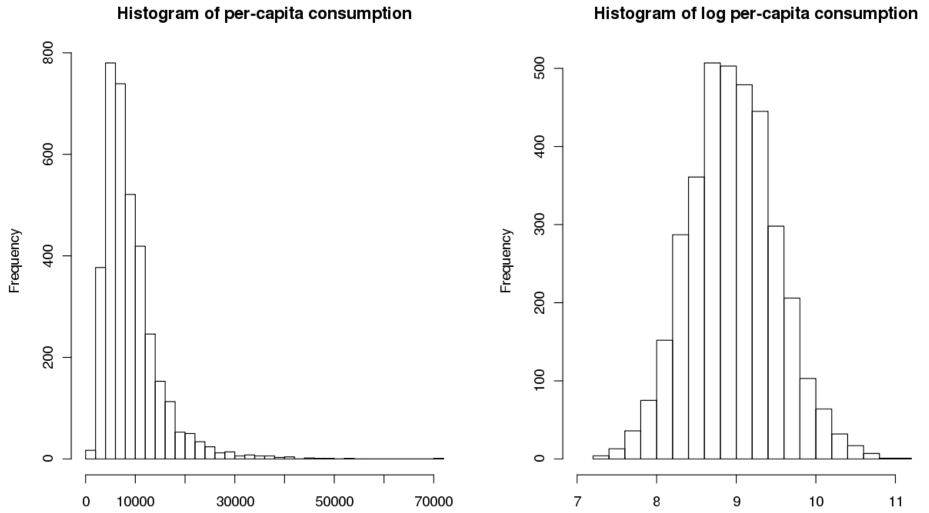

**Figure 4.1:** Distribution of the household per-capita consumption expenditure, both in original scale and in logarithmic scale.

### 4.3.2 Results

Estimates of the log per-capita consumption expenditure in each of the 36 district area are derived using the geoadditive SAE model presented in (4.4).

After the preliminary analysis of various combination of parametric and non-parametric specifications for the selected covariates, the chosen model is composed by a bivariate thin plate spline on the universal transverse Mercator (UTM) coordinates, a linear term for all the other variables and a random intercept component for the area effect. The spline knots are selected setting $K = 100$ and using the *clara* space filling algorithm of Kaufman and Rousseeuw (1990) that is available in the `R` package `cluster` (the resulting knots location is presented in Figure 4.2). The model is then fitted by REML using the `lme` function in the `R` package `nlme`.

The estimated parameters are presented in Table 4.1, along with their confidence interval at 95% and the p-values. With the exclusion of the intercept and the coordinates coefficients (that are required by the model structure), almost all the parameters are highly significant. The exceptions are the coefficients of 'marital status of the householder', 'number of children 0-5 years' and 'built with brick or stone' that are significant at 5% level, and the coefficient of 'building with 2-15 units' that is significant at 10% level.

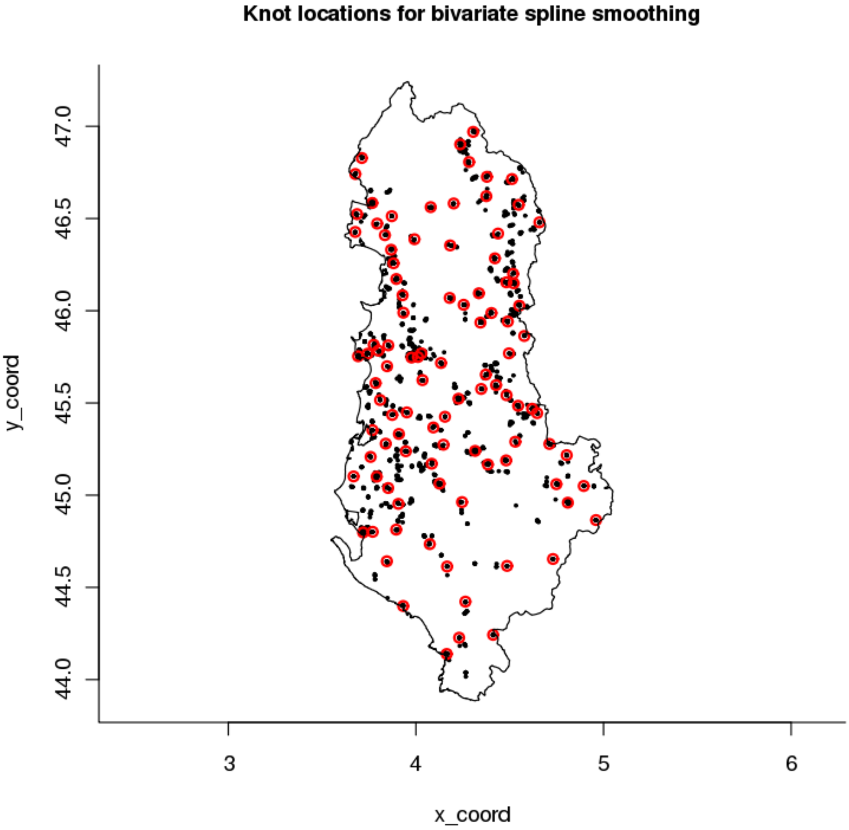

**Figure 4.2:** Knots location (in red) for the thin plate spline component. Black dots indicate the locations of the LSMS sample.

The resulting spatial smoothing of the log per-capita consumption expenditure is presented in Figure 4.3. The geoadditive SAE model (4.4) considers two random effects, once for the bivariate spline smoother and once for the small area effect, thus the estimated value of the log per-capita consumption expenditure in a specific location is obtained as sum of two components, once continuous over the space (showed in the second map) and once constant in each small area (showed in the third map). From these maps, it is evident the presence of both a spatial dynamic and a district level effect in the Albanian consumption expenditure.

The estimated parameters (presented in Table 4.1) are then combined

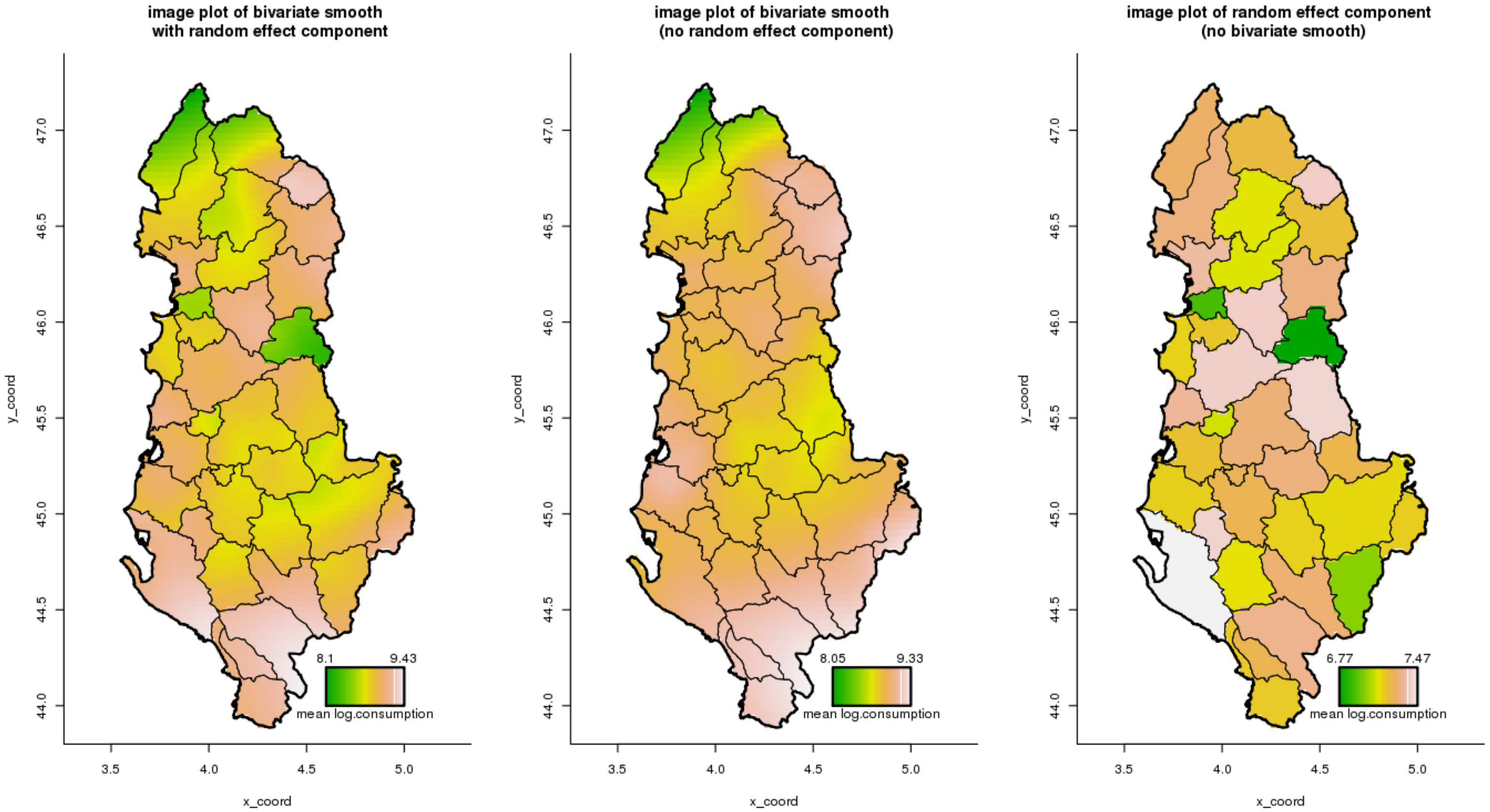
image plot of bivariate smooth
with random effect component
image plot of bivariate smooth
(no random effect component)
image plot of random effect component
(no bivariate smooth)
y_coord
x_coord
44.0
44.5
45.0
45.5
46.0
46.5
47.0
3.5
4.0
4.5
5.0
8.1
9.43
8.05
9.33
6.77
7.47
mean log.consumption

**Figure 4.3:** Spatial smoothing and district random effects of the household log per-capita consumption expenditure. The first map (a) shows the smoothing obtained with the geoadditive sae model as sum of two components: the bivariate smoothing, in map (b), and the small area random effects, in map (c).

**Table 4.1:** Estimated parameters of the geoadditive SAE model for the household log per-capita consumption expenditure at district level.

| Parameter | Estimate | Confidence Interval | p-value |
|---|---:|---:|---:|
| Fixed Effects | | | |
| Intercept | 7.11 | (-34.32;48.55) | 0.736 |
| X coordinate | -0.0594 | (-0.7807;0.6618) | 0.872 |
| Y coordinate | 0.0393 | (-0.8700;0.9487) | 0.932 |
| household size | -0.0775 | (-0.0913;-0.0638) | $< 0.001$ |
| age of the householder | 0.0029 | (0.0014;0.0044) | $< 0.001$ |
| marital status of the householder | 0.0745 | (0.0004;0.1485) | 0.049 |
| age of the spouse or husband | -0.0021 | (-0.0035;-0.0008) | 0.001 |
| number of children 0-5 years | -0.0202 | (-0.0382;-0.0023) | 0.027 |
| age of the first child | -0.0023 | (-0.0037;-0.0009) | 0.001 |
| number of components without work | -0.0661 | (-0.0784;-0.0537) | $< 0.001$ |
| high level of education | 0.0913 | (0.0648;0.1178) | $< 0.001$ |
| medium level of education | 0.2397 | (0.2007;0.2788) | $< 0.001$ |
| building with 2-15 units | 0.0261 | (-0.0034;0.0557) | 0.083 |
| built with brick or stone | 0.0342 | (0.0001;0.0684) | 0.049 |
| built before 1960 | -0.0442 | (-0.0734;-0.0151) | 0.003 |
| number of rooms per person | 0.1364 | (0.1037;0.1690) | $< 0.001$ |
| house surface $< 40$ m$^2$ | -0.0518 | (-0.0932;-0.0105) | 0.014 |
| house surface $40 - 69^2$ | -0.0365 | (-0.0625;-0.0105) | 0.006 |
| wc inside | 0.0511 | (0.0190;0.0833) | 0.002 |
| TV | 0.1066 | (0.0510;0.1623) | $< 0.001$ |
| parabolic | 0.0768 | (0.0473;0.1062) | $< 0.001$ |
| refrigerator | 0.1183 | (0.0827;0.1539) | $< 0.001$ |
| washing machine | 0.1140 | (0.0843;0.1438) | $< 0.001$ |
| air conditioning | 0.2434 | (0.1593;0.3275) | $< 0.001$ |
| computer | 0.2403 | (0.1668;0.3138) | $< 0.001$ |
| car | 0.3233 | (0.2846;0.3621) | $< 0.001$ |
| ownership of agricultural land | 0.0484 | (0.0153;0.0815) | 0.004 |
| Random Effects | | | |
| $\sigma_\gamma$ | 0.4096 | (0.2700;0.6214) | $< 0.001$ |
| $\sigma_u$ | 0.1756 | (0.1290;0.2389) | $< 0.001$ |
| $\sigma_e$ | 0.3285 | (0.3208;0.3363) | $< 0.001$ |

with the census mean values as in (4.9) to obtain the district level estimates of the average household log per-capita consumption expenditure. Due to the unavailability of the geographical coordinates for the PHC dataset, the true $\overline{\mathbf{z}}_t$ cannot be calculated. We approximate the missing information by using the centroids of each small area to locate all the units belonging to the same area[2].

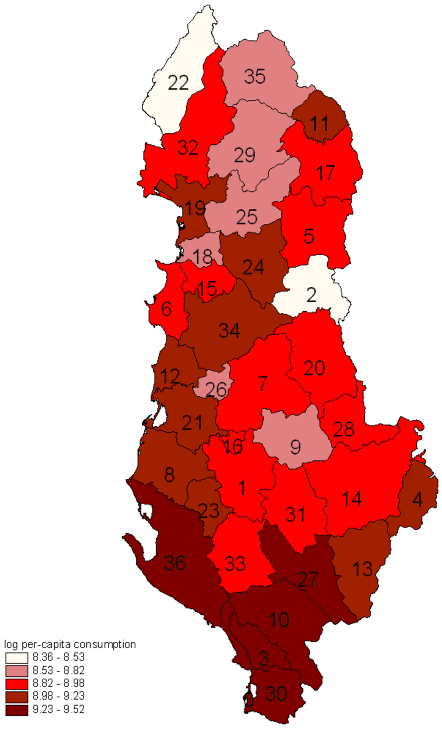

**Figure 4.4:** District level estimates of the mean of household log per-capita consumption expenditure.

The district level estimates are showed in Figure 4.4 and in Table 4.2. All the coefficients of variation[3] (CV) are less that 2%, with a mean value of 0.91%, thus the estimates have low variability. The higher values are

[2] As we have discussed in Section 3.5.6, we are confident that a measurement error approach considering a more realistic hypothesis on spatial distribution for the households can improve the estimates, with respect to the centroids approach. Further investigations will be done in this direction.

[3] The MSE - and consequently the CV - is calculated using the robust MSE estimator of Salvati et al. (2008). For a discussion about MSE estimation see Section 4.4.

**Table 4.2:** District level estimates of the mean of household log per-capita consumption expenditure. The root mean squared error (RMSE) and the coefficient of variation (CV%) are obtained with the robust MSE estimator of Salvati et al. (2008).

| Code | District Name | Estimate | RMSE | CV% |
|---|---|---|---|---|
| 1 | Berat | 8.91 | 0.0472 | 0.53 |
| 2 | Bulqize | 8.35 | 0.0514 | 0.62 |
| 3 | Delvine | 9.46 | 0.1552 | 1.64 |
| 4 | Devoll | 9.17 | 0.1529 | 1.67 |
| 5 | Diber | 8.96 | 0.0542 | 0.60 |
| 6 | Durres | 8.98 | 0.0601 | 0.67 |
| 7 | Elbasan | 8.93 | 0.0368 | 0.41 |
| 8 | Fier | 9.13 | 0.0441 | 0.48 |
| 9 | Gramsh | 8.82 | 0.0426 | 0.48 |
| 10 | Gjirokast | 9.52 | 0.1130 | 1.19 |
| 11 | Has | 9.15 | 0.1046 | 1.14 |
| 12 | Kavaje | 9.22 | 0.0535 | 0.58 |
| 13 | Kolonje | 9.05 | 0.1608 | 1.78 |
| 14 | Korce | 8.92 | 0.0630 | 0.71 |
| 15 | Kruje | 8.91 | 0.0758 | 0.85 |
| 16 | Kucove | 8.96 | 0.0449 | 0.50 |
| 17 | Kukes | 8.97 | 0.0753 | 0.84 |
| 18 | Kurbin | 8.67 | 0.0549 | 0.63 |
| 19 | Lezhe | 9.21 | 0.0773 | 0.84 |
| 20 | Librazhd | 8.88 | 0.0450 | 0.56 |
| 21 | Lushnje | 9.10 | 0.0576 | 0.63 |
| 22 | Malesi e Madhe | 8.53 | 0.1661 | 1.95 |
| 23 | Mallakaster | 9.11 | 0.0654 | 0.72 |
| 24 | Mat | 9.15 | 0.0969 | 1.06 |
| 25 | Mirdite | 8.79 | 0.1049 | 1.19 |
| 26 | Peqin | 8.74 | 0.0864 | 0.99 |
| 27 | Permet | 9.34 | 0.1365 | 1.46 |
| 28 | Pogradec | 8.88 | 0.0626 | 0.70 |
| 29 | Puke | 8.70 | 0.1388 | 1.60 |
| 30 | Sarande | 9.34 | 0.0809 | 0.87 |
| 31 | Skrapar | 8.93 | 0.0999 | 1.12 |
| 32 | Shkoder | 8.90 | 0.0640 | 0.72 |
| 33 | Tepelene | 8.95 | 0.0871 | 0.97 |
| 34 | Tirane | 9.23 | 0.0441 | 0.48 |
| 35 | Tropoje | 8.78 | 0.0679 | 0.77 |
| 36 | Vlore | 9.37 | 0.0635 | 0.68 |

registered in those districts where the sample size is quite low (see Table 4.3). In addition, district 22 suffers particularly from the centroid approximation due to its geographical morphology: it is mostly mountainous and the urban area is mainly in the south.

The map presents a clear geographical pattern, with the higher values in the south and south-west of the country and the lower value in the mountainous area (north and north-east). These results are consistent with previous applications on the same datasets presented in literature (Neri et al., 2005; Tzavidis et al., 2008).

## 4.4 MSE Estimation

Along with the definition of the non-parametric SAE model, Opsomer et al. (2008) study the theoretical properties of the mean squared error (MSE) of the small area mean estimator and propose both an analytic and a bootstrap estimator for the MSE quantity. Alternatively, Salvati et al. (2008) propose a robust estimator of the conditional MSE of the same non-parametric SAE model, based on the pseudo-linearization approach to MSE estimation described in Chambers et al. (2007).

We decided to apply both the analytic estimator of Opsomer et al. (2008) and the robust estimator of Salvati et al. (2008) and, in order to evaluate their performance, a desing-based simulation study is implemented.

We build a fixed pseudo-population of $N = 689733$ households by sampling N times with replacement and with probability proportional to the unit sample weights from the LSMS dataset. A total of 500 independent stratified random samples of the same size as the original sample is then selected from this pseudo-population, with districts sample sizes fixed to be the same as in the original sample. For each sample we apply the geoadditive SAE model of the previous section and we calculate the EBLUP 4.9 for the mean household log per-capita consumption expenditure of each district and the two relative MSE estimates.

The behaviour of the empirical true root MSE and its estimators for each district is shown in Figure 4.5. It can be seen that there isn't a substantial difference in the performance of the two estimators, even if the analytic estimator of Opsomer et al. (2008) is always lower that the robust estimator of Salvati et al. (2008). However, the robust estimator seems to better track the irregular profile of the empirical RMSE, while the analytic estimator is slightly over-smoothed. The anomalous value of district 22 is due to the high value of the bias component (see Table 4.3) and both the estimators undervalue it. After these considerations, we prefer to present the MSE

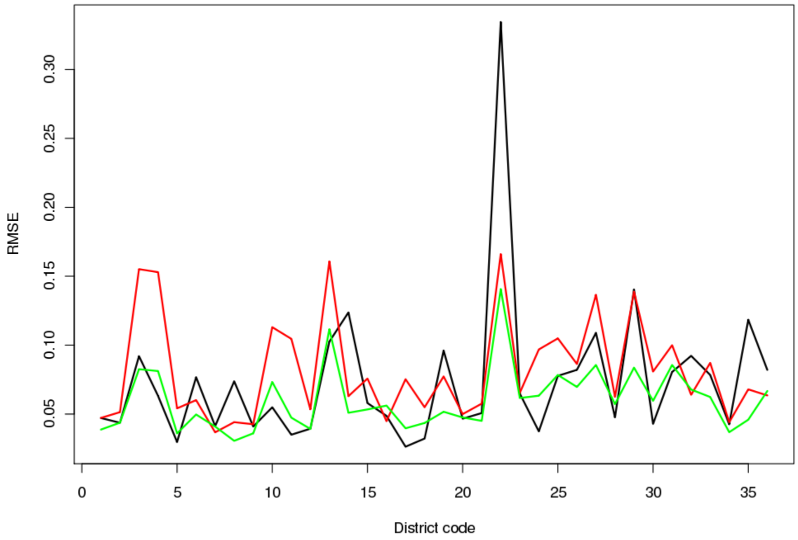

**Figure 4.5:** District level of actual design-base RMSE (black line) and average estimated RMSE. The red line indicates the robust estimator of Salvati et al. (2008) and the green line indicates the analytic estimator of Opsomer et al. (2008).

estimated with the robust estimator of Salvati et al. (2008) (see Table 4.2).

The simulation study permits also to evaluate the performance of the geoadditive SAE EBLUP. For each district we compute the *Relative Bias* (RB) and the *Relative Root MSE* (RRMSE) defined as

$$RB = \frac{1}{M}\frac{\sum_{m=1}^{M}(\hat{\bar{y}}_{tm} - \bar{y}_t)}{\bar{y}_t}$$

and

$$RRMSE = \frac{\sqrt{\frac{1}{M}\sum_{m=1}^{M}(\hat{\bar{y}}_{tm} - \bar{y}_t)^2}}{\bar{y}_t},$$

where $\bar{y}_t$ denotes the actual district mean $t$ and $\hat{\bar{y}}_{tm}$ is the predicted value at simulation $m$, $m = 1, ..., M$.

The values of RB and RRMSE are shown in Table 4.3: all the values are quite small and indicate that the geoadditive SAE EBLUP is quite stable. Once again, we note the anomalous value of district 22, that presents a relative bias of -3.64%.

**Table 4.3:** Relative bias (RB) and relative RMSE (RRMSE) of the geoadditive SAE EBLUP for the mean household log per-capita consumption expenditure of each district.

| Code | $n_t$ | $N_t$ | RB% | RRMSE% |
|---|---|---|---|---|
| 1 | 120 | 25422 | 0.0415 | 0.5253 |
| 2 | 128 | 8499 | -0.2331 | 0.5315 |
| 3 | 16 | 3211 | -0.1014 | 0.9636 |
| 4 | 16 | 6229 | 0.4183 | 0.6849 |
| 5 | 232 | 16529 | -0.2085 | 0.3340 |
| 6 | 160 | 42332 | -0.6805 | 0.8689 |
| 7 | 152 | 47709 | -0.2717 | 0.4713 |
| 8 | 224 | 45729 | 0.7459 | 0.8192 |
| 9 | 120 | 7538 | 0.2225 | 0.4748 |
| 10 | 32 | 10948 | 0.0481 | 0.5735 |
| 11 | 48 | 3450 | -0.0556 | 0.3822 |
| 12 | 88 | 18294 | -0.0752 | 0.4343 |
| 13 | 8 | 2291 | 0.8891 | 1.1532 |
| 14 | 136 | 34914 | -1.3420 | 1.3859 |
| 15 | 39 | 13477 | 0.2356 | 0.6724 |
| 16 | 32 | 11019 | -0.3109 | 0.5480 |
| 17 | 184 | 12183 | -0.0256 | 0.3023 |
| 18 | 64 | 12938 | 0.0352 | 0.3769 |
| 19 | 64 | 13538 | 0.9217 | 1.0422 |
| 20 | 200 | 14345 | -0.0589 | 0.5352 |
| 21 | 152 | 31953 | 0.3629 | 0.5639 |
| 22 | 24 | 9294 | -3.6434 | 3.8363 |
| 23 | 32 | 7067 | -0.2909 | 0.7124 |
| 24 | 32 | 11803 | -0.1271 | 0.4165 |
| 25 | 16 | 5468 | 0.1970 | 0.8839 |
| 26 | 24 | 8814 | -0.4073 | 0.9564 |
| 27 | 16 | 5377 | 0.3386 | 1.1810 |
| 28 | 48 | 17418 | 0.1240 | 0.5389 |
| 29 | 24 | 8633 | -1.1829 | 1.6128 |
| 30 | 48 | 9874 | -0.0584 | 0.4638 |
| 31 | 16 | 5453 | 0.4741 | 0.9140 |
| 32 | 140 | 43578 | -0.9057 | 1.0329 |
| 33 | 32 | 11202 | 0.1509 | 0.8862 |
| 34 | 684 | 121020 | 0.1318 | 0.4668 |
| 35 | 88 | 5876 | -1.2654 | 1.3579 |
| 36 | 152 | 36308 | 0.6493 | 0.8853 |

# Conclusions

The aim of our work was to introduce the geoadditive models and discuss their applicability in fields of statistics research that differ from their native research areas. In fact, nowadays, the geographical information is frequently available in many areas of observational sciences, and the use of specific techniques of spatial data analysis can improve our understanding of the studied phenomena.

The general review on the use of spatial information in statistics that we presented in Chapter 2, shows how in the last years the interest to the spatial data analysis is increased in every area of statistical research, from official statistics to demography to econometrics. Particular interest is given to the possible ways in which spatially referenced data can support local policy makers, especially in areas of social and economical interventions. Strictly connected with this subject are the spatial small area estimation methods that exploit the spatial information to “borrow strength” from the neighbour areas to produce more reliable estimations. As both the SAE models and the geoadditive models are formulated as linear mixed models, it seems an obvious choice to merge the two models in a geoadditive SAE model to exploit the spatial information and produce estimates at small area level.

However, the use of geostatistics methodologies for these applications is not always straightforward. If we use a geoadditive model to produce estimates of a parameter of interest for some geographical domains, like the small areas, then we need all the population units to be referenced at point locations. This requirement is not so easy to be accomplished, especially if we work with socio-economic data. Usually is much more easy to know the areas to which the units belong - like census districts, blocks, municipalities, etc.

If we have collected the required spatial information for the sampled units, from previous datasets or from specific surveys, then we can continue to use the geoadditive model with some approximations. The classic approach is to locate all the units with their corresponding area centre, however we decided to investigate a different approach, treating the lack of geographical information as a particular problem of *measurement error* and imposing a distribution for the locations inside each area.

We analyzed the performance of our approach implementing some MCMC analyses. In our experiments we consider first the case of a univariate missing variable, both with uniform distribution and with different non-uniform distribution inside each regions, modeled with a beta distribution with parameters that differ in each area. Then we implement the case of a bivariate uniform distribution. The results, presented in Chapter 3, show that, when the distribution hypothesis of the missing variable is correctly specified, our measurement error approach produce better estimates of the region level mean with respect to the same estimator under the classic approach. To observe the effect of the distribution hypothesis choice, we implement two experiments with real data. We can choose the distribution hypothesis by visual inspection of the empirical distribution of the variable for the sampled units. The results show that if the distribution hypothesis is approximatively correct - like in the California experiment - then the ME approach produces good estimations. On the other hand, if the hypothesis is completely untrue - like in the Albania experiment - then the ME approach and the classic approach are equivalent.

We want to highlight the fact that in our experiments we considered only two kind of univariate distributions, but the same model can be implemented with other distributions. The only demand is that the chosen distribution needs to be defined on a closed interval or otherwise truncated. On the other side, the beta distribution has the advantage to model many different shapes depending on the parameters value, including even the uniform distribution as a special case. Thus, we think that the beta distribution hypothesis could be the best choice. Referring to the bivariate case, until now we have implemented only the uniform case, but the next step should be the definition of other bivariate distributions, like a truncated normal, or the bivariate beta presented in Olkin and Ruixue Liu (2003) or some mixtures. A more complex issue is the definition of a bivariate distribution on an irregular polygon,like in the Albania experiment. We think that the idea proposed in the experiment of assigning a selection probability to each point inside the polygon, can provide a possible solution to such problem. On the other hand, it requires some priors information to decide how to assign the selection probability.

The final part of the thesis is devoted to the application of a geoadditive model in the field of poverty mapping at small area level. The geoadditive SAE model introduced is applied in order to estimate the district level mean of the household log per-capita consumption expenditure for the Republic of Albania. The results of our analysis shows that the consumption expenditure has both spatial dynamics and area specific effects. The map of the estimated district means presents an evident geographical pattern, with the higher values in the south and south-west of the country and the lower value in the mountainous area (north and north-east), confirming the results of previous applications on the same datasets presented in literature. Differently from other methods of analysis that exploit some spatial information, like the spatial SAE model, the geoadditive model produces not only the map of estimated mean values, but also a spatial interpolation of all the observation. Thus, with this model we can produce an estimated value in any point of the country.

When we produce estimates of a parameter of interest over some prespecified area, we should always consider the modifiable area unit problem (MAUP). With the geoadditive model we obtain a continuous surface estimation over the entire area, without define the area a priori, thus the MAUP can't occur. In our application the geoadditive model is associated with a SAE model, so in this case we need to define the areas before estimate the model, however the possible MAUP - if occurs - will be only related to the definition of the small area and not to the spatial interpolation of the studied phenomenon.

Finally, the results of the design-based simulation study, presented in Chapter 4, show that the geoadditive SAE EBLUP for the mean is quite stable, and that the performance of the robust MSE estimator of Salvati et al. (2008) is slightly better than the performance of the analytic MSE estimator of Opsomer et al. (2008). However, the two estimator are quite comparable.

In conclusion, exploiting the available geographical information, the use of a geoadditive model can improve the understanding of the phenomenon of interest describing and analyzing his spatial behaviour. Moreover, even if we don't know the exact location of all the population units, the geoadditive models could still be applied with good results using the mean estimator with the measurement error approach.

# Appendix A

# Outputs of the MCMC Experiments

## A.1 Univariate Uniform Model

`WinBUGS` **code**

```
for (i in 1:n1) {
  mu1[i] <- beta0 + betas*s1[i] + betat*t1[i] +
            inprod(u[],Z1[i,])
  y1[i] ~ dnorm(mu1[i], tauEps) }
for (i in 1:n2) {
  s2Hat[i] ~ dunif(aVals[i], bVals[i])
  y2Hat[i] <- beta0 + betas*s2Hat[i] + betat*t2[i] +
            inprod(u[],Z2Hat[i,])
  centrVals[i] <- (aVals[i] + bVals[i])/2
  y2Naive[i] <- beta0 + betas*centrVals[i] + betat*t2[i] +
            inprod(u[],Z2Naive[i,]) }
for (k in 1:numKnots) {
  for (i in 1:n2) {
    Z2Hat[i,k] <- (s2Hat[i]-knots[k])*step(s2Hat[i]-knots[k])
    Z2Naive[i,k] <- (centrVals[i]-knots[k])*step(centrVals[i]-
            knots[k]) }
  u[k] ~ dnorm(0.00000E+00,tauU) }
beta0 ~ dnorm(0.00000E+00,1.00000E-08)
betas ~ dnorm(0.00000E+00,1.00000E-08)
betat ~ dnorm(0.00000E+00,1.00000E-08)
tauU ~ dgamma(1.00000E-08,1.00000E-08)
tauEps ~ dgamma(1.00000E-08,1.00000E-08)
```

## Graphics

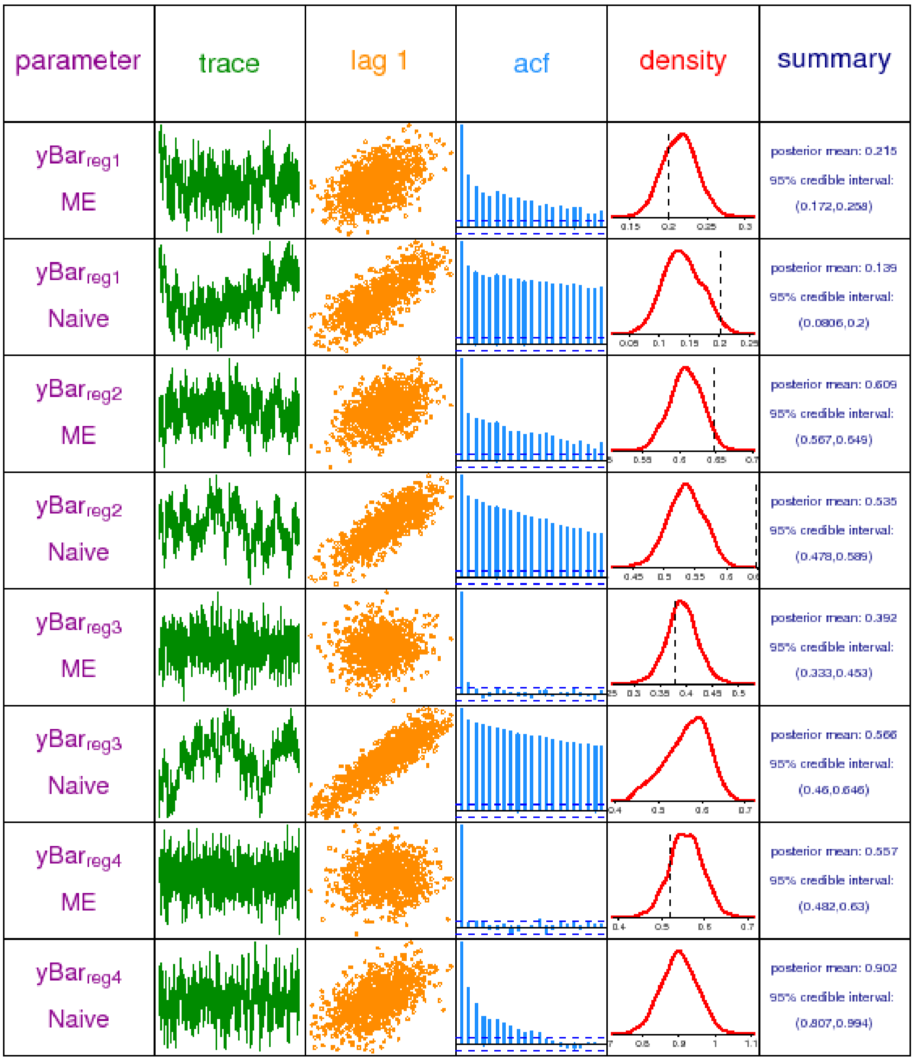

**Figure A.1:** Graphical summary of the MCMC output for region mean estimators for the univariate uniform model. The columns are: parameter, trace plot of MCMC sample, plot of sample against 1-lagged sample, sample autocorrelation function, kernel estimates posterior density and basic numerical summaries. The vertical dashed lines in the density plots correspond to the true values of the parameters according to the simulation set-up.

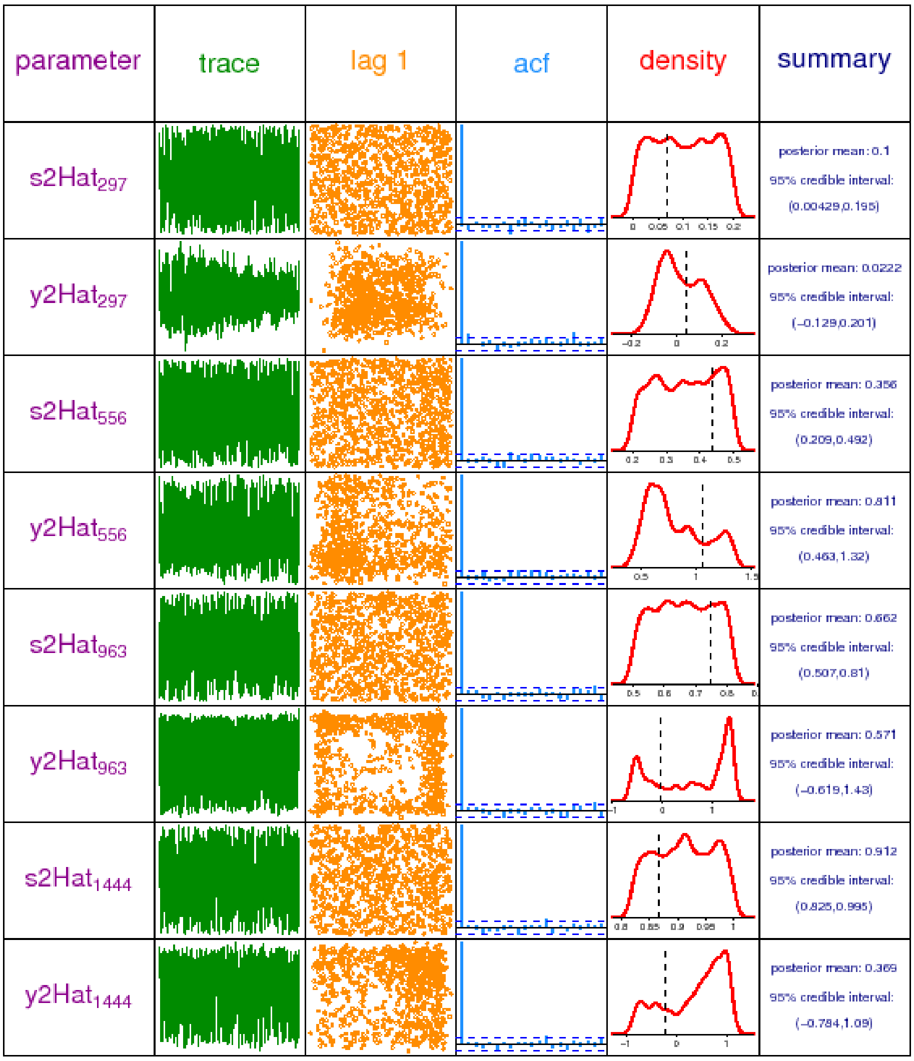

**Figure A.2:** Graphical summary of the MCMC output for four randomly chosen couples (once for each region) of non-measured and response variables for the univariate uniform model. The columns are: parameter, trace plot of MCMC sample, plot of sample against 1-lagged sample, sample autocorrelation function, kernel estimates posterior density and basic numerical summaries. The vertical dashed lines in the density plots correspond to the true values of the parameters according to the simulation set-up.

### A.1.1 Model without the dummy covariate

#### WinBUGS code - partial

```
for (i in 1:n1) {
   mu1[i] <- beta0 + betas*s1[i] + inprod(u[],Z1[i,])
   y1[i] ~ dnorm(mu1[i], tauEps) }
for (i in 1:n2) {
   s2Hat[i] ~ dunif(aVals[i], bVals[i])
   y2Hat[i] <- beta0 + betas*s2Hat[i] + inprod(u[],Z2Hat[i,])
   centrVals[i] <- (aVals[i] + bVals[i])/2
   y2Naive[i] <- beta0 + betas*centrVals[i] +
               inprod(u[],Z2Naive[i,]) }
```

#### Graphics

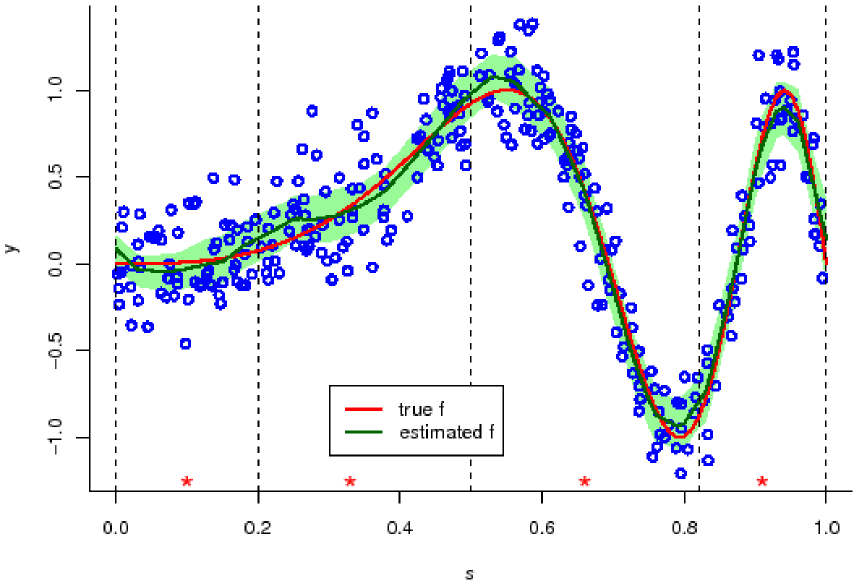

**Figure A.3:** MCMC-base fitting of the univariate uniform model without covariate $t$. The blue points are the sampled units and the pale green shaded region corresponds to pointwise 95% credible sets. The vertical dashed lines delimit the regions and the red stars indicate the centroids of each region.

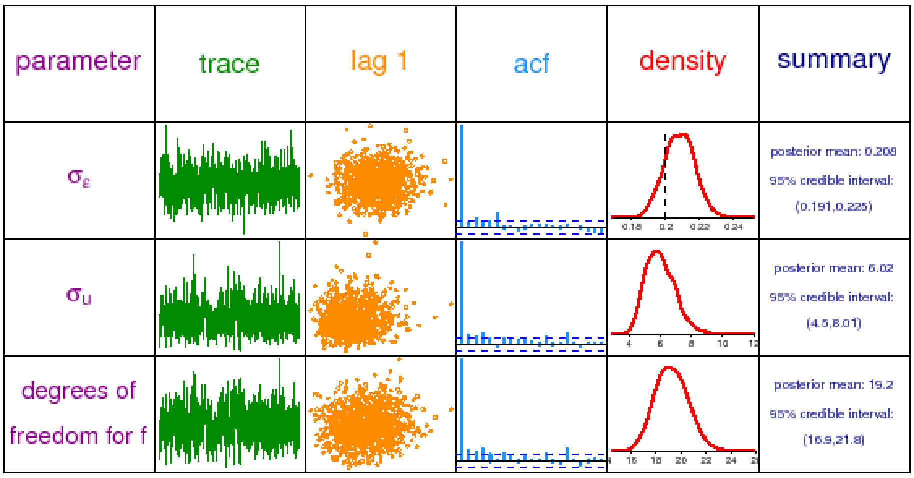

**Figure A.4:** Graphical summary of MCMC-based inference for the parameters of the univariate uniform model. The columns are: parameter, trace plot of MCMC sample, plot of sample against 1-lagged sample, sample autocorrelation function, kernel estimates posterior density and basic numerical summaries. The vertical dashed lines in the density plots, where present, correspond to the true values of the parameters according to the simulation set-up.

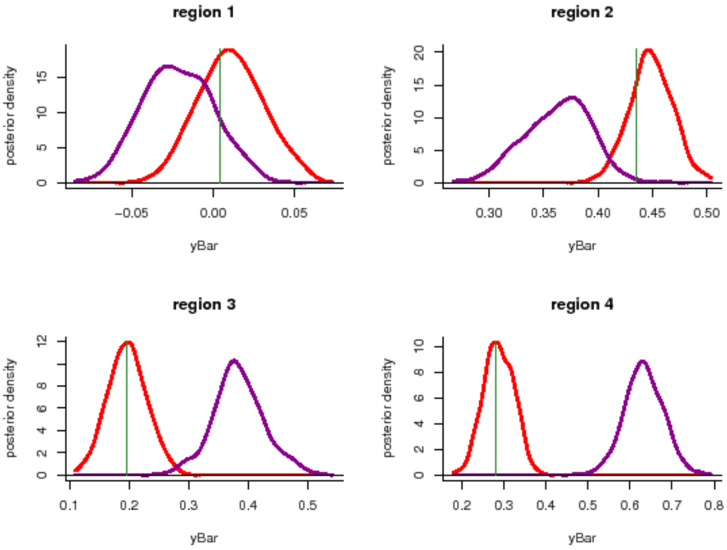

**Figure A.5:** Posterior density of the region mean estimator for the univariate uniform model without covariate $t$. The red lines correspond to the ME approach, the purple lines to the naive approach and the vertical green lines are the true mean values.

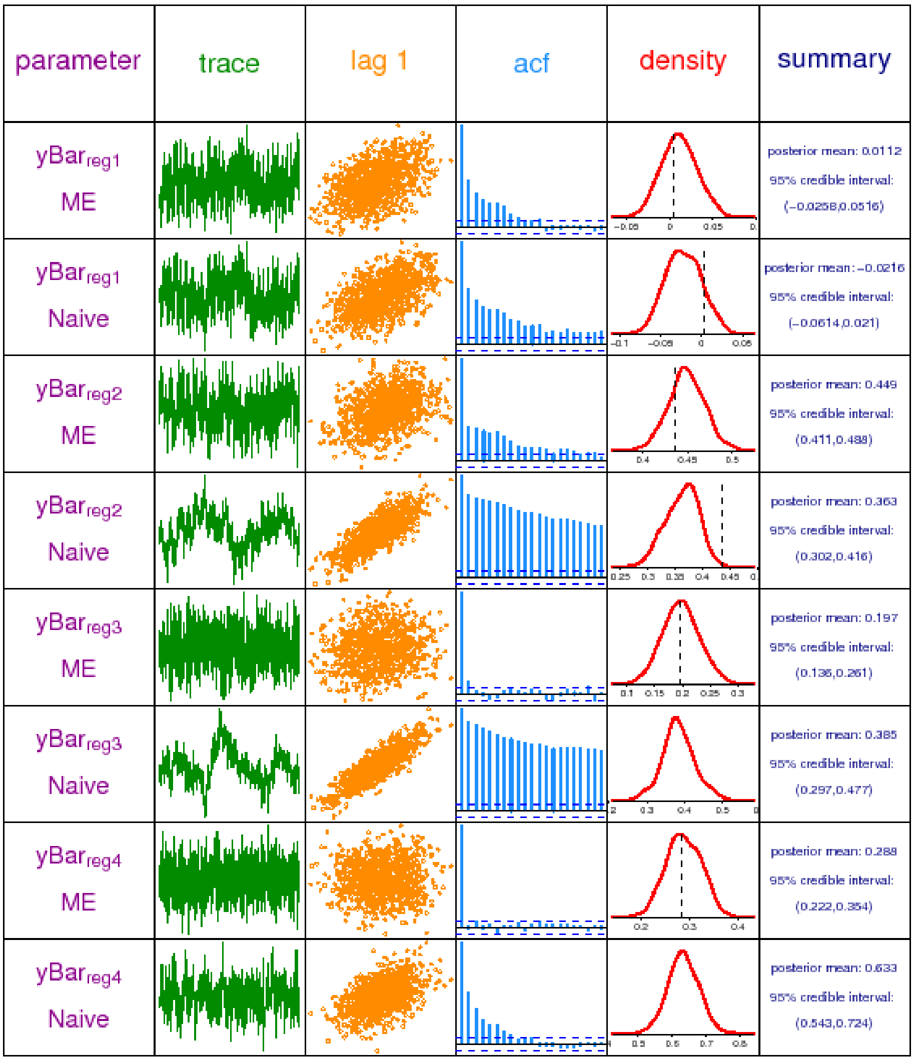

**Figure A.6:** Graphical summary of the MCMC output for region mean estimators for the univariate uniform model without covariate $t$. The columns are: parameter, trace plot of MCMC sample, plot of sample against 1-lagged sample, sample autocorrelation function, kernel estimates posterior density and basic numerical summaries. The vertical dashed lines in the density plots correspond to the true values of the parameters according to the simulation set-up.

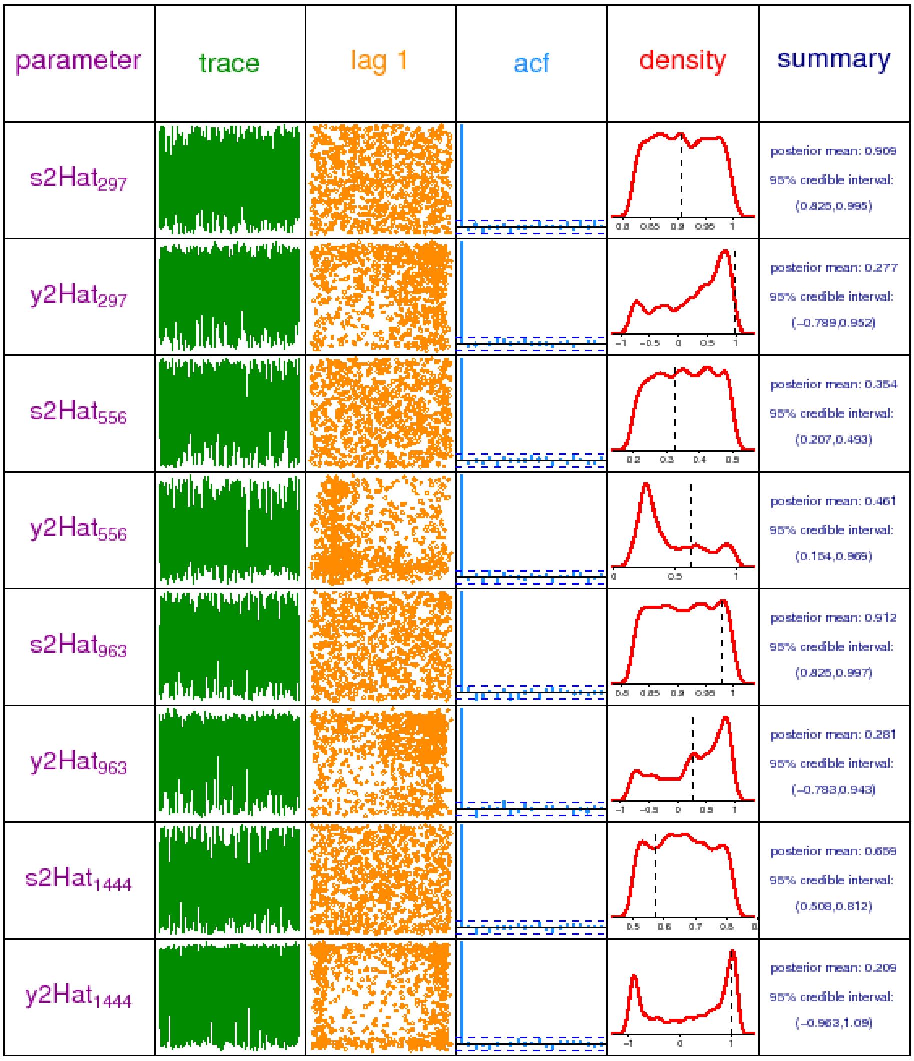

**Figure A.7:** Graphical summary of the MCMC output for four randomly chosen couples (once for each region) of non-measured and response variables for the univariate uniform model without covariate $t$. The columns are: parameter, trace plot of MCMC sample, plot of sample against 1-lagged sample, sample autocorrelation function, kernel estimates posterior density and basic numerical summaries. The vertical dashed lines in the density plots correspond to the true values of the parameters according to the simulation set-up.

## A.2 Univariate Beta Model

**WinBUGS code**

```
for (i in 1:n1) {
  ss1[i] <- (s1[i] - aVals1[i])/(bVals1[i] - aVals1[i])
  mu1[i] <- beta0 + betas*s1[i] + betat*t1[i] +
          inprod(u[],Z1[i,])
  y1[i] ∼ dnorm(mu1[i],tauEps) }
for (i in 1:n2) {
  s2Hat[i] <- (bVals2[i]-aVals2[i])*ss2Hat[i] + aVals2[i]
  y2Hat[i] <- beta0 + betas*s2Hat[i] + betat*t2[i] +
          inprod(u[],Z2Hat[i,])
  centrVals[i] <- (aVals[i] + bVals[i])/2
  y2Naive[i] <- beta0 + betas*centrVals[i] + betat*t2[i] +
          inprod(u[],Z2Naive[i,]) }
for (t in 1:numRegions) {
  for (i in (lim1[t] + 1):lim1[t + 1]) {
   ss1[i] ∼ dbeta(c[t], d[t]) }
  for (i in (lim2[t] + 1):lim2[t + 1]) {
   ss2Hat[i] ∼ dbeta(c[t], d[t]) }
  c[t] ∼ dunif(0.00000E+00, 100)
  d[t] ∼ dunif(0.00000E+00, 100) }
for (k in 1:numKnots) {
  for (i in 1:n2) {
   Z2Hat[i,k] <- (s2Hat[i]-knots[k])*step(s2Hat[i]-knots[k])
   Z2Naive[i,k] <- (centrVals[i]-knots[k])*step(centrVals[i]-
          knots[k]) }
  u[k] ∼ dnorm(0.00000E+00,tauU) }
beta0 ∼ dnorm(0.00000E+00,1.00000E-08)
betas ∼ dnorm(0.00000E+00,1.00000E-08)
betat ∼ dnorm(0.00000E+00,1.00000E-08)
tauU ∼ dgamma(1.00000E-08,1.00000E-08)
tauEps ∼ dgamma(1.00000E-08,1.00000E-08)
```

## Graphics

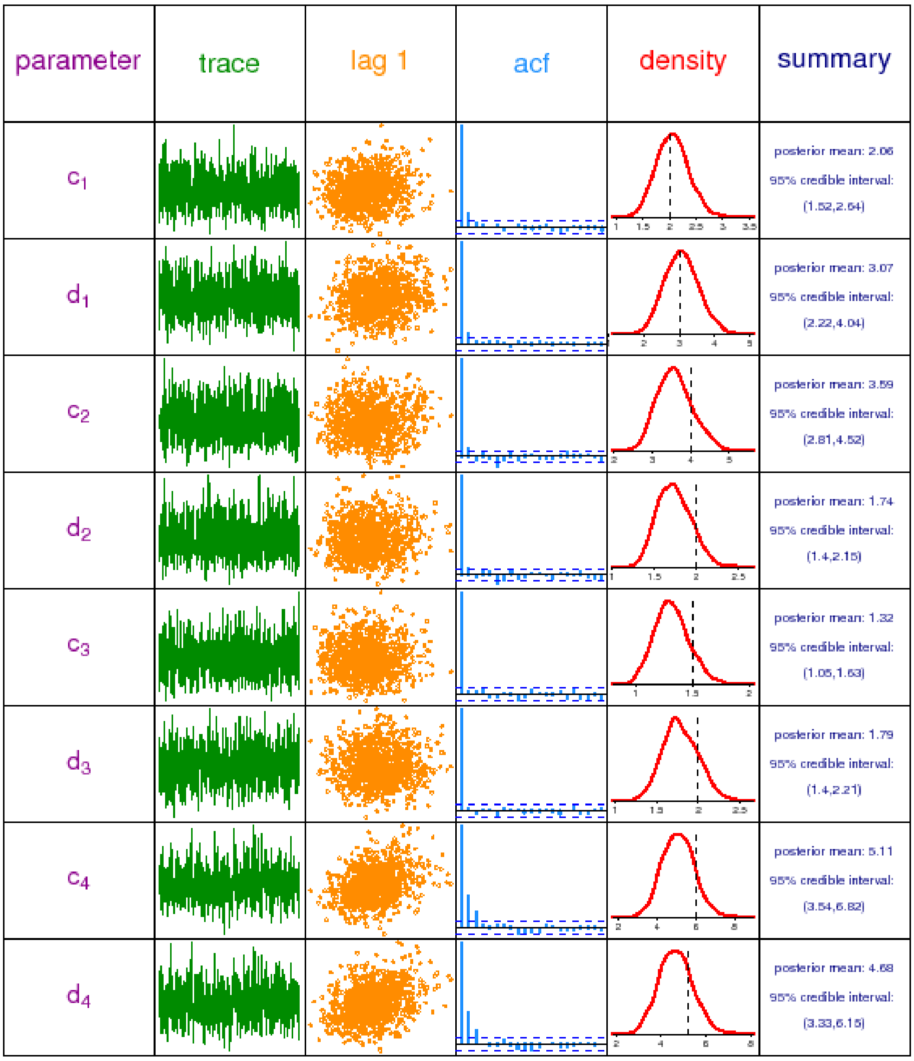

**Figure A.8:** Graphical summary of MCMC-based inference for the beta parameters of the univariate beta model. The columns are: parameter, trace plot of MCMC sample, plot of sample against 1-lagged sample, sample autocorrelation function, kernel estimates posterior density and basic numerical summaries. The vertical dashed lines in the density plots, where present, correspond to the true values of the parameters according to the simulation set-up.

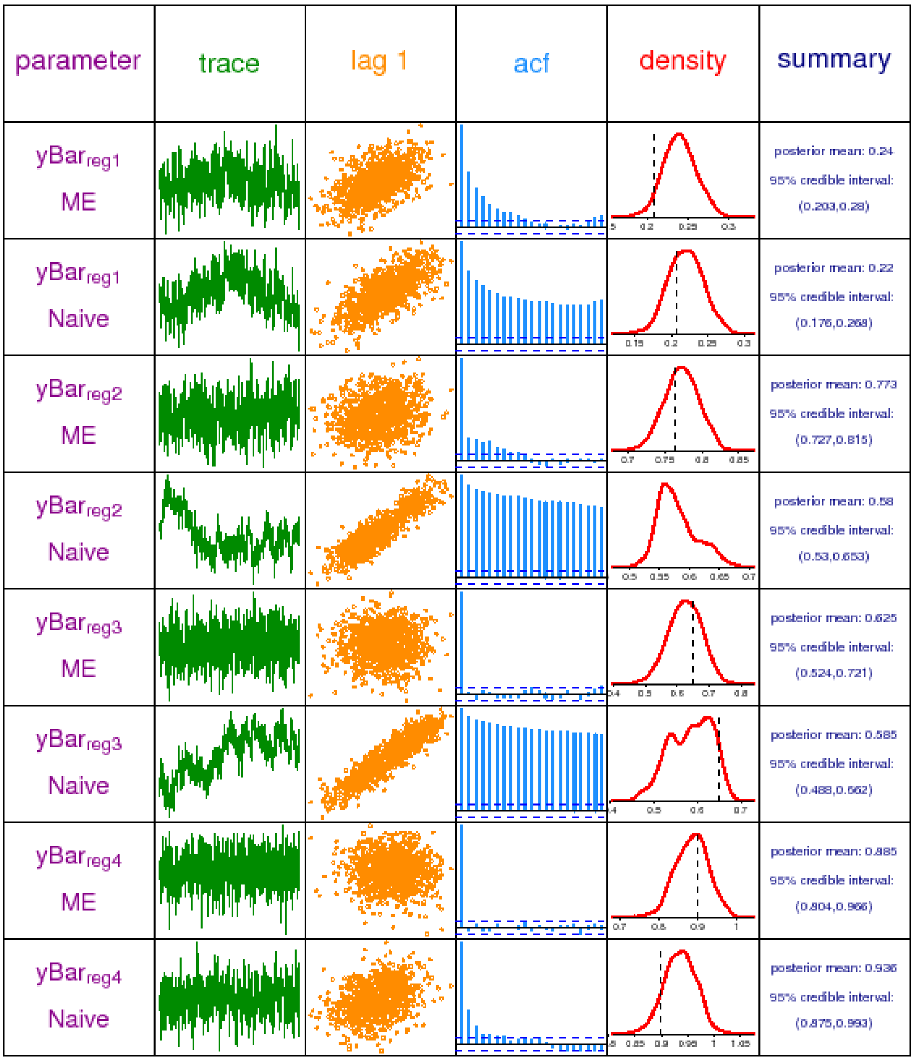

**Figure A.9:** Graphical summary of the MCMC output for region mean estimators for the univariate beta model. The columns are: parameter, trace plot of MCMC sample, plot of sample against 1-lagged sample, sample autocorrelation function, kernel estimates posterior density and basic numerical summaries. The vertical dashed lines in the density plots correspond to the true values of the parameters according to the simulation set-up.

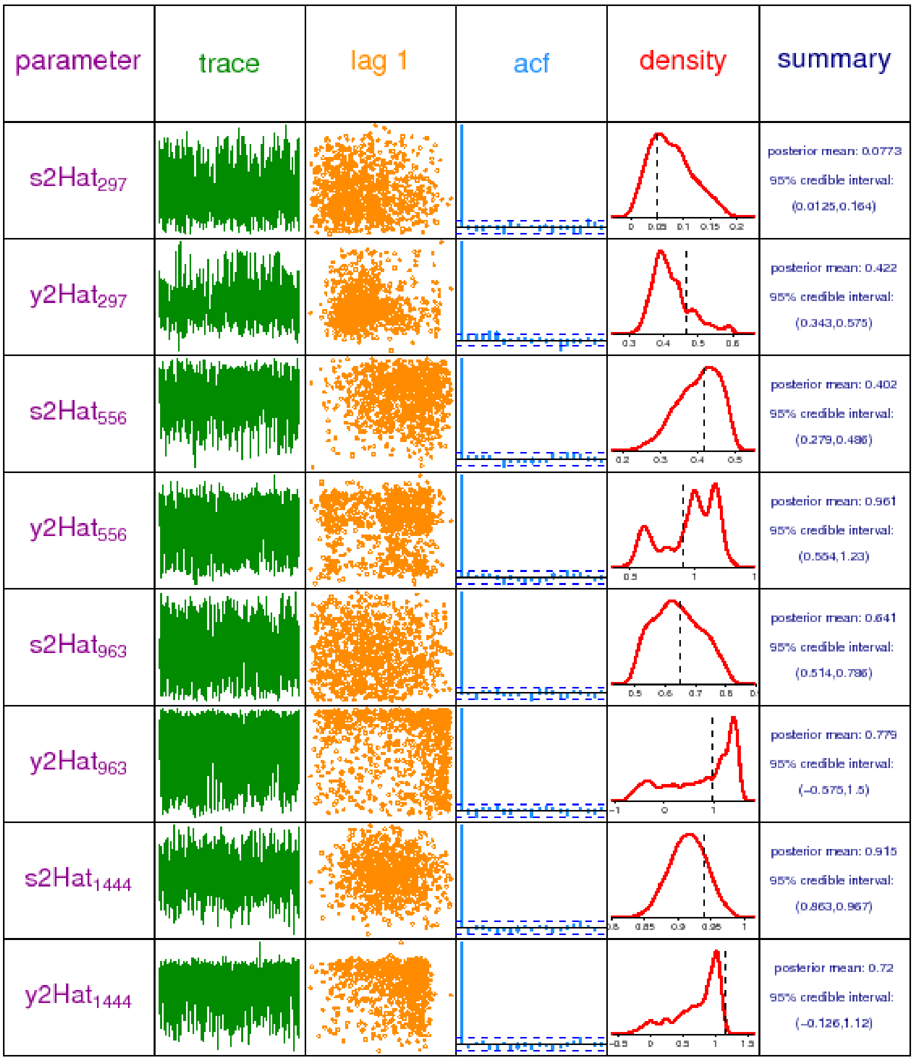

**Figure A.10:** Graphical summary of the MCMC output for four randomly chosen couples (once for each region) of non-measured and response variables for the univariate beta model. The columns are: parameter, trace plot of MCMC sample, plot of sample against 1-lagged sample, sample autocorrelation function, kernel estimates posterior density and basic numerical summaries. The vertical dashed lines in the density plots correspond to the true values of the parameters according to the simulation set-up.

## A.3 California Model

### WinBUGS code

```
for (i in 1:n1) {
  mu1[i] <- beta0+betax*x1[i]+betas*s1[i]+betat*t1[i]+inprod(Ux[],
            Zx1[i,])+inprod(Us[],Zs1[i,])+inprod(Ut[],Zt1[i,])
  y1[i] ~ dnorm(mu1[i],tauEps) }
for (i in 1:n2) {
  s2Hat[i] ~ dunif(aVals[i],bVals[i])
  y2Hat[i] <- beta0+betax*x2[i]+betas*s2Hat[i]+betat*t2[i]+
            inprod(Ux[],Zx2[i,])+inprod(Us[],Zs2Hat[i,])+
            inprod(Ut[],Zt2[i,])
  centrVals[i] <- (aVals[i] + bVals[i])/2
  y2Naive[i] <- beta0+betax*x2[i]+betas*centrVals[i]+betat*t2[i]+
            inprod(Ux[],Zx2[i,])+inprod(Us[],Zs2Naive[i,])+
            inprod(Ut[],Zt2[i,]) }
for (k in 1:numKnots) {
  for (i in 1:n2) {
   Zs2Hat[i,k] <- (s2Hat[i]-knotsS[k])*step(s2Hat[i]-knotsS[k])
   Zs2Naive[i,k] <- (centrVals[i]-knotsS[k])*
            step(centrVals[i]-knotsS[k]) }
  Ux[k] ~ dnorm(0.00000E+00,tauUx)
  Us[k] ~ dnorm(0.00000E+00,tauUs)
  Ut[k] ~ dnorm(0.00000E+00,tauUt) }
beta0 ~ dnorm(0.00000E+00,1.00000E-08)
betax ~ dnorm(0.00000E+00,1.00000E-08)
betat ~ dnorm(0.00000E+00,1.00000E-08)
betas ~ dnorm(0.00000E+00,1.00000E-08)
tauUx ~ dgamma(1.00000E-08,1.00000E-08)
tauUs ~ dgamma(1.00000E-08,1.00000E-08)
tauUt ~ dgamma(1.00000E-08,1.00000E-08)
tauEps ~ dgamma(1.00000E-08,1.00000E-08)
```

## Graphics

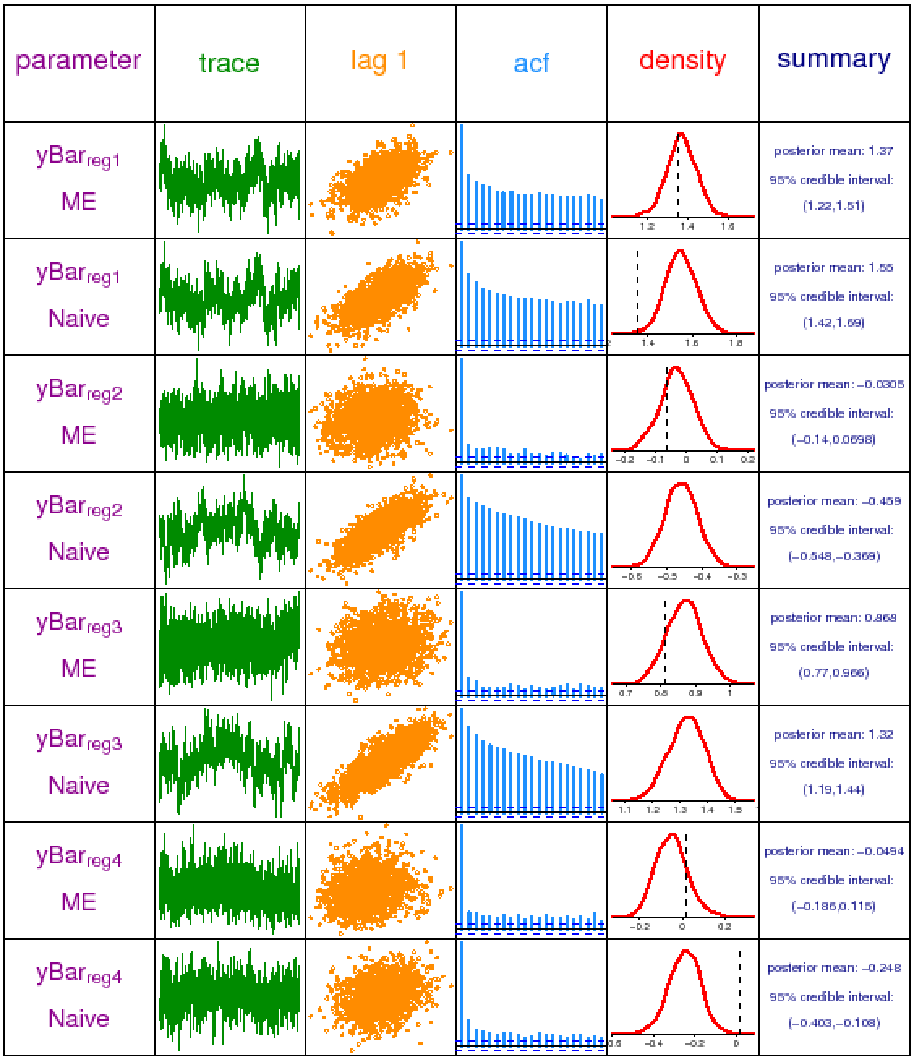

**Figure A.11:** Graphical summary of the MCMC output for region mean estimators for the univariate beta model. The columns are: parameter, trace plot of MCMC sample, plot of sample against 1-lagged sample, sample autocorrelation function, kernel estimates posterior density and basic numerical summaries. The vertical dashed lines in the density plots correspond to the true values of the parameters according to the simulation set-up.

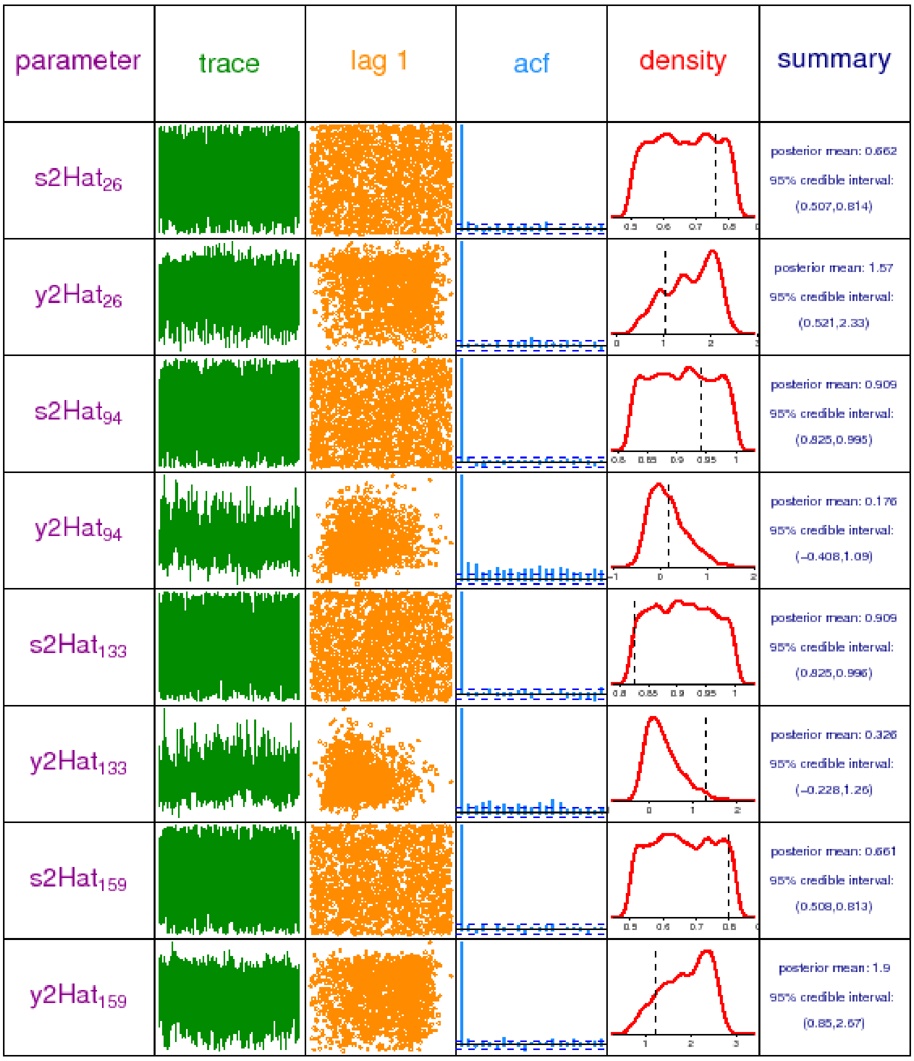

**Figure A.12:** Graphical summary of the MCMC output for four randomly chosen couples (once for each region) of non-measured and response variables for the univariate beta model. The columns are: parameter, trace plot of MCMC sample, plot of sample against 1-lagged sample, sample autocorrelation function, kernel estimates posterior density and basic numerical summaries. The vertical dashed lines in the density plots correspond to the true values of the parameters according to the simulation set-up.

## A.4 Bivariate Uniform Model

`WinBUGS` **code**

```
for (i in 1:n1) {
  mu1[i] <- beta0+betat*t1[i]+beta1s*s11[i]+beta2s*s12[i]+
            inprod(Us[],Zs1[i,])
  y1[i] ~ dnorm(mu1[i],tauEps) }
for (i in 1:n2) {
  s2Hat1[i] ~ dunif(aVals1[i],bVals1[i])
  s2Hat2[i] ~ dunif(aVals2[i],bVals2[i])
  y2Hat[i] <- beta0+betat*t2[i]+beta1s*s2Hat1[i]+beta2s*s2Hat2[i]+
            inprod(Us[],Zs2Hat[i,])
  centrVals1[i] <- (aVals1[i] + bVals1[i])/2
  centrVals2[i] <- (aVals2[i] + bVals2[i])/2
  y2Naive[i] <- beta0+betat*t2[i]+beta1s*centrVals1[i]+
            beta2s*centrVals2[i]+inprod(Us[],Zs2Naive[i,]) }
for (k in 1:numKnots) {
  for (i in 1:n2) {
    distHat[i,k] <- sqrt(pow(s2Hat1[i]-knots1S[k],2) +
            pow(s2Hat2[i]-knots2S[k],2))
    Zs2Hat[i,k] <- pow(abs(distHat[i,k]),2)*log(abs(distHat[i,k]))
    distNaive[i,k] <- sqrt(pow(centrVals1[i]-knots1S[k],2) +
            pow(centrVals2[i]-knots2S[k],2))
    Zs2Naive[i,k] <- pow(abs(distNaive[i,k]),2)*
            log(abs(distNaive[i,k])) }
  for (h in 1:numKnots) {
    MAT[h,k] <- tauUs * TAUmat[h,k] } }
Us[1:numKnots] ~ dmnorm(meanUs[], MAT[,])
beta0 ~ dnorm(0.00000E+00, 1.00000E-08)
betat ~ dnorm(0.00000E+00, 1.00000E-08)
beta1s ~ dnorm(0.00000E+00, 1.00000E-08)
beta2s ~ dnorm(0.00000E+00, 1.00000E-08)
tauUs ~ dgamma(1.00000E-08, 1.00000E-08)
tauEps ~ dgamma(1.00000E-08, 1.00000E-08)
```

## Graphics

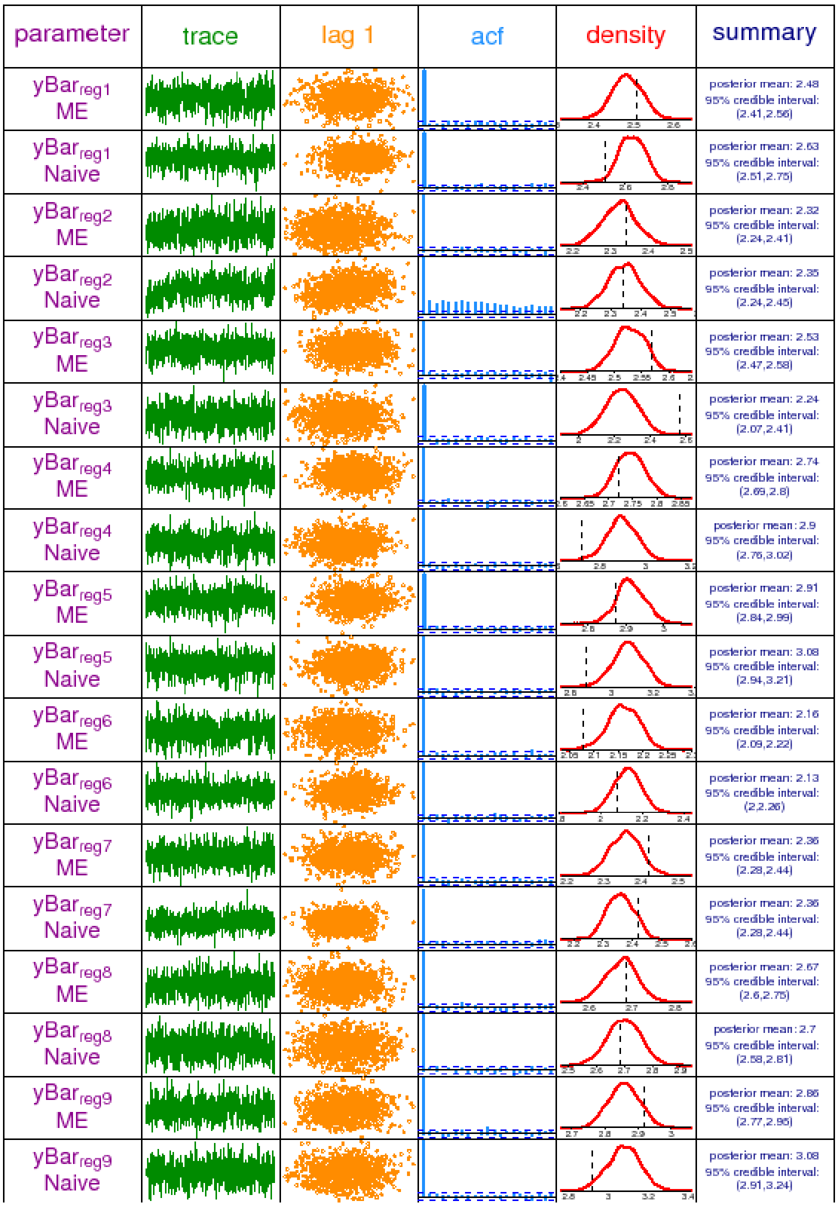

**Figure A.13:** Graphical summary of the MCMC output for region mean estimators for the bivariate uniform model. The columns are: parameter, trace plot of MCMC sample, plot of sample against 1-lagged sample, sample autocorrelation function, kernel estimates posterior density and basic numerical summaries. The vertical dashed lines in the density plots correspond to the true values of the parameters according to the simulation set-up.

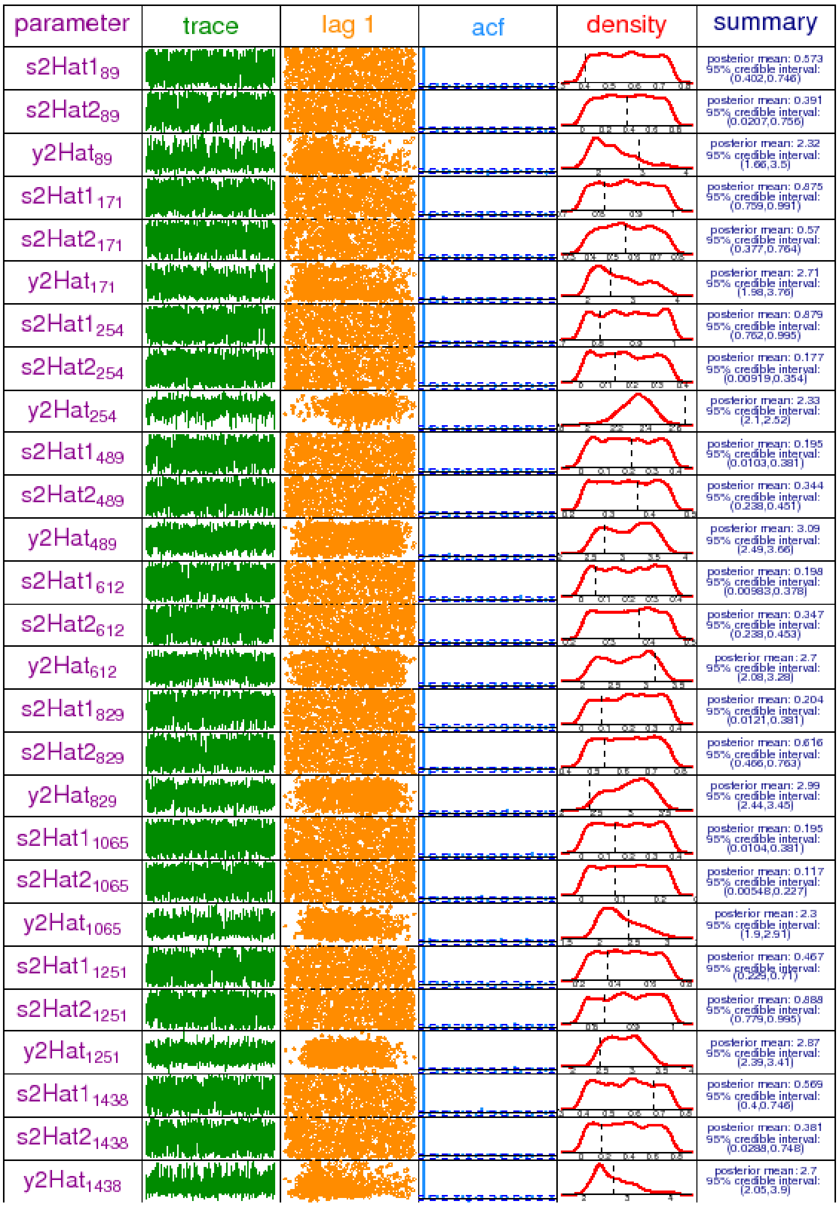

**Figure A.14:** Graphical summary of the MCMC output for four randomly chosen couples (once for each region) of non-measured and response variables for the bivariate uniform model. The columns are: parameter, trace plot of MCMC sample, plot of sample against 1-lagged sample, sample autocorrelation function, kernel estimates posterior density and basic numerical summaries. The vertical dashed lines in the density plots correspond to the true values of the parameters according to the simulation set-up.

## A.5 Albania Model

**`WinBUGS` code**

```
for (i in 1:n1) {
  mu1[i] <- beta0+beta1s*s1[i,1]+beta2s*s1[i,2]+
            inprod(Us[],Zs1[i,])
  y1[i] ~ dnorm(mu1[i],tauEps) }
for (i in 1:n2) {
  indice[i] ~ dcat(pSample[(baseSample[i]+1):(baseSample[i]+
            nSample[i])])
  ind[i] <- baseSample[i] + indice[i]
  s2Hat1[i] <- disPoints[ind[i],1]
  s2Hat2[i] <- disPoints[ind[i],2]
  y2Hat[i] <- beta0+beta1s*s2Hat1[i]+beta2s*s2Hat2[i]+
            inprod(Us[],Zs2Hat[i,])
  y2Naive[i] <- beta0+beta1s*centrVals[i,1]+beta2s*centrVals[i,2]+
            inprod(Us[],Zs2Naive[i,]) }
for (k in 1:numKnots) {
  for (i in 1:n2) {
   distHat[i,k] <- sqrt(pow(s2Hat1[i] - knotsS[k,1],2) +
            pow(s2Hat2[i] - knotsS[k,2],2))
   Zs2Hat[i,k] <- pow(abs(distHat[i,k]),2)*log(abs(distHat[i,k]))
   distNaive[i,k] <- sqrt(pow(centrVals[i,1] - knotsS[k,1],2) +
            pow(centrVals[i,2] - knotsS[k,2],2))
   Zs2Naive[i,k] <- pow(abs(distNaive[i,k]),2)*
            log(abs(distNaive[i,k])) }
  for (h in 1:numKnots) {
   MAT[h,k] <- tauUs * TAUmat[h,k] } }
Us[1:numKnots] ~ dmnorm(meanUs[],MAT[,])
beta0 ~ dnorm(0.00000E+00,1.00000E-08)
beta1s ~ dnorm(0.00000E+00,1.00000E-08)
beta2s ~ dnorm(0.00000E+00,1.00000E-08)
tauUs ~ dgamma(1.00000E-08,1.00000E-08)
tauEps ~ dgamma(1.00000E-08,1.00000E-08)
```

## Graphics

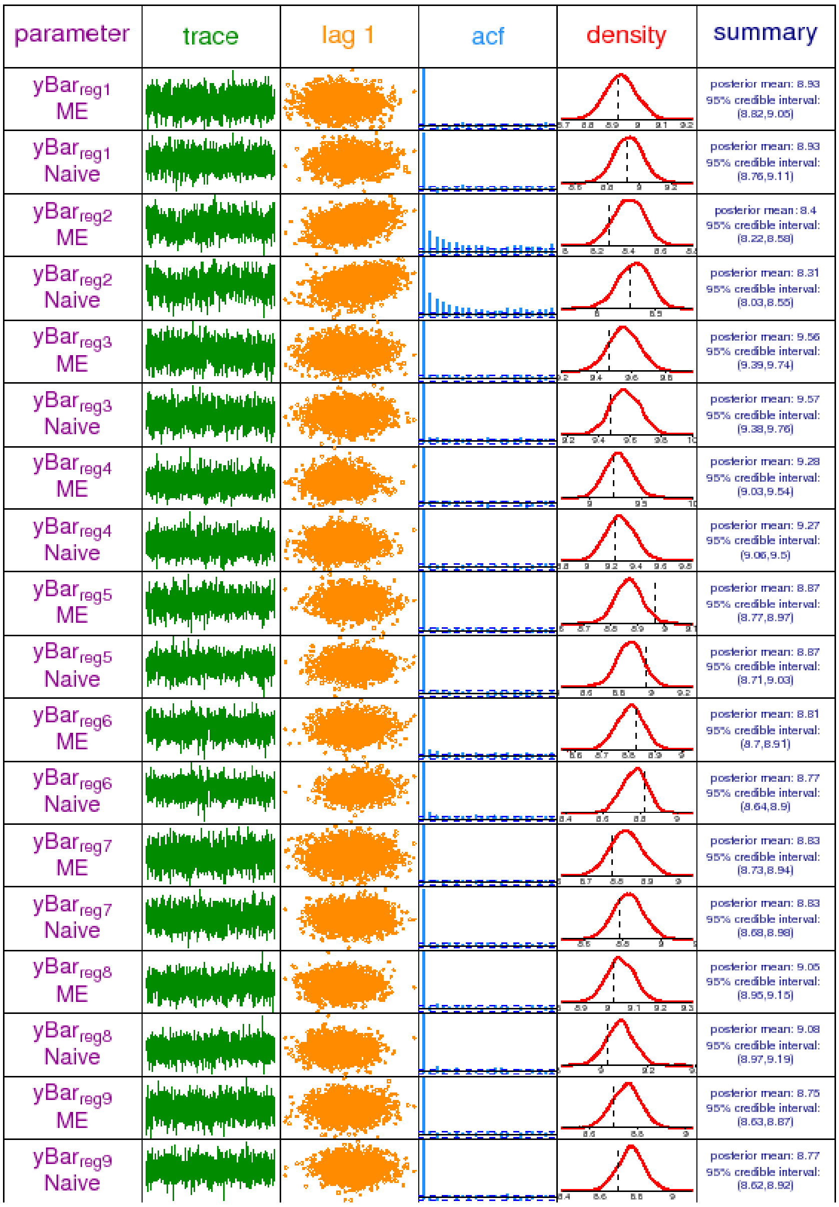

**Figure A.15:** Graphical summary of the MCMC output for region mean estimators for the Albania model. The columns are: parameter, trace plot of MCMC sample, plot of sample against 1-lagged sample, sample autocorrelation function, kernel estimates posterior density and basic numerical summaries. The vertical dashed lines in the density plots correspond to the true values of the parameters according to the simulation set-up.

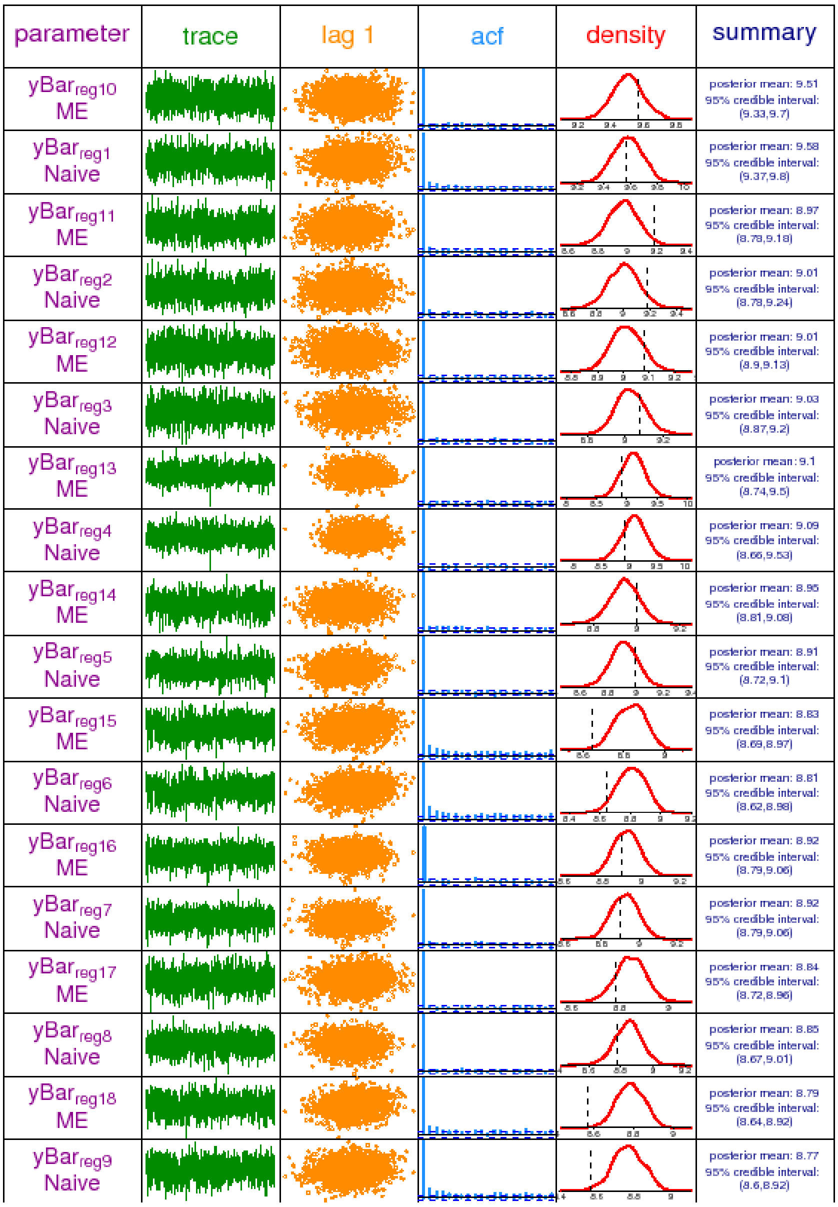

**Figure A.15:** CONTINUE - Graphical summary of the MCMC output for region mean estimators for the Albania model. The columns are: parameter, trace plot of MCMC sample, plot of sample against 1-lagged sample, sample autocorrelation function, kernel estimates posterior density and basic numerical summaries. The vertical dashed lines in the density plots correspond to the true values of the parameters according to the simulation set-up.

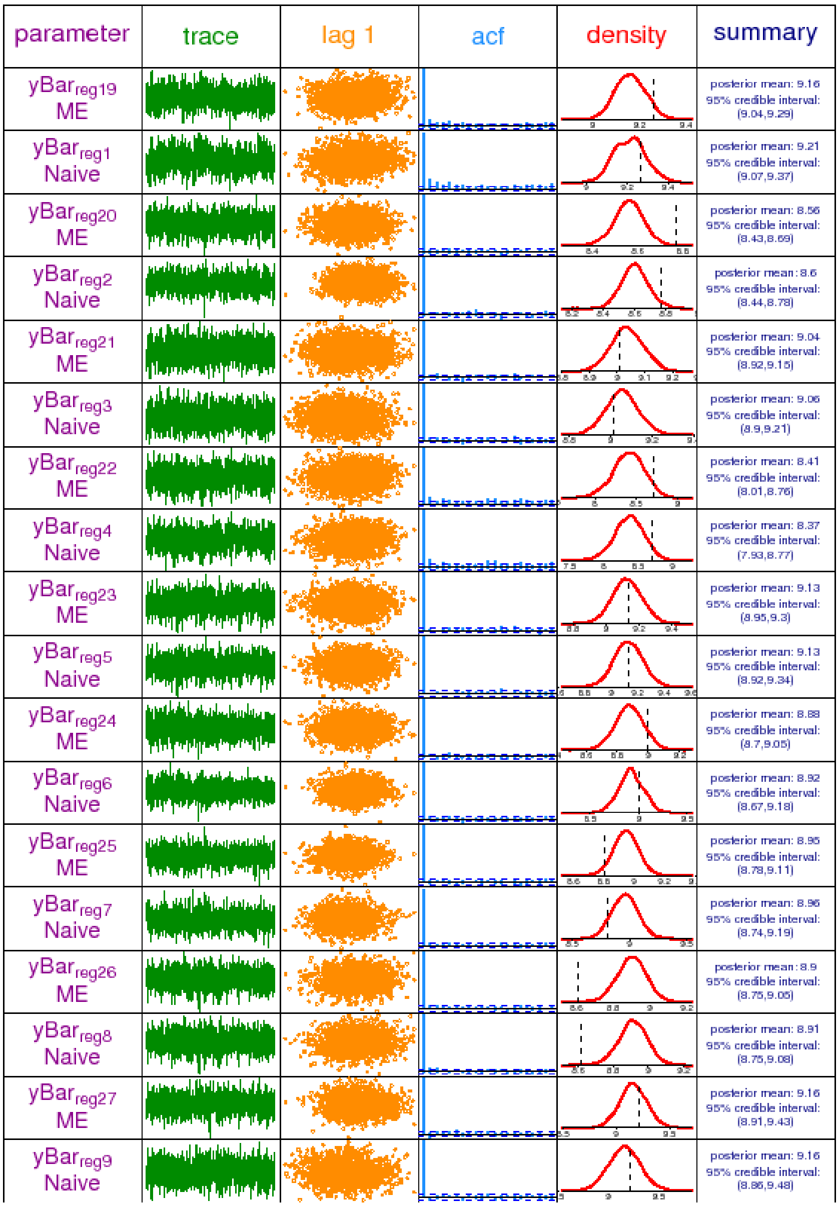

**Figure A.15:** CONTINUE - Graphical summary of the MCMC output for region mean estimators for the Albania model. The columns are: parameter, trace plot of MCMC sample, plot of sample against 1-lagged sample, sample autocorrelation function, kernel estimates posterior density and basic numerical summaries. The vertical dashed lines in the density plots correspond to the true values of the parameters according to the simulation set-up.

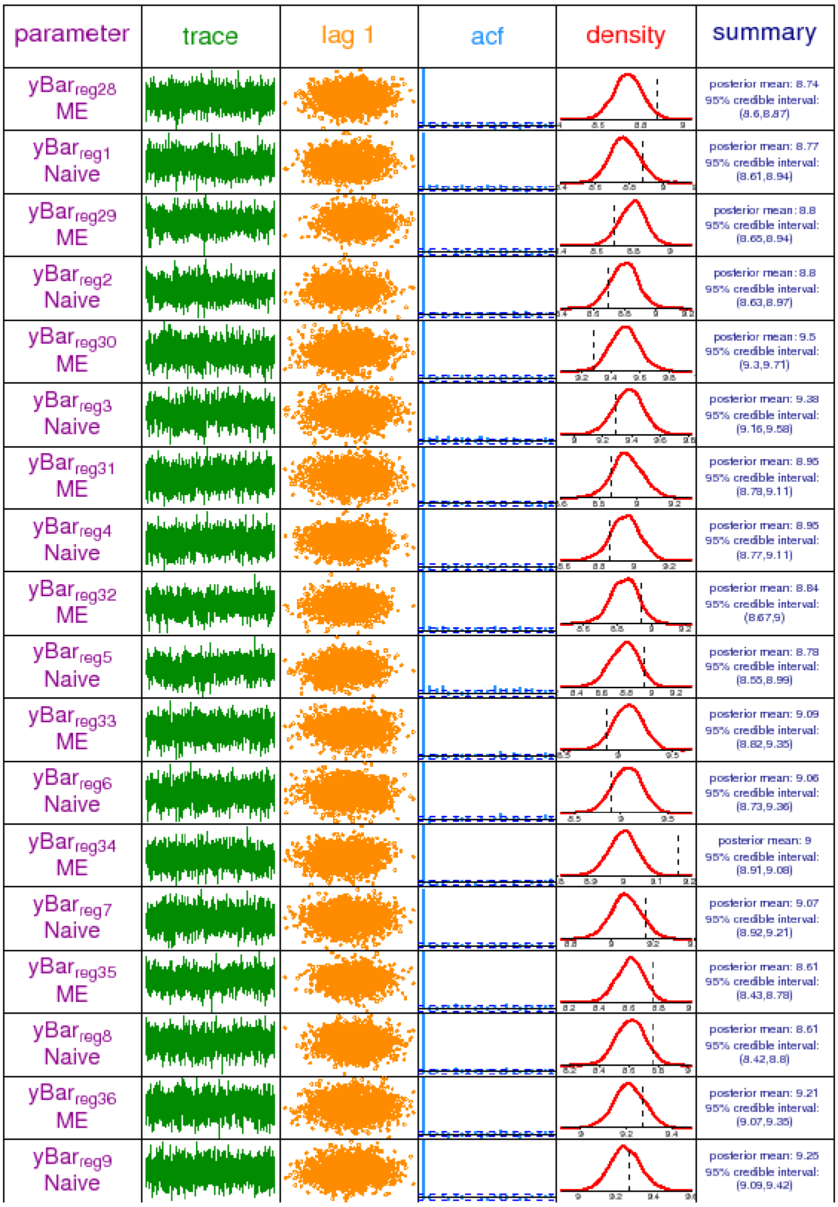

**Figure A.15:** CONTINUE - Graphical summary of the MCMC output for region mean estimators for the Albania model. The columns are: parameter, trace plot of MCMC sample, plot of sample against 1-lagged sample, sample autocorrelation function, kernel estimates posterior density and basic numerical summaries. The vertical dashed lines in the density plots correspond to the true values of the parameters according to the simulation set-up.

# Bibliography

Anselin, L. (1988), *Spatial Econometrics: Methods and Models*, Kluwer Academic, Dordrecht.

Arbia, G. (2006), *Spatial Econometrics: Statistical Foundation and Applications to Regional Convergence*, Springer, Berlin.

Arbia, G. and Baltagi, B. H., eds (2008), *Spatial Econometrics: Methods And Applications*, Physica-verlag, Heidelberg.

Bailey, T. C. and Gatrell, A. C. (1995), *Intereactive Spatial Data Analysis*, Longman, Harlow.

Banerjee, S., Carlin, B. P. and Gelfand, A. E. (2004), *Hierarchical modelling and analysis for spatial data*, Chapman & Hall/CRC, Boca Raton, Florida.

Battese, G. E., Harter, R. M. and Fuller, W. A. (1988), 'An Error Component Model for Prediction of County Crop Areas Using Survey and Satelite Data', *Journal of the American Statistical Association* **83**, 28–36.

Berkson, J. (1950), 'Are There Two Regressions?', *Journal of the American Statistical Association* **45**, 164–180.

Betti, G. and Ballini, F. and Neri, L. (2003), *Poverty and Inequality Mapping in Albania, Final Report to the World Bank.*

Boffi, M. (2004), *Scienza dell'Informazione Geografica. Introduzione ai GIS*, Zanichelli, Bologna.

Bollinger, C. R. (1998), 'Measurement Error in the Current Population Survey: A Nonparametric Look', *Journal of Labor Economics* **16**, 576–594.

Breiman, L. and Friedman, J. (1985), 'Estimating optimal transformations for multiple regression and correlation (with discussion)', *Journal of the American Statistical Association* **80**, 580–619.

Brumback, B. A., Ruppert, D. and Wand, M. P. (1999), 'Comment on "Variable Selection and Function Estimation in Additive Nonparametric Regression Using a Data-Based Prior" by T.S. Shively, R. Kohn and S. Wood', *Journal of the American Statistical Association* **94**, 794–797.

Buonaccorsi, J. P. (1995), 'Prediction in the presence of measurement error: General discussion and an example predicting defoliation', *Biometrics* **51**, 1562–1569.

Calzaroni, M. (2008), Le fonti amministrative nei processi e nei prodotti della statistica ufficiale, *in* 'Atti della Nona Conferenza Nazionale di Statistica, 15-16 dicembre 2008', Istat.

Carroll, R., Eltinge, J. L. and Ruppert, D. (1993), 'Robust linear regression in replicated measurement error models', *Statistics & Probability Letters* **16**, 169–175.

Carroll, R. J., Ruppert, D., Stefanski, L. A. and Crainiceanu, C. M. (2006), *Measurement Error in Nonlinear Models. A Modern Perspective (second edition)*, Chapman & Hall/CRC, Boca Raton, Florida.

Carroll, R., Ruppert, D., Tosteson, T., Crainiceanu, C. and Karagas, M. (2004), 'Nonparametric regression and instrumental variables', *Journal of the American Statistical Association* **99**, 736–750.

Chambers, R., Chandra, H. and Tzavidis, N. (2007), *On robust mean squared error estimation for linear predictors fro domains*, CCSR Working paper 2007-10, Cathie Marsh Centre for Census ans Survey Research, University of Manchester.

Chesher, A. and Schluter, C. (2002), 'Welfare Measurement and Measurement Error', *The Review of Economic Studies* **69**, 357–378.

Crainiceanu, C., Ruppert, D. and Wand, M. P. (2005), 'Bayesian analysis for penalized spline regression using WinBUGS', *Journal of Statistical Software* **14**(14).

Cressie, N. (1991), Small-area prediction of undercount using the general linear model, *in* 'Proceedings of Statistic Symposium 90: Measurement and Improvement of Data Quality', Statistics Canada, Ottawa, pp. 93–105.

Cressie, N. (1993), *Statistics for Spatial Data (revised edition)*, Waley, New York.

Cromley, R. G. (1996), ‘A comparison of optimal classification strategies for choropleth displays of spatially aggregated data’, *International Journal of Geographical Information Systems* **10**, 405–424.

de Castro, M. C. (2007), ‘Spatial Demography: An Opportunity to Improve Policy Making at Diverse Decision Levels’, *Population Research and Policy Review* **26**, 477–509.

de Gruijter, J., Brus, D., Bierkens, M. and Knotters, M. (2006), *Sampling for Natural Resource Monitoring*, Springer, The Netherlands.

Eilers, P. H. C. and Marx, B. D. (1996), ‘Flexible smoothing with B-splines and penalties (with discussion)’, *Statistical Science* **11**, 89–121.

Elliott, P., Wakefield, J. C., Best, N. G. and Briggs, D. J. (2000), Spatial Epidemiology: Methods and Applications, *in* P. Elliott, J. C. Wakefield, N. G. Best and D. J. Briggs, eds, ‘Spatial Epidemiology: Methods and Applications’, Oxford University Press, Oxford, pp. 3–14.

Fay, R. E. and Herriot, R. A. (1979), ‘Estimation of Income from Small Places: An Application of James-Stein Procedures to Census Data’, *Journal of the American Statistical Association* **74**, 269–277.

French, J. L., Kammann, E. E. and Wand, M. P. (2001), ‘Comment on ”Semiparametric Nonlinear Mixed-Effects Models and Their Applications” by C. Ke and Y. Wang’, *Journal of the American Statistical Association* **96**, 1285–1288.

Friedman, J. H. and Stuetzle, W. (1981), ‘Projection pursuit regression’, *Journal of the American Statistical Association* **76**, 817–823.

Fuller, W. A. (1987), *Measurement Error Models*, John Wiley & Sons, New York.

Ganguli, B., Staudenmayer, J. and Wand, M. P. (2005), ‘Additive models with predictors subject to measurement error’, *Australia and New Zealand Journal of Statistics* **47**, 193–202.

Goodchild, M. F. (1991), Issues of quality and uncertainty, *in* J. C. Muller, ed., ‘Advances in Cartography’, Elsevier Science, New York, pp. 113–139.

Goodchild, M. F. (1999), Measurement-based GIS, *in* W. Shi, M. F. Goodchild and P. F. Fisher, eds, ‘Proceedings of the International Symposium on Spatial Data Quality 99’, Hong Kong Polytechnic University, Hong Kong, pp. 1–9.

Goodchild, M. F. and Gopal, S., eds (1989), *Accuracy of spatial databases*, Taylor & Francis, London.

Goodchild, M. F. and Janelle, D. G., eds (2004), *Spatially integrated social science*, Oxford University Press, Oxford, England.

Goovaerts, P. (2009), 'Combining area-based and individual-level data in the geostatistical mapping of late-stage cancer incidence', *Spatial and Spatio-temporal Epidemiology* **1**, 61–71.

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

Grosh, M. and Glewwe, P. (2000), A Guide to Living Standards Measurement Study Surveys, *in* 'Household Accounting: Experience and Concepts in Compilation', Statistics Division. Department of Economic and Social Affairs. The United Nations.

Gryparis, A., Coull, B. A., Schwartz, J. and Suh, H. H. (2007), 'Semiparametric latent variable regression models for spatiotemporal modelling of mobile source particles in the greater Boston area', *Applied Statistics* **56**, 183–209.

Gryparis, A., Paciorek, C. J., Zeka, A., Schwartz, J. and Coull, B. A. (2009), 'Measurement error caused by spatial misalignment in environmental epidemiology', *Biostatistics* **10**, 258–274.

Haining, R. (2003), *Spatial Data Analysis: Theory and practice*, Cambridge University Press, Cambridge.

Hastie, T. J. (1996), 'Pseudosplines', *Journal of the Royal Statistical Society, Series B* **58**, 379–396.

Hastie, T. J. and Tibshirani, R. (1990), *Generalized Additive Models*, Chapman & Hall, London.

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

Heuvelink, G. B. M. (1998), *Error propagation in environmental modelling with GIS*, Taylor & Francis, London.

Holt, D., Steel, D. G. and Tranmer, M. (1996), 'Area homogeneity and the modifiable areal unit problem', *Geographical Systems* **3**, 181–200.

Jiang, J. and Lahiri, P. (2006), 'Mixed Model Prediction and Small Area Estimation (with discussion)', *Test* **15**, 1–96.

Kammann, E. E. and Wand, M. P. (2003), 'Geoadditive Models', *Applied Statistics* **52**, 1–18.

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

Kelly, C. and Rice, J. (1990), 'Monotone smoothing with application to dose-response curves and the assessment of synergism', *Biometrics* **46**, 1071–1085.

Lawson, A. B. (2009), *Bayesian Disease Mapping. Hierarchical Modeling in Spatial Epidemiology*, Chapman & Hall/CRC, Boca Raton, Florida.

Lawson, A. B. and Cressie, N. (2000), Spatial Statistical Methods for Environmental Epidemiology, *in* P. K. Sen and C. R. Rao, eds, 'Handbook of Statistics', Vol. 18, Elsevier, The Netherlands, pp. 357–396.

Leung, Y., Ma, J.-H. and Goodchild, M. F. (2004), 'A general framework for error analysis in measurement-based GIS: Part 1–4', *Journal of Geographical Systems* **6**, 325–428.

Leung, Y. and Yan, J. P. (1998), 'A locational error model for spatial features', *International Journal of Geographical Information Science* **12**, 607–620.

Ligges, U., Thomas, A., Spiegelhalter, D., Best, N., Lunn, D., Rice, K. and Sturtz, S. (2009), `BRugs 0.5-3`. R package.
**URL:** *cran.r-project.org*

Lunn, D., Thomas, A., Best, N. and Spiegelhalter, D. (2000), 'WinBUGS – a Bayesian modelling framework: concepts, structure, and extensibility', *Statistics and Computing* **10**, 325–337.

Madsen, L., Ruppert, D. and Altman, N. S. (2008), 'Regression with spatially misaligned datas', *Environmetrics* **19**, 453–467.

Marley, J. and Wand, M. P. (2010), 'Non-Standard Semiparametric Regression via BRugs', *Journal of Statistical Software* . Forthcoming.

McCulloch, C. E. and Searle, S. R. (2001), *Generalized, Linear, and Mixed Models*, Wiley, New York.

Mowrer, H. T. and Congalton, R. G., eds (2000), *Quantifying spatial uncertainty in natural resources: Theory and applications for GIS and remote sensing*, Ann Arbor Press, Chelsea.

Neri, L., Ballini, F. and Betti, G. (2005), 'Poverty and inequality mapping in transition countries', *Statistics in Transition* **7**, 135–157.

Nychka, D. and Saltzman, N. (1998), Design of air quality monitoring networks, *in* D. Nychka, W. W. Piegorsch and L. H. Cox, eds, 'Case Studies in Environmental Statistics (Lecture Notes in Statistics, vol. 132)', Springer-Verlag, New York, pp. 51–76.

Olkin, I. and Ruixue Liu, R. (2003), 'A bivariate beta distribution', *Statistics & Probability Letters* **62**, 407–412.

Opsomer, J. D., Claeskens, G., Ranalli, M. G., Kauermann, G. and Breidt, F. J. (2008), 'Non-parametric small area estimation using penalized spline regression', *Journal of the Royal Statistical Society, Series B* **70**, 265–286.

O'Sullivan, F. (1986), 'A statistical perspective on ill-posed inverse problems (with discussion)', *Statistical Science* **1**, 505–527.

O'Sullivan, F. (1988), 'Fast computation of fully automated log-density and log-hazard estimators', *SIAM Journal on Scientific and Statistical Computing* **9**, 363–379.

Paelinck, J. H. P. and Klaassen, L. H. (1979), *Spatial Econometrics*, Gower, Westmead, Farnborough.

Parker, R. L. and Rice, J. A. (1985), 'Discussion of "Some aspects of the spline smoothing approach to nonparametric curve fitting" by B.W. Silverman', *Journal of the Royal Statistical Society, Series B* **47**, 40–42.

Patil, G. P. and Rao, C. R., eds (1994), *Handbook of Statistics Vol.12 - Environmental Statistics*, Elsevier, The Netherlands.

Petrucci, A., Bocci, C., Borgoni, R., Civardi, M., Salvati, N., Salvini, S. and Vignoli, D. (2009), *Indagine sulla georeferenziazione dei dati nella statistica ufficiale*, Rapporto di Indagine, Commissione per la Garanzia dell'Informazione Statistica, Presidenza del Consiglio dei Ministri, Roma.

Petrucci, A., Pratesi, M. and Salvati, N. (2005), 'Geographic Information in Small Area Estimation: Small Area Models ans Spatially Correlated Random Area Effects', *Statistics in Transition* **7**, 609–623.

Petrucci, A. and Salvati, N. (2006), ‘Small area estimation for spatial correlation in watershed erosion assessment’, *Journal of Agricultural, Biological and Environmental Statatistics* **11**, 169–182.

Pfeffermann, D. (2002), ‘Small Area Estimation - New Developments and Directions’, *International Statistical Review* **70**, 125–143.

Pratesi, M. and Salvati, N. (2008), ‘Small area estimation: the EBLUP estimator based on spatially correlated random area effects’, *Statistical Methods & Applications* **17**, 113–141.

R Development Core Team (2009), R: A language and environment for statistical computing, R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.
**URL:** *www.R-project.org*

Rao, J. N. K. (2003), *Small area estimation*, John Wiley & Sons, New York.

Richardson, H. W. (1970), *Regional Economics*, MacMillan, London.

Richardson, S., Leblond, L., Jaussent, I. and Green, P. J. (2002), ‘Mixture models in measurement error problems, with reference to epidemiological studies’, *Journal of the Royal Statistical Society, Series A* **165**, 549–566.

Robinson, G. K. (1991), ‘That BLUP is a Good Thing: The Estimation of Random Effects (with discussion)’, *Statistical Science* **6**, 15–51.

Romei, P. and Petrucci, A. (2003), *L’analisi del territorio. I sistemi informativi geografici*, Carocci, Roma.

Ruppert, D. (2002), ‘Selecting the Number of Knots for Penalized Splines’, *Journal of Computational and Graphical Statistics* **11**, 735–757.

Ruppert, D. and Carroll, R. J. (2000), ‘Spatially-Adaptive Penalties for Spline Fitting’, *Australian and New Zealand Journalof Statistics* **42**, 205–223.

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

Ruppert, D., Wand, M. P. and Carroll, R. J. (2009), ‘Semiparametric regression during 2003–2007’, *Electronic Journal of Statistics* **3**, 1193–1256.

Saei, A. and Chambers, R. (2005), *Small area estimation under linear and generalized linear mixed models with time and area effects*, Working Paper M03/15, Southampton Statistical Sciences Research Institute, University of Southampton.

Salvati, N., Chandra, H., Ranalli, M. G. and Chambers, R. (2008), 'Small Area Estimation Using a Nonparametric Model Based Direct Estimator'. Submitted for pubblication.

Singh, B., Shukla, G. and Kundu, D. (2005), 'Spatio-temporal models in small area estimation', *Survey Methodology* **31**, 183–195.

Spiegelhalter, D., Thomas, A., Best, N., Gilks, W. and Lunn, D. (2003), BUGS: Bayesian inference using Gibbs sampling, MRC Biostatistics Unit, Cambridge, England.
**URL:** *www.mrc-bsu.cam.ac.uk/bugs*

Stanislawski, L. V., Dewitt, B. A. and Shrestha, R. S. (1996), 'Estimating positional accuracy of data layers within a GIS through error propagation', *Photogrammetric Engineering and Remote Sensing* **62**, 429–433.

Stein, M. L. (1999), *Interpolation of Spatial Data: Some Theory for Kriging*, Springer-Verlag, New York.

Tobler, W. R. (1970), 'A Computer Movie Simulating Urban Growth in the Detroit Region', *Economic Geography* **46**, 234–240.

Torabi, M., Datta, G. S. and Rao, J. N. K. (2009), 'Empirical Bayes Estimation of Small Area Means under a Nested Error Linear Regression Model with Measurement Errors in the Covariates', *Scandinavian Journal of Statistics* **36**, 355–368.

Tzavidis, N., Salvati, N., Pratesi, M. and Chambers, R. (2008), 'M-quantile models with application to poverty mapping', *Statistical Methods and Applications* **17**, 393–411.

Veregin, H. (1989), *A taxonomy of error in spatial databases*, Technical Paper 89-12, National Center for Geographic Information & Analysis, Geography Department, University of California, Santa Barbara, California.

Voss, P. R. (2007), 'Demography as a Spatial Social Science', *Population Research and Policy Review* **26**, 457–476.

Waller, A. L. and Gotway, C. A. (2004), *Applied Spatial Statistics for Public Health Data*, John Wiley & Sons, Hoboken, New Jersey.

Wand, M. P. (1999), 'On the optimal amount of smoothing in penalised spline regression', *Biometrika* **86**, 936–940.

Wand, M. P. (2003), 'Smoothing and mixed models', *Computational Statistics* **18**, 223–249.

Wand, M. P. and Jones, M. C. (1993), 'Comparison of Smoothing Parameterizations in Bivariate Kernel Density Estimation', *Journal of the American Statistical Association* **88**, 520–528.

Wang, L. (2004), 'Estimation of Nonlinear Models with Berkson Measurement Errors', *The Annals of Statistics* **32**, 2559–2579.

Wolf, P. R. and Ghilani, C. D. (1997), *Adjustment computations: Statistics and least squares in surveying and GIS*, John Wiley, New York.

Wood, S. N. (2006), *Generalized Additive Models. An introduction with R*, Chapman & Hall/CRC, Boca Raton, Florida.

Woods, R. (1984), Spatial demography, *in* J. I. Clarke, ed., 'Geography and population: Approaches and applications', Pergamon Press, New York, pp. 43–50.

World Bank and INSTAT (2003), *Albania Living Standard Measurement Survey 2002. Basic Information Document.*
**URL:** *http://go.worldbank.org/IDTKJRT8Y0*

Ybarra, L. M. R. and Lohr, S. L. (2008), 'Small area estimation when auxiliary information is measured with error', *Biometrika* **95**, 919–931.

Zhang, J. X. and Goodchild, M. F. (2002), *Uncertainty in Geographical Information*, Taylor & Francis, New York.

Zhuly, L., Carlin, B. P. and Gelfand, A. E. (2003), 'Hierarchical regression with misaligned spatial data: relating ambient ozone and pediatric asthma ER visits in Atlanta', *Environmetrics* **14**, 537–557.