Dipartimento di Statistica "Giuseppe Parenti"

W O R K I N G P A P E R 2011/01

Estimates for geographical domains through geoadditive models in presence of missing geographical information

Chiara Bocci,
Emilia Rocco

Università degli Studi
di Firenze

Dipartimento di Statistica "G. Parenti" – Viale Morgagni 59 – 50134 Firenze - www.ds.unifi.it

Estimates for geographical domains through geoadditive models in presence of missing geographical information

Chiara Bocci, Emilia Rocco
Department of Statistics "G. Parenti", University of Florence

Abstract

This paper deals with the matter of applying a geoadditive model to produce estimates for some geographical domains in the absence of point referenced geographical data. The implementation of a geoadditive model needs the statistical units to be referenced at point locations and if we use it to produce model-based estimates of a parameter of interest for some geographical domains, the spatial location is required for all the population units. This information is not always easily available. Typically, we know the coordinates for sampled units, but for the non-sampled units we only know the areas - like blocks, municipalities, etc. - to which they belong. In such situation, the classic approach is to locate all the non-sampled units by the coordinates of their corresponding area centroid. This is obviously an approximation and its effect on the estimates can be strong, depending on the level of nonlinearity in the spatial pattern and on the area dimension. We propose a different approach that, instead of using the same coordinates for all the units, imposes a distribution for the locations inside each area. Our approach is formalized under a Bayes inferential perspective and its performance is evaluated through various Markov Chain Monte Carlo experiments implemented under different scenarios.

Keywords: Hierarchical Bayesian models, Imputation, Penalized splines, Linear mixed model, Sample representativeness.

1 Introduction

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. Obviously 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.

Geostatistical methodologies are concerned with this target and typically are applied, exploiting the exact knowledge of the spatial coordinates (latitude and

longitude) of sampled units, to analyze the spatial pattern or to obtain spatial interpolations and predictions of the studied phenomenon. In particular, geoadditive models (Kammann and Wand, 2003) belong to this set of methodologies and analyze the spatial distribution of the study variable while accounting for possible linear or non-linear covariate effects by merging an additive model (Hastie and Tibshirani, 1990) and a kriging model (Cressie, 1993) and by expressing both as a linear mixed model.

To fit a geoadditive model, only the sampled statistical units are required to be referenced at points locations. If, however, we use the same geoadditive model to produce model-based estimates of a parameter of interest for some geographical domains, the spatial information is required for all the population units.

However often we don't know the exact location of all the population units, 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. For the non-sampled units we know just the areas to which they belong like census districts, blocks, municipalities, etc. How can we continue to use the geostatistical techniques under these circumstances?

In such situation, the classic approach is to locate all the units belonging to the same area by the coordinates (latitude and longitude) of the geographical centre or *centroid* of the 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.

In this paper we propose to fill the holes in the geographical information following a stochastic imputation approach (Little and Rubin, 1987) instead of the classic deterministic one with the centroids. In particular we suggest to treat the lack of geographical information imposing a distribution for the locations inside each area. This is realized through a hierarchical Bayesian formulation of the geoadditive model in which a prior distribution on the spatial coordinates is defined. The performance of our imputation approach is evaluated through various Markov Chain Monte Carlo (MCMC) experiments implemented under different scenarios: true distribution of the spatial coordinates (homogeneous Poisson process, inhomogeneous Poisson process, Beta distribution) and a-priori coordinate distribution used in the hierarchical Bayesian formulation (Centroid, Uniform and Beta).

The results shows that our approach, that includes the classical imputation approach through the centroids as a special case, is promising and that its performance depends on the properness of the hypothesis used in the definition of the a-priori distribution of spatial coordinates.

The paper is organized as follows. Section 2 briefly reviews the theory of geoadditive model. In Section 3 the matter of applying geoadditive models to produce model-based estimates for some geographical domains in the absence of point referenced auxiliary data is discussed and the complete hierarchical Bayesian

formulation of the geoadditive model under our stochastic imputation approach is presented. The performance of this approach is evaluated in Sections 4 and 5 through various MCMC experiments. Section 6 concludes with final remarks and ongoing questions.

2 Geoadditive models

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 linear or non-linear covariate effects. Under the additivity assumption they can handle such covariate effects by merging an additive model - that accounts for the 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.

Let x_i and t_i, $1 \leq i \leq n$,a linear and a non-linear predictors of y_i at spatial location $\mathbf{s}_i$, $\mathbf{s} \in \Re$. A geoadditive model for such data can be formulated as

$$y_i = \alpha + \beta_x x_i + g(t_i) + h(\mathbf{s}_i) + \varepsilon_i, \qquad \varepsilon_i \sim N(0, \sigma_\varepsilon^2), \tag{1}$$

where g is an unspecified univariate smooth function and h is an unspecified bivariate smooth function.

Representing $g(\cdot)$ with a low-rank truncated linear spline with K_t knots and $h(\cdot)$ with a low-rank thin plate spline with K_s knots

$$g(t) = \beta_{0t} + \beta_t t + \sum_{k=1}^{K_t} u_k^t (t - \kappa_k^t)_+$$

$$h(\mathbf{s}) = \beta_{0s} + \mathbf{s}^T \boldsymbol{\beta}_s + \sum_{k=1}^{K_s} u_k^s b_{tps}(\mathbf{s}, \boldsymbol{\kappa}_k^s)$$

the model (1) can be written as a mixed model (Kammann and Wand, 2003)

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \boldsymbol{\varepsilon}, \tag{2}$$

with

$$\mathrm{E}\begin{bmatrix}\mathbf{u}\\ \boldsymbol{\varepsilon}\end{bmatrix} = \begin{bmatrix}\mathbf{0}\\ \mathbf{0}\end{bmatrix}, \qquad \mathrm{Cov}\begin{bmatrix}\mathbf{u}\\ \boldsymbol{\varepsilon}\end{bmatrix} = \begin{bmatrix}\sigma_t^2\mathbf{I}_{K_t} & \mathbf{0} & \mathbf{0}\\ \mathbf{0} & \sigma_s^2\mathbf{I}_{K_s} & \mathbf{0}\\ \mathbf{0} & \mathbf{0} & \sigma_\varepsilon^2\mathbf{I}_n\end{bmatrix}.$$

where

$$\begin{aligned}\mathbf{X} &= \left[1, x_i, t_i, \mathbf{s}_i^T\right]_{1\leq i\leq n},\\ \boldsymbol{\beta} &= \left[\beta_0, \beta_x, \beta_t, \boldsymbol{\beta}_s^T\right],\\ \mathbf{u} &= \left[u_1^t, ..., u_{K_t}^r, u_1^s, ..., u_{K_s}^s\right],\end{aligned}$$

$\beta_0 = \alpha + \beta_{0t} + \beta_{0s}$ and $\mathbf{Z}$ is obtained by concatenating the matrices containing spline basis functions to handle g and h, respectively

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

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

where $C(\mathbf{v}) = \|\mathbf{v}\|^2 \log\|\mathbf{v}\|$ and $\kappa_1^t, ..., \kappa_{K_t}^t$ and $\boldsymbol{\kappa}_1^s, ..., \boldsymbol{\kappa}_{K_s}^s$ are the knots locations for the two functions.

The amount of smoothing for both the additive component and the geostatistical component of the model can be quantified through the variance components ratios $\sigma_\varepsilon^2/\sigma_t^2$ and $\sigma_\varepsilon^2/\sigma_s^2$.

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 (Ruppert et al., 2009).

The mixed model (2) could be fit in a frequentist framework using Best Linear Unbiased Predictor (BLUP) or Penalized Quasi Likelihood (PQL) estimation. In this paper we adopt a Bayesian inferential perspective, by placing priors on the model parameters and simulating their joint posterior distribution. Often the posterior density is analytically unavailable but can be simulated using Markov Chain Monte Carlo (MCMC). Moreover, the posterior distribution of any explicit function of the model parameters can be obtained as a by-product of simulation algorithm.

3 Lack of geographical information and regional mean estimation

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 (2):

$$\hat{\bar{y}}_q = \frac{1}{N_q}\left[\sum_{i\in S_q} y_i + \sum_{i\in R_q}\left(\mathbf{x}_i\hat{\boldsymbol{\beta}} + \mathbf{z}_i\hat{\mathbf{u}}\right)\right], \tag{3}$$

where N_q is the total number of units in region q and S_q and R_q indicate the sets of the sampled and non-sampled units belonging to region q.

We obtain the estimated parameters from the sampled units, but, if we don't know s for the not-sample units, we cannot use directly (3). To better show the problem, consider to have a linear predictor x_i of y_i at spatial location $\mathbf{s}_i$ and to use the following spline regression model

$$y_i = \beta_0 + \beta_x x_i + \boldsymbol{\beta}_s^T\mathbf{s}_i + \sum_{k=1}^{K_s} u_k^s b_{tps}(\mathbf{s}_i, \boldsymbol{\kappa}_k^s) + \varepsilon_i.$$

The *model-based* mean estimator becomes:

$$\hat{\bar{y}}_q = \frac{1}{N_q}\left[\sum_{i\in S_q} y_i + \sum_{i\in R_q}\left(\hat{\beta}_0 + \hat{\beta}_x x_i + \hat{\boldsymbol{\beta}}_s^T\mathbf{s}_i + \sum_{k=1}^{K_s}\hat{u}_k^s b_{tps}(\mathbf{s}_i, \boldsymbol{\kappa}_k^s)\right)\right], \tag{4}$$

How can we still apply this estimator if we don't know $\mathbf{s}_i$ for the non-sampled units R_q? In the classic approach the $\mathbf{s}_i$ values are replaced with the region centroid $\mathbf{c}_q$, that is a constant for all the units in region q.

We decided to proceed differently, treating the lack of geographical information as a particular problem of missing data: instead of use the same coordinates $\mathbf{c}_q$ for all the units in region q, which may be defined as a particular case of deterministic imputation, we choose to use a stochastic Bayesian imputation approach including in the hierarchical Bayesian formulation of the geoadditive model (Ruppert et al., 2003, Chapter 16) a prior distribution $f_s(\boldsymbol{\theta}_q)$ for $\mathbf{s}_i$ inside each region q and then using the joint posterior distribution of all parameters given the data as the basis of inference.

Thus, under stochastic imputation, our complete hierarchical Bayesian formulation of the geoadditive model (following specifications of Crainiceanu et al. (2005)) is

$$\begin{aligned}
&y_i|\boldsymbol{\beta}, \mathbf{u}, \sigma_\varepsilon^2 \overset{\text{ind}}{\sim} N\left(\beta_0 + \beta_x x_i + \boldsymbol{\beta}_s^T\mathbf{s} + \sum_{k=1}^{K_s} u_k^s b_{tps}(\mathbf{s}, \boldsymbol{\kappa}_k^s), \sigma_\varepsilon^2\right),\\
&\mathbf{u}|\sigma_s^2 \sim N(0, \sigma_s^2\mathbf{I}_{K_s}),\\
&s_i|\boldsymbol{\theta}_q \sim f_s(\boldsymbol{\theta}_q),
\end{aligned} \tag{5}$$

with non-informative priors for $\boldsymbol{\theta}_q$, here not specified as depending on the choice of f_s, and for $\boldsymbol{\beta}$, σ_s^2, σ_ε^2

$$\begin{cases}
\beta_0, \beta_x, \beta_s \overset{\text{ind}}{\sim} N(0, 10^8)\\
\sigma_s^{-2}, \sigma_\varepsilon^{-2} \overset{\text{ind}}{\sim} \text{Gamma}(10^{-8}, 10^{-8}).
\end{cases}$$

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 parameterize the inverse of the variance, that is the *precision* parameter.

We should note that if, for each region, f_s is a probability mass function that assumed value 1 when $\mathbf{s} = \mathbf{c}_q$ and 0 otherwise, then our formulation corresponds to the centroid approach.

4 MCMC Experiments

In order to evaluate the performance of our approach with respect to the classic centroid approach, various MCMC experiments are implemented under different scenarios.

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 analysis 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, 2010). 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.

4.1 Scenarios Specification

Three MCMC experiments are implemented for each of four scenarios concerning the population data. All scenarios are characterized by the following setting:

- The study variable is simulated by the model

$$y_i = \alpha + \beta_x x_i + f(\mathbf{s}_i) + \varepsilon_i$$

  where $\varepsilon_i \sim N(0, \sigma^2_\varepsilon)$, $\sigma_\varepsilon = 0.2$, $\alpha = 10$, $\beta_x = 0.4$, $x \sim \text{Ber}(0.5)$ is a dummy variable known for the whole population, $\mathbf{s}$ represents the spatial location that is generated by a different spatial point process in each scenario and function $f(\mathbf{s})$ is obtained as a bivariate normal mixture density (following Wand and Jones (1993)) and is represented in Figure 1.

- The population consisting of $N = 3000$ units is located in the unit squared $O = [0, 1] \times [0, 1]$ which is divided in $Q = 9$ rectangular regions that can be represented by their vertices $[(l_{1q}, m_{1q}), (l_{2q}, m_{1q}), (l_{2q}, m_{2q}), (l_{1q}, m_{2q})]$. The regions are obtained using a random binary splitting procedure.

Each scenario differs from the others for the spatial point process used to generate $\mathbf{s}$. We consider 4 data generating processes (presented in Figure 2): (A)

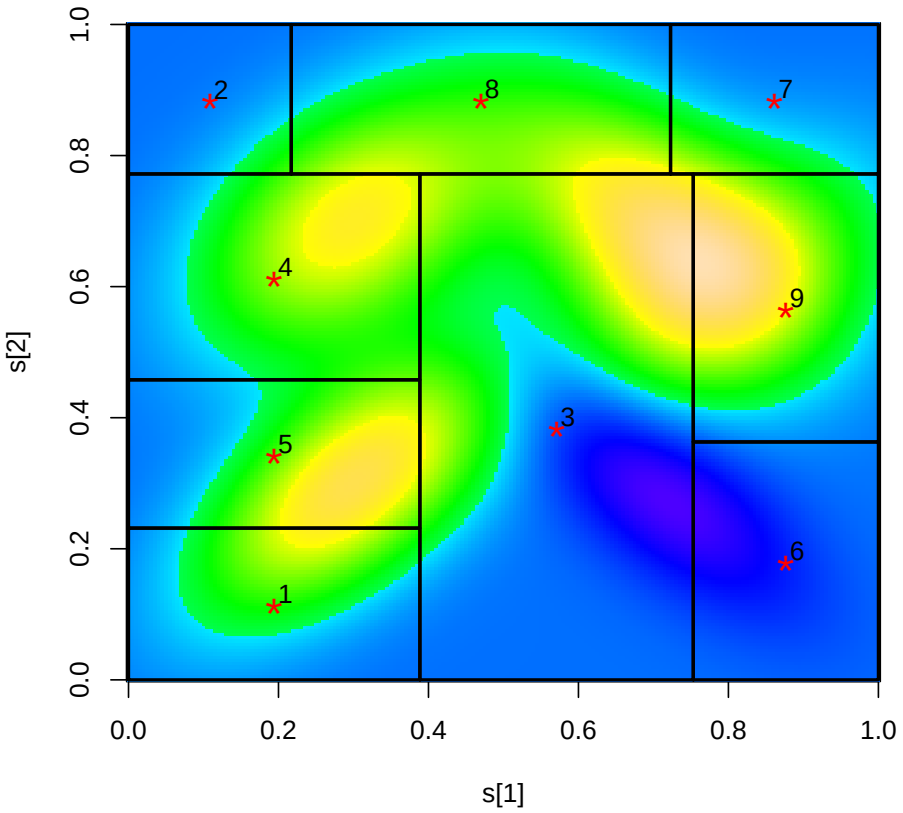
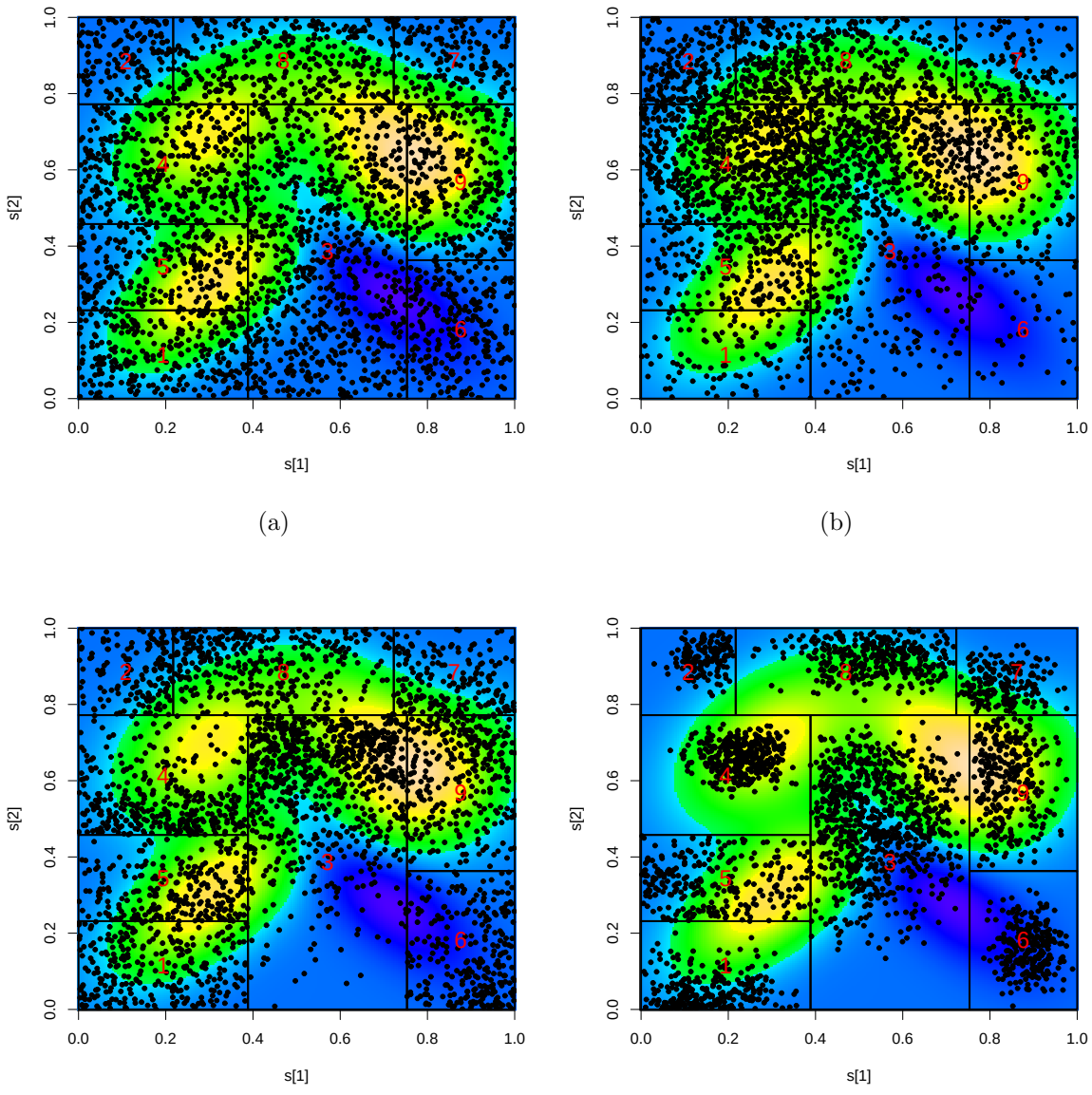


Figure 1: Bivariate normal mixture density $f(\mathbf{s})$. Rectangles are the regions and the red stars indicate the centroids.

homogeneous Poisson process on O, (B) inhomogeneous Poisson process on O, (C) inhomogeneous Poisson process on each rectangular region, (D) independent bivariate Beta distribution on each rectangular region.

For a detailed description of the Poisson processes we refer to Diggle (1983, Chapter 4), here we say that the processes parameters are set to obtain a realization of roundly 3000 units and that for scenario C, the intensity function $\lambda(\mathbf{s})$ of the inhomogeneous Poisson process changes in every area.

For the independent bivariate Beta distribution, we generate in each region a number of units proportional to the area size, with the coordinates $\mathbf{s}$ obtained as realizations of two independent different Beta distributions, defined respectively on the latitude interval $[l_{1q}; m_{1q}]$ and on the longitude interval $[l_{2q}; m_{2q}]$. The parameters of the two Beta distributions change in every area.



(a) (b)

(c) (d)

Figure 2: Spatial distributions of the population units: (a) homogeneous Poisson process, (b) inhomogeneous Poisson process, (c) inhomogeneous Poisson process on each region, (d) independent bivariate Beta distribution on each region.

For each population setting three MCMC experiments are performed to estimate the mean of y in the 9 regions applying the estimator (4) and using the model

formulation (5). They are characterized by three different choices of the prior distribution $f_s(\boldsymbol{\theta}_q)$ for $\mathbf{s}_i$ inside each region q, that is by three different imputation models:

- **Centroid Imputation**: $\mathbf{s}_i$ is replaced with the region centroid $\mathbf{c}_q$, constant in region q. f_s is the probability mass function that assumed value 1 when $\mathbf{s} = \mathbf{c}_q$ and 0 otherwise.

$$f_s(\mathbf{s}_i) = \begin{cases} 1 & \text{if } \mathbf{s}_i = \mathbf{c}_q \\ 0 & \text{if } \mathbf{s}_i \neq \mathbf{c}_q \end{cases}.$$

- **Uniform Imputation**: f_s is a bivariate Uniform distribution on $[l_{1q}; m_{1q}] \times [l_{2q}; m_{2q}]$

$$\mathbf{s}_i \overset{\text{ind}}{\sim} \text{Uniform on } [l_{1q}; m_{1q}] \times [l_{2q}; m_{2q}].$$

- **Beta Imputation**: f_s is obtained as product of two independent Beta distributions on $[l_{1q}; m_{1q}] \times [l_{2q}; m_{2q}]$

$$\mathbf{s}_i | a_{1q}, b_{1q}, a_{2q}, b_{2q} \overset{\text{ind}}{\sim} \text{Beta}\,(a_{1q}; b_{1q}) \times \text{Beta}\,(a_{2q}; b_{2q}) \text{ on } [l_{1q}; m_{1q}] \times [l_{2q}; m_{2q}].$$

with parameters $a_{1q}, b_{1q}, a_{2q}, b_{2q}$ estimated directly in the MCMC process, by adding the following priors to model (5)

$$a_{1q}, b_{1q}, a_{2q}, b_{2q} \overset{\text{ind}}{\sim} \text{Unif}(0; 100).$$

The common elements to all the MCMC experiments are:

- $f(\mathbf{s})$ is modeled considering a penalized thin plate spline function with $K_s = 64$ knots selected on a regular grid on space. We choose this type of splines since it tends to have good numerical properties and, as pointed out by Crainiceanu et al. (2005, p.2), the posterior correlation of parameters for the thin-plate splines is much smaller than for other basis, which greatly improves mixing.
- the MCMC analysis is implemented 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 $h = 1000$ retained for inference.
- the geoadditive model (5) is fitted using a stratified sample of $n = 500$ units selected from the population with strata corresponding to the rectangular regions, proportional allocation of the sample units in each strata and simple random sample selection in each strata.

In next section, for each population scenario the mean estimates obtained in correspondence of each imputation approach are compared. In order to take into account not only the model variability but also the variability due to the sampling design each MCMC experiment is repeated $m = 100$ times. In each replica a new sample is selected and used to fit the model and calculate the mean estimates. Obviously, for each population scenario the three imputation hypothesis are compared using the same 100 samples.

4.2 Results

Figures 3, 4, 5 and 6 show, for scenarios A, B, C and D, the posterior density of the regional mean estimator under the three imputation approaches. The green, red and blue lines correspond respectively to the Centroid, Uniform and Beta imputation approaches, while the vertical lines indicate the true mean values. Each posterior density is evaluated over an MCMC sample of 100000 units (unifying the $m = 100$ chains of $h = 1000$ units), thus its variability include both the model effect and the sampling design effect.

In addition to graphical representation, the estimator performance is evaluated computing the Relative Bias (RB) and the Relative Root Mean Square Error (RRMSE) defined as

$$RB_q = \frac{1}{hm} \frac{\sum_{j=1}^{hm} (\hat{\bar{y}}_{qj} - \bar{y}_q)}{\bar{y}_q}$$

and

$$RRMSE_q = \frac{\sqrt{\frac{1}{hm} \sum_{j=1}^{hm} (\hat{\bar{y}}_{qj} - \bar{y}_q)^2}}{\bar{y}_q},$$

where $\bar{y}_q$ denotes the actual mean of region q and $\hat{\bar{y}}_{jq}$ is the predicted value at simulation j, $j = 1, ..., hm$. Tables 1, 2, 3 and 4 present the values of RB and RRMSE of the regional mean estimator under the three imputation approaches, for scenarios A, B, C and D respectively.

The values reported in the last row of each table correspond to RB and RRMSE of the overall mean estimator. Moreover, the posterior densities of the overall mean estimator are presented in Figure 7.

Observing Figures 3 and 6 and the corresponding Tables 1 and 4, it is straightforward to note that if the imputation distribution f_s corresponds to the population spatial distribution, the stochastic imputation approach produces better estimates than the classic centroid approach. This is the case of the Uniform approach in scenario A (Fig. 3) and of the Beta approach in scenario D (Fig. 6).

The Beta imputation approach works well also in scenario A, due to the fact that the true spatial distribution in each region is a special case of the bivariate Beta distribution, but it produces less precise estimates than the Uniform imputation since the Beta parameters need to be estimated in the fitting process.

For scenarios B (Fig. 4 and Table 2) and C (Fig. 5 and Table 3) none of the imputation models corresponds to the population spatial distribution, but the Beta approach still presents a good performance. This depends on the fact that the Beta distribution has the advantage of modeling different shapes depending on the parameters value. In our approach these parameters are estimated directly in the MCMC process exploiting the spatial distribution of the sampled units and producing a posterior bivariate Beta distribution that is as similar as possible to the sample spatial distribution. Obviously, the good performance of this approach relies on the representativeness of the sample. This aspect is further investigated in Section 5.

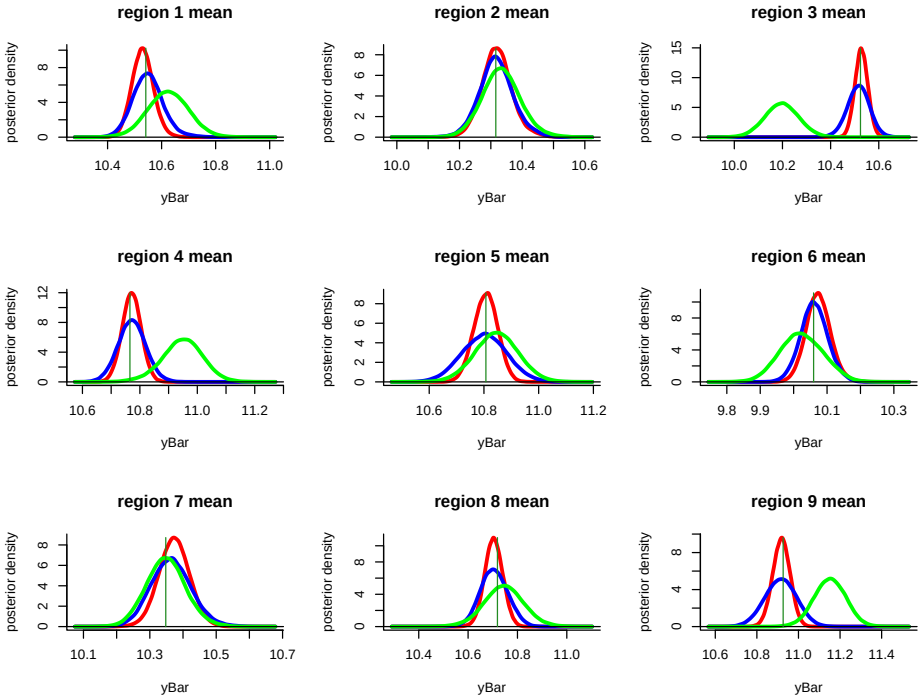


Figure 3: Posterior density of the regional model-based mean estimator under scenario A (homogeneous Poisson process) for the three imputation approaches: Centroid (green line), Uniform (red line) and Beta (blue line). The vertical lines indicate the true mean values.

Table 1: Empirical RB and RRMSE of the model-based mean estimator under scenario A (homogeneous Poisson process) for the three imputation approaches.

	Centroid Imputation		Uniform Imputation		Beta Imputation	
Region	RBias %	RRMSE %	RBias %	RRMSE %	RBias %	RRMSE %
1	0.7934	1.0726	-0.1060	0.3926	0.0748	0.5429
2	0.1790	0.6245	0.0380	0.4737	0.0169	0.5396
3	-3.1145	3.1856	0.0317	0.2641	-0.0834	0.4517
4	1.6969	1.8209	0.0565	0.3141	0.0632	0.4521
5	0.3569	0.8211	-0.0051	0.4036	-0.0537	0.7527
6	-0.3597	0.7326	0.1338	0.3857	0.0177	0.4059
7	0.0116	0.5783	0.2422	0.5127	0.1608	0.6099
8	0.2409	0.7920	-0.1407	0.3687	-0.1274	0.5344
9	2.0669	2.1835	-0.0713	0.3975	-0.1114	0.7074
Overall	-0.2861	0.3860	0.0089	0.1254	-0.0269	0.1979

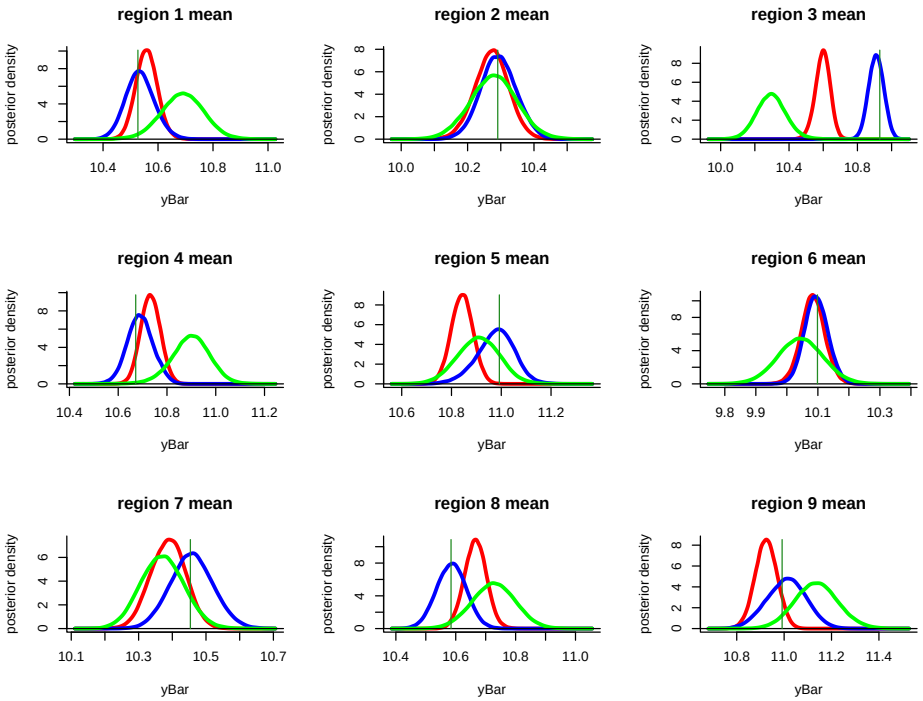


Figure 4: Posterior density of the regional model-based mean estimator under scenario B (inhomogeneous Poisson process on O) for the three imputation approaches: Centroid (green line), Uniform (red line) and Beta (blue line). The vertical lines indicate the true mean values.

Table 2: Empirical RB and RRMSE of the model-based mean estimator under scenario B (inhomogeneous Poisson process on O) for the three imputation approaches.

	Centroid Imputation		Uniform Imputation		Beta Imputation	
Region	RBias %	RRMSE %	RBias %	RRMSE %	RBias %	RRMSE %
1	1.5427	1.7086	0.3019	0.4796	0.0770	0.5067
2	-0.1183	0.6995	-0.1747	0.5283	0.0219	0.5200
3	-5.8033	5.8560	-3.0559	3.0820	-0.1910	0.4458
4	2.1125	2.2369	0.5721	0.6848	0.1740	0.5280
5	-0.7650	1.0841	-1.3633	1.4199	-0.1384	0.6976
6	-0.5418	0.9037	-0.1433	0.4055	-0.0390	0.3751
7	-0.7888	1.0004	-0.6332	0.8045	0.0422	0.6082
8	1.3448	1.5056	0.7700	0.8457	0.0124	0.4686
9	1.3178	1.5541	-0.6146	0.7459	0.1635	0.7659
Overall	-1.1534	1.1905	-0.9406	0.9535	-0.0207	0.1979

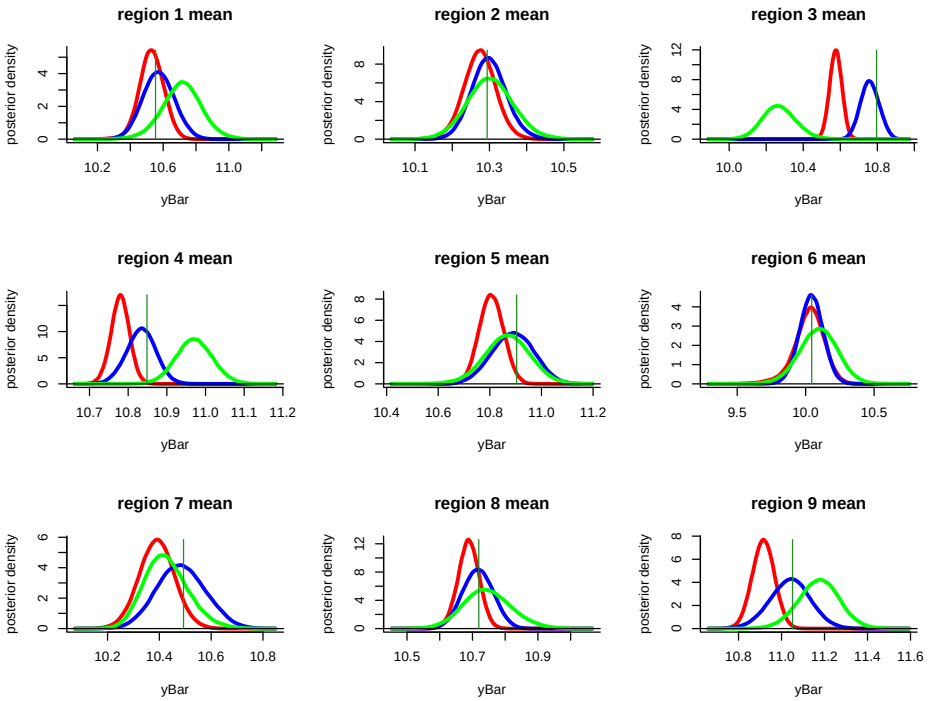


Figure 5: Posterior density of the regional model-based mean estimator under scenario C (inhomogeneous Poisson process on each region) for the three imputation approaches: Centroid (green line), Uniform (red line) and Beta (blue line). The vertical lines indicate the true mean values.

Table 3: Empirical RB and RRMSE of the model-based mean estimator under scenario C (inhomogeneous Poisson process on O) for the three imputation approaches.

	Centroid Imputation		Uniform Imputation		Beta Imputation	
Region	RBias %	RRMSE %	RBias %	RRMSE %	RBias %	RRMSE %
1	1.5210	1.9031	-0.2256	0.7465	0.1777	0.9244
2	0.0809	0.6108	-0.1959	0.4660	0.0475	0.4581
3	-4.8635	4.9351	-2.0413	2.0646	-0.3337	0.5723
4	1.1259	1.2053	-0.6395	0.6768	-0.1398	0.3759
5	-0.3290	0.8681	-0.8983	0.9951	-0.1369	0.7468
6	0.4805	1.4733	-0.2036	1.1369	-0.0393	0.8779
7	-0.6250	1.0207	-0.9846	1.1835	-0.0861	0.8967
8	0.2472	0.7276	-0.2919	0.4204	-0.0261	0.4347
9	1.1119	1.4059	-1.2302	1.3163	-0.1010	0.8539
Overall	-0.8551	0.9073	-0.9873	0.9953	-0.1425	0.2483

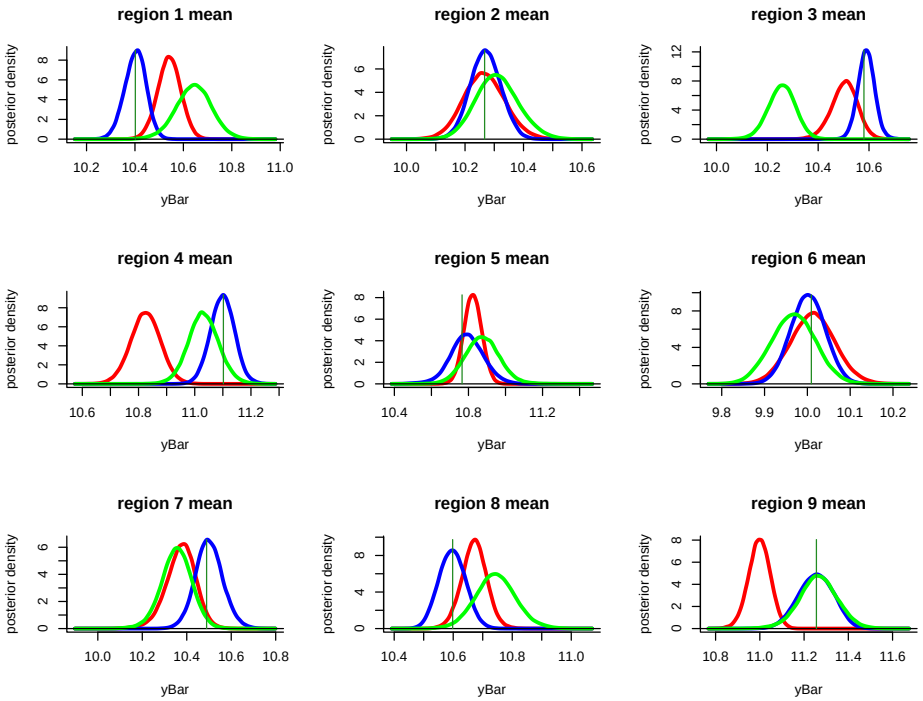
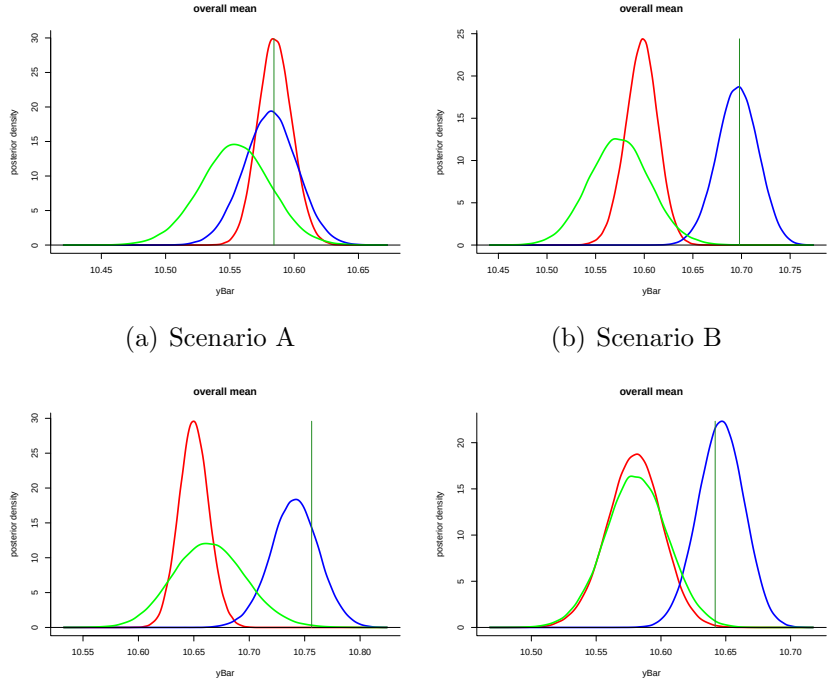


Figure 6: Posterior density of the regional model-based mean estimator under scenario D (bivariate Beta distribution on each region) for the three imputation approaches: Centroid (green line), Uniform (red line) and Beta (blue line). The vertical lines indicate the true mean values.

Table 4: Empirical RB and RRMSE of the model-based mean estimator under scenario D (bivariate Beta distribution on each region) for the three imputation approaches.

	Centroid Imputation		Uniform Imputation		Beta Imputation	
Region	RBias %	RRMSE %	RBias %	RRMSE %	RBias %	RRMSE %
1	2.3203	2.4269	1.3440	1.4220	-0.0046	0.4289
2	0.4227	0.8285	-0.0076	0.6944	0.0336	0.5121
3	-3.0580	3.1006	-0.7285	0.8775	0.0863	0.3261
4	-0.6350	0.8042	-2.4948	2.5406	-0.0298	0.3874
5	1.0137	1.3327	0.5416	0.7010	0.2488	0.8755
6	-0.4144	0.6593	0.0342	0.5208	-0.0778	0.4076
7	-1.2892	1.4489	-1.1110	1.2835	0.1120	0.6043
8	1.3736	1.5183	0.6886	0.7977	-0.0116	0.4258
9	0.0750	0.7751	-2.2920	2.3339	-0.0230	0.6983
Overall	-0.5710	0.6144	-0.5872	0.6204	0.0403	0.1710



(a) Scenario A (b) Scenario B

(c) Scenario C (d) Scenario D

Figure 7: Posterior density of the overall model-based mean estimator under the four scenarios and for the three imputation approaches: Centroid (green line), Uniform (red line) and Beta (blue line). The vertical lines indicate the true mean values.

It should be noted, however, that the sample representativeness affects estimator (4) under all the imputation approaches. The fact that the population structure should be valid for the sample too is a common assumption in almost all inferential methods. In our case we require this assumption primarily to infer the unit spatial distribution rather than to model y. In fact, since the geoadditive model is a semiparametric procedure where the spatial influence is modeled locally with a spline structure, the parameter estimation of model 5 is little sensitive to the spatial sample representativeness (unless huge amount of spatially clustered data are missing, especially near the region boundaries).

About the classical centroid approach we can observe that almost in all the cases it performs worse than the Beta imputation, even if there are some specific situation in which it seems a good choice (see for example region 7 in scenario A or region 5 in scenario C). However, this strictly depends on the specific units spatial distribution and values of y inside that region. This consideration applies also to the behavior of the Uniform distribution in Scenarios B, C and D: generally it doesn't work well but may be good in some specific situation.

The good performance of the Beta imputation under all the scenarios is reflected also in the mean estimation for the overall area O, as we observe from the posterior distribution of the mean estimator showed in Figure 7.

5 Non-representative Samples

The results presented in Section 4.2 show that the hierarchical Bayesian formulation of geoadditive models with stochastic imputation, here suggested to deal with the lack of spatial coordinates for non sampled units, produces better regional mean estimates than the classical centroid imputation approach. This happens not only when the prior distribution chosen for the spatial coordinates corresponds to the true coordinates distribution, but also when a "flexible" distribution, like the independent bivariate Beta, is used to obtain an approximate spatial distribution of $\mathbf{s}$, as long as the sample spatial distribution reflects that one of the population.

In order to show the relevance of the spatial representativeness property of the sample we have performed some other MCMC experiments. To better visualize the spatial distribution of sample units in these new experiments we assume to have a univariate s so that the regions are actually intervals.

The study variable y is simulated by the model

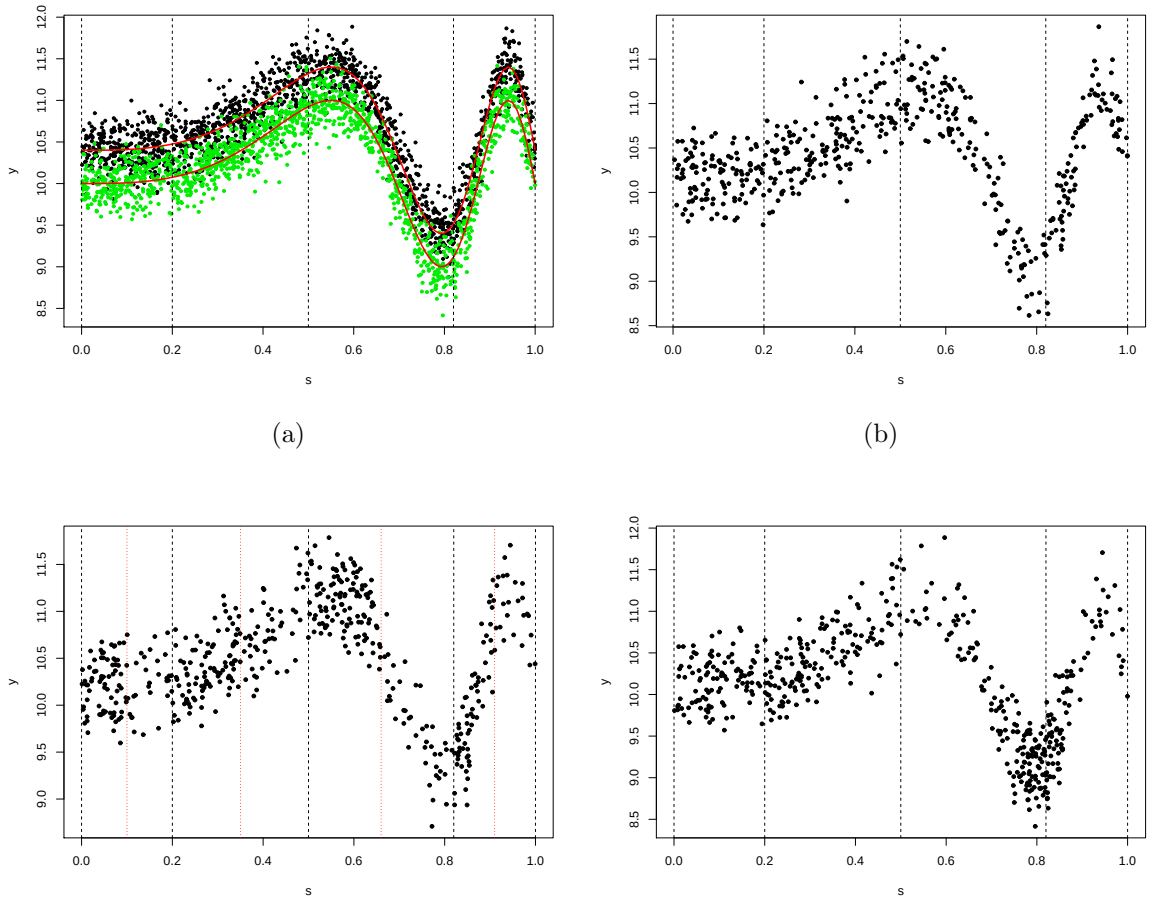
$$y_i = \alpha + \beta_x x_i + f(\mathbf{s}_i) + \varepsilon_i \tag{6}$$

where $\varepsilon_i \sim N(0, \sigma^2_\varepsilon)$, $\sigma_\varepsilon = 0.2$, $\alpha = 10$, $\beta_x = 0.4$, $x \sim \text{Ber}(0.5)$ is a dummy variable known for the whole population, s represents the spatial location and is generated by a uniform distribution in every region and function $f(s) = sin(3\pi s^3)$. The population consisting of $N = 3000$ units is located in the interval $O = [0, 1]$ which is divided in $Q = 4$ intervals $[0, 0.2], [0.2, 0.5], [0.5, 0.82], [0.82, 1]$. The obtained population is showed in Figure 8(a): the green and black dots correspond to the units with $x_i = 0$ and $x_i = 1$ respectively, the vertical dashed lines indicate the regions and the red lines indicate the deterministic component of model (6).

Given this population setting, three different scenarios are considered: varying the type of samples selected from the population. For each scenario, three MCMC experiments are performed to estimate the mean of y in the 4 regions. Each experiment corresponds to the univariate version of one of the imputation models defined in Section 4.1.

All the tree types of samples are stratified samples of $n = 500$ units with strata corresponding to the 4 regions and proportional allocation of sampled units in each strata. They differs in the sampling design adopted to select the units in each strata:

- **Representative sample**: a simple random sample is selected in each strata;
- **Type 1 Non-representative sample**: in each strata, the 70% of the sample is randomly selected among the units with s values lower than the centroid and the remaining 30% is randomly selected among the units with s values greater than the centroid;



(a) (b)

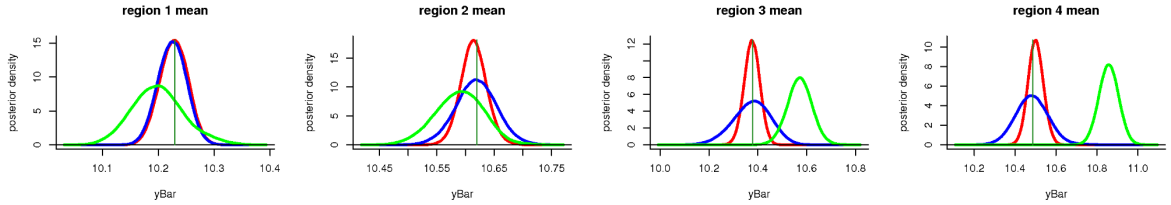
(c) (d)

Figure 8: Scenarios Settings: (a) Simulated population, with the green and black dots corresponding to the units with $x_i = 0$ and $x_i = 1$ respectively, the vertical dashed lines indicating the regions and the red lines indicating the deterministic component of model (6); (b) Distribution of a Representative sample; (c) Distribution of a Type 1 Non-representative sample; (d) Distribution of a Type 2 Non-representative sample.

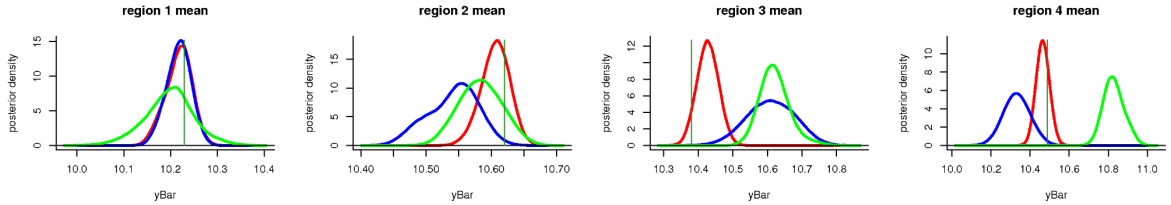
- **Type 2 Non-representative sample**: in each strata the units are selected with probability proportional to the inverse of the y values.

Examples of the three samples spatial distribution are showed in Figure 8.

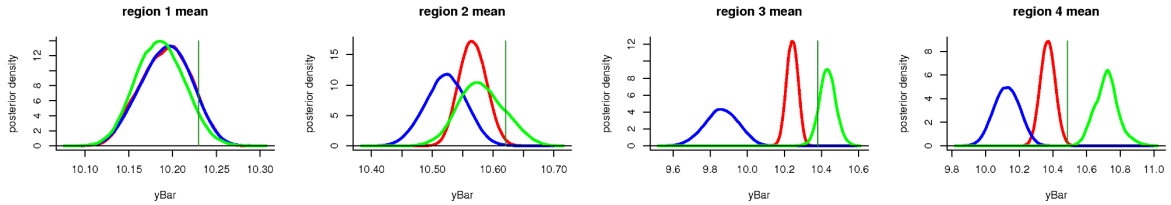
The MCMC experiments follow the same settings described in Section 4.1 and are replicated $m = 100$ times in order to take into account both the model and the the sampling design variability. Function $f(s)$ is modeled with a low-rank



(a) Representative samples



(b) Type 1 Non-representative samples



(c) Type 2 Non-representative sample

Figure 9: Posterior density of the regional model-based mean estimator under the three scenarios and for the three imputation approaches: Centroid (green line), Uniform (red line) and Beta (blue line). The vertical lines indicate the true mean values.

truncated linear spline with $K_s = 30$ knots located on the quantiles of the sample distribution of s.

The posterior densities of the regional model-based mean estimator under the three scenarios are presented in Figure 9. It is evident that when a simple random sample is selected in each strata both the uniform imputation (that correspond to the true spatial distribution) and the Beta imputation work well (analogously to the bivariate scenario A).

In the other two scenarios the performance of both the two imputation approaches get worse but the Beta imputation is more affected. This is due to the fact that the Beta imputation exploits the spatial distribution of the sampled units to estimate its parameters, and as long as the spatial sample distribution does not reflect the one of population the estimated parameters produce a posterior spatial distribution different from the true one.

On the contrary, the Uniform imputation does not exploit any sample information, thus it correctly imputes the coordinates of the non sampled units. However, since the sampled units selection depends on their location (or to their value of y which is strictly connected to s by $f(s)$) the joint spatial distribution of sampled and imputed units will not be Uniform.

Analogous considerations apply with the classic centroid imputation approach. Thus, whichever imputation approach we use the mean estimator (4) will be affected by the sample non-representativity.

It is important to note that the sample non-representativity is strictly related to the imputation step of our analysis. Due to its semiparametric splines structure, the geoadditive model is robust to sample non-representativity and the model fitting step is hardly influenced by it.

6 Final remarks

In the last years the use of geostatistical techniques to produce model-based estimates of a parameter of interest for some geographical domains is grown. This is also shown by some relative recent papers (Opsomer et al., 2008; Bocci, 2010; Salvati et al., 2010) in which the use of geoadditive small area models are investigated.

Their use however is not always straightforward as it needs for all the population units to be referenced at point location, but this requirement in not so easy to be accomplished. In this paper we have suggested a solution to this problem that propose a hierarchical Bayesian formulation of a geoadditive model in which a prior distribution for the spatial coordinates is defined to characterize the knowledge prior to data collection. The missing spatial coordinates are then extracted from their posterior distribution, obtained by MCMC simulation.

We have also shown that in absence of a prior knowledge of the spatial distribution, the spatial coordinates imputation approach with the Beta prior distribution works well as long as the sample units have the same spatial distribution of the population units. Moreover the Beta imputation is surely preferable to the classic approach that locate each units with their corresponding area centroid.

We should highlight the generality of our stochastic Bayesian imputation approach. In all situation examined in the paper the use of the Beta distribution as prior distribution is the better choice as long as the prior knowledge does not allows to specify the true distribution (such as the Uniform distribution in scenario A). If however the sample spatial distribution presents a more complex pattern (like multimodal or clustered distributions) our model formulation approach can still be used, by choosing a more flexible prior distribution (i.e. with more parameters).

Moreover, even if neither the exact units coordinates nor their true distribution are known, some auxiliary information may still be available (e.g. land use, land elevations, etc...). The advisability to use them in the formulation of the model should be a future development.

Finally, it might also be desirable to extend the Bayesian model formulation in order to include information on the sample design. This should produce a robust estimation with respect to the hypothesis of spatial representativeness of the sample. We intend to also develop this aspect in our future work.

Acknowledgements

This work is supported by the project PRIN2007 (PRIN 2007RHFBB3_004) '*Efficient use of auxiliary information at the design and at the estimation stage of complex surveys: methodological aspects and applications for producing official statistics*' awarded by the Italian Ministry for Education, University and Research to the universities of Perugia, Cassino, Firenze, Pisa and Trieste.

References

Bocci, C., 2010. Geoadditive Small Area Model for the Estimation of Consumption Expenditure in Albania. Working Paper 2010/14, Department of Statistics "G. Parenti", University of Florence.

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

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

Diggle, P.J., 1983. Statistical analysis of spatial point patterns. Academic Press, London.

Hastie, T.J., Tibshirani, R., 1990. Generalized Additive Models. Chapman & Hall, London.

Kammann, E.E., Wand, M.P., 2003. Geoadditive Models. Applied Statistics 52, 1–18.

Ligges, U., Thomas, A., Spiegelhalter, D., Best, N., Lunn, D., Rice, K., Sturtz, S., 2009. `BRugs 0.5-3`. R package.

Little, R.J.A., Rubin, D.B., 1987. Statistical analysis with Missing Data. John Wiley & Sons, New York.

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

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

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

R Development Core Team, 2010. R: A language and environment for statistical computing, R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

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

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

Salvati, N., Chandra, H., Ranalli, M.G., Chambers, R., 2010. Small area estimation using a nonparametric model-based direct estimator. Computational Statistics and Data Analysis 54, 2159–2171.

Spiegelhalter, D., Thomas, A., Best, N., Gilks, W., Lunn, D., 2003. BUGS: Bayesian inference using Gibbs sampling, MRC Biostatistics Unit, Cambridge, England.

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