An Imputation Method For Missing Covariate Data

Bocci, Chiara
University of Florence, Department of Statistics
Viale Morgagni, 59
50134 Firenze, Italy
E-mail: bocci@ds.unifi.it

Rocco, Emilia
University of Florence, Department of Statistics
Viale Morgagni, 59
50134 Firenze, Italy
E-mail: rocco@ds.unifi.it

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. Geoadditive models introduced by Kammann and Wand (2003), allow to 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. Therefore, 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, geoadditive models represent a powerful geostatistical methodology. However, their implementation needs the statistical units to be referenced at point locations and if we use them to produce model-based estimates of a parameter of interest for some geographical domains, the spatial location is required for all the population units. 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. 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 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.

Methodological Framework and Basic Assumption

Let t_i, $1 \leq i \leq n$, be a linear predictor of y_i at spatial location $\mathbf{s}_i$, $\mathbf{s} \in \mathbb{R}^2$. A geoadditive model for such data can be formulated as

$$y_i = \alpha + t_i\beta_t + h(\mathbf{s}_i) + \varepsilon_i, \qquad \varepsilon_i \sim N(0, \sigma^2_\varepsilon), \tag{1}$$

where h is an unspecified bivariate smooth functions. Representing $h(\cdot)$ with a low-rank thin plate

spline with K knots

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

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

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

with

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

where

$$\begin{aligned}\mathbf{X} &= \left[1, \mathbf{t}_i, \mathbf{s}_i^T\right]_{1\le i\le n},\\ \boldsymbol{\beta} &= \left[\beta_0, \beta_t, \boldsymbol{\beta}_s^T\right], \qquad \beta_0 = \alpha + \beta_{0s}\\ \mathbf{u} &= \left[u_1, ..., u_K\right],\end{aligned}$$

and $\mathbf{Z}$ is the matrix containing the spline basis functions, that is

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

where $C(\mathbf{v}) = \|\mathbf{v}\|^2 \log\|\mathbf{v}\|$ and $\boldsymbol{\kappa}_1, ..., \boldsymbol{\kappa}_K$ are the spline knots locations.

The amount of smoothing for the spline component of the model can be quantified through the variance components ratio $\sigma_\varepsilon^2/\sigma_s^2$.

The addition of others explicative variables is straightforward: smoothing components are added in the random effects term $\mathbf{Z}$, while linear components can be incorporated as fixed effects in the $\mathbf{X}$ 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.

Now, suppose to have a population of N units divided in Q regions, and to be interested in estimating 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 $\mathbf{s}$ 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 (1):

$$\begin{aligned}(3) \quad \hat{\bar{y}}_q &= \frac{1}{N_q}\left[\sum_{i\in S_q} y_i + \sum_{i\in R_q}\left(\mathbf{x}_i\hat{\boldsymbol{\beta}} + \mathbf{z}_i\hat{\mathbf{u}}\right)\right] =\\ &= \frac{1}{N_q}\left[\sum_{i\in S_q} y_i + \sum_{i\in R_q}\left(\hat{\beta}_0 + \hat{\beta}_t t_i + \mathbf{s}_{iq}^T\hat{\boldsymbol{\beta}}_s + \sum_{k=1}^{K}\hat{u}_k b_{tps}(\mathbf{s}_{iq}, \boldsymbol{\kappa}_k)\right)\right],\end{aligned}$$

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) as we don't know $\mathbf{s}$ for the not-sample units. In the classic approach the $\mathbf{s}_{iq}$ values are replaced with the region centroid $\mathbf{c}_q$, that is a constant for all the units in region q. Here we suggest to impute the $\mathbf{s}_i$ values through a stochastic Bayesian imputation procedure. Thus we adopt a hierarchical Bayesian formulation of model (1) (Ruppert et al, 2003) with a prior distribution $f_s(\boldsymbol{\theta}_q)$ for $\mathbf{s}_i$ inside each region q and then use the joint posterior distribution of all parameters given the data as the

basis of inference. Thus, under stochastic imputation, our complete hierarchical Bayesian formulation (following specifications of Crainiceanu at al, 2003) is

$$
\begin{aligned}
& y_i|\boldsymbol{\beta}, \mathbf{u}, \sigma_\varepsilon^2 \overset{\text{ind}}{\sim} N\left(\beta_0 + \beta_t t_i + \boldsymbol{\beta}_s^T \mathbf{s} + \sum_{k=1}^{K} u_k b_{tps}(\mathbf{s}, \boldsymbol{\kappa}_k), \sigma_\varepsilon^2\right), \\
& \mathbf{u}|\sigma_s^2 \sim N(0, \sigma_s^2 \mathbf{I}_K), \\
& s_i|\boldsymbol{\theta}_q \sim f_s(\boldsymbol{\theta}_q),
\end{aligned} \tag{4}
$$

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

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

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 the choice of the prior distribution $f_s(\boldsymbol{\theta}_q)$ identifies a specific imputation process. Anyway this choice may take into account the *a priori* knowledge about the spatial distribution of the studied phenomenon and the spatial distribution of the sampled points (that is their spatial representativeness). If the true function f_s is known, it is obviously "opportune" to use it, but this assumption is against our work hypothesis. If no information on the distribution of the true point locations is available but the sample can be assumed representative of it, a non-informative f_s which learns from the sample is a "good" choice. In particular if the spatial pattern is not too complex (like multi-modal or clustered distributions) the Beta distribution can be used. When the spatial pattern is more complex a more flexible (i.e. mixture model) prior should be used. However the use of a non-informative prior works well as long as the spatial distribution of the sampled points reflects the spatial distribution of the population points. When this condition is not satisfied, because the sampling design is not self-weighting and the inclusion probabilities depend on the spatial locations, a possible prior is a categorical distribution with each category corresponding to the coordinates of a sampled point and associated probability proportional to the corresponding sampling weights. The same solution could be adopted if the spatial non-representativeness of the sample does not depend on the sampling design but on non sampling errors as long as the survey weights can be estimated. When the relation between the spatial distribution of the sampled points and the spatial distribution of the population is unknown, the problem can only be solved making assumptions on this relation. Finally, we note that if for each region $f_s(\boldsymbol{\theta}_q)$ is a probability mass function that assumes value 1 when $\mathbf{s} =$ *centroid* and otherwise 0, then our formulation corresponds to the centroid imputation approach.

MCMC Experiments

In order to evaluate the performance of our approach with respect to the centroids classic approach, various MCMC experiments are implemented under different scenarios. All the analysis are implemented using the `OpenBUGS` Bayesian inference package. We access `OpenBUGS` using the package `BRugs` in the `R` computing environment.

All MCMC 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 $\varepsilon_i \sim N(0, \sigma_\varepsilon^2)$, $\sigma_\varepsilon = 0.2$, $\alpha = 10$, $\beta_x = 0.4$, $x \sim \text{Ber}(0.5)$ 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 and is represented in Figure 1(a).

- 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.

- $f(\mathbf{s})$ 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 (4) is fitted using a sample of $n = 500$ units and 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 100 times.

Each scenario differs from the others for the spatial point process used to generate $\mathbf{s}$ (an homogenous Poisson process, an inhomogeneous Poisson process and a cluster Poisson process) and for the sampling design (proportional stratified simple random sampling (PSRT) with strata corresponding to the regions and unequal probability (UP) design). Figures 1(b), 1(c) and 1(d) show the three different spatial patterns. Figure 2 shows the posterior density of the regional mean estimator under four different scenarios: (A) an homogeneous Poisson process for $\mathbf{s}$ and a PSTR sampling design; (B) an inhomogeneous Poisson process for $\mathbf{s}$ and a PSTR sampling design; (C) a cluster Poisson process for $\mathbf{s}$ and a PSTR sampling design; and (D) an homogenous Poisson process for $\mathbf{s}$ and an UP sampling design.

For each scenario the regional mean estimator is evaluated under three different types of imputation, that is considering three different *a priori* distributions for $\mathbf{s}$: uniform distribution, beta distribution and mass function on the centroid. In addition specifically for the scenario D we consider the imputation using a categorical distribution and the mean estimator (3) is modified in order to weight the observed y values with their corresponding sampling weights. For lack of space the Figure 2 shows the results only for 3 of the 9 geographic domains in which the study area is partitioned.

Concluding 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. 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 suggest 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. The missing spatial coordinates are then extracted from their posterior distribution, obtained by MCMC simulation.

Observing Figure 2 it is straightforward to note that, when the sampling design is self-weighting, if the imputation distribution corresponds to the population spatial distribution, the stochastic imputation approach produces better estimates than the classic centroid approach. This is the case of the

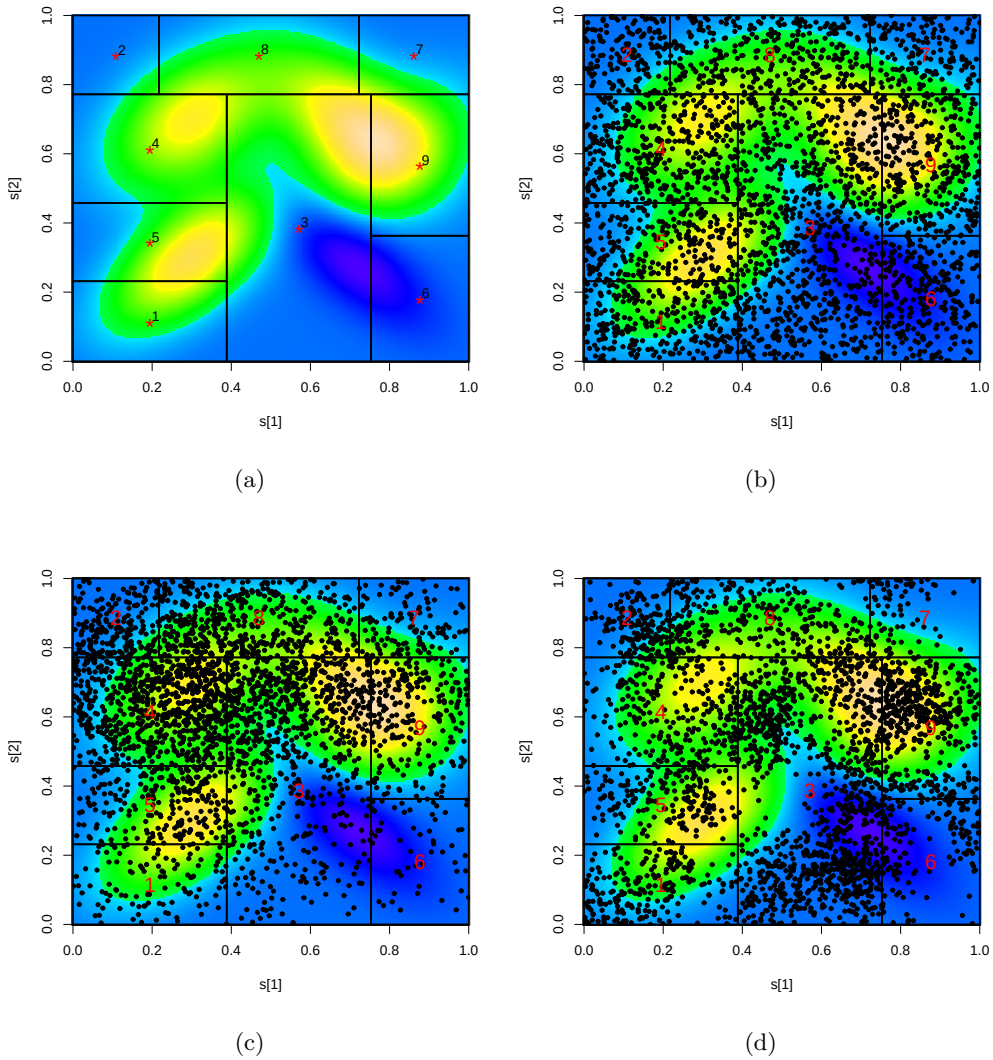
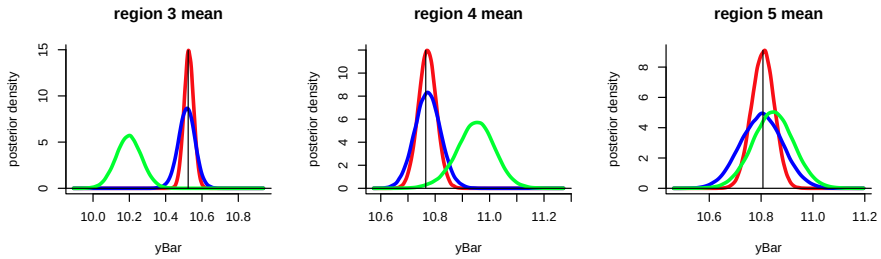


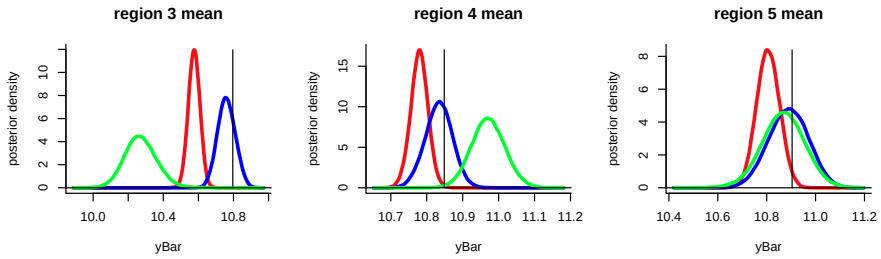
Figure 1: Spatial distributions of the population units: (a) homogeneous Poisson process, (b) inhomogeneous Poisson process, (c) inhomogeneous Poisson process on each region, (d) independent bivariate Beta distribution on each region.

Uniform approach in scenario A. Also the Beta imputation approach works well 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. When none of the imputation models corresponds to the population spatial distribution, 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, as seen in scenario D. Under such scenario, we obtain a good performance with the Categorical imputation approach that consider explicitly the sampling weights.

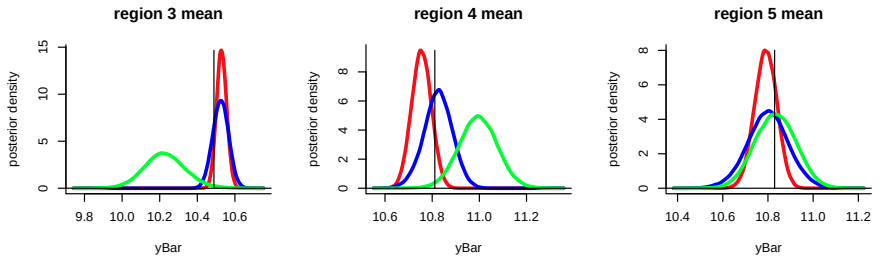
Finally when the spatial pattern is more complex the use of a more flexible prior may produce better results then the Beta imputation approach. Research in the use of a mixture of Beta distributions would certainly be our next step.



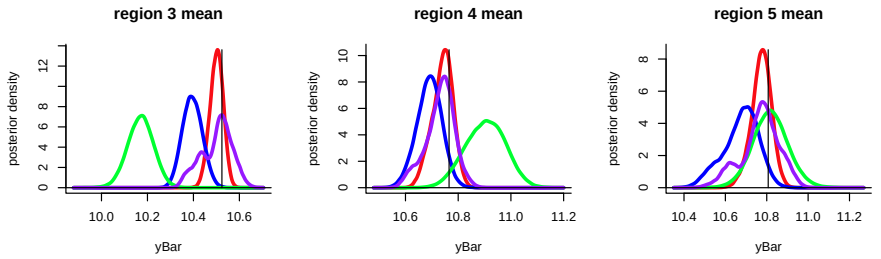
(a) Scenario A



(b) Scenario B



(c) Scenario C



(d) Scenario D

Figure 2: Posterior density of the regional model-based mean estimator under the four scenarios and for the four imputation approaches: Centroid (green line), Uniform (red line), Beta (blue line) and Categorical (purple line). The vertical lines indicate the true mean values.

REFERENCES

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.

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.

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