Bayesian Cross-Validation of Geostatistical Models

Viviana G R Lobo[1], Thaís C O Fonseca[1,*], Fernando A S Moura[1]

[1] *Department of Statistical Methods, Federal University of Rio de Janeiro, Rio de Janeiro, Brazil*

Abstract

The problem of validating or criticising models for georeferenced data is challenging as the conclusions may be sensitive to which partition of the data into training and validation cases is utilized. This is an obvious problem with the basic model validation scheme, since only a few out-of-sample locations are usually selected to validate the model. On the other hand, cross-validation approach which considers several possible configurations of data divided into training and validation observations is an appealing alternative, but it could be computationally demanding, as estimation of parameters usually requires computationally intensive methods.

This work proposes the use of cross-validation techniques to choose between competing models and to assess the goodness of fit of spatial models in different regions of the spatial domain. In particular, we consider uncertainty in the locations by assigning a probability distribution to them. To deal with the computational burden of cross-validation we propose the use of estimated discrepancy functions in a computationally efficient manner based on importance weighting posterior samples. Furthermore, we propose a stratified cross-validation scheme to take into account spatial heterogeneity, reducing the total variance of estimated predictive discrepancy measures considered for model assessment. We illustrate the advantages of our proposal with simulated examples of homogeneous and inhomogeneous spatial processes and with an application to a rainfall dataset in Rio de Janeiro.

Keywords: Bayesian inference, Data partition, Spatial processes, Model

[*] Corresponding Author: Av. Athos da Silveira Ramos. Centro de Tecnologia, Bloco C sala C114D, IM-UFRJ CEP 21941-909, Rio de Janeiro, Brazil.

Email addresses: viviana@dme.ufrj.br (Viviana G R Lobo), thais@im.ufrj.br (Thaís C O Fonseca), fmoura@im.ufrj.br (Fernando A S Moura)

criticism, Discrepancy function, Importance sampling.

1. Introduction

In many practical problems, the researcher is interested in modelling some phenomenon that occurred in space as a stochastic process. The usual model criticism is done through model comparison and prediction for a few out-of-sample observations. These model checks are often not able to assess whether the assumed model is plausible for the data in the whole spatial domain. From a theoretical viewpoint, statistical inference should go beyond parameter estimation and prediction (see Robert, 2007, page 343). Notice that if hypothesis testing is performed regarding parameters from models which are not adequate to the data then the conclusions from the tests are not meaningful. In this context, checking the goodness of fit of the assumed model is an important step. However, in the geostatistical context, this is a challenging task since only one realization of the process is available for both parameter estimation and model checking.

The usual approaches for model checking in spatial statistics are based on selecting a subset from the locations to make prediction with the assumed model. The observed values which were left out of the estimation procedure are then compared with the predictions. However, the choice of locations used for model fitting and prediction is usually ad-hoc. Some examples can be seen in the literature such as in multivariate random fields context, Majumdar and Gelfand (2007) and Apanasovich and Genton (2010) considered 68 monitoring stations used for estimation and take out 5 locations for verification purposes using pollution data. In the spatial-temporal context, Fonseca and Steel (2011) and Bueno et al. (2017) used the same idea to check the non-gaussian models using 67 locations for parameter estimation and they leave out three locations for predictive performance assessment in temperature data. Ideally, model validation techniques should allow for assessing the goodness of fit of spatial models in different regions of the spatial domain. Diggle (2014) points out that if a spatial model fits the data well, it can be used to generate datasets which are statistically similar to the observed sample. This idea suggests that cross-validation techniques are potentially useful tools for model checking.

In the literature, various authors have suggested the use of cross-validation for modelling univariate data. Burman (1989) introduces validation techniques in a study of optimal transformation of variables, based on k-fold cross-validation and repeated learning testing methods. Thall et al. (1997) demonstrated that repeated data splitting is preferred over k-fold cross-validation. They propose to apply cross-validation to a very large number of randomly generated partitions of the data. The conditional predictive ordinate (CPO) proposed by Gelfand (1996) is a very useful model assessment tool which has been widely used in the statistical literature under various contexts, such as in the detection of surprising observations. Conditional predictive ordinate is based on leave-one-out cross-validation (LOO-CV).

From a Bayesian standpoint, Marshall and Spiegelhalter (2003) and Burman (1989), amongst others, show that the cross-validation can be computationally very expensive, since a full MCMC analysis has to be repeated, leaving out each in turn validation set. Stern and Cressie (2000) considered importance weighting and re-sampling methods in the context of posterior predictive model checking via CPO and posterior predictive p value. Gelman et al. (2014) review some information criteria, such as Akaike, deviance and Watanabe-Akaike (WAIC - Watanabe (2010)) from a Bayesian perspective, using out-of-sample and LOO-CV techniques. In the same context, Li et al. (2016) discussed two predictive evaluation methods based on Importance Sampling (Gelfand et al. (1992)) and WAIC in Bayesian models with possibly correlated latent variables via LOO-CV and Vehtari et al. (2017) used the sample approach introducing an efficient computation of LOO-CV using Pareto-smoothed importance sampling to measure the predictive accuracy in Bayesian models. In the context of accounting for uncertainty in the choice of validations sets, Alqallaf and Gustafson (2001) propose Bayesian cross-validation for several data partitions sampled from the prior distribution of the possible sets of training and validation cases. The model checking is based on estimating discrepancy functions, which are statistical measures commonly used in the literature for model comparison.

Although many papers have exploited cross-validation methods for univariate data, this is not the case for spatial data analysis. For instance, the usual setup for model checking in geostatistics is to make prediction for one or a few selected validation sets. However, the choice of observation sites for validation of spatial models is not always “robust” to the considered sampling or allocation of sites. In general, it does not consider the sampling process that generated the locations. In fact, models that ignore information about sample selection can lead to biased inferences and predictions (Diggle et al., 2010; Ferreira and Gamerman, 2015). Pfeffermann et al. (2006) discuss this problem in the context of a finite superpopulation model.

The use of cross-validation techniques in a large volume of spatial data becomes a computational challenge, due to the difficulty of applying traditional prediction methods in a time-tolerant boundary. If we were to make prediction for several vectors of points, the cross-validation procedure would be repeated again for all possible selected configurations of training and validation samples. For most geostatistics problems, this scheme becomes computationally prohibitive. Thus, more sophisticated approach are useful, both to reduce the final cost and increase efficiency.

In this work, we propose to use cross-validation techniques to choose between geostatistical models and to assess the goodness of fit of spatial models in different regions of the spatial domain. Our proposal extends the work of Alqallaf and Gustafson (2001) to correlated data modelling. In this context, we allow for uncertainty in the selection of the validation sets in spatial data analysis by considering a probability distribution for the spatial locations. In particular, we propose three distributions for the spatial locations. The first proposal is uniform on the spatial domain. The second proposal is a conditional distribution which is based on distances from already selected points aiming a better coverage of the spatial region of interest. The third proposal is uniform in several strata. For that purpose, we adopt spatial stratified sampling, where the possibly heterogeneous area is divided into several subareas more homogeneous than the whole area, reducing the total variance of estimated predictive discrepancy measures. Besides, this proposal allow for identification of sub-regions where the model has a poor predictive performance. This may be used as a tool to indicate outliers or non-stationarity. To deal with the computational burden of cross-validation techniques, we propose an efficient algorithm based on importance weighting and only a handful of MCMC (Markov Chain Monte Carlo) runs.

This paper is organized as follows. In Section 2, an illustration which motivates this work is presented. Section 3 briefly reviews the main aspects of spatial data analysis, namely, basic geostatistical models, spatial arrangements and inference. Section 4 and 5 describe Bayesian cross-validation using expected discrepancy estimation via MCMC (*Markov Chain Monte Carlo*), and report a procedure for validating models based on stratified spatial data. In particular, scheme based on stratification aims to allow for: (i) spatial heterogeneity, (ii) reduction in total variance of estimated predictive discrepancy measures considered for model assessment. In Section 6, simulated examples are presented. Finally, Section 7 and 8 show an application to a rainfall dataset and conclusions, respectively.

2. Motivation

In this work, the spatial locations are assumed to be a random sample from a specified distribution and locations are sampled to compose the validation set according to the prior probabilities of the sets.

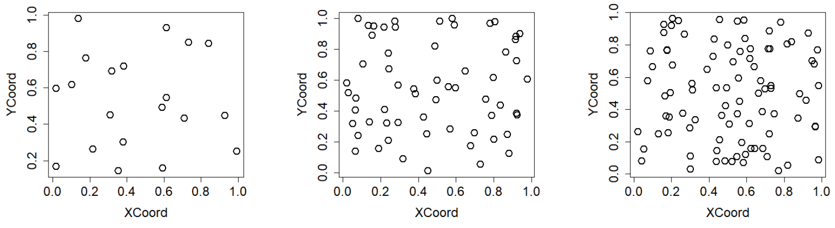
As follows we present an illustrative example that shows that the usual validation setup in spatial data analysis might select, with quite high probability, a model which is not the best option for a certain application. This happens mostly if the uncertainty in the choice of locations for model validation is not taken into account.

2.1. An illustrative example: Uncertainty of Data Partition

Consider location $x_1, \ldots, x_n$ randomly simulated in a unit square with $n = (20, 60, 90)$ within an irregular grid. Responses $Y = (Y(x_1), \ldots, Y(x_n))$ are generated from a Student-t process, that is,

$$Y \mid \sigma^2, \phi, \nu \sim ST\left(\mathbf{0}, \nu, \sigma^2 R\right), \tag{1}$$

with R the correlation matrix with elements $r_{ij} = \exp\{-||x_i - x_j||/\phi\}$ and range parameter $\phi > 0$, which determines the rate at which the correlation between observations decreases as distances grow. The decay is faster for small values for ϕ and smoother for larger values for ϕ (more details see Appendix A). In this example, the parameters are set to $\nu = 3$, $\sigma^2 = 1$ and varying values for $\phi = (0.05, 0.30, 0.70)$. Figures 1 and 2 present the spatial arrangement and the three different correlation functions considered, respectively.



(a) $n = 20$ (b) $n = 60$ (c) $n = 90$

Figure 1: Sample locations randomly generated within the unit square for three datasets with n=20, 60 and 90.

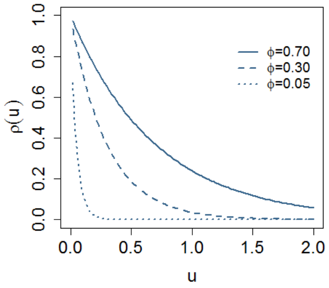


Figure 2: Exponential correlation function $\rho(u) = exp\{-u/\phi\}$ with varying range ϕ.

For prediction purpose, we randomly choose $I = 100$ validation configurations of all possible subsets of size n_V. Notice that there are $\binom{n}{n_V}$ possible validation sets, thus performing cross-validation for all possibilities is too time consuming. For this arbitrarily chosen configurations, we omitted randomly $0.05n$ and $0.25n$ points (the validation sample) and calculated the predicted value for these locations using the remaining points (the training sample). We fitted the Gaussian (GM) and Student-t (STM) models to the simulated data and used MCMC techniques for estimating the model parameters for each of the 100 sets at each data configuration (n, ϕ, n_V). The prior distributions considered for STM and GM models are $\sigma^{-2} \sim G(0.1; 0.1)$ and $\phi \sim G(1, 0.22/med(u))$, with $med(u)$ representing the median of distances in the data. The parameter ν has a Jeffreys prior distribution as proposed in Fonseca et al. (2008). The prior distribution assigned to σ^2, ϕ and ν, can be seen in Appendix C. For model assessment we consider the Mahalanobis-Distance (Mahalanobis (1936) - see more details in Appendix B) as the discrepancy measure (D) which takes into account the correlation between observations. Thus, the model choice will depends on the difference in the discrepancy functions for each model which is given by

$$\delta^{(i)} = D^{(i)}_{STM} - D^{(i)}_{GM}, \quad i = 1, \dots, I, \tag{2}$$

where I is the amount of validation configurations and D is a discrepancy measure, so that if $\delta^{(i)} < 0$ we have that STM is preferable than GM. Figure 3 presents the box-plots for 100 randomly selected validation configurations for cross-validation performance varying n, ϕ and n_V. Notice that we have different results for each choice of validation sample size and parameter ϕ. If we consider the validation set with size $n_V = 5\%n$, the percentages of wrong decisions are considerably larger than when we consider $n_V = 25\%n$. The percentages of wrong decisions are also larger when ϕ is large, for example, $\phi = 0.70$, which indicates that the larger the spatial correlation the more difficult it is to choose between Gaussian and Student-t models. According to Breusch et al. (1997), similar inferences are made about the mean μ under GM and STM, but different inferences can be made about scale, because the scale is differently represented in the two models. Thus, each model used for prediction might lead to different prediction intervals for ungauged locations. This difference in the inference and predictions also depends on the range parameter as indicated by our simulated illustration.

In the context of spatial data analysis, the influence of location arrangement in the region of interest must be taken into account. In addition, the size of the validation sample and the value of the range parameter seem to be crucial for model choice. This illustrative example motivates our work to incorporate the uncertainty in the validation set in the full Bayesian inference for the unknowns which besides the model parameter and predictions include the partition in validation and training sets.

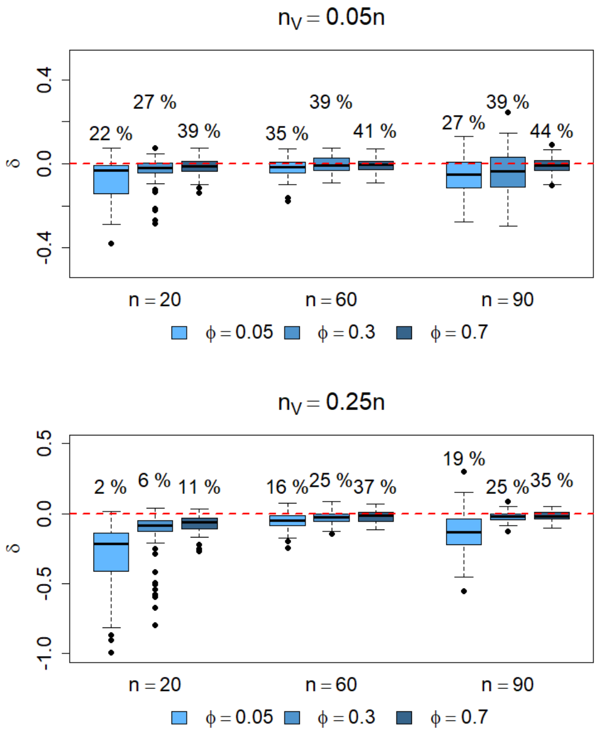


Figure 3: Cross-validation performance: box-plots of predictive discrepancy measure for GM versus STM varying n, ϕ and n_V. First row represents n_V equal to $0.05n$ and second row represents n_V equal to $0.25n$. Values of δ below the dashed line imply that the STM model is preferable. Numbers represent the percentage of times the wrong model (Gaussian) was selected when considering the discrepancy function in equation (2).

3. Basic geostatistical model

Let us consider that data are obtained by sampling a spatially continuous phenomenon $S(x)$ at a finite number of locations $x_1, \ldots, x_n$ which varies continuously within a region A. Hence, if Y_i denotes the measured value at the location x_i, a simple model for the data takes the form

$$Y_i = \mu + S(x_i) + Z_i \quad i = 1, \ldots, n, \tag{3}$$

where μ represent the mean and the $Z_i's$ are mutually independent, zero-mean random variables with variance τ^2 called nugget effect, which can be interpreted as sampling error or inherent geological variability (or both).

The underlying spatial process $\{S(x) : x \in \mathbb{R}^2\}$ is a stationary process with zero mean, constant variance σ^2 and correlation function $\rho(u; \phi)$, where ϕ is the correlation parameter and u is the distance between two locations. If Gaussianity is assumed, $Y \sim N_n(\mu\mathbf{1}, \sigma^2 R + \tau^2 I_n)$ where R represents the correlation matrix with elements $r_{ij} = \rho(||x_i - x_j||; \phi)$ and diagonal matrix I_n.

Gaussian stochastic processes are commonly used in practice for geostatistical data due to the facilities coming from the properties of the multivariate Gaussian distributions. Although the Gaussian process is mathematically convenient, its assumption can be very restrictive and the data may often present non-Gaussian characteristics, see Fonseca and Steel (2011) for details.

The next subsection presents a benchmark model for a spatial data analysis which allows for non-Gaussian behaviour of spatial data. This model is compared using the efficient cross-validation techniques schemes proposed in this paper. Inference for model parameters and predictive distributions are also described.

3.1. Spatial mixture model

As follows we consider three model specifications for spatial data analysis: the Gaussian, the Student-t and the Gaussian-log-Gaussian processes. These models might be written as spatial mixture models, with the base model being the Gaussian usual setup.

(M1) Gaussian model: As a benchmark we assume the Gaussian model. The distribution of $\mathbf{Y}$ is

$$\mathbf{y} \mid \mu, \sigma^2, \phi \sim N\left(\mathbf{1}\mu, \tau^2 I_n + \sigma^2 R\right). \tag{4}$$

(M2) Student-t model: As an alternative to Gaussianity we assume a Student-t model with ν degrees of freedom. Notice that for $\nu \to \infty$ we recover the Gaussian model. The distribution of $\mathbf{Y}$ is

$$\mathbf{y} \mid \mu, \sigma^2, \phi, \nu \sim ST\left(\mathbf{1}\mu, \nu, \tau^2 I_n + \sigma^2 R\right). \tag{5}$$

Similar to the Gaussian process, the Student-t process has the advantage of depending on the mean and covariance functions. Details about the Student-t process in a non-Bayesian context can be seen in Roislien and Omre (2006).

(M3) Gaussian-Log-Gaussian model: As proposed by Palacios and Steel (2006), this process is able to capture heterogeneity in space through a mixing process used to increase the Gaussian process variability,

$$\mathbf{y} \mid \mu, \sigma^2, \phi, \boldsymbol{\Delta} \sim N\left(\mathbf{1}\mu, \tau^2 I_n + \sigma^2(\boldsymbol{\Delta}^{-1/2}\, R\, \boldsymbol{\Delta}^{-1/2})\right). \tag{6}$$

This model assumes $\boldsymbol{\Delta} = diag(\delta(x_1), \ldots, \delta(x_n))$ and $ln(\boldsymbol{\delta}) \sim N_n\left(-\frac{\upsilon}{2}\mathbf{1}, \upsilon\, R\right)$. This mixing generates a multivariate scale mixture of Normals. Properties, estimation and prediction for the GLG model are introduced in Palacios and Steel (2006) and extended to the space-time case in Fonseca and Steel (2011). The $\upsilon \in \mathbb{R}^+$ is a scalar parameter introduced into the distribution $ln(\boldsymbol{\delta})$ and variation inflation is achieved when it is close to zero.

3.2. Inference for geostatistical models

In this work, we follow the Bayesian approach to inference and prediction. In the context of geostatistical models, the posterior distribution of model parameters θ, $p(\theta \mid \mathbf{y}) \propto f(\mathbf{y} \mid \theta)\pi(\theta)$, is not obtained in closed form, and stochastic simulation methods are often considered (Gamerman and Lopes, 2006).

In the simulated study and in our application, we assume the exponential correlation function given by

$$\rho(||u||, \phi) = exp\left\{-\frac{||u||}{\phi}\right\}, \tag{7}$$

where $\phi > 0$ represents the range parameter which controls the rate of decay with distance u, i.e., the distance at which there is essentially no spatial correlation.

For all models we assign the same independent non-informative priors to μ, σ^2 and ϕ. In particular, $\mu \sim N_n\left(0, \tau_\mu^2\right)$ with large value of τ_μ^2, $\sigma^{-2} \sim Ga(a, b)$ and $\tau^{-2} \sim Ga(a, b)$ with small values of a and b. For the range parameter ϕ we take into account that the prior is critically dependent on the scale of distances between locations. So, $\phi \sim Ga\left(1, c/med(d)\right)$, with $med(d)$ representing the median of distances in the data.

If Gaussianity is assumed for $S(x)$ as in equation (4), then the likelihood function for the spatial model is given by

$$f(\mathbf{y} \mid \mu, \sigma^2, \phi) = (2\pi)^{-n/2}\, |\Sigma|^{-1/2}\, exp\left\{-\frac{1}{2}(\mathbf{y} - \mathbf{1}\mu)'\Sigma^{-1}(\mathbf{y} - \mathbf{1}\mu)\right\}, \tag{8}$$

that is, $\mathbf{y} = (y_1, \ldots, y_n)'$ follows an n-variate Normal distribution with mean μ and covariance matrix $\Sigma = \tau^2 I_n + \sigma^2 R$. The posterior samples for model parameters μ, σ^2, ϕ are obtained by the Gibbs algorithm with Metropolis-Hastings steps considering random walk proposals. Further details about the MCMC scheme are deferred in Appendix C.

The likelihood function of the Student-t spatial process is given by

$$f(\mathbf{y} \mid \mu, \sigma^2, \phi, \nu) = \frac{\Gamma(\frac{\nu+n}{2})}{\Gamma(\frac{\nu}{2})(\nu\pi)^{n/2}|\Sigma|^{1/2}} \left[1 + \frac{(\mathbf{y} - \mathbf{1}\mu)'\Sigma^{-1}(\mathbf{y} - \mathbf{1}\mu)}{\nu}\right]^{-(\nu+n)/2}, \tag{9}$$

with $\Gamma(\cdot)$ the gamma function, mean μ and covariance matrix $\Sigma = \tau^2 I_n + \sigma^2 R$. For the degrees of freedom parameter ν we assign a Jeffreys prior distribution, as proposed in (Fonseca et al., 2008) and detailed in Appendix C.1. The posterior samples for model parameters μ, σ^2, ϕ, ν are obtained by the Gibbs algorithm with Metropolis-Hastings steps considering random walk proposals.

For the Gaussian-log-Gaussian spatial process, we assume a mixing variable $\delta_i \in \mathbb{R}_+$ assigned to each observation $i = 1, \ldots, n$, yielding to a multivariate Gaussian distribution for $\mathbf{y}$ conditional on $\boldsymbol{\delta} = (\delta_1, \ldots, \delta_n)$. The resulting likelihood function looks like equation (8) with Σ replaced by $\Sigma = \tau^2 I_n + \sigma^2(\boldsymbol{\Delta}^{-1/2} R \boldsymbol{\Delta}^{-1/2})$, with $\boldsymbol{\Delta} = Diag(\delta_1, \ldots, \delta_n)$. For the parameter υ we set a $GIG(0, \delta, \iota)$ (generalized inverse Gaussian) prior. Notice that very small values of υ (around 0.01) correspond to near Normality while large values of υ (of the order of say 3) indicate very thick tails and $ln(\boldsymbol{\delta}) \sim N_n(-\frac{\upsilon}{2}\mathbf{1}, \upsilon R)$. The posterior samples for model parameters are obtained by the Gibbs algorithm with Metropolis-Hastings steps for ϕ, υ and $\boldsymbol{\delta}$ which are based on random walk proposals. For a more elaborate algorithm, see Palacios and Steel (2006). Appendix C presents the prior distributions and posterior inference for the model parameters.

In the context of prediction, let $\mathbf{y} = (y_0, \mathbf{y}_s)$ where y_0 represents out-of-sample observations for which we want to obtain predictions and $\mathbf{y}_s$ represents the observations used for parameter estimation. Conditional predictive distributions are obtained in closed form for all considered models. For the Gaussian model the conditional distributions remain Gaussian with mean and variance given by

$$E[Y_0 \mid \mathbf{y}_s] = \mu_0 + \Sigma_{0s}\Sigma_{ss}^{-1}(\mathbf{y}_s - \mathbf{1}\mu_s) \tag{10}$$

$$Var[Y_0 \mid \mathbf{y}_s] = \Sigma_{00} - \Sigma_{0s}\Sigma_{ss}^{-1}\Sigma_{s0}, \tag{11}$$

where we have partitioned

$$\mu = \begin{pmatrix} \mu_0 \\ \mu_s \end{pmatrix} \qquad \Sigma = \begin{pmatrix} \Sigma_{00} & \Sigma_{0s} \\ \Sigma_{s0} & \Sigma_{ss} \end{pmatrix}$$

For the Student-t model the conditional distributions remain Student-t with degrees of freedom $\nu_{0|s} = \nu + d_s$, with mean and variance given by

$$E[Y_0 \mid \mathbf{y}_s] = \mu_0 + \Sigma_{0s}\Sigma_{ss}^{-1}(\mathbf{y}_s - \mathbf{1}\mu_s), \tag{12}$$

$$Var[Y_0 \mid \mathbf{y}_s] = \xi(s)\left[\Sigma_{00} - \Sigma_{0s}\Sigma_{ss}^{-1}\Sigma_{s0}\right], \tag{13}$$

with

$$\xi(s) = \frac{\nu + (\mathbf{y}_s - \mathbf{1}\mu_s)'\Sigma_{ss}^{-1}(\mathbf{y}_s - \mathbf{1}\mu_s)}{\nu + d_s},$$

which d_s represents the dimension of vector $\mathbf{y}_s$. Note that by letting ν go to infinity, we can recover the Gaussian conditional covariance structure. See Roislien and Omre (2006) for details.

For the Gaussian-Log-Gaussian model case and conditional on the mixing variables $\boldsymbol{\delta}$, the predictive distributions are analogous to (10) and (11) with $\Sigma = \tau^2 I_n + \sigma^2(\boldsymbol{\Delta}^{-1/2}\, R\, \boldsymbol{\Delta}^{-1/2})$. The mixing variables $\boldsymbol{\delta}$ are considered latent variables and are sampled in the MCMC algorithm, details are deferred to Appendix C.1.

4. Cross-validation of Bayesian models for spatially correlated data

We consider the uncertainty in the choice of data split into validation and training sets by defining a prior distribution of such sets. In the Bayesian analysis of spatial data, the fit of the model usually requires MCMC sampling from the posterior distribution. We extend the technique proposed by Alqallaf and Gustafson (2001) to spatially correlated data, so the validation measure does not require a separate posterior sample for each training sample.

4.1. General Development

Suppose that the observed data consist of responses $\mathbf{y} = (y_1, y_2, \ldots, y_n)$ arising from the process $Y(x)$ and locations $x = (x_1, \ldots, x_n)$ described in equation (3). For a given model, let θ and $\mathbf{y}^{rep}$ be the *vector of model parameters* and the *replicated response* of a hypothetical realization of the response vector, respectively. We define *split* $\mathbf{s}$ as a $0-1$ vector, which divides the n cases into training and validation vectors. We adopt $T[\mathbf{s}]$ and $V[\mathbf{s}]$ to denote the training and validation cases in each split vector considered.

For the purpose of building the split vectors, we denote the sample sizes of training and validation by n_T and n_V, respectively. In this case, n_T and n_V are fixed and $n = n_T + n_V$. In spatially correlated data, we define the specific split vector as

$$s_k = \begin{cases} 0, & x_k \text{ is a training location} \\ 1, & \text{otherwise,} \end{cases}$$

and the split vector $\mathbf{s} = (s_1, \ldots, s_n)$ of the same dimension of observed data, indicating for each locations x_k, $k = 1, \ldots, n$, if y_k is used for training ($k \in T[\mathbf{s}]$), or y_k is used for validation ($k \in V[\mathbf{s}]$).

Our goal is to average of cross-validation results considering many data partitions. Indeed, this averaging is done with respect to $p(\mathbf{s})$, the distribution of $\mathbf{s}$, that is, $p(\mathbf{s}) = p(s_1, s_2, \ldots, s_n)$. A first approach would be to consider the distribution $p(\mathbf{s})$ to be the uniform distribution over such splits. Under this assumption,

$$p(\mathbf{s}) = \binom{n}{n_T}^{-1}, \quad \text{if} \quad \sum_{j=1}^{n} s_j = n_T.$$

Notice that this choice of prior might not be reasonable if there is a pattern in $x = (x_1, \ldots, x_n)$ as in the case of inhomogeneous processes. To account for this possible feature of spatial data, we consider an extension of this prior in Section 5. An alternative is to assume a probability distribution for the sets based on Euclidean distances considering a finite set of spatial sample locations $\tilde{x}_k = (x_0, x_1, \ldots, x_{k-1})$, for $k = 1, \ldots, n$ within a region of interest. This idea derives from the cluster selection via the K-means ++ method (Arthur and Vassilvitskii (2007)). A first location (x_0) is sampled based in an unconditional prior with probability $p(x_0) = \frac{1}{n}$ and the other locations are sequentially sampled based on a conditional prior over the already sampled locations, that is, $p(x_k \mid \tilde{x}_k)$, for $k = 1, \ldots, n$. The prior via distances can be obtained as

$$p(\mathbf{s}) = p(x_0)p(x_1 \mid \tilde{x}_1)p(x_2 \mid \tilde{x}_2) \ldots p(x_k \mid \tilde{x}_k),$$

where n_T is the training sample size and x_0 is the starting point selected with probability $p(x_0) = \frac{1}{n}$. We select the locations x_k with probability given by

$$p(x_k \mid \tilde{x}_k) = \prod_{j=1}^{k} \left\{ \frac{min\left\{|x_j - x_0|, \ldots, |x_j - x_{k-1}|\right\}}{\sum_{x_j \in \tilde{x}_k} min\left\{|x_j - x_0|, \ldots, |x_j - x_{k-1}|\right\}} \right\},$$

as in the case of the K-means ++ method. This prior assumes different probabilities for the sample selection locations and could be potentially useful in irregular spatial regions as often seen in data applications.

After the choice of a specific split vector $\mathbf{s}$, let $\mathbf{y}_{T[\mathbf{s}]}$ and $\mathbf{y}_{V[\mathbf{s}]}$ be defined as the observed training and validation cases. Given the split $\mathbf{s}$, $p(\theta \mid \mathbf{y}_{T[\mathbf{s}]})$ is defined as the posterior distribution of θ given the training data only. Thus, using Bayes theorem, the posterior distribution is given by

$$p(\theta \mid \mathbf{y}_{T[\mathbf{s}]}) \propto f(\mathbf{y}_{T[\mathbf{s}]} \mid \theta)\pi(\theta), \tag{14}$$

where for each split vector $\mathbf{s}$ there is a single corresponding data vector $\mathbf{y}_{T[\mathbf{s}]}$.

Notice that $\mathbf{y}^{rep}$ is simply distributed according to the sampling model assumed for the data, i.e., $[\mathbf{y}^{rep} \mid \theta, \mathbf{y}_{T[\mathbf{s}]}]$, which represents the predictive distribution given the training data, for a specific split vector $\mathbf{s}$. This distribution is used to obtain samples from the marginal predictive density $p(\mathbf{y}^{rep} \mid \mathbf{y}_{T[\mathbf{s}]})$ in a composition sampling algorithm.

4.2. Expected discrepancy estimation

Our cross-validation assessments are based on $r(\mathbf{y}^{rep}, \mathbf{y}_{V[\mathbf{s}]})$, called a discrepancy function for checking model adequacy, whose expectation under $f(\mathbf{y}^{rep} \mid \mathbf{y}_{T[\mathbf{s}]})$ is evaluated. It requires the distribution of the replicated response vector $\mathbf{y}^{rep}$. In particular, we are interested in computing the expectation bellow

$$\Psi = E\left\{r(\mathbf{y}^{rep}, \mathbf{y}_{V[\mathbf{s}]})\right\}. \tag{15}$$

The expected value in (15) represents a statistical measure for comparing Bayesian models and r represents a discrepancy function. Notice that r

depends on two unknown quantities $\mathbf{y}^{rep}$ and $\mathbf{s}$, thus Ψ can be computed as

$$\begin{aligned}\Psi &= \int \sum_{\mathbf{s}\in S} r(\mathbf{y}^{rep}, \mathbf{y}_{V[\mathbf{s}]}) f(\mathbf{y}^{rep} \mid \mathbf{y}_{T[\mathbf{s}]}) p(\mathbf{s})\, d\mathbf{y}^{rep} \\ &= \sum_{\mathbf{s}\in S} E\left[r(\mathbf{y}^{rep}, \mathbf{y}_{V[\mathbf{s}]}) \mid \mathbf{y}_{T[\mathbf{s}]}\right]\, p(\mathbf{s})\end{aligned}$$

The Monte Carlo estimator for the expected discrepancy is given by

$$\hat{\Psi} = \frac{1}{I}\sum_{i=1}^{I} E\left\{r(\mathbf{y}^{rep}, \mathbf{y}_{V[\mathbf{s}^{(i)}]}) \mid \mathbf{y}_{T[\mathbf{s}^{(i)}]}\right\}. \tag{16}$$

The split vectors $\mathbf{s}^{(1)}, \mathbf{s}^{(2)} \ldots, \mathbf{s}^{(I)}$ are simulated independently from $p(\mathbf{s})$ and I represents the number of splits. The Monte Carlo estimator for the expectancy of interest is well known to be unbiased. If the posterior distribution $f(\mathbf{y}^{rep} \mid \mathbf{y}_{T[\mathbf{s}]})$ is not available analytically, then methods based on stochastic simulation can be employed to obtain samples from the posterior of interest. Notice that the expected discrepancy of interest may be rewritten as

$$\Psi = \int\int \sum_{\mathbf{s}\in S} r(\mathbf{y}^{rep}, \mathbf{y}_{V[\mathbf{s}]}) f(\mathbf{y}^{rep} \mid \theta, \mathbf{y}_{T[\mathbf{s}]}) p(\theta \mid \mathbf{y}_{T[\mathbf{s}]}) p(\mathbf{s})\, d\theta\, d\mathbf{y}^{rep}$$

Let $(\theta_{ij}, \mathbf{y}_{ij}^{rep})$, $i = 1, \ldots, I$ and $j = 1, \ldots, J$ be samples from the joint conditional distribution of θ and $\mathbf{y}^{rep}$, $f(\mathbf{y}^{rep} \mid \theta, \mathbf{y}_{T[\mathbf{s}]}) p(\theta \mid \mathbf{y}_{T[\mathbf{s}]})$, then the Algorithm 1 describes how to compute (16) by simulating from the posterior distribution of model parameters via MCMC. This approach is based on obtaining one MCMC sample for each split. We call this estimator by the *monte carlo* (MC) estimator

$$\hat{\Psi}_{mc} = \frac{1}{I}\sum_{i=1}^{I}\frac{1}{J}\sum_{j=1}^{J} r(y_{ij}^{rep}, \mathbf{y}_{V[\mathbf{s}^{(i)}]}), \tag{17}$$

where I and J represent the number of splits and size of the posterior sample, respectively. The MC estimator is an unbiased estimator of expression (15). Notice that (17) requires a MCMC sample for each validation set sampled from $p(\mathbf{s})$. This if often very expensive. Next, we present some alternatives

```
Algorithm 1: Monte Carlo (MC) estimator
  1. Simulate independent split vectors $\mathbf{s}^{(1)}, \mathbf{s}^{(2)}, \ldots, \mathbf{s}^{(I)}$ from $p(\mathbf{s})$;
  2. for each $\mathbf{s}^{(i)}$ do
      use a MCMC run to draw a sample $\theta_{i1}, \ldots, \theta_{iJ}$ from $p(\theta \mid \mathbf{y}_{T[\mathbf{s}^{(i)}]})$;
  end
  3. for each $(i, j)$ do
      simulate $y_{ij}^{rep}$ from $p(\mathbf{y}^{rep} \mid \theta = \theta_{ij}, \mathbf{y}_{T[\mathbf{s}^{(i)}]})$;
  end
```

to (15), which are computed from samples from the posterior $p(\theta \mid \mathbf{y})$ and $p(\mathbf{y}^{rep} \mid \theta, \mathbf{y})$. The number of splits I and the size of the posterior sample for each split J must be specified.

For illustrating the method, we applied algorithm 1 to the motivation example of subsection 2.1. Table 1 presents the Monte Carlo estimate based on the 100 validation sets, using the discrepancy measure D and varying n, ϕ and n_V for each model. We can verify that the STM best fits the dataset for all cases. Indeed, this is the model that was used to generate the data. Notice, however, that particularly for this illustration, it was run 100 MCMC chains for each data configuration $(n \times \phi \times n_V)$ and each competing model. This is too time consuming even for these small spatial datasets.

Table 1: MC estimate based on algorithm 1 for 100 validation sets using the Mahalanobis-Distance for each model and varying n, ϕ and n_V. The model that best fits the data is the one which presents smaller values in the measure.

		ϕ				ϕ		
	5%n	0.05	0.30	0.70	25%n	0.05	0.30	0.70
$n = 20$	GM	1.287	1.121	1.850	GM	3.519	3.080	3.173
	STM	1.214	1.092	1.168	STM	3.216	2.961	3.102
$n = 60$	GM	2.363	2.233	2.133	GM	5.653	5.284	5.179
	STM	2.344	2.226	2.125	STM	5.599	5.256	5.156
$n = 90$	GM	2.697	6.517	2.552	GM	6.777	6.518	6.466
	STM	2.642	6.471	2.546	STM	6.637	6.497	6.449

Aiming to reducing the computational cost, we consider the *importance* *sample estimator* (SIR), which requires only a handful of MCMC runs as an alternative estimate of expression (15). The idea is to approximate the posterior density of a given training sample by a distribution based heuristically on the same amount of data, but which does not depend on the specific split $\mathbf{s}$. In particular, this distribution is used as an importance function and is defined as

$$g(\theta) \propto f(\mathbf{y} \mid \theta)^{\alpha}\pi(\theta), \tag{18}$$

where $f(\mathbf{y} \mid \theta)$ denotes the likelihood function for the complete data, $\pi(\theta)$ is the prior distribution and $\alpha = n_T/n$ with n_T fixed. Notice that if both n and n_T are large, the likelihood $f(\mathbf{y}_T \mid \theta)$ based only on the training sample $\mathbf{y}_T$ will approximate to the full likelihood $f(\mathbf{y} \mid \theta)$ raised to the power α.

Alqallaf and Gustafson (2001) claim that raising the whole-data likelihood to the power α has the effect of flattening the posterior to a degree commensurate with conditioning only on a fraction α of the data. The function $g(\theta)$ is the same function employed in fractional Bayes factor for model comparison. In that context, O'Hagan (1995) proposed using a fractional part of the entire likelihood, $f(\mathbf{y} \mid \theta)^{\alpha}$, instead of the training sample. Sampling importance resampling is considered to obtain a sample from the desired posterior distribution using the approximation in (18). The Algorithm 2 describes how to compute the SIR estimator. The SIR estimator is defined as the average of importance sampling estimate of $E\left[r(\mathbf{y}^{rep}, \mathbf{y}_{V[\mathbf{s}^{(i)}]}) \mid \mathbf{y}_{T[\mathbf{s}^{(i)}]}\right]$, across the I independent splits and the H independent samples from $g(\theta)$,

```
Algorithm 2: Sampling Importance Resampling (SIR) estimator
  1. Simulate independent split vectors $\mathbf{s}^{(1)}, \mathbf{s}^{(2)}, \ldots, \mathbf{s}^{(I)}$ from $p(\mathbf{s})$;
  2. Let $\theta_{h1}, \ldots, \theta_{hJ}$ be the $h$th of $H$ independent MCMC samples
     simulated from $g(\theta)$ ;
  3. Draw $y_{hj}^{rep}$ from $p(\mathbf{y}^{rep} \mid \theta = \theta_{hj}, \mathbf{y})$, for $h = 1, \ldots, H$ and
     $j = 1, \ldots, J$ ;
  4. Each of these $H$ samples yields an importance sampling estimate
     of $E[r(\mathbf{y}^{rep}, \mathbf{y}_{V[\mathbf{s}^{(i)}]}) \mid \mathbf{y}_{T[\mathbf{s}^{(i)}]}]$.;
```

$$\hat{\Psi}_{sir} = \frac{1}{H}\sum_{h=1}^{H}\frac{1}{I}\sum_{i=1}^{I}\frac{\sum_{j=1}^{J} r\left(y_{hj}^{rep}, \mathbf{y}_{V[\mathbf{s}^{(i)}]}\right) w_{hj}}{\sum_{j=1}^{J} w_{hj}}, \tag{19}$$

where each weight term $w_{hj} = p(\theta_{hj} \mid \mathbf{y}_{T[\mathbf{s}]})/g(\theta_{hj})$ has simple form[1]

$$log(w_{hj}) = log f(\mathbf{y}_{T[\mathbf{s}]} \mid \theta_{hj}) - \alpha \, log f(\mathbf{y} \mid \theta_{hj}).$$

The number of splits I, the size of the posterior sample for each split J and the H independent MCMC samples must be specified.

Note that if the simulation standard error is not required, then in fact this estimator can be based on a single MCMC run, i.e., $H = 1$, otherwise $H > 1$ and it is expected to be quite small. Appendix D shows how to determine a standard error of $\hat{\Psi}_{mc}$ and $\hat{\Psi}_{sir}$ estimators.

So far we have considered prior distributions for the validation sets which do not accommodate spatial heterogeneity. In the next section, we propose an uniform prior in sub-regions to take into account spatial heterogeneity and to accommodate possible preferential sampling in the locations.

5. Accounting for heterogeneity in the spatial domain

In this section, we propose a stratified sampling scheme to obtain training and validation sets. In stratified sampling, the region of n locations is first divided into sub-regions which are called *strata* of sizes $n_1, n_2, \ldots, n_K$, respectively. These sub-regions are non-overlapping, and together they comprise the whole region, so that, $n = \sum_{k=1}^{K} n_k$. The full training data size is denoted by n_T, i.e., $n_T = \sum_{k=1}^{K} n_{Tk}$ where each term of this summation represents the training data size in each stratum $k = 1, \ldots, K$. Analogously, the full validation data size is $n_V = \sum_{k=1}^{K} n_{Vk}$. If a simple random sample is taken in each stratum, the whole procedure is described as stratified random sampling.

The following notations in Table 2 refer to stratum k.

[1] The weights are obtained in Appendix D.1.

Table 2: Stratified sampling notation.

Notation
n_k: total number of spatial points in stratum k
n_{T_k}: number of spatial points of training data in stratum k
n_{V_k}: number of spatial points of validation data in stratum k
$w_k = \frac{n_{V_k}}{n_V}$: stratum weight
$f_V = \frac{n_V}{n}$: sampling fraction, i.e., the ratio of validation sample size to the total sample size.
$f_{T_k} = \frac{n_{T_k}}{n_k}$: training sampling fraction in the k^{th} stratum
$f_{V_k} = \frac{n_{V_k}}{n_k}$: validation sampling fraction in the k^{th} stratum

Stratification might produce a gain in precision in the estimates of characteristics of the whole region, if the variability inside each stratum is small and the variability between strata is large (Cochran, 1999). It may be possible to divide a heterogeneous region into sub-regions, where each sub-region is internally homogeneous in the context of spatial cross-validation.

The following steps should be carried out to perform cross-validation using a stratified sampling scheme :

1. Stratify the study region into k strata.
2. Sample in each stratum k, assuming a uniform prior on the splits, to obtain the split vectors $\mathbf{s}^{(i,k)}$, where $k = 1, \ldots, K$ represents the stratum and $i = 1, 2, \ldots, I_k$ the sizes of the split vectors generated in each stratum k.

For the sake of simplicity, we set the sizes of the split vectors I_k equal for all strata, I_k, $k = 1, 2, \ldots, K$. Note that the sizes of split vectors I_k do not need to be the same.

The split vector in each stratum $\mathbf{s}^{(1,k)}, \ldots, \mathbf{s}^{(I_k,k)}$ is jointly generated from $p(\mathbf{s})$. Thus, we define the vector $\mathbf{s}^{(i)}$ as the *i-th* split vector of all strata.

$$\mathbf{s}^{(i)} = \left(\mathbf{s}^{(i,1)}, \mathbf{s}^{(i,2)}, \ldots, \mathbf{s}^{(i,K)}\right), \;\; i = 1, \ldots, I.$$

Notice that,

$$\mathbf{s}^{(i,k)} = (s_1^{(i,k)}, s_2^{(i,k)}, \ldots, s_{n_k}^{(i,k)}).$$

The proposed stratification changes the sampling of spatial locations for validation and training sets, however, the sampling model is conditional on $\mathbf{s}$ and does not change with our proposal. Thus, the likelihood function is not affected. The vector of all observations can be written as $\mathbf{y} = (y_{1,1}, \ldots, y_{1,n_1}, \ldots, y_{k,i}, \ldots, y_{K,n_K})$. The splits $\mathbf{s}$ are not uniformly distributed over the entire spatial because they are jointly generated from a uniform prior in each stratum. The proposed prior for the stratification design is given by

$$p(\mathbf{s}) \quad = \quad \binom{n_1}{n_{T_1}}^{-1} \binom{n_2}{n_{T_2}}^{-1} \cdots \binom{n_K}{n_{T_K}}^{-1} \quad \text{if} \quad \sum_{j=1}^{n_k} s_j^{(\cdot,k)} = n_{T_k}, \qquad (20)$$

where each term of the product in equation (20) is the probability of choosing a sample of size n_{T_k} in each stratum k. The expectations are computed with respect to the discrepancy function for each stratum, denoted generically as

$$\Psi_k = E\left\{r_k(\mathbf{y}^{rep}, \mathbf{y}_{V[\mathbf{s}]})\right\}, \quad k = 1, \ldots, K, \qquad (21)$$

where the expression (21) represents the expectation with respect to the discrepancy measure in each stratum k.

5.1. Stratified Estimators

To compute the stratified estimators, we jointly simulate the split vectors $\mathbf{s}^{(1)}, \ldots, \mathbf{s}^{(I)}$ from $p(\mathbf{s})$ as defined in (20). Following the same steps as in Section 4.2, the stratified MC estimator is obtained as

$$\begin{aligned} \hat{\Psi}^{st}_{mc} &= \sum_{k=1}^{K} w_k \left\{ \frac{1}{I_k} \sum_{i=1}^{I_k} \frac{1}{J} \sum_{j=1}^{J} r_k\left(y^{rep}_{ij}, \mathbf{y}_{V[\mathbf{s}^{(i)}]}\right) \right\} \qquad (22) \\ &= \sum_{k=1}^{K} w_k \hat{\Psi}_{mc_k} = \sum_{k=1}^{K} \hat{\Psi}^{st}_{mc_k}, \end{aligned}$$

and the stratified SIR estimator as,

$$\hat{\Psi}^{st}_{sir} \quad = \quad \sum_{k=1}^{K} w_k \left\{ \frac{1}{H} \sum_{h=1}^{H} \frac{1}{I_k} \sum_{i=1}^{I_k} \Psi^{(k)}_{hi} \right\} = \sum_{k=1}^{K} w_k \hat{\Psi}_{sir_k} = \sum_{k=1}^{K} \hat{\Psi}^{st}_{sir_k}, \qquad (23)$$

where,

$$\Psi^{(k)}_{hi} = \frac{\sum_{j=1}^{J} r_k\left(y^{rep}_{hj}, \mathbf{y}_{V[\mathbf{s}^{(i)}]}\right) w^*_{hj}}{\sum_{j=1}^{J} w^*_{hj}}, \;\; k = 1, \ldots, K,$$

and $w_k = \frac{n_{V_k}}{n_V}$ is the stratified weight.Each weight term of the stratified SIR estimator is given by $w^*_{hj} = p(\theta_{hj} \mid \mathbf{y}_{T[\mathbf{s}]})/g(\theta_{hj})$. The properties about unbiased estimator are available for stratified estimators. See Appendix D.2 for further details about the computation of the weights.

5.2. The choice of stratification in spatial context

The strata can be defined based on prior knowledge of the degree of homogeneity of regions or defined arbitrarily according to easily specified spatial boundaries, such as, latitude or longitude. Considerations about stratum sizes and their shapes and sample sizes need to be made to reduce the sampling variance of the expected discrepancy estimator.

According to (see Diggle, 2014, page 99), when the occurrence of an event at a particular location makes it more likely than other events will be located nearby, the resulting patterns display a kind of pattern. In this context, the local knowledge of the underlying process could suggest the shape of the strata (see Cressie, 1993, page 317).

Clustering methods may be used to obtain the strata. For instance, the well-known K-means ++ (Arthur and Vassilvitskii (2007)) is an algorithm that optimizes the criteria of grouping by using an iterative technique. The initial step is to create an initial partition. The objects are then attributed to the cluster with the closest mean. This procedure is done repeatedly until achieved convergence. Notice that it is necessary to specify the number of clusters to be analyzed.

According to Katzfuss et al. (2014) in context of Gaussian random fields, the choice of partitions should be independent of the observed data, but it should depend on the application under consideration. In their applications, they consider a suitable general partitioning strategy using auxiliary variables or the latitude for produce subsets. Although the authors considered procedures for create subsets, they do not take into account the restriction of contiguity of geographic neighbourhood locations.

Another way of stratifying is to consider plausible strata and the possibility of modifying them to take into account geographical features of the site, for example, mountains could influence the contiguity of spatially located data. Gordon (1996) consider the selection of contiguity graphs.

In this work, stratified sampling was use to divide a possible heterogeneous spatial domain in subregions more homogeneous to achieve smaller variances for the estimated discrepancy functions. In many analysis, the interest focuses on identifying regions that rapid change (lines or curves) in the spatial surface. Identify locations or curves that fastly change is referred to as *wombling* or *boundary analysis*. Banerjee and Gelfand (2006) developed an inferential method for boundaries on Gaussian process surfaces, but this had been applied only to usual geostatistical models. Liang et al. (2009) present wombling methods for estimated intensity surfaces within a hierarchical point-process setting. This direction would be pursued in a future research.

6. Simulated Study

To illustrate the usefulness of our cross-validation proposal, we consider homogeneous and inhomogeneous processes scenarios under geostatistical modelling. We simulated data from different scenarios, considering different configurations for the location sampling. For each scenario, we first simulated a realization of a stationary Gaussian process on the unit square, treating the spatially continuous process $S(\cdot)$ as constant within each lattice cell. Then, we simulated non-preferentially or preferentially scenarios according to each of the sampling designs presented in Figure 4 (complete spatial randomness and inhomogeneous process). The data were generated from equation (3) with:

(i) S is stationary Gaussian process with mean 0, variance σ^2 and correlation function $\rho(u, \phi) = Corr(S(x), S(x'))$ for any x and x' from a distance u apart.

(ii) $X \mid S$ is an inhomogeneous Poisson process with log-linear intensity function

$$\lambda(x) = exp\left\{\alpha + \beta S(x)\right\}. \tag{24}$$

(iii) $Y \mid S, X$ is a set of mutually independent Gaussian variables with

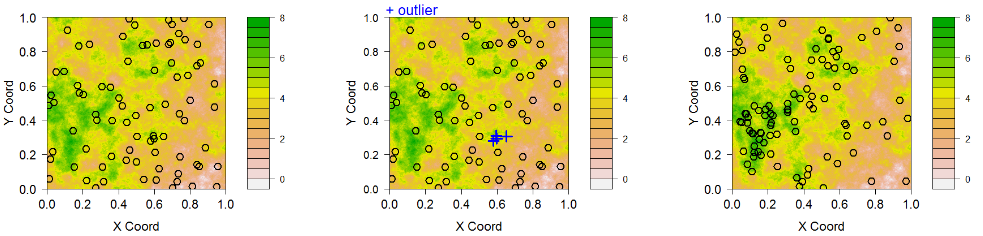
$$Y_i \sim N(\mu + S(x_i), \tau^2).$$

Note that if $\beta = 0$, the sampling is done completely at random, i.e., we have a homogeneous Poisson process. The simulated surface in (i) is given by a Gaussian process with the following parameters: $\mu = 4, \sigma^2 = 1.5, \phi = 0.15, \kappa = 0.5$ and $\tau^2 = 0.25$. We adopted the exponential correlation function in all scenarios.

Scenario 1 – CSR (Complete spatial randomness), we considered the case where the intensity function $\lambda(x)$ is a contanst. A dataset was simulated with sample size equal to $n = 82$ and intensity parameters $\beta = 0, \alpha = 4.605$. This is presented in Figure 4 (a).

Scenario 2 – CSR with outliers, we study the same surface of Figure 4 (a) with observations contaminated by summing a random increment $u\sigma$, such that σ is the observational standard deviation and $u \sim U(6, 8)$ for observations 10, 48, 50 and 82. The contaminated locations considered are neighbours in space. This is presented in Figure 4 (b).

Scenario 3 – Preferential Sampling, we choose the configuration with highest concentration of points in a given region. The point process represents the inhomogeneous Poisson process, with intensity $\lambda(x)$, $\alpha = 2.996$, $\beta = 1.0$ and $n = 100$. This is presented in Figure 4 (c).



(a) (b) (c)

Figure 4: Sample locations and underlying realizations of the signal process for the three models considered in the simulation study: (a) CSR (complete spatial randomness) ; (b) CSR with outliers; (c) preferential sampling.

For all sample designs presented above, we made a cross-validation comparison of the three geostatistical models presented in section 3.2.Consider the observations $y_1, \ldots, y_n$ at locations $x_1, \ldots, x_n$. In this study we will compare three models using the cross-validation technique:
(M1) Gaussian model:

$$\mathbf{y} \mid \mu, \sigma^2, \phi \sim N\left(\mathbf{1}\mu, \tau^2 I_n + \sigma^2 R\right)$$

(M2) Student-t model:

$$\mathbf{y} \mid \mu, \sigma^2, \phi, \nu \sim ST\left(\mathbf{1}\mu, \nu, \tau^2 I_n + \sigma^2 R\right)$$

(M3) Gaussian-Log-Gaussian model:

$$\mathbf{y} \mid \mu, \sigma^2, \phi, \boldsymbol{\Delta} \sim N\left(\mathbf{1}\mu, \tau^2 I_n + \sigma^2(\boldsymbol{\Delta}^{-1/2}\, R\, \boldsymbol{\Delta}^{-1/2})\right)$$

Parameter estimation and prediction follow the Bayesian approach as presented in subsection 3.2 using the three proposed distributions for $\mathbf{s}_y$. We sampled from the posterior of the model parameters using Metropolis-Hastings with random walk proposals, which led to reasonable acceptance rates in the vicinity of 30% to 50% for each parameter. The chains for the simulated parameters have burn-in of 10000 and lag of 10 with resulting posterior sample size of 6981. Convergence was checked using `coda package` (Plummer et al. (2006)) through `R` software. For all models the nugget effect was fixed in the true value so that the focus of this study could be the spatial surface estimation and prediction. The prior distributions used for all models were $\mu \sim N(0; 10^4)$, $\sigma^{-2} \sim G(0.1; 0.1)$, $\phi \sim G(1; 2.3/med(u))$. For the M3 model, $\upsilon \sim GIG(0; 0.75; 6)$ and $\boldsymbol{\delta} \mid \phi, \upsilon \sim LG(-\frac{\upsilon}{2}; \upsilon\sigma^2 R)$.

For the CSR and CSR with outlier scenarios were arbitrarily chosen $n_T = 77, n_V = 5$. MC and SIR estimators are based on averaging over the same $I = 100$ splits. The parameter ν is fixed at 3 for the M2 model in the CSR scenario so that we are actually fitting a wrong model. For the preferential scenario, we considered $n_T = 95$ and $n_V = 5$.

As can be seen in Table 3, the execution time (minutes) for the SIR estimator using $H = 5$ is smaller than that for the MC estimator, if it is considered the uniform prior. Analogously, we verified the computational time considering a prior via distances. The time was similar to the previous case and we omitted from the text. The computational cost is approximately: 5 to 6 times smaller for the Gaussian model, 5 to 8 times smaller for the Student-t model and 6 to 10 times smaller for the GLG model when using the SIR estimator. The high computational cost of the MC estimator is due to the need of calculating the covariance matrix for each sampled split vector.

Table 3: Computational times (in minutes) for three competing models.

	M1		M2		M3	
	MC	SIR	MC	SIR	MC	SIR
CSR	672	139	926.4	140	2028	210
CSR with outlier	672	120	828	183	1212	208.8
Preferential	967.2	163	1423.2	187	2481.6	298.8

We adopted the discrepancy measures based on the MC and SIR estimators with their respective standard errors, assuming the uniform prior and the prior via distances for the split vectors. In these examples, both prior distributions led to similar conclusions. For clarity of analysis exposition, we omit the results of uniform prior and Table 4 presents the discrepancy measures based on the MC and SIR estimators with their respective standard errors adopting the prior via distances. We use Mahalanobis Distance (MH), average Interval Score (IS) and Log Predictive Score (LPS) for predictive performance evaluation. As expected the SIR estimator variability is greater than that of the MC estimator for the three scenarios, because the SIR estimator is a heuristic approximation based on the same amount of data. However, the point estimator obtained by SIR is a good approximation of the original estimator.

Table 4 (CSR) presents predictive measure estimates for the complete random scenario. It indicates that M3 and M1 models have similar values, although it still correctly chooses the Gaussian model as the best model. Model M2 with $\nu = 3$ has much worse performance than the other models as it is not able to recover Gaussian tails. This example indicates that the proposed cross-validation approach is leading to correct indications of best model for this scenario.

Table 4 (CSR with outlier) correctly indicates that the M3 model is the best choice for this scenario. This is due to the fact that this model tends to detect sub-regions with larger variability. On the other hand, M1 and M2 models overestimate the variance in the whole spatial domain. Although the Student-t process has heavier tails than the Gaussian, it does not have the flexibility to model georeferenced data. The Student-t process inflates the variance of the whole process in the presence of outliers and does not allow for both individual or regional outlier detection and different kurtosis behaviours across space (see Lobo and Fonseca (2019) for a more detailed discussion).

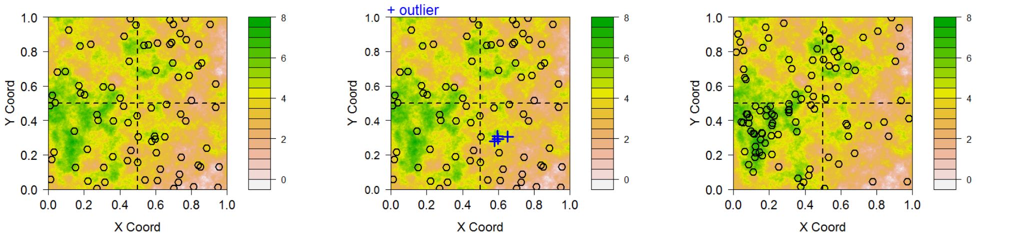
Table 4 (Preferential Sampling) indicates similar results for M1 and M3 models for all adopted measures. We emphasize that although our dataset is under effect of preferential sampling, we fit usual geostatistical models which do not take this effect into account.

Table 4: Cross-validation for M1 and M3 models in each scenario with prior via distances. The same splits are considered for all models.

			CSR	
		MH	average IS	LPS
M1	MC	3.113 (1.3×10^{-6})	4.422 (1.2×10^{-9})	7.352 (2.5×10^{-6})
	SIR	3.057 (0.004)	4.191 (0.014)	7.420 (0.013)
M2	MC	4.030 (8.7×10^{-6})	6.468 (5.4×10^{-10})	7.917 (2.5×10^{-6})
	SIR	3.828 (0.004)	6.491 (0.003)	7.958 (0.012)
M3	MC	3.477 (3.4×10^{-6})	4.763 (7.2×10^{-8})	6.907 (1.5×10^{-11})
	SIR	3.394 (0.020)	4.990 (0.253)	7.551 (0.062)
			CSR with Outlier	
		MH	average IS	LPS
M1	MC	2.760 (1.3×10^{-6})	7.158 (7.2×10^{-8})	14.503 (4.1×10^{-6})
	SIR	2.829 (0.004)	7.114 (0.004)	14.536 (0.029)
M2	MC	4.499 (8.7×10^{-6})	11.352 (3.0×10^{-9})	18.795 (3.2×10^{-6})
	SIR	4.505 (0.002)	9.933 (0.015)	18.489 (0.023)
M3	MC	2.795 (4.1×10^{-6})	4.966 (2.5×10^{-8})	9.409 (9.9×10^{-5})
	SIR	2.062 (0.010)	4.699 (0.125)	7.751 (0.031)
			Preferential Sampling	
		MH	average IS	LPS
M1	MC	3.154 (1.7×10^{-6})	5.691 (9.4×10^{-8})	7.441 (6.8×10^{-5})
	SIR	3.116 (0.007)	5.470 (0.020)	7.520 (0.059)
M2	MC	3.039 (1.3×10^{-6})	6.781 (7.1×10^{-8})	8.380 (3.4×10^{-6})
	SIR	3.762 (0.008)	6.241 (0.004)	8.160 (0.036)
M3	MC	4.052 (4.9×10^{-6})	5.498 (6.4×10^{-7})	7.573 (5.9×10^{-5})
	SIR	3.957 (0.005)	5.558 (0.131)	7.422 (0.102)

6.1. Analysing heterogeneity in the spatial domain

The same data presented in section 6 were stratified into four strata for all scenarios. In this application, we do not apply any stratification method to define the strata. Figure 5 presents the strata of the study region in A.



(a) (b) (c)

Figure 5: Sample locations and underlying realizations of the signal process for the considered model in the simulation study: (a) CSR ; (b) CSR with outliers; (c) preferential sampling. Strata are divided as: stratum 1 (bottom left), stratum 2 (top left), stratum 3 (bottom right) and stratum 4 (top right).

Table 5 shows the strata and selection of training and validation data for the respective stratum via the sampling process for all scenarios. In this study, we set the number of locations sampled for validation proportional to the number of locations in each stratum. Notice that in scenario CRS, a homogeneous process is considered, therefore it is expected the number of events to be similar in each stratum, as shown in Table 5 (a). The I_k is arbitrarily chosen in each stratum.

The execution time (in minutes) for the SIR estimator using $H = 5$ is smaller than of the MC estimator, see Table 6. The computational cost for SIR estimator is approximately: 4 to 6 times lower than MC estimator for the Gaussian model, 4 times smaller than MC estimator for the Student-t model and 6 to 8 times smaller than MC estimator for the GLG model.

Table 5: Stratified sample for all scenarios.

	CSR / Outlier					Preferential			
strata	n_k	n_{Tk}	n_{Vk}	w_k	*strata*	n_k	n_{Tk}	n_{Vk}	w_k
1	21	19	2	0.250	1	47	42	5	0.500
2	17	15	2	0.250	2	20	18	2	0.200
3	24	22	2	0.250	3	13	12	1	0.100
4	20	18	2	0.250	4	20	18	2	0.200
total	82	74	8	1	*total*	100	90	10	1

Table 6: Computational time (in minutes) for three competing models.

	M1		M2		M3	
	MC	SIR	MC	SIR	MC	SIR
CSR	880.2	205.2	861.6	196.8	1533	208
CSR with outlier	808.8	183	928.8	231	1823.22	282.6
Preferential	1195.2	207.6	1101.6	248.4	2259.6	286.80

Tables 7, 8 and 9 show that stratification reduces the variability of discrepancy estimates for all scenarios and discrepancy measures. The hypothesis of *complete spatial randomness* implies that the number of events per unit area is constant (λ) over all considered region. In the homogeneous case, the estimates are approximately the same for each stratum. Regarding the performance of the stratified estimator, their variability is smaller, mainly for the SIR estimator.

The use of Mahalanobis distance, average Interval Score and Log Predictive Score discrepancies leads to adequate model discrimination by indicating the Gaussian model as the best model for scenario CRS. This is an expected result, since the data are generated by the Gaussian model.

Clearly there is an increasing in the accuracy of the estimator by stratifying the spatial region. Furthermore, the stratification allows the identification of lack of fit for all models in region 3 for the scenario with outliers. All models have much larger values of the discrepancy function for stratum 3 (bottom right in Figure 5 (b)), that contains the contaminated observations, as presented in Tables 8 (CSR with outlier) for all discrepancy measures considered. Note, however the reduction of the variability of estimator. The performances of the models are similar, except for M3 model. In this case, the GLG model has better performance, indicating that if the region are divided in sub-regions, a better predictive performance assessment of this model for all sub-regions are provided .

The M3 model again stands out, because of its ability to capture heterogeneity in space. This is an appealing feature in the non-homogeneous setup, because strata with a high concentration of events might present larger variability.

Table 9 presents the preferential sampling scenario. The performance of M1 and M3 models are similar, while M2 has the worst performance of all three models. Stratified estimator shows the poor predictive performance in region 1 for all models. In fact this is expected, as the fitted models do not consider preferential sampling in its specification. We omitted the results of M2 model for the stratified study, since it has a worse performance than M1 and M3 models for all scenarios.

Table 7: Stratified cross-validation for M1 and M3 models for the complete spatial randomness (CSR) scenario. The same splits are considered for all models.

		M1		M3	
	strata	MC	SIR	MC	SIR
MH	1	1.81 (8.0×10^{-4})	1.82 (0.002)	2.05 (1.8×10^{-6})	2.08 (0.002)
	2	1.78 (1.2×10^{-6})	1.79 (0.002)	2.09 (1.9×10^{-6})	2.04 (0.002)
	3	1.73 (1.1×10^{-6})	1.69 (0.002)	1.92 (1.6×10^{-6})	1.96 (0.002)
	4	1.66 (1.2×10^{-6})	1.67 (0.003)	1.85 (1.7×10^{-6})	1.91 (0.003)
	$\hat{\Psi}^{st}$	**1.74** (3.0×10^{-7})	**1.74** (5.6×10^{-4})	**1.98** (4.5×10^{-7})	**1.99** (5.6×10^{-4})
average IS	1	4.72 (1.9×10^{-9})	4.94 (0.007)	4.25 (1.0×10^{-10})	4.63 (0.001)
	2	4.56 (1.8×10^{-9})	4.48 (0.026)	5.79 (2.7×10^{-8})	6.28 (0.409)
	3	4.04 (1.2×10^{-10})	4.76 (0.011)	4.34 (5.0×10^{-10})	4.57 (0.000)
	4	4.27 (1.7×10^{-9})	5.09 (0.038)	5.37 (6.7×10^{-9})	5.57 (0.089)
	$\hat{\Psi}^{st}$	**4.39** (3.4×10^{-10})	**4.82** (5.1×10^{-3})	**4.94** (2.2×10^{-9})	**5.26** (0.031)
LPS	1	3.27 (1.4×10^{-6})	3.44 (0.007)	3.34 (1.5×10^{-6})	3.43 (0.004)
	2	2.89 (9.7×10^{-7})	3.39 (0.009)	3.46 (2.4×10^{-6})	3.38 (0.003)
	3	2.59 (6.0×10^{-7})	3.33 (0.009)	3.06 (1.4×10^{-6})	3.21 (0.002)
	4	2.82 (1.4×10^{-6})	3.17 (0.011)	2.97 (2.4×10^{-6})	3.21 (0.005)
	$\hat{\Psi}^{st}$	**2.89** (2.7×10^{-7})	**3.33** (2.3×10^{-3})	**3.21** (4.7×10^{-7})	**3.31** (8.7×10^{-4})

Table 8: Stratified cross-validation for M1 and M3 models for the complete spatial randomness (CSR) with outlier scenario. The same splits are considered for all models.

		M1		M3	
	strata	MC	SIR	MC	SIR
MH	1	2.41 (3.9×10^{-6})	2.14 (0.002)	1.98 (3.9×10^{-6})	1.52 (0.003)
	2	2.43 (4.1×10^{-6})	2.48 (0.003)	1.20 (4.1×10^{-6})	1.55 (0.006)
	3	3.42 (6.8×10^{-6})	5.29 (0.070)	2.67 (6.8×10^{-6})	2.68 (0.059)
	4	2.33 (3.7×10^{-6})	1.97 (0.004)	1.83 (3.7×10^{-6})	1.47 (0.003)
	$\hat{\Psi}^{st}$	**2.64** (1.7×10^{-6})	**2.97** (0.004)	**1.92** (1.8×10^{-5})	**1.81** (0.004)
average IS	1	4.12 (1.7×10^{-10})	4.49 (0.003)	2.658 (3.6×10^{-10})	2.012 (0.022)
	2	4.01 (2.5×10^{-10})	4.64 (0.002)	2.29 (3.1×10^{-10})	2.17 (0.012)
	3	6.51 (2.8×10^{-7})	7.73 (0.055)	5.62 (2.6×10^{-10})	6.45 (0.079)
	4	3.44 (4.0×10^{-10})	4.87 (0.003)	3.66 (3.6×10^{-10})	2.82 (0.014)
	$\hat{\Psi}^{st}$	**4.52** (1.7×10^{-8})	**5.43** (0.003)	**3.56** (1.3×10^{-6})	**3.36** (0.008)
LPS	1	3.35 (2.5×10^{-5})	3.46 (0.004)	2.42 (1.5×10^{-5})	3.46 (8.0×10^{-4})
	2	3.26 (2.6×10^{-5})	3.72 (0.003)	1.92 (1.9×10^{-5})	2.64 (0.002)
	3	4.64 (8.4×10^{-5})	5.98 (0.293)	3.92 (8.8×10^{-4})	4.27 (0.053)
	4	3.08 (2.1×10^{-5})	3.36 (0.004)	3.15 (1.4×10^{-5})	3.33 (5.0×10^{-4})
	$\hat{\Psi}^{st}$	**3.58** (7.7×10^{-7})	**4.13** (0.019)	**2.85** (9.3×10^{-4})	**3.42** (0.003)

Table 9: Stratified cross-validation for M1 and M3 models in preferential sampling scenario. The same splits are considered for all models.

		M1		M3	
	strata	MC	SIR	MC	SIR
MH	1	3.57 (1.7×10^{-6})	3.63 (0.008)	3.38 (1.5×10^{-6})	3.16 (0.004)
	2	1.56 (9.9×10^{-6})	1.58 (3.0×10^{-4})	1.68 (1.2×10^{-6})	1.84 (0.001)
	3	0.97 (7.7×10^{-7})	0.95 (1.0×10^{-4})	1.16 (1.0×10^{-6})	1.15 (0.001)
	4	1.64 (1.0×10^{-6})	1.65 (2.0×10^{-4})	1.73 (1.2×10^{-6})	1.95 (0.003)
	$\hat{\Psi}^{st}$	**2.53** (2.8×10^{-7})	**2.56** (8.7×10^{-4})	**2.48** (5.0×10^{-6})	**2.46** (0.001)
average IS	1	6.98 (1.9×10^{-7})	6.50 (0.021)	6.89 (1.7×10^{-11})	6.72 (0.142)
	2	4.27 (1.1×10^{-9})	4.83 (0.009)	6.87 (8.5×10^{-9})	7.37 (0.128)
	3	4.35 (1.2×10^{-10})	4.60 (0.000)	6.91 (1.3×10^{-5})	6.08 (0.210)
	4	4.72 (2.6×10^{-9})	4.32 (0.004)	6.87 (2.3×10^{-11})	6.47 (0.139)
	$\hat{\Psi}^{st}$	**5.73** (1.2×10^{-8})	**5.54** (0.002)	**6.88** (7.9×10^{-7})	**6.74** (0.048)
LPS	1	8.17 (1.9×10^{-5})	9.36 (0.134)	9.02 (8.2×10^{-5})	9.55 (0.018)
	2	2.50 (1.0×10^{-6})	3.23 (0.013)	4.42 (8.8×10^{-6})	3.41 (0.010)
	3	1.20 (1.6×10^{-7})	1.36 (0.001)	2.67 (2.3×10^{-3})	1.77 (0.004)
	4	2.98 (1.6×10^{-6})	2.95 (0.012)	4.71 (9.8×10^{-6})	3.46 (0.012)
	$\hat{\Psi}^{st}$	**5.30** (1.4×10^{-6})	**6.06** (0.034)	**5.21** (1.4×10^{-4})	**6.32** (0.005)

7. Application to a rainfall data

The dataset used in this application contains the total rainfall (in *mm*) recorded in October 2010 in 32 locations in the city of Rio de Janeiro, Brazil, obtained from *Instituto Pereira Passos*, known for offering one of the largest collections of maps and statistical data of Rio de Janeiro available in *Armazem de Dados*. Stations with missing information were removed from the study. Ferreira and Gamerman (2015) analyzed the same kind of data for October 2005 in the context of optimal design using preferential sampling.

Figure 6 presents the spatial arrangement of rainfall stations in the city of Rio de Janeiro. Note that the spatial arrangement of the monitoring stations seems to indicate a higher concentration in places where precipitation levels are very large. It appears that the point pattern associated with the stations has been observed from an inhomogeneous process. Besides that, Figure 7 presents the altitude according to stations installed in the city of Rio de Janeiro. Note that, higher altitudes are concentrated where there is a higher rainfall intensity.

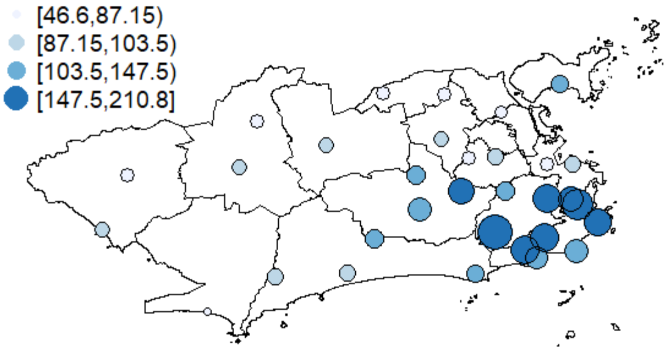


Figure 6: Rainfall data: stations installed in the city of Rio de Janeiro (the monitoring stations are separated according to the intensity of rainfall).

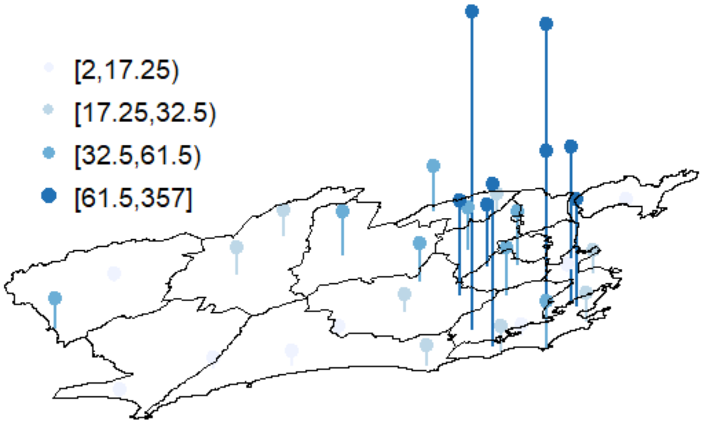


Figure 7: Rainfall data: altitude according to stations installed in the city of Rio de Janeiro (the monitoring stations are separated according to the altitude).

For statistical inference purposes, the spatial mean was adjusted considering latitude and longitude as covariates. For this analysis, the models fitted were the Gaussian (M1) and Gaussian-Log-Gaussian (M3) models presented in Section 3.2.

We evaluated both models with exponential covariance structures for spatial dependence. Parameter estimation and prediction follow the Bayesian paradigm as presented in subsection 3.2 using the three proposed distributions for $\mathbf{s}_y$. The chains for the simulated parameters have burn-in of 10000 and lag of 10 with resulting posterior sample size of 3981. Convergence was checked using `coda package` (Plummer et al. (2006)). The prior distributions used for all models were $(\beta_0, \beta_1, \beta_2) \sim N(\mathbf{0}, 10^4 I_\beta)$, $\sigma^{-2} \sim G(0.1; 0.1)$, $\phi \sim G(1; 2.3/med(u))$. For M3 $\upsilon \sim GIG(0; 0.75; 6)$ and $\boldsymbol{\delta} \mid \phi, \upsilon \sim LG(-\frac{\upsilon}{2}; \upsilon\sigma^2 R)$. The nugget effect τ^2 was set to 0 in this application. The analysis of the posterior distribution of spatial mean shows significantly different estimates for both models. The spatial mean for M3 is significantly lower than the spatial mean estimated by M1. Actually, this is plausible since the process for the data is inhomogeneous and model M3 compensates this heterogeneity by estimating different variances across space.

An important issue in using cross-validation is the training dataset size. If we have an acceptable amount of training data, the model is sufficiently informed by the training set. First, we arbitrarily choose $n_T = 84\%n$ and $n_V = 16\%n$, for the training and validation samples, respectively. We also considered an extreme sampling setup with a small training sample, $n_T = 32\%n$ and $n_V = 68\%n$. It is expected that using a reduced training sample size might cause some impact on the estimation of model parameters. For both scenarios, we set $I = 500$ split vectors and $H = 3$ independent MCMC samples simulated from SIR estimator. Note that the SIR estimator can produce estimates close enough to the MC estimator.

Table 10 displays the performance of both models according to the Mahalanobis distance, average Interval Score and Log Predictive Score when it is assigned a uniform prior to the splits. As expected, the results of our analysis suggest it is best to use a relatively large training sample for making cross-validation under our approach. The estimates obtained for M3 model are smaller than for M1 model for both estimators and measures. This is due to the fact that the Gaussian-Log-Gaussian process proposed by Palacios and Steel (2006) is able to capture heterogeneity in space through a mixing process used to increase the Gaussian process variability, although it does not take into account dependence between the monitoring stations arrangement and the total rainfall. Table 11 shows the performance of both models using prior via distances. We obtained the same conclusions compared to uniform prior. Both models do not perform well when the training sample size is very small.

Table 10: Rainfall Data: cross-validation using discrepancy measures for M1 and M3 models using uniform prior. The same splits are considered for both models.

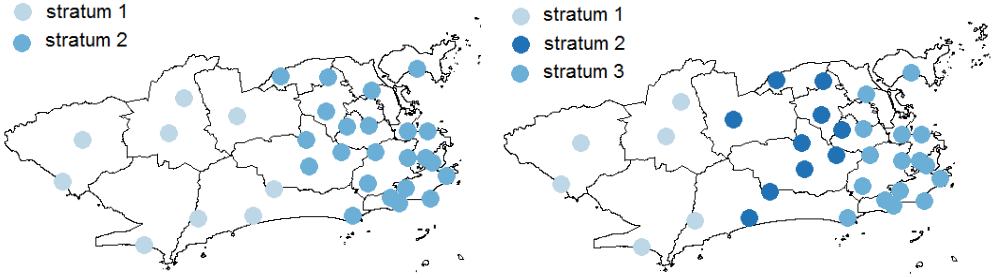
$n_V = 16\%n$		MH	average IS	LPS
M1	MC	6.983 (1.9×10^{-6})	246.732 (1.2×10^{-4})	28.868 (0.024)
	SIR	6.419 (0.003)	232.934 (3.314)	27.941 (0.080)
M3	MC	4.552 (6.9×10^{-6})	149.468 (1.6×10^{-6})	26.765 (7.7×10^{-6})
	SIR	4.805 (0.001)	131.533 (0.013)	25.463 (0.022)
$n_V = 68\%n$		MH	average IS	LPS
M1	MC	15.257 (7.4×10^{-6})	376.952 (8.3×10^{-4})	161.751 (0.001)
	SIR	14.699 (0.006)	355.724 (4.093)	131.105 (0.456)
M3	MC	7.829 (5.6×10^{-7})	188.203 (4.6×10^{-4})	114.341 (3.6×10^{-6})
	SIR	7.746 (0.020)	192.63 (0.112)	102.925 (0.626)

Table 11: Rainfall Data: cross-validation using discrepancy measures for M1 and M3 models using prior via distances. The same splits are considered for both models.

$n_V = 16\%n$		MH	average IS	LPS
M1	MC	5.382 (1.6×10^{-6})	423.291 (2.4×10^{-4})	33.032 (0.076)
	SIR	6.419 (0.005)	469.677 (0.498)	29.787 (0.062)
M3	MC	4.552 (6.9×10^{-6})	209.612 (1.3×10^{-6})	26.765 (7.7×10^{-6})
	SIR	4.805 (0.011)	189.610 (0.057)	27.564 (0.575)
$n_V = 68\%n$		MH	average IS	LPS
M1	MC	14.784 (1.2×10^{-5})	584.878 (0.002)	185.215 (0.082)
	SIR	14.699 (0.004)	620.874 (0.103)	173.387 (0.407)
M3	MC	7.829 (5.6×10^{-7})	238.587 (1.6×10^{-4})	117.092 (0.118)
	SIR	7.746 (0.008)	216.337 (0.249)	111.636 (0.181)

In addition, we take into account spatial heterogeneity using stratified cross-validation techniques. The choice of strata was performed after observing the sample locations. We divide the spatial region into $k = 2$ and $k = 3$

strata via K-means ++ criteria. Although K-means ++ algorithm does not take into accounting spatial contiguity constraints, define by the boundaries between regions, the procedure indicated that locations belong to the same strata if there is a contiguous spatial representation between these locations. Figure 8 presents the two proposals for stratification considering K-means ++ as criteria. Notice that in the two cases in Figure 8 (i) and (ii), there is a specific stratum where the monitoring stations are closer together and there is a higher concentration of total rainfall data. The others strata are defined by the remainder of the locations, that is, more distant locations with lower values of total precipitation and altitudes, as can be seen in Figures 6 and 7.



(i)

(ii)

Figure 8: Proposals for stratification via K-means ++ algorithm: (i) $k = 2$ strata and (ii) $k = 3$ strata.

Table 12 presents the sample arrangement considering the two proposal with $n_V = 12.5\%n$ taking into account the weights, w, for each stratum, respectively. The results for stratified scenario can been seen in Table 13.

Table 12: Stratified scenarios: training and validation samples.

	$k = 2$				$k = 3$		
strata	n_T	n_V	w	*strata*	n_T	n_V	w
1	8	1	0.250	1	5	1	0.250
2	20	3	0.750	2	9	1	0.250
-	-	-	-	3	14	2	0.500
total	28	4	1.000	*total*	28	4	1.000

Table 13: Rainfall Data: stratified cross-validation using discrepancy measures for M1 and M3 models. The same splits are considered for both models.

		M1		M3	
$k = 2$	*strata*	MC	SIR	MC	SIR
MH	1	1.549 (9.2×10^{-7})	2.264 (0.002)	1.437(6.7×10^{-6})	1.391 (0.003)
	2	3.668 (1.2×10^{-6})	3.334 (0.003)	3.194(4.11×10^{-5})	3.123 (0.013)
	$\hat{\Psi}^{st}$	**3.139** (1.1×10^{-6})	**3.067** (0.001)	**2.755** (2.3×10^{-5})	**2.690** (0.004)
average IS	1	233.694 (2.8×10^{-5})	236.143 (0.066)	171.451 (2.5×10^{-6})	199.013 (0.002)
	2	255.229 (7.2×10^{-5})	253.440 (0.074)	219.911 (1.3×10^{-5})	222.920 (0.014)
	$\hat{\Psi}^{st}$	**249.845** (5.0×10^{-5})	**240.467** (0.035)	**207.796** (6.7×10^{-5})	**204.99** (0.002)
LPS	1	5.360 (5.0×10^{-6})	6.989 (0.029)	8.772 (7.7×10^{-4})	8.869 (0.076)
	2	17.612 (1.3×10^{-5})	16.603 (0.023)	10.585(0.036)	11.470 (0.213)
	$\hat{\Psi}^{st}$	**14.549** (8.8×10^{-6})	**14.200** (0.013)	**9.225** (0.018)	**10.820** (0.032)
		M1		M3	
$k = 3$	*strata*	MC	SIR	MC	SIR
MH	1	1.764 (1.1×10^{-6})	1.325 (0.002)	1.057 (3.9×10^{-7})	1.652 (0.003)
	2	1.336 (4.9×10^{-7})	1.441 (0.001)	1.731 (2.5×10^{-6})	1.189 (0.005)
	3	3.020 (1.1×10^{-6})	3.763 (0.004)	2.816 (9.2×10^{-6})	2.975 (0.019)
	$\hat{\Psi}^{st}$	**2.285** (8.8×10^{-7})	**2.573** (0.002)	**2.105** (4.0×10^{-5})	**2.198** (0.006)
average IS	1	310.195 (3.9×10^{-5})	297.982 (0.003)	167.967(3.0×10^{-7})	161.333 (0.026)
	2	98.212 (1.1×10^{-6})	104.186 (0.130)	158.002 (4.1×10^{-6})	146.484 (0.019)
	3	282.455 (5.2×10^{-5})	262.927 (1.867)	318.476 (2.6×10^{-5})	314.289 (0.930)
	$\hat{\Psi}^{st}$	**243.329** (3.1×10^{-5})	**240.769** (0.222)	**240.730** (1.1×10^{-5})	**234.099** (0.108)
LPS	1	5.906 (6.0×10^{-6})	5.019 (0.009)	7.582 (2.0×10^{-6})	8.886 (0.049)
	2	4.624 (1.0×10^{-6})	4.999 (0.004)	4.633 (4.7×10^{-4})	5.685 (0.307)
	3	12.111 (8.0×10^{-6})	15.236 (0.053)	9.275 (0.003)	12.330 (0.306)
	$\hat{\Psi}^{st}$	**8.688** (5.1×10^{-6})	**10.122** (0.007)	**7.692** (0.001)	**9.807** (0.073)

Observe that the different discrepancy estimates in Table 13 between strata considering MH, average IS and LPS measure seen in both scenarios, $k = 2$ and $k = 3$, respectively. For $k = 2$ scenario, observe that M1 produce high estimates in stratum 2 considering MH, average IS and LPS measures, as can be seem from Table 13. In fact, this might be a result of the number of neigbours that are far or more close apart. Furthermore, the stratification allows the identification of lack of fit for both models in stratum 2 compared to stratum 1 for $k = 2$ and all strata considering $k = 3$. Besides that, M3 model apparently performs better for both choices of training sample, indicating better predictive performance considered in the two stratification proposed in this application.

8. Conclusions

This work considers Bayesian model comparison and criticism for spatially correlated data analysis. Cross-validation techniques are considered to evaluate the model predictive performances and we allow for uncertainty in the choice of validation sets through the prior distribution on the possible sets.

The proposed split vector prior distributions allow to accommodate the uncertainty in the validation and trainning set choice. This addresses important issues that have not been completely dealt with in the literature, such as the ad hoc choice of validation sets in spatial data analysis.

The prior via distances choose the location to compose the training sample according to their respective probabilities. These probabilities depending on their respective distance between of the previous selected point and all the others candidates points to the training sample. Since irregular spatial regions as often occur in data applications, the prior via distances is an useful alternative to the uniform prior.

The SIR estimator is a good approximation of the MC estimator and requires only a few MCMC runs for the parameter estimation step, besides overcoming the computational limitation of Bayesian cross-validation techniques.

The proposed stratified scheme contributes to reducing the global variability of SIR estimators. Futhermore, it indicates regions with lack of fit in the spatial domain textcolorredsuch as presence of outliers and preferentiability in the point pattern.

Our stratification approach relies on the definition of strata in the spatial domain. As pointed out by Cochran (1999), there are important issues related to the building of the strata, such as: the potential variables used to determine them; the determination of their boundaries; and the number of strata.

Moreover, the question of choosing the training sample size should also be considered and it is not trivial. We considered two different scenarios in the application to rainfall data to accommodate the possible effect of choosing either a too small or too large training set through the three proposed prior distributions: uniform, distances and stratified. Further research would consider the validation size effects in the context of spatial model choice.

Acknowledgements

This work was part of the Ph.D. research of V. G. R. Lobo under the supervision of T. C. O. Fonseca and F. A. S. Moura. V. G. R. Lobo benefited from a scholarship from *Conselho de Aperfeiçoamento de Pessoal de Nível Superior* (CAPES), Brazil. T. C. O. Fonseca was partially supported by *Conselho Nacional de Desenvolvimento Científico e Tecnológico* (CNPq). F. A. S. Moura was partially supported by CNPq.

Appendix A. Student-t process

An n-dimentional random vector $\mathbf{y} = (y(x_1), \ldots, y(x_n))$ follows a Student-t spatial process (Roislien and Omre (2006)) with degrees of freedom $\nu \in \mathbb{R}_+$, mean vector $\mu\mathbf{1} \in \mathbb{R}^n$ and covariance matrix $\Sigma = \sigma^2 R$ if its joint probabilidade density function is given by

$$f(\mathbf{y} \mid \mu, \sigma^2, \phi, \nu) = \frac{\Gamma(\frac{\nu+n}{2})}{\Gamma(\frac{\nu}{2})(\nu\pi)^{n/2}|\Sigma|^{1/2}} \left[1 + \frac{(\mathbf{y} - \mathbf{1}\mu)'\Sigma^{-1}(\mathbf{y} - \mathbf{1}\mu)}{\nu}\right]^{-(\nu+n)/2}, \tag{A.1}$$

where $\Gamma(\cdot)$ denotes the gamma function and R is the correlation matrix with elements $r_{ij} = \exp\{-||x_i - x_j||/\phi\}$, with range parameter $\phi > 0$, that determines the rate at which the correlation between observations decreases as distances grow.

The Gaussian process is a special case of Student-t proces with degrees of freedom $\nu \to \infty$.

Appendix B. The choice of discrepancy function

One way to evaluate a spatial model is through the accuracy of its predictions. In particular, we are interested in using the predictions to measure the performance of a model and to compare several models.

It is very common to use the *sum of squared prediction errors* as a measure of discrepancy, because it is type of cross-validation that provides a measure of model fitness for those observations left out of the estimation procedure.

Alqallaf and Gustafson (2001) and Thall et al. (1997) adopt this measure for fitting a univariate dataset. As follows some common used discrepancy measures are presented.

Mean squared prediction errors It can be written as

$$r(\mathbf{y}^{rep}, \mathbf{y}_{V[\mathbf{s}]}) = \frac{1}{n_V} \mid\mid (\mathbf{y}_{V[\mathbf{s}]}^{rep} - \mathbf{y}_{V[\mathbf{s}]}) \mid\mid^2 . \tag{B.1}$$

Gelman et al. (2014) explain that equation (B.1) has the advantage of being easy to compute and to interpret, but the disadvantage of being less appropriate for models that are far from the normal distribution. In addition, this measure does not take into account the correlation between observations.

Mahalanobis Distance This measure takes into account the covariance matrix of the common distribution of the two random vectors (Mahalanobis (1936)).

$$r(\mathbf{y}^{rep}, \mathbf{y}_{V[\mathbf{s}]}) = \sqrt{(\mathbf{y}_{V[\mathbf{s}]}^{rep} - \mathbf{y}_{V[\mathbf{s}]})'\Sigma^{-1}(\mathbf{y}_{V[\mathbf{s}]}^{rep} - \mathbf{y}_{V[\mathbf{s}]})}, \tag{B.2}$$

where $\Sigma = \tau^2 I_\tau + \sigma^2 R$ is the predictive covariance matrix of the regions formed by the locations that belong to both vectors. Therefore, using the Mahalanobis distance, we can compare the validation data sample, taking into account the spatial dependence. Extreme values for the Mahalanobis distance indicate a conflict between the validation data and predictive data. Bastos and O'Hagan (2008) adopt this measure to validate and assess the adequacy of a Gaussian processes emulator.

We adopt as discrepancy measures the predictive accuracy for probabilistic forecasts known as scoring rules A review of the most common scoring rules and properties are presented by Gneiting and Raftery (2007). In this direction, Gelman et al. (2014) presents different ways of defining the accuracy or error of model's predictions, and show methods for estimating predictive accuracy or error from data. Vehtari et al. (2017) consider these measures in the context of leave-one-out cross-validation (LOO-CV).

Interval Score Interval forecasts is a crucial special case of quantile prediction (Gneiting and Raftery (2007)). It compares the predictive credibility interval with the true value one, and consider the uncertainty in the predictions such that the model is penalized if an interval is too narrow and misses the true value (validation observation). The Interval Score is given by

$$r(\mathbf{y}^{rep}, \mathbf{y}_{V[\mathbf{s}]}) = (u - l) + \frac{2}{\gamma}(l - \mathbf{y}_{V[\mathbf{s}]})I_{\left[\mathbf{y}_{V[\mathbf{s}]}<l\right]} + \frac{2}{\gamma}(\mathbf{y}_{V[\mathbf{s}]} - u)I_{\left[\mathbf{y}_{V[\mathbf{s}]}>u\right]}, \quad \text{(B.3)}$$

where l and u represent for the forecaster quoted $\frac{\gamma}{2}$ and $1 - \frac{\gamma}{2}$ quantiles and $\mathbf{y}_{V[\mathbf{s}]}$ the sample validation vector. A natural choice for γ is 0.05 resulting in a range of 95% of credibility. Note that for each element of $\mathbf{y}_{V[\mathbf{s}]}$ we have one interval score measure. The global measure is obtained taking the average of the interval scores for all validation cases.

Log Predictive Score This measure evaluates the accuracy of the density forecasts using predictive log-scores. It is based on the predictive distribution q and on the observed $\mathbf{y}_{V[\mathbf{s}]}$,

$$r(\mathbf{y}^{rep}, \mathbf{y}_{V[\mathbf{s}]}) = -\,log\, \left[q(\mathbf{y}_{V[\mathbf{s}]})\right]. \quad \text{(B.4)}$$

Note that under the Gaussian model assumption, it is similar to the Mahalanobis distance in (B.2).

Appendix C. Markov chain Monte Carlo sampler

The prior distributions considered for the parameters in Section 2 and proposal densities used in the MCMC algorithm are detailed as follows.

1. $\sigma^2 \sim GI(a, b)$, $a, b > 0$. The proposed density in the MCMC sampler is:
$$ln(\sigma^2) \sim Normal(ln(\sigma^{2(k-1)}), \sigma^2_{(\sigma^2)}).$$
2. $\mu \sim Normal_n(\mathbf{0}, \tau^2_\mu)$, $\tau^2_\mu > 0$. The proposal density in the MCMC sampler is:
$$\mu \sim Normal(\mu^{(k-1)}, \sigma^2_{(\mu)}).$$
3. $\phi \sim Gama(1, c/med(u_s))$, with $c > 0$ and $med(u_s)$ denoting the median distance in the observed data. The proposal density in the MCMC sampler is:
$$ln(\phi) \sim Normal(ln(\phi^{(k-1)}), \sigma^2_{(\phi)}).$$
4. Jeffreys independent prior distribution Fonseca et al. (2008):
$$p(\nu) \propto \left(\frac{\nu}{\nu+3}\right)^{1/2} \left\{\psi'\left(\frac{\nu}{2}\right) - \psi'\left(\frac{\nu+1}{2}\right) - \frac{2(\nu+3)}{\nu(\nu+1)^2}\right\}^{1/2},$$

with $\psi'(a) = \frac{d\{\psi(a)\}}{da}$ the trigamma function. In the context of regression models, this prior distribution guarantees that the posterior distribution for ν is proper. The proposal density in the MCMC sampler is:

$$ln(\nu) \sim Normal(ln(\nu^{(k-1)}), \sigma^2_{(\nu)}).$$

Appendix C.1. GLG Bayesian model

We follow Palacios and Steel (2006) to obtain the posterior distribution of parameters in the GLG model. The vector $\mathbf{y}$ has conditional distribution given by

$$\mathbf{y} \mid \mu, \phi, \sigma^2, \mathbf{\Delta}, \tau^2 \sim Normal_n(\mu, \tau^2 I_n + \sigma^2(\mathbf{\Delta}^{-1/2}\, R\, \mathbf{\Delta}^{-1/2}))$$

with $\mathbf{\Delta} = diag(\delta_1, \ldots, \delta_n)$ and ϕ the spatial range parameter. Define $\Sigma = \tau^2 I_n + \sigma^2(\mathbf{\Delta}^{-1/2} R \mathbf{\Delta}^{-1/2})$.

1. $\sigma^2 \sim GI(a, b)$, $a, b > 0$. The proposed density in the MCMC sampler is:
$$ln(\sigma^2) \sim Normal(ln(\sigma^{2(k-1)}), \sigma^2_{(\sigma^2)}).$$
2. $\mu \sim Normal_n(0, \tau^2_\mu)$, $\tau^2_\mu > 0$. The proposal density in the MCMC sampler is:
$$\mu \sim Normal(\mu^{(k-1)}, \sigma^2_{(\mu)}).$$
3. $\phi \sim Gama(1, c/med(u_s))$, with $c > 0$ and $med(u_s)$ denoting the median distance in the observed data. The proposal density in the MCMC sampler is:
$$ln(\phi) \sim Normal(ln(\phi^{(k-1)}), \sigma^2_{(\phi)}).$$
4. $\upsilon \sim GIG(\zeta, \delta, \iota)$, $\zeta \in \mathbb{R}$, $\delta \in \mathbb{R}$ and $\iota \in \mathbb{R}$. The proposal density in the MCMC sampler is:
$$ln(\upsilon) \sim Normal(ln(\upsilon^{(k-1)}), \sigma^2_{(\upsilon)}).$$
5. $ln(\boldsymbol{\delta}) \mid \upsilon, \phi \sim Normal_n\left(-\frac{\upsilon}{2}\mathbf{1}, \upsilon\, R\right)$. The proposal in the MCMC sampler is:
The spatial region is divided in subregions and a random walk proposal density is used for each subregion. Palacios and Steel (2006) propose a independent sampler which might be more efficient than random walk proposals in the case of large datasets.

In Gaussian case, we run the sampler without the steps for υ (next to zero) and $\boldsymbol{\delta}$ (which is equal to $\mathbf{1}$).

Appendix D. Variance estimator

According to Robert and Casella (2009), the generic problem involves evaluating the integral

$$E_f(h(X)) = \int_{\chi} h(x)f(x)dx, \tag{D.1}$$

where χ denotes the set where the random variable X takes its values, which is usually equal to the support of the density f.

The principle of the Monte Carlo method for approximating equation (D.1) is to generate a sample $X_1, \ldots, X_n$ from the density f and proposed as an approximation to the empirical average

$$\bar{h}_n = \frac{1}{n}\sum_{j=1}^{n} h(x_j)$$

since $\bar{h}_n$ converges almost surely to $E_f(h(X))$ by the strong law of large numbers.

When $h^2(X)$ has a finite expectation under f the speed of convergence of $\bar{h}_n$ can be assessed, since the convergence takes place at a speed $O(\sqrt{n})$ and the asymptotic variance of the approximation is

$$var(\bar{h}_n) = \frac{1}{n}\int_{\chi} \left[h(x) - E_f(h(X))\right]^2 f(x)dx, \tag{D.2}$$

which can also be estimated from the sample $(X_1, \ldots, X_n)$ through

$$v_n = \frac{1}{n^2}\sum_{j=1}^{n} \left[h(x_j) - \bar{h}_n\right]^2.$$

Analogously to equation (D.2), we can obtain the variance of the estimators $\hat{\Psi}_{mc}$ and $\hat{\Psi}_{sir}$. Notice that from the equation (17) we obtain,

$$var(\hat{\Psi}_{mc}) = \frac{1}{I^2}\frac{1}{J^2}\sum_{i=1}^{I}\sum_{j=1}^{J}\left[r\left(y_{ij}^{rep}, \mathbf{y}_{V[\mathbf{s}^{(i)}]}\right) - \hat{\Psi}_{mc}\right]^2 \tag{D.3}$$

Thus,

$$var(\hat{\Psi}_{sir}) = \frac{1}{H^2}\frac{1}{I^2}\sum_{h=1}^{H}\sum_{i=1}^{I}\left[\Psi_{hi} - \hat{\Psi}_{sir}\right]^2. \tag{D.4}$$

is the SIR estimator variance, obtained from equation (19), where,

$$\Psi_{hi} = \frac{\sum_{j=1}^{J} r\left(y_{hj}^{rep}, \mathbf{y}_{V[\mathbf{s}^{(i)}]}\right) w_{hj}}{\sum_{j=1}^{J} w_{hj}}$$

According to Alqallaf and Gustafson (2001) to determine the variance of $\hat{\Psi}_{sir}$, consider the terms Ψ_{hi} as elements of an H by I matrix, and note that each element has the same distribution. We can consider a term that denote the variance of this distribution, a term the common covariance of any pair of distinct elements from the same row, and another term be the common covariance of any pair of distinct elements from the same column. Notice that any two elements from different rows and columns are uncorrelated. Therefore,

$$\begin{aligned} var(\hat{\Psi}_{sir}) &= \frac{1}{H^2}\left\{\frac{1}{I^2}\sum_{h=1}^{H}\sum_{i=1}^{I}(\Psi_{hi}-\hat{\Psi}_{sir})^2 + 2\sum_{h=1}^{H}\sum_{i=1}^{I}\sum_{j=1}^{i-1}(\Psi_{hi}-\hat{\Psi}_{sir})(\Psi_{hj}-\hat{\Psi}_{sir})\right. \\ &+ \left. 2\sum_{i=1}^{I}\sum_{h=1}^{H}\sum_{j=1}^{h-1}(\Psi_{hi}-\hat{\Psi}_{sir})(\Psi_{ji}-\hat{\Psi}_{sir})\right\}. \end{aligned}$$

Appendix D.1. SIR estimator details

We draw a MCMC sample from $g(\theta)$, which is then reweighted using importance sampling to obtain $p(\theta \mid \mathbf{s})$. The same posterior sample is used for every split $\mathbf{s}$ considered, saving computational time.

The equation weighting term $w_{hj} = \frac{p(\theta_{hj}|\mathbf{y}_{T[\mathbf{s}]})}{g(\theta_{hj})}$ can be obtained by applying the logarithm of the ratio as follows

$$\begin{aligned} log(w_{hj}) &= log\left\{\frac{p(\theta_{hj} \mid \mathbf{y}_{T[\mathbf{s}]})}{g(\theta_{hj})}\right\} = log\left\{\frac{f(\mathbf{y}_{T[\mathbf{s}]} \mid \theta_{hj})}{f(\mathbf{y} \mid \theta_{hj})^{\alpha}}\right\} \\ &= log\, f(\mathbf{y}_{T[\mathbf{s}]} \mid \theta_{hj}) - \alpha\, logf(\mathbf{y} \mid \theta_{hj}) \end{aligned} \tag{D.5}$$

Appendix D.2. Stratified Variance $var(\hat{\Psi}_k)$

For the MC estimator, we have each $r\left(y_{ij}^{rep}, \mathbf{y}_{V[\mathbf{s}^{(i)}]}\right)$ as the discrepancy distribution. Then

$$\hat{\Psi}_{mc_k} = \frac{1}{I_k}\sum_{i=1}^{I_k}\frac{1}{J}\sum_{j=1}^{J} r_k\left(y_{ij}^{rep}, \mathbf{y}_{V[\mathbf{s}^{(i)}]}\right)$$

is the MC estimator in each stratum. We can obtain the variance of the stratified MC estimator as

$$\begin{aligned} var(\hat{\Psi}^{st}) &= \frac{1}{n^2}\sum_{k=1}^{K} n_k(n_k - n_{V_k})\frac{s_k^2}{n_{V_k}} \\ &= \sum_{k=1}^{K}\frac{w_k}{n}(n_k - n_{V_k})\frac{s_k^2}{n_{V_k}} \\ &= \sum_{k=1}^{K}\frac{w_k}{n}(1 - f_{V_k})\frac{s_k^2}{n_{V_k}} \end{aligned} \tag{D.6}$$

where

$$s_k^2 = \frac{1}{(n_{V_k} - 1)}\sum_{i=1}^{n_k}(r_{ki} - \hat{\Psi}_k)^2$$

and r_k denotes any discrepancy function. Note that equation (D.6) can be written as

$$\begin{aligned} var(\hat{\Psi}^{st}) &= var\left(\sum_{k=1}^{K}\hat{\Psi}_k^{st}\right) \\ &= var\left(\sum_{k=1}^{K} w_k\hat{\Psi}_k\right) \\ &= \sum_{k=1}^{K} w_k^2\, var(\hat{\Psi}_k) \end{aligned} \tag{D.7}$$

Therefore, $var(\hat{\Psi}_k^{st}) = var(w_k\hat{\Psi}_k) = w_k^2\, var(\hat{\Psi}_k), \forall\, k = 1, \ldots, K$. Analogously, we have a similar result for the SIR estimator variance.

References

C. P. Robert, The Bayesian Choice, second edition ed., Springer, New York, 2007.

A. Majumdar, A. E. Gelfand, Multivariate spatial modeling for geostatistical data using convolved covariance functions, Mathematical Geology 39 (2007) 225–245.

T. V. Apanasovich, M. G. Genton, Cross-covariance functions for multivariate random fields based on latent dimensions, Biometrika 97 (2010) 15–30.

T. C. O. Fonseca, M. F. J. Steel, Non- gaussian spatiotemporal modelling through scale mixing, Biometrika 98 (2011) 761–774.

R. S. Bueno, T. C. O. Fonseca, A. M. Schmidt, Accounting for covariate information in the scale component of spatial-temporal mixing models, Spatial Statistics 22 (2017) 196–218.

P. J. Diggle, Statistical Analysis of Spatial and Spatio-Temporal Point Patterns, Chapman and Hall/CRC, 2014.

P. Burman, A comparative study of ordinary cross-validation, v-fold cross-validation and the repeated learning-testing methods, Biometrika 76 (1989).

P. Thall, K. Russel, R. Simon, Variable selection in regression via repeated data splitting 6 (1997) 416–434.

A. Gelfand, Model Determination Using Samplings Based Methods., Chapman & Hall, Boca Raton, FL., 1996.

E. C. Marshall, D. J. Spiegelhalter, Approximate cross-validatory predictive checks in disease mapping models, Statistics in medicine 22 (2003) 1649–1660.

H. S. Stern, N. Cressie, Posterior predictive model checks for disease mapping models, Statistics in medicine 19 (2000) 2377–2397.

A. Gelman, J. Hwang, A. Vehtari, Understanding predictive information criteria for bayesian models, Statistics and Computing 24 (2014) 997–1016. URL: http://dx.doi.org/10.1007/s11222-013-9416-2. doi:10.1007/s11222-013-9416-2.

S. Watanabe, Asymptotic equivalence of bayes cross validation and widely applicable information criterion in singular learning theory, J. Mach. Learn. Res. 11 (2010) 3571–3594. URL: http://dl.acm.org/citation.cfm?id=1756006.1953045.

L. Li, S. Qiu, B. Zhang, C. Feng, Approximating cross-validatory predictive evaluation in bayesian latent variable models with integrated is and waic, Statistics and Computing 26 (2016) 881–897.

A. E. Gelfand, D. K. Dey, H. Chang, Model determination using predictive distributions with implementation via sampling-based methods, Technical Report, Department of Statistics Stanford University, 1992.

A. Vehtari, A. Gelman, J. Gabry, Practical bayesian model evaluation using leave-one-out cross-validation and waic, Statistics and Computing 27 (2017) 1413–1432. URL: https://doi.org/10.1007/s11222-016-9696-4. doi:10.1007/s11222-016-9696-4.

F. Alqallaf, P. Gustafson, On cross-validation of Bayesian models, The Canadian Journal of Statistis 29 (2001) 333–340.

P. J. Diggle, R. Menezes, T.-l. Su, Geostatistical inference under preferential sampling, Journal of the Royal Statistical Society: Series C (Applied Statistics) 59 (2010) 191–232. URL: http://dx.doi.org/10.1111/j.1467-9876.2009.00701.x. doi:10.1111/j.1467-9876.2009.00701.x.

G. S. Ferreira, D. Gamerman, Optimal design in geostatistics under preferential sampling, Bayesian Analysis 10 (2015) 711–735.

D. Pfeffermann, F. Moura, P. Silva, Multi-level modeling under informative sampling, Biometrika 93 (2006) 943–959.

T. C. O. Fonseca, M. A. R. Ferreira, H. S. Migon, Objective bayesian analysis for the student-t regression model, Biometrika 95 (2008) 325–333.

P. C. Mahalanobis, On the generalized distance in statistics, Proceedings of the National Institute of Sciences (Calcutta) 2 (1936) 49–55.

T. S. Breusch, J. C. Robertson, A. H. Welsh, The emperor's new clothes: a critique of the multivariate t regression model, Statistica Neerlandica 51 (1997) 269–286. URL: http://dx.doi.org/10.1111/1467-9574.00055. doi:10.1111/1467-9574.00055.

J. Roislien, H. Omre, T-distributed random fields: A parametric model for heavy-tailed well-log data, Mathematical Geology 38 (2006) 821–849.

M. B. Palacios, M. F. J. Steel, Non-gaussian bayesian geostatistical modeling, Journal of the American Statistical Association 101 (2006) 604–618.

D. Gamerman, H. Lopes, Markov Chain Monte Carlo: Stochastic Simulation for Bayesian Inference, Texts in Statistical Science, Taylor & Francis, 2006.

D. Arthur, S. Vassilvitskii, K-means++: the advantages of careful seeding, Technical Report, Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms. Society for Industrial and Applied Mathematics Philadelphia, PA, USA, 2007.

A. O'Hagan, Fractional bayes factors for model comparison, Journal of the Royal Statistical Society. Series B 57 (1995) 99–138.

W. G. Cochran, Sampling Techniques, Wiley Student Edition, 3 ed., Wiley, 1999.

N. Cressie, Statistics for Spatial Data, New York: Wiley, 1993.

M. M. J. Katzfuss, J. Hu, V. E. Johnson, Assessing fit in bayesian models for spatial processes, Environmetrics 25 (2014) 584–595.

A. D. Gordon, A survey of constrained classification, Comput. Stat. Data Anal. 21 (1996) 17–29.

S. Banerjee, A. E. Gelfand, Bayesian wombling: Curvilinear gradient assessment under spatial process models, Journal of the American Statistical Association 101 (2006) 1487–1501. URL: http://EconPapers.repec.org/RePEc:bes:jnlasa:v:101:y:2006:p:1487-1501.

S. Liang, S. Banerjee, B. P. Carlin, Bayesian wombling for spatial point processes, Biometrics 65 (2009) 1243–1253. URL: http://EconPapers.repec.org/RePEc:bla:biomet:v:65:y:2009:i:4:p:1243-1253.

M. Plummer, N. Best, K. Cowles, K. Vines, Coda: Convergence diagnosis and output analysis for mcmc, R News 6 (2006) 7–11. URL: https://journal.r-project.org/archive/.

V. G. R. Lobo, T. C. O. Fonseca, Bayesian residual analysis for spatially correlated data, Statistical Modelling (2019).

L. S. Bastos, A. O'Hagan, Diagnostics for gaussian process emulators., Technometrics 51 (2008) 425–438.

T. Gneiting, A. E. Raftery, Strictly proper scoring rules, prediction and estimation, Journal of the American Statistical Association 102 (2007) 360–378.

C. P. Robert, G. Casella, Introducing Monte Carlo Methods with R (Use R), 1st ed., Springer-Verlag, Berlin, Heidelberg, 2009.