

**Manuscript version: Author's Accepted Manuscript**

The version presented in WRAP is the author's accepted manuscript and may differ from the published version or Version of Record.

**Persistent WRAP URL:**

<http://wrap.warwick.ac.uk/130116>

**How to cite:**

Please refer to published version for the most recent bibliographic citation information. If a published version is known of, the repository item page linked to above, will contain details on accessing it.

**Copyright and reuse:**

The Warwick Research Archive Portal (WRAP) makes this work by researchers of the University of Warwick available open access under the following conditions.

© 2019 Elsevier. Licensed under the Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International <http://creativecommons.org/licenses/by-nc-nd/4.0/>.



**Publisher's statement:**

Please refer to the repository item page, publisher's statement section, for further information.

For more information, please contact the WRAP Team at: [wrap@warwick.ac.uk](mailto:wrap@warwick.ac.uk).

# Bayesian Cross-Validation of Geostatistical Models

Viviana G R Lobo<sup>1</sup>, Thaís C O Fonseca<sup>1,\*</sup>, Fernando A S Moura<sup>1</sup>

<sup>1</sup>*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 much as conclusions may be sensitive to the partition of data into training and validation cases. This is an obvious issue related to the basic validation scheme which selects a subset of the data to leave out of estimation and to make predictions with an assumed model. In this setup, only a few out-of-sample locations are usually selected to validate the model. On the other hand, the 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 the estimation of parameters usually requires computationally intensive methods. The purpose of this work is to use 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. We consider the sampling design for selecting the training and validation sets by assigning a probability distribution to the possible data partitions. To deal with the computational burden of cross-validation, we estimate discrepancy functions in a computationally efficient manner based on the importance weighting of 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. We also illustrate the advantages of our proposal with simulated examples of homogeneous and inhomogeneous spatial processes and with an application to rainfall dataset in Rio de Janeiro.

---

\*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](mailto:viviana@dme.ufrj.br) (Viviana G R Lobo), [thais@im.ufrj.br](mailto:thais@im.ufrj.br) (Thaís C O Fonseca), [fmoura@im.ufrj.br](mailto:fmoura@im.ufrj.br) (Fernando A S Moura)

*Keywords:* Validation samples, Data partition, Spatial processes, Model criticism, Discrepancy function, Importance sampling.

---

## 1. Introduction

In many practical problems, a researcher is interested in modelling a phenomenon which 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, model adequacy checking should not be based on model parameter estimation, hypothesis testing 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 an 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. This paper proposes a model comparison approach, based on predictive discrepancies for out-of-sample observations, which aim to be representative of the spatial process as a whole. In this setting, cross-validation techniques are considered and feasible computation is proposed.

The usual approaches for model checking in spatial statistics are based on selecting a subset from the locations to make prediction with an assumed model. The observed values, which were left out of the estimation procedure, are then compared with the predictions. However, the choice of locations is often based on a small subset of the data and random sampling does not guarantee spatial coverage of the region of interest. 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 5 locations were taken out 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 left out 3 locations for predictive performance assessment in temperature data. This validation procedure might fail in 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 spatial model checking.

Several 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. \(1997\)](#) demonstrate that repeated data splitting is preferred over  $k$ -fold cross-validation. They propose the application of cross-validation to a large number of randomly generated partitions of data.

[Gelfand \(1996\)](#) proposes the conditional predictive ordinate (CPO), which represents a useful model assessment tool widely used in the statistical literature under various contexts, such as the detection of surprising observations. CPO is based on leave-one-out cross-validation (LOO-CV) approach.

From a Bayesian standpoint, [Marshall and Spiegelhalter \(2003\)](#) and [Burman \(1989\)](#), amongst others, show that cross-validation can be computationally very expensive, since usually a full Markov chain Monte Carlo (MCMC) analysis has to be repeated, several times, leaving out each validation set. [Stern and Cressie \(2000\)](#) consider importance weighting and re-sampling methods for the posterior predictive model checking via CPO and posterior predictive p-value. [Gelman et al. \(2014\)](#) review Akaike, deviance and Watanabe-Akaike (WAIC - [Watanabe, 2010](#)) information criteria, from a Bayesian perspective. [Li et al. \(2016\)](#) discuss two predictive evaluation methods based on Importance Sampling ([Gelfand et al., 1992](#)) and WAIC, with possibly correlated latent variables via LOO-CV. [Vehtari et al. \(2017\)](#) apply the same approach and introduce 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 validation 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. Model checking is based on estimating discrepancy functions, which are statistical measures commonly used in the literature for model comparison.

Many works have exploited cross-validation methods for univariate and multivariate data analysis. [Arlot and Celisse \(2010\)](#) review some cross-validation strategies. [Burman et al. \(1994\)](#) consider cross-validation for correlated observations of stationary processes. [Bergmeir and Benitez \(2012\)](#)

discuss cross-validation techniques applied to time series data. Recently, [Bergmeir et al. \(2018\)](#) have investigated k-fold cross-validation applied to autoregressive models. [Roberts et al. \(2017\)](#) have considered blocking designs to account for the data dependency in hierarchical, temporal and spatial data. In particular for spatial processes, few proposals deal with cross-validation. For instance, the usual setup for model checking in geostatistics is to evaluate the prediction performance for a single or a few selected validation sets.

However, the choice of observation sites for validation of spatial models is not always robust enough to the considered sampling or allocation of sites. In general, it does not consider the sampling design for selecting the training and validation sets. 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 to large spatial data is a computational challenge, due to the difficulty in applying traditional prediction methods in a time-tolerant boundary. In applications involving high-resolution geocoded data analysis, large covariance matrices need to be inverted in the prediction and estimation steps. The computational effort is of cubic order on the number of locations. In this scenario, likelihood, covariance or process approximations could be used to overcome the computational burden. Some examples can be found in [Vecchia \(1988\)](#) and [Stein et al. \(2004\)](#), where they use conditional distributions that depends on the nearest neighbours only. [Furrer et al. \(2006\)](#) propose the tapering approach, which sets the covariance to zero beyond a certain range. [Banerjee et al. \(2008\)](#) propose the predictive processes based on low-rank models, which achieve the computational feasibility by writing the spatial component as a linear combination of spatial basis functions. More recently, [Datta et al. \(2015\)](#) have propose a spatial process called the Nearest Neighbor Gaussian Process, where the sparse matrices of covariance depends on the definition of a set of nearest neighbours. Notice that, if we were to make prediction for several vectors of points, the cross-validation procedure would become computationally prohibitive even for moderate and small datasets such as the ones considered in this paper. Thus, more sophisticated approach are useful, both to reduce the final cost and increase efficiency.

In this paper, our proposal extends the work of [Alqallaf and Gustafson \(2001\)](#) to correlated data modelling. We allow for uncertainty in the choice of the validation sets in spatial data analysis by considering a probability dis-

tribution for the possible data partitions into validation and training cases. In particular, we propose three distributions for selecting spatial locations. The first proposal is a uniform prior which results in the split vectors being independently generated and uniformly distributed over the entire spatial domain. This approach is useful if the area under study is homogeneous in space and distribution of locations is not clustered in subregions. The second proposal is to set 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 a uniform prior in several strata. For that purpose, we adopt spatially 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 allows for identification of subregions 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 runs.

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

## 2. Geostatistical modelling

Let us consider that data are obtained by sampling a spatially continuous phenomenon  $S(x)$  at a finite number of locations  $x_1, \dots, x_n$  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

$$Y_i = \mu + S(x_i) + Z_i \quad i = 1, \dots, n, \quad (1)$$

where  $\mu$  represents the mean and the  $Z_i$ 's are mutually independent, zero-mean random variables with variance  $\tau^2$  called nugget effect, which can be

interpreted as sampling error or inherent variability (or both). The underlying spatial process  $\{S(x) : x \in \mathbb{R}^2\}$  is a stationary process with zero mean, constant variance  $\sigma^2$  and correlation function  $\rho(u; \phi)$ , where  $\phi$  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 convenient properties of the multivariate Gaussian distributions. However, Gaussian processes are not able to accommodate common characteristics of spatial applications such as the presence of outliers, skewness and non-constant variance over space. Some discussion and examples of departures from normality may be found at [Palacios and Steel \(2006\)](#), [Fonseca and Steel \(2011\)](#) and [Bueno et al. \(2017\)](#).

The next subsection considers a more general model for spatial data analysis which allows for non-Gaussian behaviour of spatial data. This model is compared to the usual Gaussian model using our cross-validation techniques schemes. Inference for model parameters and predictive distributions are also described.

### 2.1. Spatial mixture model

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

#### (GM) Gaussian model:

As a benchmark we assume the Gaussian model, where the distribution of  $\mathbf{Y}$  is given by

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

#### (STM) Student-t model:

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

$$\mathbf{y} \mid \mu, \sigma^2, \phi, \nu \sim ST(\mathbf{1}\mu, \nu, \tau^2 I_n + \sigma^2 R). \quad (3)$$

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 \(2006\)](#).

**(GLGM) 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,

$$\mathbf{y} \mid \mu, \sigma^2, \phi, \mathbf{\Delta} \sim N(\mathbf{1}\mu, \tau^2 I_n + \sigma^2(\mathbf{\Delta}^{-1/2} R \mathbf{\Delta}^{-1/2})). \quad (4)$$

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

## 2.2. Inference for geostatistical models

We follow the Bayesian approach to inference and prediction. The posterior distribution of model parameters  $\theta$ ,  $p(\theta \mid \mathbf{y}) \propto f(\mathbf{y} \mid \theta)\pi(\theta)$ , 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  $\rho(\|u\|, \phi) = \exp\{-\|u\|/\phi\}$ , where  $\phi > 0$  is the range parameter which controls the rate of decay with distance  $u$ .

For all models we assign the same independent non-informative priors to  $\mu$ ,  $\sigma^2$  and  $\phi$ . In particular,  $\mu \sim N(0, \tau_\mu^2)$  with large value of  $\tau_\mu^2$ ,  $\sigma^{-2} \sim \text{Gamma}(a, b)$  and  $\tau^{-2} \sim \text{Gamma}(a, b)$  with small values of  $a$  and  $b$ . For the range parameter  $\phi$  we take into account that the prior is critically dependent on the scale of distances between locations. So,  $\phi \sim \text{Gamma}(1, c/\text{med}(d))$ , with  $\text{med}(d)$  representing the median of distances in the data. Notice that we assumed a simple mean function  $\mu(x) = \mu, \forall x \in A$ , however, alternative models may be considered. For instance,  $\mu(x) = \boldsymbol{\beta}\mathbf{u}'(x)$ , with  $\mathbf{u}(x)$  a vector of spatial covariates and  $\boldsymbol{\beta}$  regression coefficients. In this case,  $\boldsymbol{\beta} \sim N(\mathbf{0}, I_d\tau_\beta^2)$ , with  $d = \text{dim}(\boldsymbol{\beta})$ .

If Gaussianity is assumed for  $S(x)$  as in equation (2), then the likelihood function for the spatial model is given by

$$f(\mathbf{y} \mid \mu, \sigma^2, \phi) = (2\pi)^{-n/2} |\Sigma|^{-1/2} \exp\left\{-\frac{1}{2}(\mathbf{y} - \mathbf{1}\mu)' \Sigma^{-1}(\mathbf{y} - \mathbf{1}\mu)\right\}, \quad (5)$$

that is,  $\mathbf{y} = (y_1, \dots, y_n)'$  follows an n-variate Normal distribution with mean  $\mu$  and covariance matrix  $\Sigma = \tau^2 I_n + \sigma^2 R$ . The posterior samples for model parameters  $\mu, \sigma^2, \phi$  are obtained by the Gibbs algorithm with Metropolis-Hastings steps considering random walk proposals.

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

$$f(\mathbf{y} \mid \mu, \sigma^2, \phi, \nu) = \frac{\Gamma(\frac{\nu+n}{2})}{\Gamma(\frac{\nu}{2})(\nu\pi)^{n/2}|\Sigma|^{1/2}} \left[ 1 + \frac{(\mathbf{y} - \mathbf{1}\mu)' \Sigma^{-1}(\mathbf{y} - \mathbf{1}\mu)}{\nu} \right]^{-(\nu+n)/2}, \quad (6)$$

with  $\Gamma(\cdot)$  the gamma function, mean  $\mu$  and covariance matrix  $\Sigma = \tau^2 I_n + \sigma^2 R$ . For the degrees of freedom parameter  $\nu$  we assign a Jeffreys prior distribution, as proposed in (Fonseca et al., 2008). The posterior samples of the model parameters  $\mu, \sigma^2, \phi, \nu$  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  $i = 1, \dots, n$ , yielding to a multivariate Gaussian distribution for  $\mathbf{y}$  conditional on  $\boldsymbol{\delta} = (\delta_1, \dots, \delta_n)$ . The resulting likelihood function resembles equation (5) with  $\Sigma$  replaced with  $\Sigma = \tau^2 I_n + \sigma^2 (\boldsymbol{\Delta}^{-1/2} R \boldsymbol{\Delta}^{-1/2})$ , with  $\boldsymbol{\Delta} = \text{Diag}(\delta_1, \dots, \delta_n)$ . For the parameter  $\nu$  we set a  $GIG(0, \delta, \nu)$  (generalized inverse Gaussian) prior. Notice that very small values of  $\nu$  (around 0.01) correspond to near Normality while large values of  $\nu$  (of the order of say 3) indicate very thick tails and  $\ln(\boldsymbol{\delta}) \sim N_n(-\frac{\nu}{2}\mathbf{1}, \nu R)$ . The posterior samples of the model parameters are obtained by the Gibbs algorithm with Metropolis-Hastings steps for  $\phi, \nu$  and  $\boldsymbol{\delta}$  which are based on random walk proposals.

For prediction, let  $\mathbf{y} = (\mathbf{y}_V, \mathbf{y}_T)$  with  $\mathbf{y}_V$  and  $\mathbf{y}_T$  representing out-of-sample and training observations, respectively. Conditional predictive distributions are obtained in closed form for all considered models. For the Gaussian model, the conditional distributions remain Gaussian with  $E[Y_V \mid \mathbf{y}_T] = \mathbf{1}\mu + \Sigma_{0T} \Sigma_{TT}^{-1}(\mathbf{y}_T - \mathbf{1}\mu)$  and  $\text{Var}[Y_V \mid \mathbf{y}_T] = \Sigma_{VV} - \Sigma_{VT} \Sigma_{TT}^{-1} \Sigma_{TV}$ .

$$\Sigma = \begin{pmatrix} \Sigma_{VV} & \Sigma_{VT} \\ \Sigma_{TV} & \Sigma_{TT} \end{pmatrix}$$

For the Student-t model the conditional distributions remain Student-t with degrees of freedom  $\nu_{V|s_j} = \nu + d_T$ , with mean as in the Gaussian model and variance the same as in the Gaussian case scaled by  $\xi(T) = \frac{\nu + (\mathbf{y}_T - \mathbf{1}\mu_T)' \Sigma_{TT}^{-1} (\mathbf{y}_T - \mathbf{1}\mu_T)}{\nu + d_T}$ ,

and  $d_s$  the dimension of vector  $\mathbf{y}_T$ . For the Gaussian-Log-Gaussian model case and conditional on the mixing variables  $\boldsymbol{\delta}$ , the predictive distributions are analogous to the Gaussian case with  $\Sigma = \tau^2 I_n + \sigma^2 (\boldsymbol{\Delta}^{-1/2} R \boldsymbol{\Delta}^{-1/2})$ .

### 3. Cross-validation of Bayesian models for spatially correlated data

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.

#### 3.1. Accounting for uncertainty in the data partition

Consider observations  $\mathbf{y} = (y_1, y_2, \dots, y_n)$  arising from the process  $Y(x)$  and locations  $x = (x_1, \dots, x_n)$  described in equation (1). Let  $\theta$  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 sets in each split vector considered, respectively.

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  $n = n_T + n_V$ . 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  $\mathbf{s} = (s_1, \dots, s_n)$  of the same dimension of observed data, indicating for each location  $x_k$ ,  $k = 1, \dots, n$ , 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 cross-validation results considering many data partitions. Indeed, this averaging is done with respect to the distribution of  $\mathbf{s}$ , that is,  $p(\mathbf{s}) = p(s_1, s_2, \dots, s_n)$ .

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

The second alternative is to assume a probability distribution for the sets based on Euclidean distances considering a finite set of spatial sample locations  $\tilde{x}_k = (x_0, x_1, \dots, x_{k-1})$ , for  $k = 1, \dots, n$  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 on 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,  $p(x_k | \tilde{x}_k)$ , for  $k = 1, \dots, n$ . The prior via distances can be obtained as

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

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

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

This prior assumes different probabilities for the sample selection locations and may be potentially useful in irregular spatial regions as often seen in data applications.

Notice that these two choices of prior might not be reasonable if there is preferentiability in the selection of  $x = (x_1, \dots, x_n)$ . To account for this possible feature of spatial locations in the prediction evaluation, we consider an extension of this prior in Section 4.

After choosing 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  $\mathbf{s}$ ,  $p(\theta | \mathbf{y}_{T[\mathbf{s}]})$  is defined as the posterior distribution of  $\theta$  given the training data only. Thus, using the Bayes theorem, the posterior distribution is given by

$$p(\theta | \mathbf{y}_{T[\mathbf{s}]}) \propto f(\mathbf{y}_{T[\mathbf{s}] | \theta})\pi(\theta), \quad (7)$$

where for each split vector  $\mathbf{s}$  there is a single corresponding data vector  $\mathbf{y}_{T[\mathbf{s}]}$ . Conditional on model parameters,  $\mathbf{y}^{rep}$  is simply distributed according to the sampling model assumed for the data, i.e.,  $[\mathbf{y}^{rep} | \theta, \mathbf{y}_{T[\mathbf{s}]}]$ , 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  $f(\mathbf{y}^{rep} | \mathbf{y}_{T[\mathbf{s}]})$  in a composition sampling algorithm.

### 3.2. Expected discrepancy estimation

Our cross-validation assessments are based on  $r(\mathbf{y}_{V[\mathbf{s}]}^{rep}, \mathbf{y}_{V[\mathbf{s}]})$ , called a discrepancy function for checking model adequacy, whose expectation under

$f(\mathbf{y}_{V[s]}^{rep} \mid \mathbf{y}_{T[s]})$  is evaluated. It requires the distribution of the replicated response vector  $\mathbf{y}_{V[s]}^{rep}$  for validation sets. In particular, we are interested in computing the expectation bellow

$$\Psi = E \left\{ r(\mathbf{y}_{V[s]}^{rep}, \mathbf{y}_{V[s]}) \right\}. \quad (8)$$

The expected value in (8) represents a statistical measure for comparing Bayesian models and  $r$  represents a discrepancy function (see [Appendix A](#)). Notice that  $r$  depends on two unknown quantities  $\mathbf{y}_{V[s]}^{rep}$  and  $\mathbf{s}$ , thus  $\Psi$  can be computed as

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

The Monte Carlo estimator for the expected discrepancy is given by

$$\hat{\Psi} = \frac{1}{I} \sum_{i=1}^I E \left\{ r(\mathbf{y}_{V[s^{(i)}}]^{rep}, \mathbf{y}_{V[s^{(i)}}]) \mid \mathbf{y}_{T[s^{(i)}}] \right\}. \quad (9)$$

The split vectors  $\mathbf{s}^{(1)}, \mathbf{s}^{(2)} \dots, \mathbf{s}^{(I)}$  are simulated independently from  $p(\mathbf{s})$  and  $I$  represents the number of splits. If the posterior predictive distribution  $f(\mathbf{y}^{rep} \mid \mathbf{y}_{T[s]})$  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

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

Let  $(\theta_{ij}, \mathbf{y}_{V[s^{(i)}}]_j^{rep})$ ,  $i = 1, \dots, I$  and  $j = 1, \dots, J$  be samples from the joint conditional distribution of  $\theta$  and  $\mathbf{y}_{V[s]}^{rep}$ ,  $f(\mathbf{y}_{V[s]}^{rep} \mid \theta, \mathbf{y}_{T[s]}) p(\theta \mid \mathbf{y}_{T[s]})$ , then Algorithm 1 describes how to compute (9) by simulating from the posterior distribution of model parameters via MCMC. This approach is based on obtaining one MCMC sample for each split  $s$ . The *monte carlo* (MC) estimator is given by

$$\hat{\Psi}_{mc} = \frac{1}{I} \sum_{i=1}^I \frac{1}{J} \sum_{j=1}^J r(\mathbf{y}_{V[s^{(i)}}]_j^{rep}, \mathbf{y}_{V[s^{(i)}}]), \quad (10)$$

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 (8). Notice that (10) requires a MCMC sample for each validation set sampled from  $p(\mathbf{s})$ . This is often very expensive.

---

**Algorithm 1:** Monte Carlo (MC) estimator

---

1. Simulate independent split vectors  $\mathbf{s}^{(1)}, \mathbf{s}^{(2)}, \dots, \mathbf{s}^{(I)}$  from  $p(\mathbf{s})$ ;
  2. **for** each  $\mathbf{s}^{(i)}$  **do**
    - | use a MCMC run to draw a sample  $\theta_{i1}, \dots, \theta_{iJ}$  from  $p(\theta \mid \mathbf{y}_{T[\mathbf{s}^{(i)})}$ );
  - end**
  3. **for** each  $(i, j)$  **do**
    - | simulate  $\mathbf{y}_{V[\mathbf{s}^{(i)}]j}^{rep}$  from  $f(\mathbf{y}_{V[\mathbf{s}^{(i)}]}^{rep} \mid \theta_{ij}, \mathbf{y}_{T[\mathbf{s}^{(i)})}$ );
  - end**
- 

Aiming to reduce the computational cost, we consider the *importance sample estimator* (SIR), which requires only a handful of MCMC runs as an alternative estimate of expression (8). 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

$$g(\theta) \propto f(\mathbf{y} \mid \theta)^\alpha \pi(\theta), \quad (11)$$

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. Alqallaf and Gustafson (2001) claim that raising the whole-data likelihood to the power  $\alpha$  has the effect of flattening the posterior to a degree commensurate with conditioning only on a fraction  $\alpha$  of the data. The function  $g(\theta)$  is the same function employed in fractional Bayes factor (O’Hagan, 1995).

The SIR estimator is defined as the average of importance sampling estimate of  $E \left[ r(\mathbf{y}_{V[\mathbf{s}^{(i)}]}^{rep}, \mathbf{y}_{V[\mathbf{s}^{(i)})} \mid \mathbf{y}_{T[\mathbf{s}^{(i)})} \right]$ , across the  $I$  independent splits and the  $H$  independent samples from  $g(\theta)$ ,

$$\hat{\Psi}_{sir} = \frac{1}{H} \sum_{h=1}^H \frac{1}{I} \sum_{i=1}^I \frac{\sum_{j=1}^J r \left( \mathbf{y}_{V[\mathbf{s}^{(i)}]hj}^{rep}, \mathbf{y}_{V[\mathbf{s}^{(i)})} \right) w_{ihj}}{\sum_{j=1}^J w_{ihj}}, \quad (12)$$

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

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

Importance weighting is considered to obtain the desired weights using the importance distribution in (11). It is worth noting that the chosen importance function  $g(\theta)$  has the same support as the function  $p(\theta \mid \mathbf{y}_{V[\mathbf{s}^{(i)}]})$ . Thus, the expectation of interest exists due to the support assumption being satisfied. Furthermore, the rate of convergence of the proposed estimator depends on the ratio between the importance and the target distribution. As the importance function selected is a flattened version of the target, the convergence of the SIR estimator is guaranteed. For details see [Geweke \(1989\)](#) and [Robert and Casella \(2004\)](#). Our proposal is described in Algorithm 2.

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 B](#) shows how to determine a standard error of  $\hat{\Psi}_{mc}$  and  $\hat{\Psi}_{sir}$  estimators.

---

**Algorithm 2:** Sampling Importance Resampling (SIR) estimator

---

1. Simulate independent split vectors  $\mathbf{s}^{(1)}, \mathbf{s}^{(2)}, \dots, \mathbf{s}^{(I)}$  from  $p(\mathbf{s})$ ;
  2. Let  $\theta_{h1}, \dots, \theta_{hJ}$  be the  $h$ th of  $H$  independent MCMC samples simulated from  $g(\theta)$  ;
  3. **for each split vector  $i$  do**
    - Draw  $y_{V[\mathbf{s}^{(i)}]_{hj}}^{rep}$  from  $f(\mathbf{y}_{V[\mathbf{s}^{(i)}]}^{rep} \mid \theta_{hj}, \mathbf{y}_{T[\mathbf{s}^{(i)}]})$ , for  $h = 1, \dots, H$ ,
    - $j = 1, \dots, J$ ;
  - end**
  4. Each of these  $H$  samples yields an importance sampling estimate of  $E[r(\mathbf{y}_{V[\mathbf{s}^{(i)}]}^{rep}, \mathbf{y}_{V[\mathbf{s}^{(i)}]}) \mid \mathbf{y}_{T[\mathbf{s}^{(i)}]}]$ ;
- 

So far we have considered uniform prior distributions on the possible sets. This assumption would not be adequate if there are clusters in locations, heterogeneity or preferential sampling. In the next section, we propose a uniform prior on subregions which allow for more accurate estimation of discrepancy functions in non-homogeneous spatial domains.

---

<sup>1</sup>The weights are obtained in [Appendix B.1](#).

#### 4. Accounting for heterogeneity in the spatial domain

This section proposes a stratified sampling scheme to improve the accuracy of our cross-validation estimator and identify much more precisely the regions where departures from the assumed model are more evident. In stratified sampling, the region of  $n$  locations is first divided into subregions which are called *strata* of sizes  $n_1, n_2, \dots, n_K$ , respectively. These subregions 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.,  $n_T = \sum_{k=1}^K n_{T_k}$  where each term of this summation represents the training data size in each stratum  $k = 1, \dots, K$ . Analogously, the full validation data size is  $n_V = \sum_{k=1}^K n_{V_k}$ . If a simple random sample is taken in each stratum, the whole procedure is described as stratified random sampling. The following notations in Table 1 refer to stratum  $k$ .

Table 1: 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 subregions, where each subregion 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  $k = 1, \dots, K$  represents the stratum and  $i = 1, 2, \dots, I_k$  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,  $I_k$ ,  $k = 1, 2, \dots, K$ . Note that the sizes of split vectors  $I_k$  do not need to be the same. The split vector in each stratum  $\mathbf{s}^{(1,k)}, \dots, \mathbf{s}^{(I_k,k)}$  is jointly generated from  $p(\mathbf{s})$ . Thus, the  $i$ -th split vector of all strata  $\mathbf{s}^{(i)}$  is given by  $\mathbf{s}^{(i)} = (\mathbf{s}^{(i,1)}, \mathbf{s}^{(i,2)}, \dots, \mathbf{s}^{(i,K)})$ ,  $i = 1, \dots, I$ . Notice that,  $\mathbf{s}^{(i,k)} = (s_1^{(i,k)}, s_2^{(i,k)}, \dots, s_{n_k}^{(i,k)})$ .

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

$$p(\mathbf{s}) = \binom{n_1}{n_{T_1}}^{-1} \binom{n_2}{n_{T_2}}^{-1} \dots \binom{n_K}{n_{T_K}}^{-1} \quad \text{if} \quad \sum_{j=1}^{n_k} s_j^{(\cdot,k)} = n_{T_k}, \quad (13)$$

where each term of the product in equation (13) 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

$$\Psi_k = E \left\{ r_k(\mathbf{y}_{V[\mathbf{s}]}^{rep}, \mathbf{y}_{V[\mathbf{s}]}) \right\}, \quad k = 1, \dots, K, \quad (14)$$

where the expression (14) represents the expectation with respect to the discrepancy measure in each stratum  $k$ .

Notice that the proposed stratification changes the sampling of spatial locations for validation and training sets, however, the sampling model is conditional on locations  $\mathbf{x}$  and does not change with our proposal. Thus, the likelihood function is not affected. The vector of all observations can be written as  $\mathbf{y} = (y_{1,1}, \dots, y_{1,n_1}, \dots, y_{k,i}, \dots, y_{K,n_K})$ .

#### 4.1. Stratified Estimators

To compute the stratified estimators, we jointly simulate the split vectors  $\mathbf{s}^{(1)}, \dots, \mathbf{s}^{(I)}$  from  $p(\mathbf{s})$  as defined in (13). Following the same steps as in Section 3.2, the stratified MC estimator is obtained as

$$\hat{\Psi}_{mc}^{st} = \sum_{k=1}^K w_k \left\{ \frac{1}{I_k} \sum_{i=1}^{I_k} \frac{1}{J} \sum_{j=1}^J r_k \left( \mathbf{y}_{V[\mathbf{s}^{(i)]j}}^{rep}, \mathbf{y}_{V[\mathbf{s}^{(i)]}} \right) \right\} = \sum_{k=1}^K w_k \hat{\Psi}_{mc_k} = \sum_{k=1}^K \hat{\Psi}_{mc_k}^{st}, \quad (15)$$

and the stratified SIR estimator as,

$$\hat{\Psi}_{sir}^{st} = \sum_{k=1}^K w_k \left\{ \frac{1}{H} \sum_{h=1}^H \frac{1}{I_k} \sum_{i=1}^{I_k} \Psi_{hi}^{(k)} \right\} = \sum_{k=1}^K w_k \hat{\Psi}_{sir_k} = \sum_{k=1}^K \hat{\Psi}_{sir_k}^{st}, \quad (16)$$

where,

$$\Psi_{hi}^{(k)} = \frac{\sum_{j=1}^J r_k \left( y_{V[s^{(i)]h_j}^{rep}, \mathbf{y}_{V[s^{(i)]}} \right) w_{hj}^*}{\sum_{j=1}^J w_{hj}^*}, \quad k = 1, \dots, K,$$

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

#### 4.2. 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, their shapes and sample sizes need to be made to reduce the sampling variance of the expected discrepancy estimator.

According to ([Diggle, 2014](#), page 99), when the occurrence of an event at a particular location makes it more likely than other events 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.

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 producing subsets. Although the authors considered procedures for creating 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 used to divide a possible heterogeneous spatial domain in subregions more homogeneous to achieve smaller variances for the estimated discrepancy functions.

## 5. Simulation Studies

In subsection 5.1, we present an illustrative example that shows that the usual validation setup in spatial data analysis may 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. In subsection 5.2, to illustrate the usefulness of our cross-validation proposal, we consider two different scenarios: homogeneous and inhomogeneous processes.

### 5.1. An illustrative example: Uncertainty of Data Partition

Consider locations  $x_1, \dots, x_n$  randomly simulated in a unit square with  $n$  within an irregular grid. Responses  $Y = (Y(x_1), \dots, Y(x_n))$  are generated from the Student-t process as specified in subsection 3.

The parameters are set to  $\nu = 3$ ,  $\sigma^2 = 1$  and varying values for  $\phi = (0.05, 0.30, 0.70)$ , with larger  $\phi$  indicating stronger spatial correlation. We randomly chose  $I = 100$  validation configurations of all  $\binom{n}{n_V}$  possible subsets of size  $n_V$ . For the chosen configurations, we randomly omitted  $0.05n$  and  $0.25n$  points for validation and made predictions for these locations using the remaining training points. We fitted the Gaussian (GM) and Student-t (STM) models to the simulated data and used MCMC techniques to estimate the model parameters for each of the 100 sets at each data configuration  $(\phi, n_V)$ .

For model assessment, we consider the Mahalanobis distance (Mahalanobis, 1936, details in Appendix A) as the discrepancy measure ( $D$ ). Thus, the model choice depends on

$$\delta^{(i)} = D_{STM}^{(i)} - D_{GM}^{(i)}, \quad i = 1, \dots, I, \quad (17)$$

where  $I$  is the number of validation configurations and  $D$  is a discrepancy measure, so that if  $\delta^{(i)} < 0$  we have that STM is preferable to GM. Figure 1 presents the box-plots for 100 randomly selected validation configurations for cross-validation performance varying  $\phi$  and  $n_V$ . If we consider the validation set with size  $n_V = 5\%n$ , the percentages of wrong decisions are considerably

larger than if  $n_V = 25\%n$ . The percentages of wrong decisions are also larger when  $\phi$  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. Two sample sizes were considered:  $n = 90$  and  $n = 200$ . Notice that as the sample size increases, the easier it is to distinguish between Gaussian and Student-t models. According to [Breusch et al. \(1997\)](#), similar inferences are made about the mean  $\mu$  under GM and STM, but different inferences can be made about scale resulting in different prediction intervals for ungauged locations for each model. This difference in the inference and predictions also depends on the range parameter as indicated by our motivation examples. The percentage of data used for validation and the value of the spatial range seem to be crucial for discriminating between competing models.

In this illustration, we have considered exponential covariance function. A popular alternative model is the Matérn covariance function, which has a parameter that controls the smoothness of the process. Note that in this case some issues have been pointed out in the literature regarding parameter identifiability. [Stein \(1999\)](#) discusses that the likelihood function for the smoothness parameter might not possess a maximum in the interior of the parameter space, with the supremum often being obtained when this parameter tends to infinity.

To complete our illustration, we apply Algorithm 1 seen in Section 3.2. Table 2 presents the Monte Carlo estimate based on the 100 validation sets, using the discrepancy measure  $D$  and varying  $\phi$  and  $n_V$  for each model. Opposed to the results in Figure 1, the use of expected discrepancies indicated that the STM, which generated the data, best fits the dataset for all scenarios. Notice, however, that particularly for this illustration, 100 MCMC chains for each data configuration ( $\phi \times n_V$ ) and each competing model were run. This is too time consuming even for these small spatial datasets.

### 5.2. Simulated study: Homogeneous and Inhomogeneous processes

We simulated data scenarios with different configurations for the location sampling. We first simulated a realization of a stationary Gaussian process on the unit square, treating the process  $S(\cdot)$  as constant within each lattice cell. Next, we simulated non-preferentially or preferentially scenarios according to each of the sampling designs presented in Figure 2. The data were generated from equation (1) with:

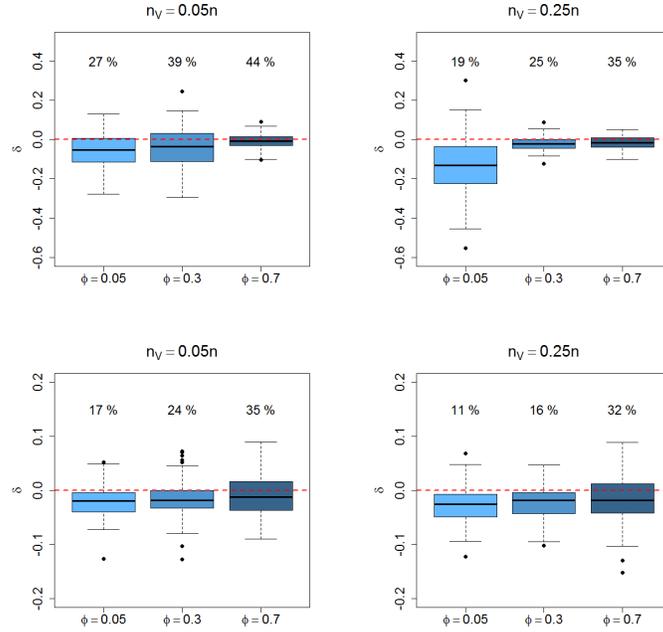


Figure 1: Cross-validation performance: box-plots of predictive discrepancy  $\delta$  for GM versus STM for  $n = 90$  (first row) and  $n = 200$  (second row), and varying  $\phi$  and  $n_V$ . Values of  $\delta$  below the dashed line indicate that the data generating model (STM) is preferable. Numbers represent the percentage of times the wrong model (Gaussian) was selected.

Table 2: MC estimate based on Algorithm 1 for 100 validation sets using the Mahalanobis distance for each model,  $n = 90$  and  $n = 200$  and varying  $\phi$  and  $n_V$ . The model that best fits the data is the one which presents smaller values of the measure.

		$\phi$			$\phi$			
		5% $n$	0.05	0.30	0.70	25% $n$	0.05	0.30
$n = 90$	GM	2.70	6.52	2.55	GM	6.78	6.52	6.47
	STM	2.64	6.47	2.55	STM	6.64	6.50	6.45
$n = 200$	GM	4.37	4.36	4.25	GM	9.95	10.0	9.76
	STM	4.35	4.33	4.23	STM	9.91	9.98	9.74

(i)  $S$  is stationary Gaussian process with mean 0, variance  $\sigma^2$  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

$$\lambda(x) = \exp\{\alpha + \kappa S(x)\}. \quad (18)$$

(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  $\kappa = 0$ , the sampling is done at random, resulting in a homogeneous Poisson process. The simulated surface in (i) is given by a Gaussian process with the following parameters:  $\mu = 4, \sigma^2 = 1.5, \phi = 0.15$  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 constant. A dataset was simulated considering intensity parameters  $\kappa = 0, \alpha = 4.605$  return the sample size  $n = 82$ . This is presented in Figure 2 (a).

*Scenario 2* – CSR with outliers, we study the same surface of Figure 2 (a) with observations contaminated by summing a random increment  $u\sigma$ , such that  $\sigma$  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 2 (b).

*Scenario 3* – Preferential Sampling, we chose the configuration with the highest concentration of points in a given region. The point process represents the inhomogeneous Poisson process, with intensity  $\lambda(x)$ ,  $\alpha = 2.996, \kappa = 1.0$  and  $n = 100$ . This is presented in Figure 2 (c).

For all sample designs presented above, we made a cross-validation comparison of the three geostatistical models presented in subsection 2.1. Parameter estimation and prediction follow the Bayesian approach as presented in subsection 2.2 using the three proposed distributions for  $\mathbf{s}_y$ . For all models, the nugget effect was fixed in the true value so that the focus of this study is on the spatial surface estimation and prediction. The prior distributions used for all models were  $\mu \sim N(0; 10^4), \sigma^{-2} \sim Gamma(0.1; 0.1)$ ,

$\phi \sim \text{Gamma}(1; 2.3/\text{med}(u))$ . For the GLGM,  $v \sim \text{GIG}(0; 0.75; 6)$  and  $\delta \mid \phi, v \sim \text{LogGaussian}(-\frac{v}{2}; v\sigma^2 R)$ . 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 10,000 and lag of 10 with resulting posterior sample size of 6,981. Convergence was checked using `coda` package (Plummer et al., 2006) through R software.

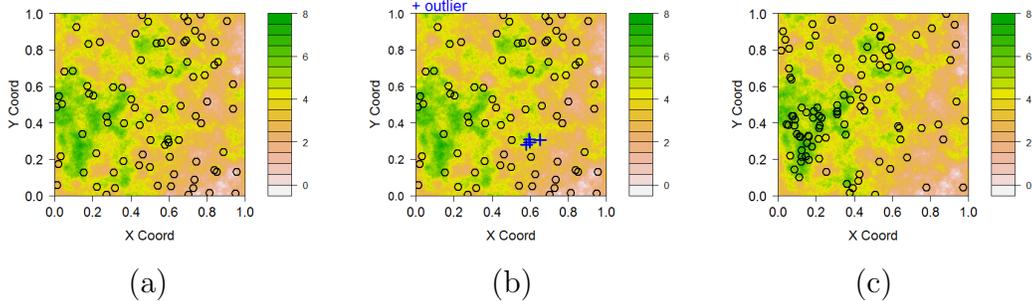


Figure 2: Sample locations and underlying realizations of the signal process for the three models considered in the simulation study: (a) CSR (complete spatial randomness) ; (b) CSR with outliers; (c) preferential sampling.

As for the CSR and CSR with outlier scenarios,  $n_T = 77, n_V = 5$  were arbitrarily chosen. MC and SIR estimators are based on averaging over the same  $I = 100$  splits. Parameter  $\nu$  is fixed at 3 for the Student-t 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 it 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 the uniform prior is considered. Analogously, we verified the computational time considering a prior via distances. The time was similar to the previous case and we omitted it from the text. The computational cost is approximately 5 to 6 times smaller for the GM, 5 to 8 times smaller for the STM and 6 to 10 times smaller for the GLGM 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: Gaussian (GM), Student-t (STM) and Gaussian-log-Gaussian (GLGM).

	GM		STM		GLGM	
	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 simplicity of exposition, we omitted the results of uniform prior and Figure 3 presents the discrepancy measures based on the MC and SIR estimators with their respective standard errors adopting the prior via distances. We used 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.

Figure 3 (a) presents predictive measure estimates for the complete random scenario. It indicates that GLGM and GM models have similar values, although it still correctly chooses the GM as the best model. Model STM with  $\nu = 3$  presents a 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.

Figure 3 (b) correctly indicates that the GLGM is the best choice for this scenario. This is due to the fact that this model tends to detect subregions with larger variability. On the other hand, GM and STM models overestimate the variance in the whole spatial domain. Nevertheless, 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 and regional outlier detection and different kurtosis behaviours across space (see [Lobo and Fonseca, 2019](#), for a more detailed discussion).

Figure 3 (c) indicates similar results for GM and GLGM models for all adopted measures. We emphasize that despite our dataset is under effect of preferential sampling, we fit usual geostatistical models which do not take this effect into account. The next subsection proposes a cross-validation scheme which potentially identifies lack of fit in subregions of the spatial domain.

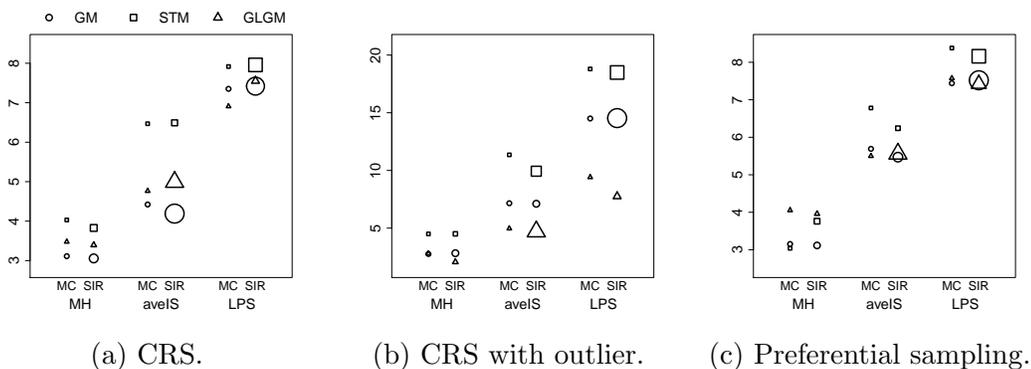


Figure 3: Cross-validation for GM, STM and GLGM models in each scenario with prior via distances. The symbols are proportional to the variance estimator size.

### 5.3. Analysing heterogeneity in the spatial locations

The data presented in Section 5 were stratified into four strata for all scenarios as presented in figure 4. Table 4 details the strata and selection of training and validation cases. The number of locations sampled for validation are proportional to the number of locations in each stratum. Observe that in the homogeneous scenario CRS it is expected that the number of events to be similar in each stratum (Table 4 and Figure 4 (a)).

Figure 5 shows that stratification reduces the variability of discrepancy estimates for all scenarios and discrepancy measures when compared to results in Figure 3. This result is even more evident for the SIR estimator. We omitted the results of STM for the stratified study, since it had a worse performance than GM and GLGM models in all scenarios. In the homogeneous case (Figure 5 (a) – (c)), the estimates are approximately the same for each stratum, as expected. The use of Mahalanobis distance, average Interval Score and Log Predictive Score discrepancies leads to adequate model discrimination by indicating the data generating model (Gaussian) as the best model in this scenario.

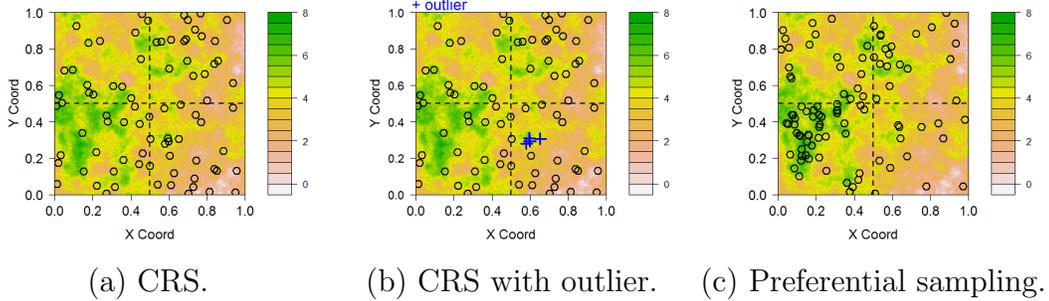


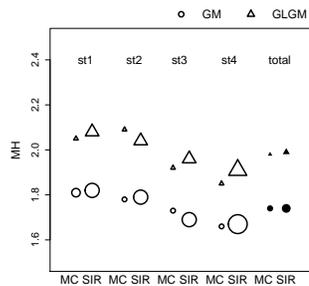
Figure 4: Sample locations and underlying signal process for all scenarios. Strata are divided as: stratum 1 (bottom left), stratum 2 (top left), stratum 3 (bottom right) and stratum 4 (top right).

Table 4: Stratified sample for all scenarios.

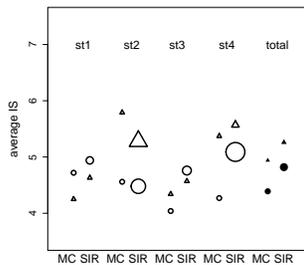
	CSR / Outlier				Preferential				
	$n_k$	$n_{T_k}$	$n_{V_k}$	$w_k$	$n_k$	$n_{T_k}$	$n_{V_k}$	$w_k$	
<i>strata</i> 1	21	19	2	0.250	1	47	42	5	0.500
2	17	15	2	0.250	2	20	18	2	0.200
3	24	22	2	0.250	3	13	12	1	0.100
4	20	18	2	0.250	4	20	18	2	0.200
<i>total</i>	82	74	8	1	<i>total</i>	100	90	10	1

For the scenario with outliers (Figure 5 (c) – (e)), the stratification allows the identification of lack of fit for all models in region 3 (bottom right in Figure 4 (b)) which contains the contaminated observations. All models have larger values of the discrepancy function for stratum 3. The GLGM has better performance, indicating that if the region is divided in subregions, a better predictive performance assessment of this model for the subregions is obtained. Figure 5 (f) – (i) presents the results for the preferential sampling scenario. The performance of GM and GLGM models are similar, while STM has the worst performance of all models. The stratified estimator shows the poor predictive performance in region 1 (bottom left in Figure 4 (c)) for all models. This subregion is indeed the one with higher values of spatial surface and also the subregion with larger intensity of points (locations).

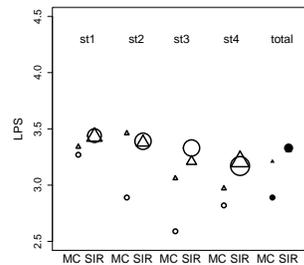
This indicates lack of fit of the fitted models in this scenario as pursued by our proposed cross-validation scheme.



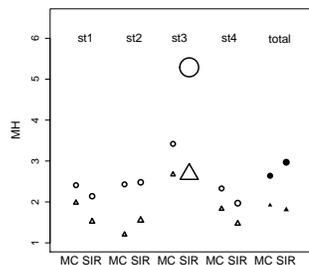
(a) MH.



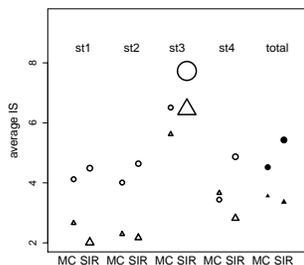
(b) IS.



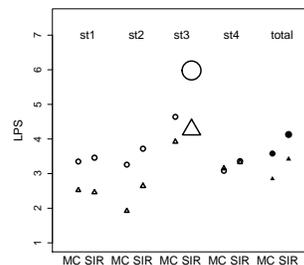
(c) LPS.



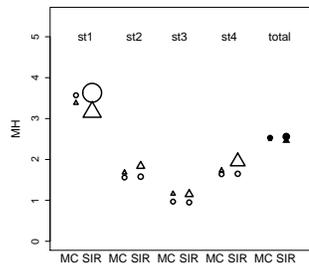
(d) MH.



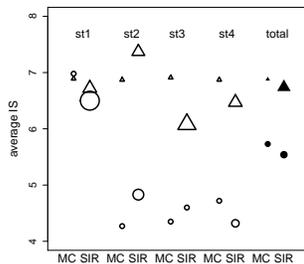
(e) IS.



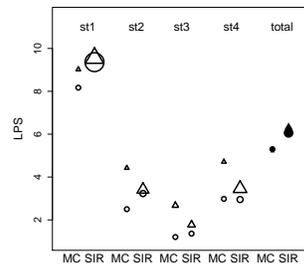
(f) LPS.



(g) MH.



(h) IS.



(i) LPS.

Figure 5: Stratified cross-validation for Gaussian (GM) and Gaussian-log-Gaussian (GLGM) models for CRS (a-c), CRS with outlier (d-f) and Preferential Sampling (g-i) scenarios. The empty circles and triangles represent the Gaussian and Gaussian-log-Gaussian models, respectively. Solid circles and triangles represent the global measure.

## 6. 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 large. It appears that the point pattern associated with the stations has been observed from an inhomogeneous process as discussed in [Ferreira and Gamerman \(2015\)](#).

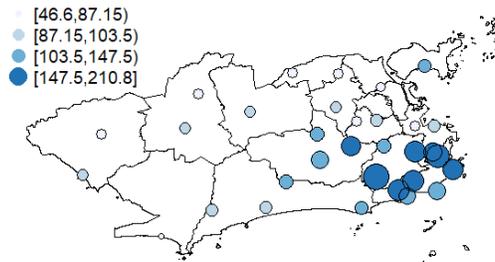


Figure 6: Rainfall data: stations installed in the city of Rio de Janeiro (the monitoring stations are separated according to the intensity of rainfall).

For statistical inference purposes, the spatial mean was adjusted considering latitude and longitude as covariates, thus  $\mu(x) = \beta \mathbf{u}'(x)$ ,  $u_1(x) = \text{lat}(x)$ ,  $u_2(x) = \text{long}(x)$ . For this analysis, the fitted models were the GM and GLGM models presented in Section 2.1. For both models an exponential covariance structures was considered to account for spatial dependence. Parameter estimation and prediction follow the Bayesian paradigm as presented in Subsection 2.2.

The analysis of the posterior distribution of spatial mean shows significantly different estimates for both models. The spatial mean for GLGM is significantly lower than the spatial mean estimated by GM. Actually, this is

plausible since the process for the data is inhomogeneous and model GLGM 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. In this context, we choose two different setups for training set size: the first has  $n_T = 84\%n$  and  $n_V = 16\%n$ , for the training and validation samples, respectively and the second is a more 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 for the discrepancy variability estimation.

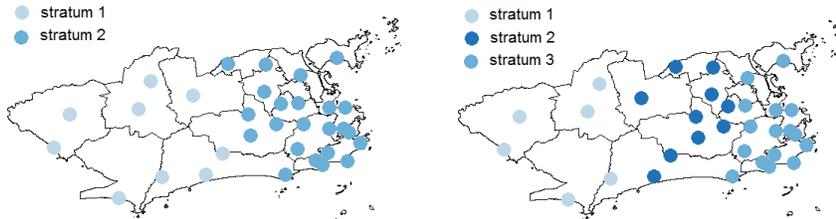
Table 5 displays the performance of both models according to the Mahalanobis distance, average Interval Score and Log Predictive Score when it is assigned a uniform prior to the splits. The SIR estimator is considered for discrepancy estimation as our simulated study has shown that it can produce estimates close enough to the MC estimator.

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 GLGM are smaller than for GM 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. The prior via distances resulted in the same conclusions and it is omitted from the text.

In addition, we take into account spatial heterogeneity using the proposed stratified cross-validation approach to model comparison and goodness of fit checking. The choice of strata was performed dividing the spatial region into 2 and 3 strata via K-means ++ criteria. Although K-means ++ algorithm does not take into account spatial contiguity constraints, defined 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 7 presents the two proposals for stratification. Notice that in the two cases in Figure 7 (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, stratum 2 and stratum 3, respectively.

Table 5: Rainfall Data: cross-validation using discrepancy measures for GM and GLGM using uniform prior.

$n_V = 16\%n$	MH	average IS	LPS
GM	6.42 (0.00)	232.93 (3.31)	27.94 (0.08)
GLGM	4.81 (0.00)	131.53 (0.01)	25.46 (0.02)
$n_V = 68\%n$	MH	average IS	LPS
GM	14.70 (0.01)	355.72 (4.09)	131.11 (0.46)
GLGM	7.75 (0.02)	192.63 (0.11)	102.93 (0.63)



(i) 2 strata ( $n_{V_1} = 1; n_{V_2} = 3$ ). (ii) 3 strata ( $n_{V_1} = 1; n_{V_2} = 1; n_{V_3} = 2$ ).

Figure 7: Proposals for stratification via K-means ++.

Figure 8 presents discrepancy estimates obtained for all strata considering MH, average IS and LPS measure for both stratification scenarios,  $k = 2$  (a) – (c) and  $k = 3$  (d) – (f). For  $k = 2$  scenario, stratum 2 has larger measures for all models, indicating worse predictive performance in this region. For  $k = 3$  scenario, stratum 3 has larger measures for all models. Thus, our proposed stratified cross-validation allows the identification of regions of poor predictive performance for both models in this application.

Overall GM produces higher discrepancy estimates  $\hat{\Psi}^{st}$  for both stratification scenarios indicating that GLGM is the best model for this data. If we analyze the results by subregion then model GLGM is the best model for most measures and scenarios, except for 2-stratum scenario and LPS, which indicates that the Gaussian process is the best model for subregion 1 and for 3-stratum scenario, which most measures indicate that the Gaussian process is the best model for subregion 1 and 2. GLGM outperforms the GM for

all measures in subregion 2 (2-stratum scenario) and subregion 3 (3-stratum scenario). These results indicate that the Gaussian model is not adequate for the whole region, highlighting the lack of fit of this model which could be due to nonstationarity or preferential sampling as indicated by other study in the literature.

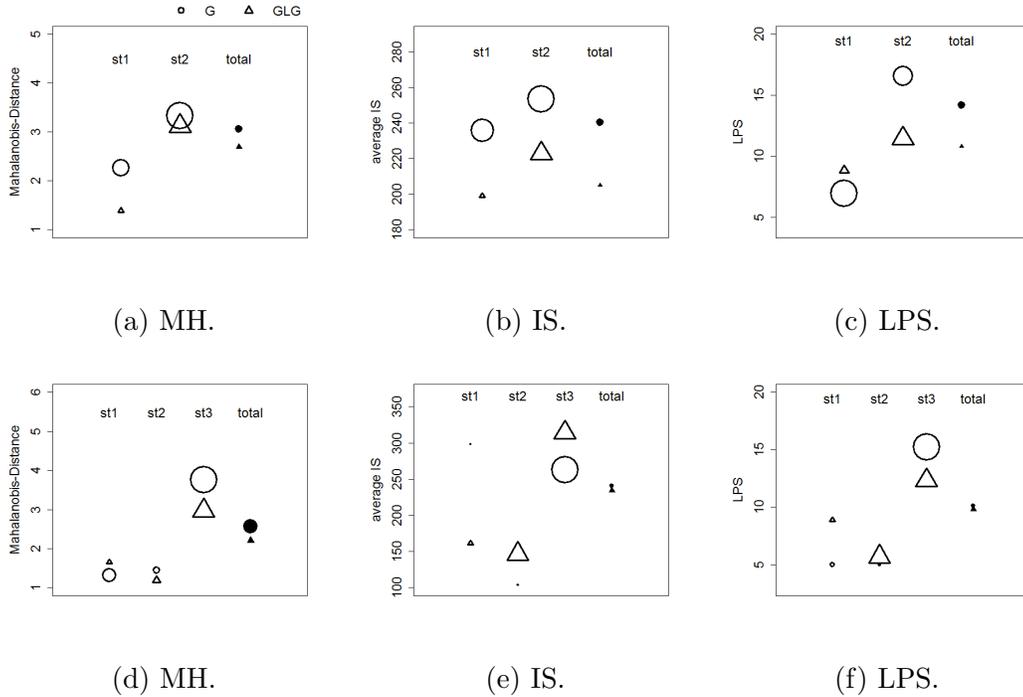


Figure 8: Rainfall Data: stratified cross-validation for GM (circles) and GLGM (triangles). First row represents  $k = 2$  and second row represents  $k = 3$ .

## 7. Discussion

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

The proposed split vector prior distributions allow to accommodate the uncertainty in the validation and training 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 distance to previous selected points and all the others candidates points to the validation sample. Since irregular spatial regions often occur in data applications, the prior via distances is a potentially 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, overcoming the computational limitation of Bayesian cross-validation techniques. The proposed stratified scheme contributes to reducing the global variability of SIR estimators. Furthermore, it indicates regions with worse predictive performance in the spatial domain such as in the 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. As a simple solution to this matter, K-means ++ was used to automatically select the stratum in rainfall data analysis. Other possible solutions will be investigated in a future work.

Moreover, the issue of choosing the training sample size is not trivial. The quality of cross-validation methods typically depends on training data size. [Berger and Pericchi \(2004\)](#) discuss the importance of minimum training samples in model selection. While in many applications it is desirable to select minimum training samples, this might not be suitable for spatial data. A small sample might lead to poor predictions in regions of spatial domain where no data was used for training. We considered two different scenarios in the application to rainfall data to accommodate the possible effect of choosing either a too small or a too large training set.

Furthermore, we have not considered cross-validation uncertainty. For instance, we have fixed the number of splits ( $I = 100$  in the motivation example) when the total number of splits is 2,555,190 for  $n = 90$  and  $n_V = 4$ . What would happen if  $I = 50$  or  $I = 200$ ? What we believe would help in that manner would be to consider different validation sizes as well. In such case, if one varies the sizes and sample a different size from a prior distribution on the sizes, then the number of possible splits  $\binom{n}{n_V}$  would also change. Thus, the two uncertainties (from the split size and from the validation size) could be taken care of jointly. Future research would consider the effect of validation sizes and the split sizes 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 and F. A. S. Moura were partially supported by *Conselho Nacional de Desenvolvimento Científico e Tecnológico* (CNPq).

## Appendix A. The choice of discrepancy function

As follows the discrepancy functions used for predictive comparison are detailed.

*Mahalanobis distance* This measure takes into account the spatial dependence (Mahalanobis, 1936).

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

with  $\Sigma = \tau^2 I_\tau + \sigma^2 R$  the predictive covariance matrix. Extreme values for the Mahalanobis distance indicate a conflict between the validation and predicted data. Bastos and O'Hagan (2008) adopt this measure to validate and assess the adequacy of Gaussian processes emulators.

As follows we detail proper scoring rules (Gneiting and Raftery, 2007) considered for model comparison and validation. In this direction, other proposal for measuring predictive accuracy are Gelman et al. (2014) and Vehtari et al. (2017).

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

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

where  $l$  and  $u$  represent for the forecaster quoted  $\frac{\gamma}{2}$  and  $1 - \frac{\gamma}{2}$  quantiles based on the predictive distribution  $f(y_{V[s]}^{rep} | y_{T[s]})$  and  $\mathbf{y}_{V[s]}$  the sample validation

vector. If  $\gamma = 0.05$  the resulting interval has 95% of credibility. For each element of  $\mathbf{y}_{V[s]}$  we have one interval score measure and 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  $\mathbf{y}_{V[s]}$ ,

$$r(\mathbf{y}_{V[s]}^{rep}, \mathbf{y}_{V[s]}) = -\log [q(\mathbf{y}_{V[s]})]. \quad (\text{A.3})$$

Note that under the Gaussian model assumption, it is similar to the Mahalanobis distance in (A.1).

## Appendix B. Variance estimator

According to [Robert and Casella \(2009\)](#), the generic problem involves evaluating the integral

$$E_f(h(X)) = \int_{\chi} h(x)f(x)dx, \quad (\text{B.1})$$

where  $\chi$  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 (B.1) is to generate a sample  $X_1, \dots, X_n$  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

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

which can be estimated from  $(X_1, \dots, X_n)$  through  $v_n = \frac{1}{n^2} \sum_{j=1}^n [h(x_j) - \bar{h}_n]^2$ .

Analogously to equation (B.2), we can obtain the variance of the estimators  $\hat{\Psi}_{mc}$  and  $\hat{\Psi}_{sir}$ . Notice that from the equation (10) we obtain,

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

Thus,

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

is the SIR estimator variance, obtained from equation (12), where,

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

According to [Alqallaf and Gustafson \(2001\)](#) to determine the variance of  $\hat{\Psi}_{sir}$ , consider the terms  $\Psi_{hi}$  as elements of an  $H$  by  $I$  matrix, and note that each element has the same distribution. We consider the variance of this distribution, the common covariance of any pair of distinct elements from the same row, and 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,

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

#### Appendix B.1. SIR estimator details

We draw a MCMC sample from  $g(\theta)$ , which is then reweighted using importance sampling to obtain  $p(\theta | \mathbf{s})$ . The same posterior sample is used for every split  $\mathbf{s}$  considered, saving computational time. The weights  $w_{ihj} = p(\theta_{hj} | \mathbf{y}_{T[\mathbf{s}^{(i)}]})/g(\theta_{hj})$  can be obtained as

$$\log(w_{ihj}) = \log \left\{ \frac{p(\theta_{hj} | \mathbf{y}_{T[\mathbf{s}^{(i)}]})}{g(\theta_{hj})} \right\} = \log \left\{ \frac{f(\mathbf{y}_{T[\mathbf{s}^{(i)}]} | \theta_{hj})}{f(\mathbf{y} | \theta_{hj})^\alpha} \right\} \quad (\text{B.5})$$

#### Appendix B.2. Stratified Variance

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

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

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

$$var(\hat{\Psi}^{st}) = \frac{1}{n^2} \sum_{k=1}^K n_k(n_k - n_{V_k}) \frac{s_k^2}{n_{V_k}} = \sum_{k=1}^K \frac{w_k}{n} (n_k - n_{V_k}) \frac{s_k^2}{n_{V_k}} = \sum_{k=1}^K \frac{w_k}{n} (1 - f_{V_k}) \frac{s_k^2}{n_{V_k}},$$

with  $s_k^2 = \frac{1}{(n_{V_k} - 1)} \sum_{i=1}^{n_k} (r_{ki} - \hat{\Psi}_k)^2$  and  $r_k$  denotes any discrepancy function. Note that equation (B.6) can be written as

$$var(\hat{\Psi}^{st}) = var \left( \sum_{k=1}^K \hat{\Psi}_k^{st} \right) = var \left( \sum_{k=1}^K w_k \hat{\Psi}_k \right) = \sum_{k=1}^K w_k^2 var(\hat{\Psi}_k) \quad (\text{B.6})$$

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

## References

- Alqallaf, F., Gustafson, P., 2001. On cross-validation of Bayesian models. *The Canadian Journal of Statistics* 29, 333–340.
- Apanasovich, T.V., Genton, M.G., 2010. Cross-covariance functions for multivariate random fields based on latent dimensions. *Biometrika* 97, 15–30.
- Arlot, S., Celisse, A., 2010. A survey of cross-validation procedures for model selection. *Statistics surveys* 4, 40–79.
- Arthur, D., Vassilvitskii, S., 2007. 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.
- Banerjee, S., Gelfand, A.E., Finley, A.O., Sang, H., 2008. Gaussian predictive process models for large spatial data sets. *Journal of the Royal Statistical Society Series B* 70, 825–848.

- Bastos, L.S., O’Hagan, A., 2008. Diagnostics for gaussian process emulators. *Technometrics* 51, 425–438.
- Berger, J.O., Pericchi, L.R., 2004. Training samples in objective Bayesian model selection. *The Annals of Statistics* 32, pp. 841–869.
- Bergmeir, C., Benitez, J.M., 2012. On the use of cross-validation for time series predictor evaluation. *Information Sciences* 191, 192–213.
- Bergmeir, C., J.Hyndman, R., Koo, B., 2018. A note on the validity of cross-validation for evaluating autoregressive time series prediction. *Computational Statistics and Data Analysis* 120, 70–83.
- Breusch, T.S., Robertson, J.C., Welsh, A.H., 1997. The emperor’s new clothes: a critique of the multivariate t regression model. *Statistica Neerlandica* 51, 269–286.
- Bueno, R.S., Fonseca, T.C.O., Schmidt, A.M., 2017. Accounting for covariate information in the scale component of spatial-temporal mixing models. *Spatial Statistics* 22, 196–218.
- Burman, P., 1989. A comparative study of ordinary cross-validation,  $v$ -fold cross-validation and the repeated learning-testing methods. *Biometrika* 76.
- Burman, P., Chow, E., Nolan, D., 1994. A cross-validatory method for dependent data. *Biometrika* 81, 351–358.
- Cochran, W.G., 1999. *Sampling Techniques*. Wiley Student Edition. 3 ed., Wiley.
- Cressie, N., 1993. *Statistics for Spatial Data*. New York: Wiley.
- Datta, A., Banerjee, S., Finley, A.O., Gelfand, A.E., 2015. Hierarchical nearest-neighbor gaussian process models for large geostatistical datasets. *Journal of the American Statistical Association* .
- Diggle, P.J., 2014. *Statistical Analysis of Spatial and Spatio-Temporal Point Patterns*. Chapman and Hall/CRC.
- Diggle, P.J., Menezes, R., Su, T.I., 2010. Geostatistical inference under preferential sampling. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 59, 191–232.

- Ferreira, G.S., Gamerman, D., 2015. Optimal design in geostatistics under preferential sampling. *Bayesian Analysis* 10, 711–735.
- Fonseca, T.C.O., Ferreira, M.A.R., Migon, H.S., 2008. Objective bayesian analysis for the student-t regression model. *Biometrika* 95, 325–333.
- Fonseca, T.C.O., Steel, M.F.J., 2011. Non- gaussian spatiotemporal modelling through scale mixing. *Biometrika* 98, 761–774.
- Furrer, R., Genton, M.G., Nychka, D., 2006. Covariance tapering for interpolation of large spatial datasets. *Journal of Computational and Graphical Statistics* 15, 502–523.
- Gamerman, D., Lopes, H., 2006. *Markov Chain Monte Carlo: Stochastic Simulation for Bayesian Inference*. Texts in Statistical Science, Taylor & Francis.
- Gelfand, A., 1996. *Model Determination Using Samplings Based Methods*. Chapman & Hall, Boca Raton, FL.
- Gelfand, A.E., Dey, D.K., Chang, H., 1992. Model determination using predictive distributions with implementation via sampling-based methods. Technical Report. Department of Statistics Stanford University.
- Gelman, A., Hwang, J., Vehtari, A., 2014. Understanding predictive information criteria for bayesian models. *Statistics and Computing* 24, 997–1016.
- Geweke, J., 1989. Bayesian inference in econometric models using monte carlo integration. *Econometrica* 57, pp. 1317–1339.
- Gneiting, T., Raftery, A.E., 2007. Strictly proper scoring rules, prediction and estimation. *Journal of the American Statistical Association* 102, 360–378.
- Gordon, A.D., 1996. A survey of constrained classification. *Comput. Stat. Data Anal.* 21, 17–29.
- Katzfuss, M.M.J., Hu, J., Johnson, V.E., 2014. Assessing fit in bayesian models for spatial processes. *Environmetrics* 25, 584–595.

- Li, L., Qiu, S., Zhang, B., Feng, C., 2016. Approximating cross-validators predictive evaluation in bayesian latent variable models with integrated is and waic. *Statistics and Computing* 26, 881–897.
- Lobo, V.G.R., Fonseca, T.C.O., 2019. Bayesian residual analysis for spatially correlated data. *Statistical Modelling* .
- Mahalanobis, P.C., 1936. On the generalized distance in statistics. *Proceedings of the National Institute of Sciences (Calcutta)* 2, 49–55.
- Majumdar, A., Gelfand, A.E., 2007. Multivariate spatial modeling for geostatistical data using convolved covariance functions. *Mathematical Geology* 39, 225–245.
- Marshall, E.C., Spiegelhalter, D.J., 2003. Approximate cross-validators predictive checks in disease mapping models. *Statistics in medicine* 22, 1649–1660.
- O’Hagan, A., 1995. Fractional bayes factors for model comparison. *Journal of the Royal Statistical Society. Series B* 57, 99–138.
- Palacios, M.B., Steel, M.F.J., 2006. Non-gaussian bayesian geostatistical modeling. *Journal of the American Statistical Association* 101, 604–618.
- Pfeffermann, D., Moura, F., Silva, P., 2006. Multi-level modeling under informative sampling. *Biometrika* 93, 943–959.
- Plummer, M., Best, N., Cowles, K., Vines, K., 2006. Coda: Convergence diagnosis and output analysis for mcmc. *R News* 6, 7–11.
- Robert, C.P., 2007. *The Bayesian Choice*. Second edition ed., Springer, New York.
- Robert, C.P., Casella, G., 2009. *Introducing Monte Carlo Methods with R (Use R)*. 1st ed., Springer-Verlag, Berlin, Heidelberg.
- Robert, C.R., Casella, G., 2004. *Monte Carlo Statistical Methods*. Springer, New York.
- Roberts, D.R., Bahn, V., Ciuti, S., Boyce, M.S., Elith, J., Guillaera-Arroita, G., Hauenstein, S., Lahoz-Monfort, J.J., Schroder, B., Thuiller, W.,

- Warton, D.I., Wintle, B.A., Hartig, F., Dormann, C.F., 2017. Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. *Ecography* 40, 913–929.
- Roislien, J., Omre, H., 2006. T-distributed random fields: A parametric model for heavy-tailed well-log data. *Mathematical Geology* 38, 821–849.
- Stein, M.L., 1999. *Interpolation of Spatial Data: Some Theory for Kriging*. Springer New York.
- Stein, M.L., Chi, Z., Welty, L.J., 2004. Approximating likelihoods for large spatial data sets. *Journal of the Royal Statistical Society Series B* 66, 275–296.
- Stern, H.S., Cressie, N., 2000. Posterior predictive model checks for disease mapping models. *Statistics in medicine* 19, 2377–2397.
- Thall, P., Russel, K., Simon, R., 1997. Variable selection in regression via repeated data splitting. *Journal of Computational and Graphical Statistics* 6, 416–434.
- Vecchia, A.V., 1988. Estimation and model identification for continuous spatial processes. *Journal of the Royal Statistical Society. Series B (Methodological)* 50, 297–312.
- Vehtari, A., Gelman, A., Gabry, J., 2017. Practical bayesian model evaluation using leave-one-out cross-validation and waic. *Statistics and Computing* 27, 1413–1432.
- Watanabe, S., 2010. Asymptotic equivalence of bayes cross validation and widely applicable information criterion in singular learning theory. *J. Mach. Learn. Res.* 11, 3571–3594.