Bayesian Spatial Predictive Synthesis

Danielle Cabel[1], Shonosuke Sugasawa[2], Masahiro Kato[3],
Kōsaku Takanashi[4], and Kenichiro McAlinn[1]

[1]Fox School of Business, Temple University, Philadelphia, USA
[2]Center for Spatial Information Science, The University of Tokyo, Tokyo, Japan
[3]Cyberagent. Inc, Tokyo, Japan
[4]Center for Advanced Intelligence Project, Riken, Tokyo, Japan

Abstract

Spatial data are characterized by their spatial dependence, which is often complex, non-linear, and difficult to fully capture with a single model. Significant levels of model uncertainty– arising from these characteristics– cannot be resolved by model selection or simple ensemble methods, as performances are not homogeneous. We address this issue by proposing a novel methodology that captures spatially-varying model uncertainty, which we call Bayesian spatial predictive synthesis. Our proposal is defined by specifying a latent factor spatially-varying coefficient model as the synthesis function, which enables model coefficients to vary over the region to achieve flexible spatial model ensembling. We show that our proposal is derived from the theoretically best approximation of the data generating process and that it provides a finite sample theoretical guarantee for its predictive performance, specifically that the predictions are exact minimax. Two MCMC strategies are implemented for full uncertainty quantification, as well as a variational inference strategy for fast point inference. We also extend the estimation strategy for general responses. Through simulation examples and two real data applications, we demonstrate that our proposed Bayesian spatial predictive synthesis outperforms standard spatial models and ensemble methods, and advanced machine learning methods, in terms of predictive accuracy, while maintaining interpretability of the prediction mechanism.

Key words: Bayesian predictive synthesis; Markov Chain Monte Carlo; variational inference; spatial process; spatially-varying coefficient model

1 Introduction

The modeling of spatial data– data that are dispersed and linked to a geographical location– has received considerable interest due to its abundance and relevance in numerous fields, such as economics, epidemiology, and climatology. Typically, these data are characterized by their spatial dependence and correlation, where "neighbors" share features and may be clustered within certain regions. As the features and outcomes are spatially dependent, taking into account that characteristic is critical to capture spatial heterogeneity and to predict unobserved locations. Several methods have been developed in response, such as geographically weighted regressions (Brunsdon et al., 1998), simultaneous autoregressive models (Anselin, 1988), and a class of hierarchical models based on latent spatial Gaussian process, e.g. spatial generalized linear mixed models (Diggle et al., 1998), spatially varying coefficient models (Gelfand et al., 2003), and spatial factor models (Wang and Wall, 2003).

Given the abundance and variety of spatial models, dealing with model uncertainty– either through model selection or averaging– is essential to achieve improved predictive accuracy and decision making. One approach is model selection, where the "best" model is selected using some information criteria, cross validation, and so on. Alternatively, ensemble methods (including model averaging) have been used to mitigate model uncertainty by averaging over several candidate models. For example, one simple approach is to take the arithmetic mean over the models, which is surprisingly effective in many applications (Genre et al., 2013). A notable drawback of existing approaches, however, is the implicit assumption of spatial homogeneity in the selection or combination of multiple models, which may not be appropriate in many spatial applications. To illustrate this point, consider real-estate data over a large space (e.g. an entire state or country). It is often the case that there are drastic differences across regions (e.g. urban and suburban areas), in terms of their characteristics. Under this setting, we can expect to observe each candidate model to have different predictive characteristics associated with each region; one model may be potent in urban areas and another in suburban areas, etc. Performing model selection under this context may– and often will– select a single model that is not the best anywhere, but has the best average performance. On the other hand, using conventional ensemble methods will assign one weight to each model, taking the average over the entire space and averaging out the region-specific performances. Despite this, only "homogeneous" model averaging methods have been considered in spatial data analysis (e.g. Debarsy and LeSage, 2020; Greenaway-McGrevy and Sorensen, 2021; LeSage and Parent, 2007; Liao et al., 2019; Zhang and Yu, 2018). Thus, existing approaches fail to take into account the critical fact about spatial data; different models may perform better or worse in different regions.

A notable recent trend in spatial modeling is to leverage machine learning techniques, such as gradient boosting trees (e.g. Chen and Guestrin, 2016), random forests (Breiman, 2001; Sekulić et al., 2020), and neural networks (e.g. Bishop et al., 1995), which could potentially offer more precise spatial predictions than conventional spatial models. In particular, Saha et al. (2021) modified the standard random forest to take into account spatial correlation. However, the main drawback of these machine learning approaches is that it does not provide uncertainty measures of predictions, such as prediction intervals or standard errors.

We contribute to this field by introducing a general framework to deal with model uncertainty in spatial data. Our approach works within the framework of Bayesian predictive synthesis (see, e.g. McAlinn and West, 2019; McAlinn et al., 2020), which is a coherent

Bayesian framework for synthesizing multiple sources of information, based on agent opinion analysis (see, e.g. Genest and Schervish, 1985; West and Crosse, 1992). Using this framework, we develop a coherent ensemble method for spatial data that explicitly takes into account the spatially dependent biases and dependencies among models. To develop an ensemble method for spatial modeling with desirable theoretical properties, we first identify the best approximate model, under the assumption that the data generating process and predictions to be synthesized follow Gaussian processes. The best approximate model, we show, is equivalent to a latent factor spatially varying coefficient model, used as a synthesis function, within the Bayesian predictive synthesis framework. We call this method Bayesian spatial predictive synthesis (BSPS). In this method, predictions from spatial models are treated as latent factors and model weights are spatially varying. This allows for spatial heterogeneity of model importance to be captured across regions. BSPS, thus, effectively learns the spatially varying coefficients that, in turn, improves predictive accuracy. We further show that spatial predictions produced by BSPS are exact minimax, providing finite sample guarantees for its predictive performance.

Several computationally efficient algorithms are developed and implemented to produce posterior and predictive analysis, depending on the scale of the dataset. The first two are MCMC based algorithms for full posterior analysis, one employing the nearest neighbor Gaussian process (Datta et al., 2016), which reduces the dimension for faster computation. For larger datasets, we also develop a variational Bayes approximation (Blei et al., 2017) that produces even faster, accurate point predictions. We also extend the estimation algorithm to deal with general responses, and specifically develop an efficient computation algorithm for the binary case, using the Pólya-gamma augmentation (Polson et al., 2013). Unlike machine learning methods that only produce point prediction, the proposed BSPS gives full posterior predictive distributions, enabling us to compute point prediction with associated uncertainty measures.

A series of simulated data and two real world applications, involving the occurrence of Tsuga canadensis and real-estate prices in Tokyo, Japan, illustrate the efficacy of our proposed method. Through these applications, we show that our method has distinct advantages over competing methods, including statistical and machine learning methods for spatial data and ensemble methods. Notably, we show that BSPS, synthesizing conventional spatial models, delivers better predictive accuracy than state-of-the-art machine learning methods, owing to the flexible ensemble BSPS provides through spatially varying model weights.

The rest of the article will proceed as follows. Section 2 introduces a fundamental theory of BSPS by identifying the best approximate model used as a synthesis function, and shows that predictive distribution derived by BSPS is exact minimax. Details of implementation of the proposed BSPS including its MCMC computational strategy, and extensions to general responses are given in Section 3. We also develop alternative computational strategies for scalable inference. Simulation studies are presented in Section 4. Real world applications with the occurrence of Tsuga canadensis and apartment prices in Tokyo are presented in Section 5. The paper concludes with additional comments and closing remarks in Section 6. Some technical details are presented in the Supplementary Material.

2 Bayesian Spatial Predictive Synthesis: Fundamental Theory

2.1 General framework of Bayesian predictive synthesis

Consider predicting a univariate outcome, $y(s)$, at some unobserved site, $s \in \mathcal{S} \subseteq \mathbb{R}^d$. Suppose that a Bayesian decision maker, $\mathcal{D}$, uses information (predictive distributions) from J models for $y(s)$, each of them denoted by the density function, $h_j(\cdot)$, for $j = 1, \ldots, J$. These forecast densities represent the individual inferences from the models and the collection of these forms the information set, $\mathcal{H}(s) = \{h_1(f_1(s)), \ldots, h_J(f_J(s))\}$, where $f_j(s)$ is a variable at site s. Thus, formal Bayesian analysis indicates that $\mathcal{D}$ will predict $y(s)$ using its implied posterior predictive distribution, $p(y(s)|\mathcal{H}(s))$. However, the set of $\mathcal{H}(s)$ is non-trivially complex, given its spatially varying structure of J density functions. As these models are not "independent"– with information overlap among models– there will be spatial dependencies and biases making straightforward Bayesian updating difficult.

The Bayesian predictive synthesis (BPS) framework (McAlinn and West, 2019; Genest and Schervish, 1985; West and Crosse, 1992; West, 1992) provides a general and coherent way for Bayesian updating, given multiple predictive distributions. Specifically, the Bayesian posterior is given as,

$$\Pi_{\text{BPS}}\big(y(s)|\Phi(s), \mathcal{H}(s)\big) = \int \alpha\big(y(s)|\boldsymbol{f}(s), \Phi(s)\big) \prod_{j=1}^{J} h_j\big(f_j(s)\big) df_j(s), \tag{1}$$

where $\alpha\big(y(s)|\boldsymbol{f}(s), \Phi(s)\big)$ is a synthesis function, $\Phi(s)$ represents the spatially varying parameters, and $\boldsymbol{f}(s) = (f_1(s), \ldots, f_J(s))$ is a vector of latent variables. Here, $\alpha\big(y(s)|\boldsymbol{f}(s), \Phi(s)\big)$ determines how the predictive distributions are synthesized and it includes a variety of existing combination methods, such as Bayesian model averaging and simple averaging (e.g. Hoeting et al., 1999; Geweke and Amisano, 2011; Aastveit et al., 2018), as special cases. The representation of (1) does not require a full specification of the joint distribution of $y(s)$ and $\mathcal{H}(s)$, and it does not restrict the functional form of the synthesis function, $\alpha\big(y(s)|\boldsymbol{f}(s), \Phi(s)\big)$. This allows $\mathcal{D}$ to flexibly specify how they want the information to be synthesized. Note that (1) is only a valid posterior if it satisfies the consistency condition. This condition states that, prior to observing $\mathcal{H}(s)$, $\mathcal{D}$ specifies their own prior predictive, $\Pi\{y(s)\}$, as well as their prior expectation of the model prediction, $\mathrm{E}[\prod_{j=1}^{J} h_j\big(f_j(s)\big)]$. Then $\Pi\{y(s)\} = \int \alpha\big(y(s)|\boldsymbol{f}(s), \Phi(s)\big) \prod_{j=1}^{J} h_j\big(f_j(s)\big) df_j(s)$ must hold, meaning that the two priors that $\mathcal{D}$ specifies must be consistent with each other.

2.2 Specification of synthesis function as best approximation

To cast the general representation (1) to synthesize multiple spatial predictions, we need to identify a synthesis function, $\alpha\big(y(s)|\boldsymbol{f}(s), \Phi(s)\big)$, that is justifiable. In this subsection, we derive the specific form of the synthesis function as the best approximation of the unknown data generating process. We consider the task of predicting the data generating process, $y(s)$, with the predictive values from the J models (predictive distributions), $\boldsymbol{f}(s) = (f_1(s), \ldots, f_J(s))$. Assume that both the data generating process and predictive values have second moments. We define the data generating process as

$$y(s) = \mu(s) + \omega(s) + \varepsilon(s), \tag{2}$$

for all $s \in \mathcal{S}$, where $\mu(s) = \mathrm{E}\,[y(s)]$, $\omega(s)$ is an isotropic and second order stationary process with $\mathrm{E}\,[\omega(s)] = 0$ and covariance function $C_\omega(\cdot, \cdot)$, and $\varepsilon(s)$ is an independent noise with $\mathrm{E}\,[\varepsilon(s)] = 0$ and $\mathrm{Var}\,[\varepsilon(s)] = \sigma^2$ with an error variance σ^2. We first assume

that $y(s)$ and $f_j(s)$ all follow Gaussian processes. Then, using series expansions of both $y(s)$ and $f_j(s)$, random variables following the jth predictive distribution, by the complete orthogonal system, we obtain the form of best approximation of (2) given $\boldsymbol{f}(s)$.

Theorem 1. *Given J-dimensional predictive variable $\boldsymbol{f}(s)$ at finite number of locations, the best approximation of the data generating process of $y(s)$ at arbitrary location $s \in \mathcal{S}$ is expressed as*

$$y(s) = \beta_0(s) + \sum_{j=1}^{J} \beta_j(s) f_j(s) + \varepsilon(s),$$

where $\mathrm{E}\left[f(s)\varepsilon(s)\right] = 0$ and $\beta_0(s), \beta_1(s), \ldots, \beta_J(s)$ are spatially varying coefficients.

Detailed proof is given in the Supplementary Material. From Theorem 1, it follows that using the following model,

$$\alpha(y(s)|\boldsymbol{f}(s), \Phi(s)) = \phi\Big(y(s); \beta_0(s) + \sum_{j=1}^{J} \beta_j(s) f_j(s), \sigma^2\Big), \tag{3}$$

as the synthesis function in (1), is the best approximation of the data generating process, and thus should be used, where $\Phi(s) = (\beta_0(s), \beta_1(s), \ldots, \beta_J(s), \sigma^2)$ and $\phi(\cdot; a, b^2)$ is the univariate normal density with mean a and variance b^2. Note that the above synthesis function includes unknown spatially varying parameters $\Phi(s)$, which are calibrated by the data. In Section 3, we will discuss Bayesian inference on these parameters. Throughout this paper, we denote BPS that uses the synthesis function (3) in (1), *Bayesian spatial predictive synthesis (BSPS).*

2.3 Exact minimaxity of predictive distributions

We next provide a theoretical property of the resulting predictive distribution based on the synthesis function (3), by showing that BSPS is exact minimax under Kullback-Leibler (KL) loss. To this end, we assume the same data generating process (2). Further, we assume that $\mu(s)$ takes a value within a compact region, $[-b, a]$, that includes the origin. Let $\boldsymbol{s} = \{s_1, \ldots, s_n\}$ be n observed locations, and s_{n+1} be an unobserved location. We consider discrete models for the $n+1$ locations, where the set of unknown parameters in the data generating model is $\theta = (\mu(s_1), \ldots, \mu(s_n), \mu(s_{n+1}), C_w, \sigma^2)$. The joint distribution of $\boldsymbol{y} = (y(s_1), \ldots, y(s_n))$ and $y(s_{n+1})$ given θ, induced from the data generating process (2), is denoted by $p^*_\theta(\boldsymbol{y}, y(s_{n+1}))$, and the marginal distribution of $y(s_{n+1})$ is $p^*_\theta(y(s_{n+1})) = \int_{\mathbb{R}^n} p^*_\theta(\boldsymbol{y}, y(s_{n+1}))d\boldsymbol{y}$. For a predictive distribution $q(y(s_{n+1})|\boldsymbol{y})$, we define the KL loss as

$$\mathsf{KL}\left(p^*_\theta | \, q\right) = \int_{y(s_{n+1})} \log \frac{p^*_\theta(y(s_{n+1}))}{q(y(s_{n+1})|\boldsymbol{y})} p^*_\theta(y(s_{n+1}))dy(s_{n+1}),$$

where the loss value depends on $\boldsymbol{y}$ and θ. The KL risk is then defined as the expectation of the loss, given by $\mathrm{E}_y\left[\mathsf{KL}\left(p^*_\theta | \, q\right)\right]$ with regard to the observation $\boldsymbol{y}$. Let $q_{\mathrm{BSPS}}(y(s_{n+1})|\boldsymbol{y}, \boldsymbol{f}(s_{n+1}))$ be the predictive distribution of $y(s_{n+1})$ obtained from BSPS, namely

$$q_{\mathrm{BSPS}}(y(s_{n+1})|\boldsymbol{y}, \boldsymbol{f}(s_{n+1})) = \frac{\int \Pi(\Phi_{n+1}) \prod_{i=1}^{n+1} \Pi_{\mathrm{BSPS}}\big(y(s_i)|\Phi(s_i), \mathcal{H}(s_i)\big) d\Phi_{n+1}}{\int \Pi(\Phi_n) \prod_{i=1}^{n} \Pi_{\mathrm{BSPS}}\big(y(s_i)|\Phi(s_i), \mathcal{H}(s_i)\big) d\Phi_n}, \tag{4}$$

with

$$\Pi_{\mathrm{BSPS}}\big(y(s)|\Phi(s), \mathcal{H}(s)\big) = \int \phi\Big(y(s); \beta_0(s) + \sum_{j=1}^{J} \beta_j(s) f_j(s), \sigma^2\Big) \prod_{j=1}^{J} h_j\big(f_j(s)\big) df_j(s),$$

where Φ_n is the collection of spatially varying parameters in the synthesis function (3) in the n observed locations, and $\Pi(\Phi_n)$ is a prior distribution of Φ_n. We then show that the following minimax property of the predictive distribution, $q_{\mathrm{BSPS}}(y(s_{n+1})|\boldsymbol{y}, \boldsymbol{f}(s_{n+1}))$.

Theorem 2. *Under a suitable class of priors on Φ_n and $\Phi(s_{n+1})$, the predictive distribution $q_{\mathrm{BSPS}}(y(s_{n+1})|\boldsymbol{y}, \boldsymbol{f}(s_{n+1}))$ of the form (4) is exact minimax in terms of the KL risk.*

The proof and detailed conditions for prior distributions are given in the Supplementary Material.

3 Implementation of SBPS

3.1 Synthesis model, posterior computation and spatial prediction

To synthesize multiple spatial predictions, we fit the synthesis model (3) to the observed data. We employ a Gaussian process to estimate unknown synthesis coefficients $\beta_j(s)$, which are independent for $j = 0, \ldots, J$, namely $\beta_j(s) \sim \mathrm{GP}(\overline{\beta}_j, \theta_j)$, where $\mathrm{GP}(\overline{\beta}_j, \theta_j)$ denotes a Gaussian process with mean $\overline{\beta}_j$ and covariance parameters θ_j. In what follows, we assume that $\overline{\beta}_j$ is fixed and $\theta_j = (\tau_j, g_j)$, where τ_j and g_j are unknown scale and spatial range parameters. A reasonable choice is $\overline{\beta}_j = 1/J$, meaning that the prior synthesis is simple averaging of J prediction models for all the locations, so that we use $\overline{\beta}_j = 1/J$ as a default choice.

Suppose we observe samples at n locations, $s_1, \ldots, s_n \in \mathcal{S}$. Let $y_i = y(s_i)$, $f_{ji} = f_j(s_i)$, $\varepsilon_i = \varepsilon(s_i)$, and $\beta_{ji} = \beta_j(s_i)$. Then, the model in (3) at the sampled locations is written as

$$\begin{aligned} y_i &= \beta_{0i} + \sum_{j=1}^{J} \beta_{ji} f_{ji} + \varepsilon_i, \quad \varepsilon_i \sim N(0, \sigma^2), \quad i = 1, \ldots, n, \\ \boldsymbol{\beta}_j &\equiv (\beta_{j1}, \ldots, \beta_{jn})^\top \sim N(0, \tau_j \boldsymbol{G}(g_j)), \quad j = 0, \ldots, J, \end{aligned} \tag{5}$$

where the (i, i')-element of $\boldsymbol{G}(g_j)$ is $C(\|s_i - s_{i'}\|; g_j)$ with valid correlation function $C(\cdot; g_j)$ and spatial range parameter g_j as defined above. The model in (3) is quite similar to the spatially varying coefficient model (Gelfand et al., 2003), but the difference is that the latent factor $f_j(s)$ in (3) is a random variable rather than fixed covariates, as in the standard varying coefficient model. For the prior distributions of the unknown parameters, we use $\sigma^2 \sim \mathrm{IG}(a_\sigma, b_\sigma)$, $\tau_j \sim \mathrm{IG}(a_\tau, b_\tau)$, and $g_j \sim U(\underline{g}, \overline{g})$, independently for $j = 1, \ldots, J$. We obtain the joint posterior distribution

$$\pi(\sigma^2) \prod_{j=0}^{J} \pi(\tau_j)\pi(g_j)\phi_n(\boldsymbol{\beta}_j; \boldsymbol{0}, \tau_j \boldsymbol{G}(g_j)) \times \prod_{j=1}^{J}\prod_{i=1}^{n} h_j(f_{ij}) \times \prod_{i=1}^{n} \phi\Big(y_i; \beta_{0i} + \sum_{j=1}^{J} \beta_{ji} f_{ji}, \sigma^2\Big)$$

where $\pi(\sigma^2), \pi(\tau_j)$ and $\pi(g_j)$ are prior distributions, and $\phi_n(\cdot; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ denotes a n-dimensional normal distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$.

At location s, the BSPS analysis will include inferences on the latent factor states, $f_j(s)$, as well as the spatially varying BSPS model parameters $\Phi(s)$. We first provide a computation algorithm using Markov chain Monte Carlo (MCMC). Suppose that $f_{ji} \sim N(a_{ji}, b_{ji})$ is received independently for $j = 1, \ldots, J$ and $i = 1, \ldots, n$, where a_{ji} and b_{ji} are provided by the J models. The MCMC algorithm to generate posterior samples of $\{f_{ji}\}, \{\beta_j\}, \{\tau_j\}, \{h_j\}$ and σ^2 is given as follows:

- (Sampling of f_{ji}) Generate f_{ji} from $N(A_{ji}^{(f)}B_{ji}^{(f)}, A_{ji}^{(f)})$, where

$$A_{ji}^{(f)} = \left(\frac{\beta_{ji}^2}{\sigma^2} + \frac{1}{b_{ji}}\right)^{-1}, \qquad B_{ji}^{(f)} = \frac{\beta_{ji}}{\sigma^2}\left(y_i - \beta_{0i} - \sum_{k\neq j}\beta_{ki}f_{ki}\right) + \frac{a_{ji}}{b_{ji}}$$

- (Sampling of $\boldsymbol{\beta}_j$) Generate $\boldsymbol{\beta}_j$ from $N(\boldsymbol{A}_j^{(\beta)}\boldsymbol{B}_j^{(\beta)}, \boldsymbol{A}_j^{(\beta)})$, where

$$\boldsymbol{A}_j^{(\beta)} = \left\{\sigma^{-2}\boldsymbol{\Omega}_j + \tau_j^{-1}\boldsymbol{G}(g_j)^{-1}\right\}^{-1},$$
$$\boldsymbol{B}_j^{(\beta)} = \frac{1}{\sigma^2}\boldsymbol{f}_j \circ \left(y - \boldsymbol{\beta}_0 - \sum_{k\neq j}\boldsymbol{f}_k \circ \boldsymbol{\beta}_k\right) + \frac{\overline{\beta}_j}{\tau_j}\boldsymbol{G}(g_j)^{-1}\mathbf{1}_n,$$

  with $\boldsymbol{\Omega}_j = \mathrm{diag}(f_{j1}^2, \ldots, f_{jn}^2)$ and $\boldsymbol{f}_j = (f_{j1}, \ldots, f_{jn})$. Note that $\circ$ denotes the Hadamard product.

- (Sampling of τ_j) Generate τ_j from $\mathrm{IG}(a_\tau + n/2, b_\tau + (\boldsymbol{\beta}_j - \overline{\beta}_j\mathbf{1}_n)^\top\boldsymbol{G}(g_j)^{-1}(\boldsymbol{\beta}_j - \overline{\beta}_j\mathbf{1}_n)/2)$.

- (Sampling of g_j) The full conditional of g_j is proportional to

$$|\boldsymbol{G}(g_j)|^{-1/2}\exp\left(-\frac{1}{2\tau_j}(\boldsymbol{\beta}_j - \overline{\beta}_j\mathbf{1}_n)^\top\boldsymbol{G}(g_j)^{-1}(\boldsymbol{\beta}_j - \overline{\beta}_j\mathbf{1}_n)\right), \quad g_j \in (\underline{g}, \overline{g}).$$

  A random-walk Metropolis-Hastings is used to sample from this distribution.

- (Sampling of σ^2) Using the conditionally conjugate prior $\sigma^2 \sim \mathrm{IG}(a_\sigma, b_\sigma)$, the full conditional is $\sigma^2 \sim \mathrm{IG}(a_\sigma + n/2, b_\sigma + \sum_{i=1}^n(y_i - \beta_{0i} - \sum_{j=1}^J\beta_{ji}f_{ji})^2/2)$.

Each item is sampled for $j = 1, \ldots, J$, per MCMC iteration. The information is then updated with each iteration to be used throughout the algorithm. Note that, in practice, $h_j(\cdot)$ is very likely to be a conditional density depending on some covariates. Extension to such a case is trivial.

Turning to predictions, let s_{n+1} be a new location where we are interested in predicting $y_{n+1} \equiv y(s_{n+1})$, assuming that the predictive distributions of $\boldsymbol{f}_{n+1} = (f_1(s_{n+1}), \ldots, f_J(s_{n+1}))$, namely, predictive distributions of the J models, are available. Then, the predictive distribution of y_{n+1} is obtained as

$$\begin{aligned} p(y_{n+1}|\boldsymbol{y}, \boldsymbol{f}_{n+1}) = \int \phi\Big(y_{n+1}; \beta_{0,n+1} + \sum_{j=1}^J \beta_{j,n+1}f_{j,n+1}), \sigma^2\Big)\prod_{j=1}^J h_j(f_{j,n+1})df_{j,n+1} \\ \times \prod_{j=0}^J p(\beta_{j,n+1}|\boldsymbol{\beta}_j; \tau_j, g_j)d\beta_{j,n+1} \times \pi(\Theta|\boldsymbol{y})d\Theta, \end{aligned} \tag{6}$$

where Θ is a collection of $\{f_{ji}\}, \{\beta_j\}, \{\tau_j\}, \{h_j\}$ and σ^2, $p(\beta_{j,n+1}|\boldsymbol{\beta}_j; \tau_j, g_j)$ is the conditional distribution of $\beta_{j,n+1}$ given $\boldsymbol{\beta}_j$, and $\pi(\Theta|\boldsymbol{y})$ is the posterior distribution of Θ. Under the assumption of Gaussian process on $\beta_j(s)$, the conditional distribution of $\beta_{j,n+1}$ is given by $N(\boldsymbol{G}_{n+1}(g_j)^\top\boldsymbol{G}(g_j)^{-1}\boldsymbol{\beta}_j, \{\tau_j - \tau_j\boldsymbol{G}_{n+1}(g_j)^\top\boldsymbol{G}(g_j)^{-1}\boldsymbol{G}_{n+1}(g_j)\}^{-1})$, where $\boldsymbol{G}_{n+1}(g_j) = (C(\|s_{n+1} - s_1\|; g_j), \ldots, C(\|s_{n+1} - s_n\|; g_j))^\top$. Sampling from the predictive distribution (6) can be easily carried out by using the posterior samples of Θ. First, independently generate $f_j(s_{n+1})$ from the predictive distribution of the jth model and generate $\beta_{j,n+1}$ from its conditional distribution given Θ. Then, we can generate y_{n+1} from $N(\beta_{0,n+1} + \sum_{j=1}^J\beta_{j,n+1}f_{j,n+1}, \sigma^2)$.

The full Gaussian process is known to be computationally prohibitive under large spatial data, since it requires computational cost $O(Jn^3)$ for each MCMC iteration of BSPS. To overcome the difficulty, we employ an m-nearest neighbor Gaussian process (Datta et al., 2016) for $\beta_j(s)$, which uses a multivariate normal distribution with a sparse precision matrix for $\beta_j(s_1), \ldots, \beta_j(s_n)$. Then, the computational cost at each iteration is reduced to $O(Jnm^2)$, which is a drastic reduction from the original computation cost $O(Jn^3)$, since m can be set to a small value (e.g. $m = 5$ or 10), even under $n \approx 10^4$. The detailed sampling steps under the nearest neighbor Gaussian process are provided in the Supplementary Material.

3.2 Variational Bayes approximation for fast point prediction synthesis

While the MCMC algorithm does provide full posterior estimation, it can also be prohibitively slow when the number of sampled locations or predictors is large. As such, we also develop an approximation algorithm using mean field variational Bayes (MFVB) approximation that is significantly more efficient than its MCMC counterpart. In applying the MFVB approximation, we assume that the prior distributions of the spatial range parameters, $g_0, g_1, \ldots, g_J$, are the uniform distribution on $\{\eta_1, \ldots, \eta_L\}$. The MFVB approximates the posterior distributions through the form

$$q(\{f_{ji}\}, \{\boldsymbol{\beta}_j\}, \{\tau_j\}, \{g_j\}, \sigma^2) = q(\sigma^2) \prod_{j=0}^{J} q(\boldsymbol{\beta}_j) q(\tau_j) q(g_j) \prod_{i=1}^{n} q(f_{ji}),$$

and each variational posterior can be iteratively updated by computing, for example, $q(\boldsymbol{\beta}_j) \propto \exp(E_{-\beta_j}[\log p(y, \Theta)])$, where $\Theta = (\{f_{ji}\}, \{\boldsymbol{\beta}_j\}, \{\tau_j\}, \{g_j\}, \sigma^2)$, and $E_{-\beta_j}$ denotes the expectation with respect to the marginal variational posterior of the parameters other than $\boldsymbol{\beta}_j$. From the forms of full conditional posterior distributions given in Section 3.1, the following distributions can be used as variational distributions:

$$q(f_{ji}) \sim N(\widetilde{m}_{ji}, \widetilde{s}^2_{ji}), \quad q(\boldsymbol{\beta}_j) \sim N(\widetilde{\mu}_j, \widetilde{\Sigma}_j), \quad q(\tau_j) \sim \mathrm{IG}(\widetilde{a}_{\tau_j}, \widetilde{b}_{\tau_j}),$$
$$q(g_j) \sim \mathrm{D}(\widetilde{p}_{j1}, \ldots, \widetilde{p}_{jL}), \quad q(\sigma^2) \sim \mathrm{IG}(\widetilde{a}_\sigma, \widetilde{b}_\sigma),$$

where $\mathrm{D}(\widetilde{p}_{j1}, \ldots, \widetilde{p}_{jL})$ is a discrete distribution on $\{\eta_1, \ldots, \eta_L\}$, such that $P(g_j = \eta_\ell) = \widetilde{p}_{j\ell}$. The MFVB algorithm is described as follows:

Algorithm 1. *Starting with* $\widetilde{m}^{(0)}_{ji}$, $\widetilde{s}^{2(0)}_{ji}$, $\widetilde{\mu}^{(0)}_j$, $\widetilde{\Sigma}^{(0)}_j$, $\widetilde{a}^{(0)}_{\tau_j}$, $\widetilde{b}^{(0)}_{\tau_j}$, $\widetilde{p}^{(0)}_{j\ell}$, $\widetilde{a}^{(0)}_\sigma$, $\widetilde{b}^{(0)}_\sigma$ *and* $r = 0$, *repeat the following process until numerical convergence: for* $j = 1, \ldots, J$, *update* $\widetilde{m}_{ji}$ *and* $\widetilde{s}^2_{ji}$ *as*

$$\widetilde{s}^{2(t+1)}_{ji} \leftarrow \left\{ \frac{1}{b_{ji}} + (\widetilde{\mu}^{(s_0)2}_{ji} + \widetilde{\Sigma}^{(s_0)}_{jii}) \frac{\widetilde{a}^{(s_0)}_\sigma}{\widetilde{b}^{(t)}_\sigma} \right\}^{-1},$$

$$\widetilde{m}^{(t+1)}_{ji} \leftarrow \frac{a_{ji}}{b_{ji}} + \widetilde{\mu}^{(t)}_{ji} \frac{\widetilde{a}^{(t)}_\sigma}{\widetilde{b}^{(t)}_\sigma} \left(y_i - \widetilde{\mu}^{(t)}_{0i} - \sum_{k<j} \widetilde{\mu}^{(t)}_{ki} \widetilde{m}^{(t+1)}_{ki} - \sum_{k>j} \widetilde{\mu}^{(t)}_{ki} \widetilde{m}^{(t)}_{ki} \right) \Big/ \widetilde{s}^{2(t+1)}_{ji}.$$

For $j = 1, \ldots, J$, *update* $\widetilde{\mu}_j$ *and* $\widetilde{\Sigma}_j$ *as*

$$\widetilde{\Sigma}^{(t+1)}_j \leftarrow \left\{ \Omega^{(t+1)}_j \frac{\widetilde{a}^{(t)}_\sigma}{\widetilde{b}^{(t)}_\sigma} + \sum_{\ell=1}^{L} \widetilde{p}^{(t)}_{j\ell} G(\eta_\ell)^{-1} \frac{\widetilde{a}^{(t)}_{\tau_j}}{\widetilde{b}^{(t)}_{\tau_j}} \right\}^{-1},$$

$$\widetilde{\mu}^{(t+1)}_j \leftarrow \{\widetilde{\Sigma}^{(t+1)}_j\}^{-1} \frac{\widetilde{a}^{(t)}_\sigma}{\widetilde{b}^{(t)}_\sigma} \widetilde{m}^{(t+1)}_j \circ \left(y - \widetilde{\mu}^{(t+1)}_0 - \sum_{k<j} \widetilde{\mu}^{(t+1)}_k \circ \widetilde{m}^{(t+1)}_k - \sum_{k>j} \widetilde{\mu}^{(t)}_k \circ \widetilde{m}^{(t+1)}_k \right),$$

where $\Omega_j^{(t+1)} = \text{diag}(\widetilde{m}_{j1}^{(t+1)2} + \widetilde{s}_{j1}^{2(t+1)}, \ldots, \widetilde{m}_{jn}^{(t+1)2} + \widetilde{s}_{jn}^{2(t+1)})$. *For* $j = 0, \ldots, J$, *set* $\widetilde{a}_{\tau_j}^{(t+1)} = a_\tau + n/2$ *and update* $\widetilde{b}_{\tau_j}$ *as*

$$\widetilde{b}_{\tau_j}^{(t+1)} \leftarrow b_\tau + \frac{1}{2}\text{tr}\left\{\left(\widetilde{\mu}_j^{(t+1)}\widetilde{\mu}_j^{(t+1)\top} + \widetilde{\Sigma}_j^{(t+1)}\right)\sum_{\ell=1}^{L}\widetilde{p}_{j\ell}^{(t)}G(\eta_\ell)^{-1}\right\}.$$

For $j = 0, \ldots, J$, *update* $\widetilde{p}_{j\ell}$ *as*

$$\widetilde{p}_{j\ell}^{(t+1)} \leftarrow \frac{|G(\eta_\ell)|^{-1/2}\exp\left(-\widetilde{a}_{\tau_j}^{(t+1)}\text{tr}\left\{(\widetilde{\mu}_j^{(t+1)}\widetilde{\mu}_j^{(t+1)\top} + \widetilde{\Sigma}_j^{(t+1)})G(\eta_\ell)^{-1}\right\}/2\widetilde{b}_{\tau_j}^{(t+1)}\right)}{\sum_{\ell'=1}^{L}|G(\eta_{\ell'})|^{-1/2}\exp\left(-\widetilde{a}_{\tau_j}^{(t+1)}\text{tr}\left\{(\widetilde{\mu}_j^{(t+1)}\widetilde{\mu}_j^{(t+1)\top} + \widetilde{\Sigma}_j^{(t+1)})G(\eta_{\ell'})^{-1}\right\}/2\widetilde{b}_{\tau_j}^{(t+1)}\right)}.$$

Set $\widetilde{a}_\sigma^{(t+1)} = a_\sigma + n/2$ *and update* $\widetilde{b}_\sigma$ *as*

$$\begin{aligned}\widetilde{b}_\sigma^{(t+1)} \leftarrow &\left(y - \widetilde{\mu}_0^{(t+1)} - \sum_{j=1}^{J}\widetilde{\mu}_j^{(t+1)} \circ \widetilde{m}_j^{(t+1)}\right)^\top\left(y - \widetilde{\mu}_0^{(t+1)} - \sum_{j=1}^{J}\widetilde{\mu}_j^{(t+1)} \circ \widetilde{m}_j^{(t+1)}\right) + \text{tr}(\widetilde{\Sigma}_0^{(t+1)}) \\ &+ \sum_{j=1}^{J}\text{tr}\left\{(\widetilde{\mu}_j^{(t+1)}\widetilde{\mu}_j^{(t+1)\top} + \widetilde{\Sigma}_j^{(t+1)}) \circ (\widetilde{m}_j^{(t+1)}\widetilde{m}_j^{(t+1)\top} + \widetilde{S}_j^{(t+1)})\right\} - \sum_{i=1}^{n}\sum_{j=1}^{J}\widetilde{m}_{ji}^{(t+1)2}\widetilde{\mu}_{ji}^{(t+1)2}.\end{aligned}$$

A reasonable starting value for Algorithm 1 is the posterior mean of a small number of MCMC samples. We note that the updating step may contain calculations of the inverse of $n \times n$ matrices, as in the MCMC algorithm, which could be computationally prohibitive when n is large. Alternatively, we can also develop a variational approximation algorithm for the nearest neighbor Gaussian process.

3.3 BSPS under general types of response variables

The proposed BSPS framework (1) can be extended to situations with general types of outcomes, by generalizing the synthesis model (3) to generalized spatially varying models (e.g. Gelfand et al., 2003; Kim and Wang, 2021). Here, we consider a specific situation where y_i is a binary response, which will be treated in Section 5.1. The linear latent factor model (3) for continuous response can be modified as

$$y_i|\psi_i \sim \text{Ber}\left(\frac{\exp(\psi_i)}{1+\exp(\psi_i)}\right), \quad \psi_i = \beta_{0i} + \sum_{j=1}^{J}\beta_{ji}f_{ji}, \quad i = 1, \ldots, n \tag{7}$$

where β_{ji} follows the same Gaussian process given in (3). Suppose that f_{ji} is a predictive distribution of binary response, that is, $f_{ji} \sim \text{Ber}(a_{ji})$.

To enhance the efficiency of posterior computation, we employ the following Pólya-gamma data augmentation (Polson et al., 2013):

$$\frac{\exp(\psi_i y_i)}{1+\exp(\psi_i)} = \frac{1}{2}\exp\left\{\left(y_i - \frac{1}{2}\right)\psi_i\right\}\int_0^\infty \exp\left(-\frac{1}{2}\omega_i\psi_i^2\right)p(\omega_i; 1, 0)d\omega_i,$$

where $p(\cdot; b, c)$ denotes the Pólya-gamma density with parameters b and c. Then, the full conditional distribution of $\boldsymbol{\beta}_j$ $(j = 0, \ldots, J)$ is a normal distribution, $N(\boldsymbol{A}_j^{(\beta)}\boldsymbol{B}_j^{(\beta)}, \boldsymbol{A}_j^{(\beta)})$, where

$$\begin{aligned}\boldsymbol{A}_j^{(\beta)} &= \text{diag}(\omega_1 f_{j1}^2, \ldots, \omega_n f_{jn}^2) + \tau_j^{-1}\boldsymbol{G}(g_j)^{-1}, \\ \boldsymbol{B}_j^{(\beta)} &= \boldsymbol{F}_j \circ \left\{\boldsymbol{y}^* - \boldsymbol{\omega} \circ \left(\boldsymbol{\beta}_0 + \sum_{k\neq j}\boldsymbol{f}_k \circ \boldsymbol{\beta}_k\right)\right\},\end{aligned}$$

where $\boldsymbol{y}^* = (y_1 - 1/2, \ldots, y_n - 1/2)$ and $\boldsymbol{\omega} = (\omega_1, \ldots, \omega_n)$. The full conditional distribution of f_{ji} is $\text{Ber}(a_{ji}^{(f)})$, where

$$a_{ji}^{(f)} = \frac{a_{ji}}{a_{ji} + (1 - a_{ji})L_{ji}}, \qquad L_{ji} = \frac{\exp(\psi_i^{(0)} y_i)}{\exp(\psi_i^{(1)} y_i)} \cdot \frac{1 + \exp(\psi_i^{(1)})}{1 + \exp(\psi_i^{(0)})}.$$

Here, $\psi_i^{(k)}$ is the value of ψ_i with $f_{ji} = k$. Finally, the full conditional distribution of ω_i is $\text{PG}(1, \psi_i)$.

4 Simulation Studies

This section provides a simulation study to illustrate the efficacy of our proposed BSPS compared to other methods for spatial data.

4.1 Empirical behavior of BSPS

We first illustrate how our proposed BSPS synthesizes candidate models. We set $n = 300$ (training sample size) and generated two-dimensional location information s_i (for $i = 1, \ldots, n$) from the uniform distribution on $[-1, 1]^2$. Let $z_1(s_i)$ and $z_2(s_i)$ be the two independent realizations of a spatial Gaussian process with mean zero and a covariance matrix defined from an isotropic exponential function: $\text{Cov}(z_k(s_i), z_k(s_j)) = \exp(-\|s_i - s_j\|/0.5)$ for $k = 1, 2$. Then, we define two covariates $x_1(s_i)$ and $x_2(s_i)$ via linear transformations, $x_1(s_i) = z_1(s_i)$ and $x_2(s_i) = r z_1(s_i) + \sqrt{1 - r^2} z_2(s_i)$ with $r = 0.2$, which allows dependence between $x_1(s_i)$ and $x_2(s_i)$. The response variable $y(s_i)$ at each location is generated from the following process:

$$y(s_i) = \begin{cases} w(s_i) + x_1(s_i) - 0.5x_2^2(s_i) + \varepsilon(s_i), & s_i \in D_1, \\ w(s_i) + x_1^2(s_i) + x_2^2(s_i) + \varepsilon(s_i), & s_i \in D_2. \end{cases}$$

where $D_1 = \{s_i = (s_{i1}, s_{i2}) \mid s_{i1} \leq 0\}$ and $D_2 = \{s_i = (s_{i1}, s_{i2}) \mid s_{i1} > 0\}$. Here $w(s_i)$ is a spatial random effect following a mean-zero Gaussian process with spatial covariance function, $\text{Cov}(s_i, s_{i'}) = (0.3)^2 \exp(-\|s_i - s_{i'}\|/0.3)$, and $\varepsilon(s_i)$ is an independent error term distributed as $\varepsilon(s_i) \sim N(0, 1)$. Note that, in the above setting, the spatial region is divided into two sub-regions, where the mean structure of the response, as a function of covariates, is different. For the data generated from the process, we apply a quadratic regression model without spatial effects, $y(s_i) = \beta_0 + \beta_1 x_1(s_i) + \beta_2 x_1(s_i)^2 + \beta_3 x_2(s_i) + \beta_4 x_2(s_i)^2 + \varepsilon_i$, to subsamples in D_1 and compute the predictive mean and variance of all the samples, which are denoted by a_{1i} (mean) and b_{1i} (variance), respectively. We conduct the same procedure using subsamples in D_2 to obtain a_{2i} and b_{2i}. We apply BSPS using the two prediction models, $f_{ji} \sim N(a_{ji}, b_{ji})$ with $j = 1, 2$, and an exponential kernel, $(G(g_j))_{ii'} = \exp(\|s_i - s_{i'}\|/g_j)$ in 10-nearest neighbor Gaussian processes. We generate 1000 posterior samples after discarding the first 1000 samples as burn-in.

We first evaluate the predictive performance in non-sampled locations. We generated 200 additional locations, as with $w(s_i)$, $x_1(s_i)$, $x_2(s_i)$, and $y(s_i)$, according to the same data generating process. Using the generated posterior samples from BSPS, posterior samples of the spatially varying coefficients in the non-sampled locations were generated to get posterior predictive distributions of the response in non-sampled locations. We evaluate the mean squared error (MSE) of the posterior predictive means of BSPS, as well as predictors of the two quadratic regressions. For comparison, we employ two methods of prediction synthesis, Bayesian model averaging (BMA) and simple averaging (SA). In

the former method, we compute the Bayesian information criterion for the two quadratic models to approximate the marginal likelihood. In the latter method, the prediction results of QR1 and QR2 are simply averaged. The MSE of non-sampled locations are 1.33 (BSPS), 2.61 (BMA), and 4.19 (SA), while the MSE of the two models are 2.61 (QR1) and 10.85 (QR2). Since QR1 is estimated using data only in D_1, its predictive performance in D_2 is not expected to be good, due to the difference in true regression structures between D_1 and D_2, which leads to QR1 having a large MSE. The same explanations can be given for QR2 and its MSE. We found that the model weight for QR1 in BMA is almost 1, so the performance of BMA and QR1 is almost identical. It is reasonable that SA performs worse than QR1, since it gives equal weight to QR2, which does not perform well in this example. Comparatively, BSPS provides much better prediction results than the two ensemble methods in terms of MSE. The main reason is that BSPS can combine the two models with spatially varying weights, and such design of weights is essential in this example, since the usefulness of the two models are drastically different in D_1 and D_2. Furthermore, although both QR1 and QR2 do not take into account the existence of the spatial random effect, the intercept term in BSPS could successfully capture the remaining spatial variation, which increases the prediction accuracy in this example.

To see how BSPS works in this example, we compute the ratio of two coefficients, $|\widehat{\beta}_{1i}|/(|\widehat{\beta}_{1i}| + |\widehat{\beta}_{2i}|)$, where $\widehat{\beta}_{1i}$ and $\widehat{\beta}_{2i}$ are posterior means of β_{1i} (weight for QR1) and β_{2i} (weight for QR2), which shows the importance of the prediction made by QR1. The result is shown in the left panel of Figure 1, which clearly shows that the model weight for QR1 is large in D_1 (left region) and is close to 0 in D_2, where QR1 is not expected to predict well. This means that BPS can automatically detect the effective model at local regions through Bayesian updating. We also note that the model weight smoothly changes over the region and two prediction models.

We finally evaluate the coverage accuracy of the 95% interval prediction. In the right panel of Figure 1, we present the mean values (point prediction) of predictive distributions with their associated 95% prediction intervals. The result shows that the prediction intervals mostly cover the true values with reasonable interval lengths. In fact, the coverage proportion is 0.97, which is practically equivalent to the nominal level, illustrating how well-calibrated BSPS is. For comparison, we also present the point prediction made by BMA and SA in the right panel of Figure 1. It can be seen that BMA and SA fail to predict observations having large absolute values.

4.2 Performance comparison

We next compare the performance of BSPS with other methods through Monte Carlo simulations. Let $s \in [0,1]^2$ be the spatial location generated from the uniform distribution of the region. The covariates $x_1 \equiv x_1(s)$ and $x_2 \equiv x_2(s)$ in the same way as in Section 4.1. Here we also generated $x_3, \ldots, x_p$ ($p \geq 5$) independently from $N(0,1)$. We then consider the following two scenarios of data generating process:

$$\textbf{Scenario 1:} \quad y(s) = w(s) + x_3^2 \exp(-0.3\|s\|^2) + s_2 \sin(2x_2) + \varepsilon,$$

$$\textbf{Scenario 2:} \quad y(s) = 2w(s) + \frac{1}{2}\sin(\pi x_1 x_2) + (x_3 - 0.5)^2 + \frac{1}{2}x_4 + \frac{1}{4}x_5 + \varepsilon,$$

where $\varepsilon \sim N(0, (0.7)^2)$ is an error term and $\omega(s)$ is an unobserved spatial effect following a zero-mean Gaussian process with covariance $\mathrm{Cov}(s, s') = (0.3)^2 \exp(-\|s - s'\|/0.3)$. Note that the mean function in Scenario 1 changes according to location s and the mean function in Scenario 2 is the well-known Friedman function (Friedman, 1991).

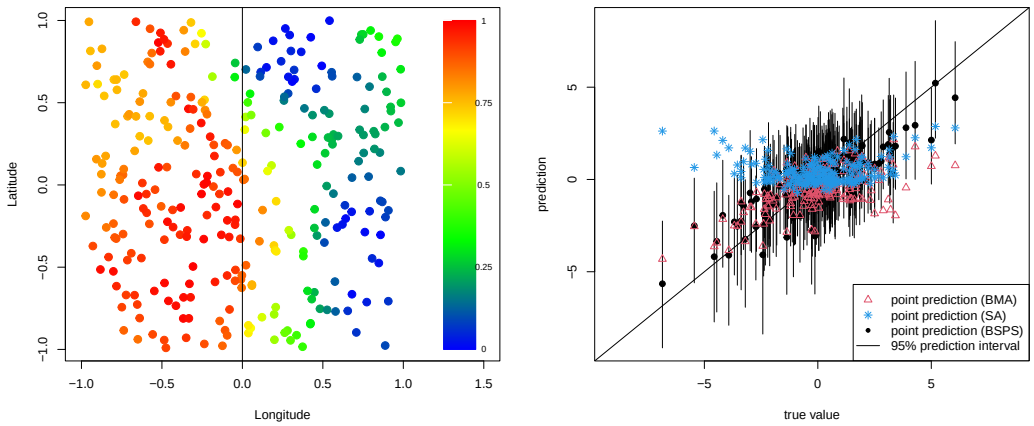

Figure 1: Left: Spatial plot of the ratio of two coefficients, $|\widehat{\beta}_{2i}|/(|\widehat{\beta}_{2i}| + |\widehat{\beta}_{3i}|)$. Right: point prediction made by BMA, SA, and BPS, and 95% prediction intervals (vertical lines) obtained from the predictive distributions of BSPS.

We generated 300 training samples and 100 test samples of (y, x, s) from each data generating process. To predict test samples, we employed the following prediction models:

- GWR (geographically weighted regression): We used the R package "spgwr" to fit the GWR model (Brunsdon et al., 1998), where the optimal bandwidth is selected by cross-validation.

- AM (additive model): We fitted the LAM model (Hastie and Tibshirani, 1987) with covariates x_i by using the R package "gam" with default settings for tuning parameters.

- SPR (spatial regression): We fitted spatial linear regression with unobserved spatial effects modeled by 5-nearest neighbor Gaussian process, using R package "spNNGP" (Finley et al., 2022), in which we draw 1000 posterior samples after discarding 1000 samples.

- SRF (spatial random forest): We applied the recently proposed SRF (Saha et al., 2021) using the R package "RandomForestsGLS" with 50 trees.

We then synthesized the three prediction results by GWR, AM, and SPR, noting that SRF does not produce uncertainty measures (e.g. prediction variance) and is kept as a competitor. We implement BSPS with 10-nearest neighbor Gaussian process and normal latent factors based on prediction values and their variances obtained from the three models. We generated 1000 posterior samples of the model coefficients, as well as the unknown parameters, after discarding the first 1000 samples as burn-in, to obtain the posterior samples of the test data. We also applied the variational Bayes (VB) algorithm described in Section 3.2 to efficiently compute approximate posterior distributions. For comparison, we synthesized the three models via BMA and SA, as considered in Section 4.1.

We compute the mean squared error (MSE) of the test data for 500 Monte Carlo replications, and present nine empirical quantiles $(10\%, 20\%, \ldots, 90\%)$ of MSE values in Figure 2. It is observed that BSPS provides the most accurate predictions in all the scenarios except for Scenario 2 with $p = 5$, but the performance of BSPS is comparable

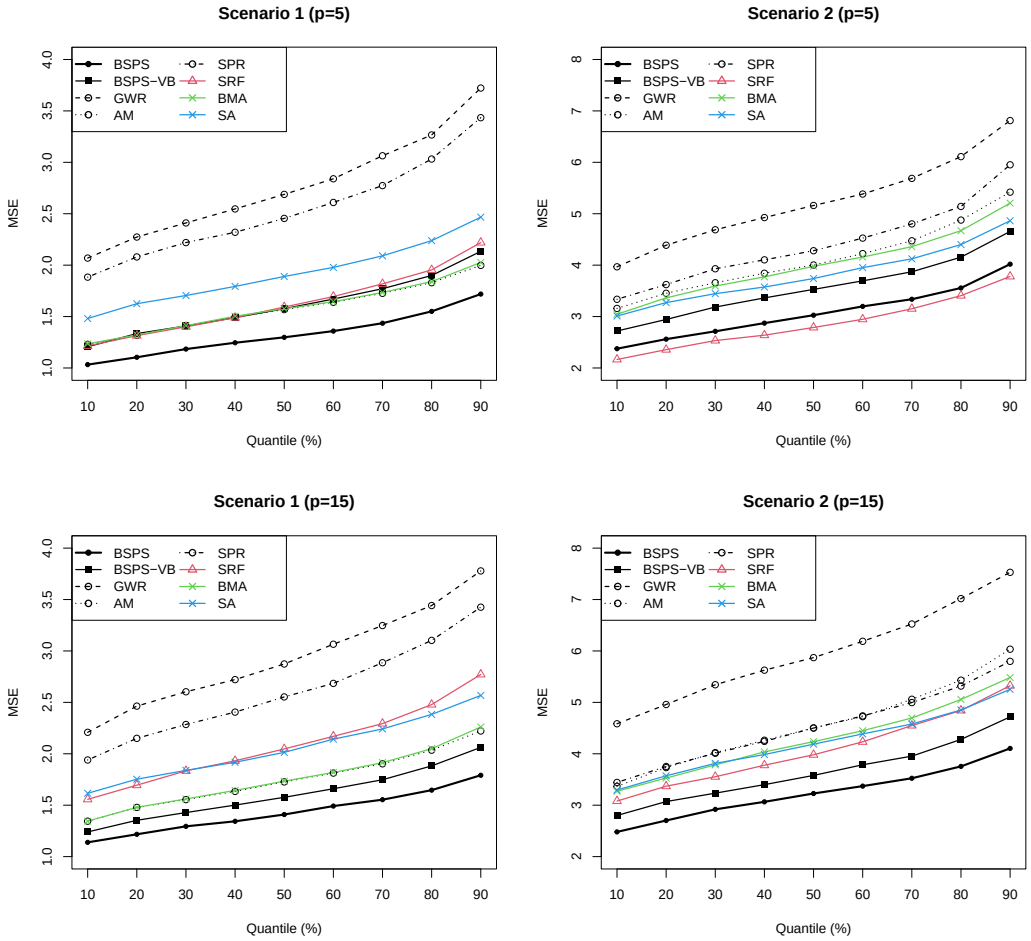

Figure 2: Empirical 9 quantiles $(10\%, 20\%, \ldots, 90\%)$ of MSE values under 500 replications under two scenarios of data generating process with $p \in \{5, 15\}$.

with the best method, SRF. It should be noted that the performance of BSPS is much better than the other ensemble methods, BMA and SA, which could be attributed to the spatially varying model weight in BSPS. Furthermore, BSPS improves the prediction accuracy of the three agent methods (GWR, AM, and SPR) in all scenarios, even though BMA and SA do not necessarily improve the performance, as confirmed in Scenario 1. The fast prediction method by BSPS-VB performs slightly worse than BSPS based on MCMC, though it still performs better than the standard ensemble methods, BMA and SA, in all scenarios.

We next evaluated the performance of 95% interval prediction. Here we focus on BSPS, GWR, AM, and SPR, since SRF, BMA, and SA do not produce interval predictions. The empirical coverage probability (CP) and average length (AL) under four scenarios are presented in Table 1. While CPs of GWR, AM, and SPR are not necessarily around the nominal level, CPs of BSPS are fairly close to the nominal level. Moreover, ALs of BSPS tend to be shorter than the other methods, indicating both accuracy and efficiency of prediction intervals of BSPS.

Table 1: Coverage probability (CP) and average interval length (AL) of 95% prediction intervals of test samples, averaged over 500 Monte Carlo replications.

		CP				AL			
Scenario	p	BSPS	GWR	AM	SPR	BSPS	GWR	AM	SPR
1	5	96.9	91.7	94.0	95.3	4.98	5.61	4.75	6.35
1	15	95.7	91.4	91.9	95.3	4.88	5.74	4.65	6.39
2	5	95.0	88.4	94.0	96.0	6.88	7.05	7.66	8.69
2	15	94.5	89.2	92.4	96.0	6.89	7.69	7.51	8.82

5 Real Data Applications

We consider two distinct real world applications to highlight the predictive performance of BSPS. The first dataset is ecological: predicting the occurrence of Tsuga canadensis in Michigan, USA. The second dataset is real-estate: predicting apartment prices in Tokyo, Japan. Both datasets are distinct, in that the ecological dataset is binary and deals with natural processes, while the real-estate dataset is continuous and deals with human economic activity. This is done to illustrate the efficacy of BSPS and compare different methods under distinctly different situations, to provide a more holistic assessment.

5.1 Occurrence of Tsuga canadensis in Michigan

The first real world application concerns the occurrence of Tsuga canadensis (Eastern hemlock) in Michigan, USA, analyzed in Lany et al. (2020). The data comprise hemlock occurrence (binary outcome) on 17743 forest stands across the state of Michigan. A set of covariates were also observed at each stand and can be used to explain the probability of hemlock occurrence. Covariates include minimum winter temperature (MIN), maximum summer temperature (MAX), total precipitation in the coldest quarter of the year (WIP), total precipitation in the warmest quarter of the year (SUP), annual actual evapotranspiration (AET), and annual climatic water deficit (DEF). Spatial coordinates are recorded in longitude (lon) and latitude (lat).

There are several reasons why the prediction of the occurrence of Tsuga canadensis is relevant for this application. As a long-lived, foundational species in Michigan, conservation is critical due to it being threatened by the hemlock woolly adelgid Adelges tsugae, an invasive sap-feeding insect. Thus, predicting the occurrence is key in protecting the hemlock from this invasive species, by proactively making preventative measures. Further, since the mechanism for hemlock habitat is not known, as hemlock does not occur in all suitable habitats, the interpretability of the prediction is relevant for future conservation.

To investigate the predictive performance of BSPS and compare it to the other methods, we randomly omitted 2000 spatial locations as the validation set, and used the remaining $n = 15743$ samples as the training set. For the models to be synthesized in BPS, we consider three Bernoulli models, $y_i \sim \text{Ber}(e^{\psi_i}/(1 + e^{\psi_i}))$, for $i = 1, \ldots, n$, with the following specifications on the linear predictor, p_i, based on generalized linear models (GLM)

and generalized additive models (GAM):

$$\text{(GLM)} \quad \psi_i = \beta_0 + \sum_{k=1}^{p} \beta_k x_{ik},$$

$$\text{(GAM1)} \quad \psi_i = g_1(\text{lon}_i) + g_2(\text{lat}_i) + \sum_{k=1}^{p} \beta_k x_{ik},$$

$$\text{(GAM2)} \quad \psi_i = g_1(\text{lon}_i) + g_2(\text{lat}_i) + \sum_{k=1}^{p} f_k(x_{ik}),$$

where $x_i = (x_{i1}, \ldots, x_{ip})$ with $p = 6$ is the vector of covariates, and g_1, g_2 and f_k are unknown functions. We compute the occurrence probability in the validation dataset using the covariates and location information. To synthesize these predictors through BSPS, we apply the logistic synthesis model (7) with $J = 3$ latent factors corresponding to the above three predictors, and employed the nearest neighbor Gaussian process for the spatially varying model coefficients with $m = 10$ nearest neighbors and an exponential covariance function. We generated 7000 posterior samples after discarding the first 3000 samples as burn-in, and generated random samples for the coefficient vectors in the validation set to compute predictions of binomial probability. For comparison, we applied Bayesian model averaging (BMA) and simple averaging (SA) to combine the three models. Furthermore, we also applied a spatial logistic regression (SPR) model with spatial random effects modeled by the $m = 5$ nearest neighbor Gaussian process, which was fitted using the R package "spNNGP." We generated 7000 posterior samples after discarding the first 3000 samples. Spatial random forest (Saha et al., 2021) was not considered for this application, since it does not support binary responses.

In Figure 3, we present the spatial distributions of the posterior means of the spatially varying model coefficients, $\beta_j(s)$ $(j = 0, 1, 2, 3)$, which shows how the importance of the three models change over regions. Particularly, it is interesting to see that the simplest GLM model is found to be more relevant for synthesis than the other models in some locations. This exemplifies how predictive performances vary spatially, where even simple models can be effective and relevant depending on the region. To compare the prediction performance in the test data, we compute the receiver operating characteristic (ROC) curves for the predicted binomial probabilities. The results are presented in the left panel of Figure 4, where the resulting values of area under the curve (AUC) are given in parenthesis. Moreover, we repeat the process, splitting the data and predicting the test data, 20 times, and report the boxplots of AUC values in the right panel of Figure 4. The figures show the superiority of BSPS to all other methods, including the existing model averaging methods, BMA and SA, in terms of AUC values. An interesting phenomenon, though consistent across the studies in this paper, is that the AUC values of GLM, GAM1, and GAM2 are lower than that of SPR, but the AUC value of the synthesized prediction, through BSPS, is higher. On the other hand, the AUC values of the ensemble methods, BMA and SA, are at most the best performing model, GAM2. This indicates the effectiveness of BSPS in synthesizing simple models to provide more accurate predictions.

5.2 Apartment prices in Tokyo

Our second application is to apply BSPS to spatial predictions of apartment prices in the 23 wards in Tokyo, Japan. We used rent information using the "Real Estate Data Library Door Data Nationwide 2013-2017 Data Set" (At Home Co., Ltd.) stored in the collaborative research system at the Center for Spatial Information Science, The University

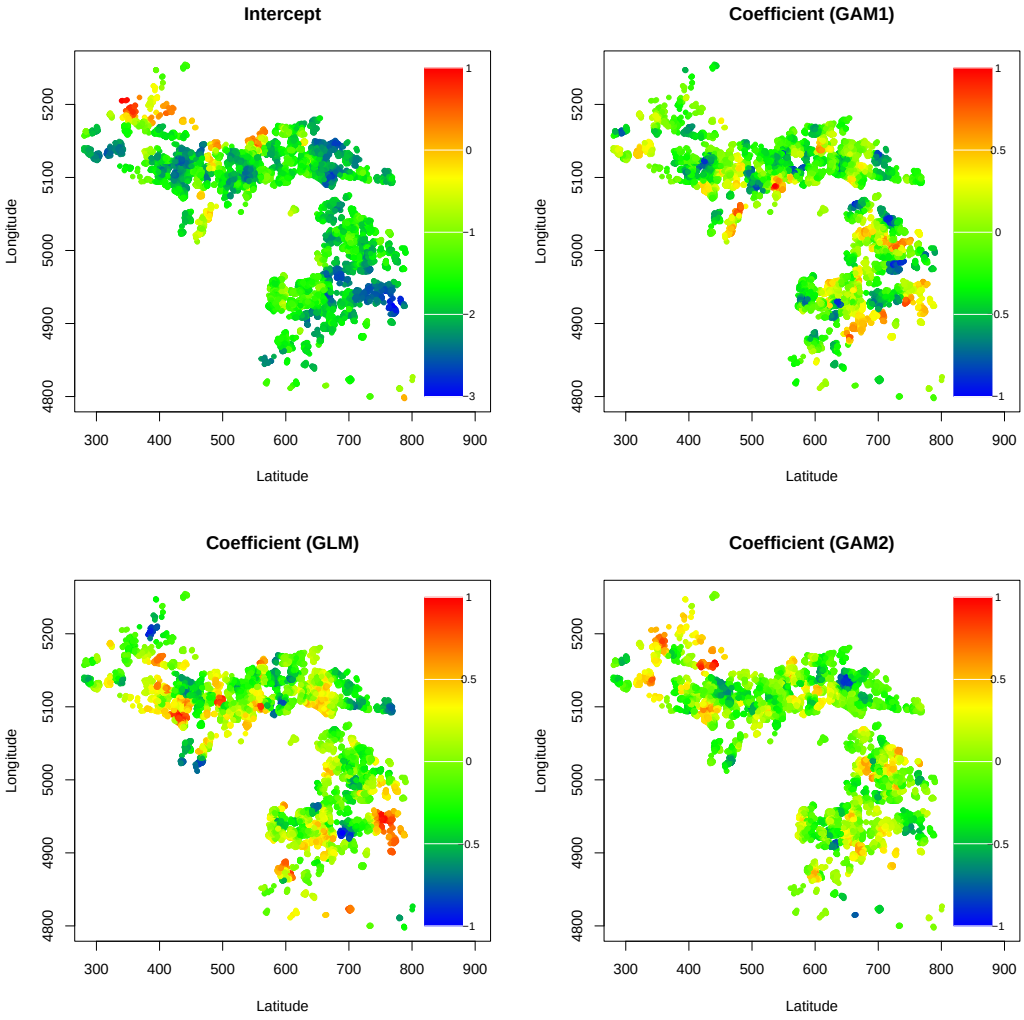

Figure 3: Posterior means of spatially varying intercept and coefficients in the logistic BSPS model, eq. (7).

of Tokyo (`https://joras.csis.u-tokyo.ac.jp`). The dataset contains the prices (yen), as well as auxiliary information on each room, for apartments handled by At Home, Inc. from 2013 to 2017. In this study, we used the samples collected in 2017, resulting in 22817 samples in total. We adopted 11 covariates, five dummy variables of room arrangement, room area (m^2), balcony area (m^2), walking minutes from the nearest train station, age of building (month), indicator of newly-built room, and location floor. For location information, the longitude and latitude information of each building, the name of the nearest train station, and the name of the ward are available. Since rooms in the same building share the same geographical information, we added a very small noise generated from $N_2(0, 10^{-3}I_2)$ to such rooms to avoid numerical instability. The room prices are log-transformed.

Similar to the ecological application in Section 5.1, this application requires both accurate and interpretable predictions. In terms of predictions, this is relevant for buyers, sellers, and real estate companies, but also for local governments to enact well-informed housing policies. As apartment prices are not always reported, in the sense that they are not listed or prices are outdated, the prediction of these prices is crucial. In terms

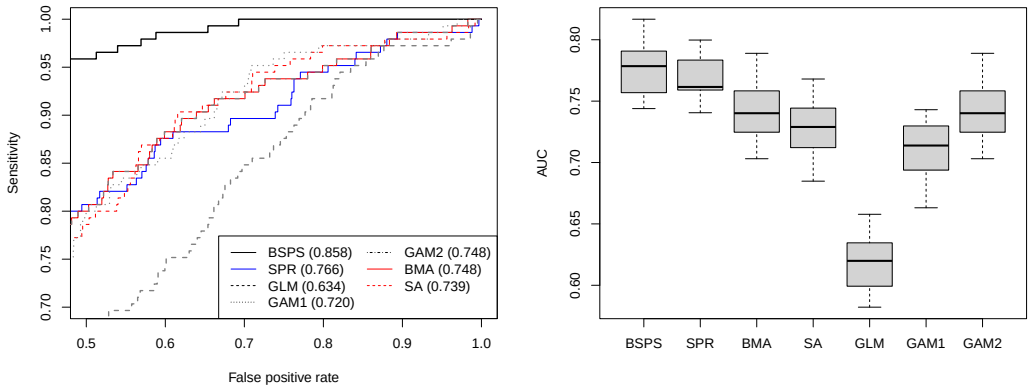

Figure 4: Receiver operating characteristic (ROC) curves for various prediction methods (left), where the resulting values of area under the curve (AUC) are given in parenthesis. AUC values for the six methods under 20 replications (right).

of interpretation, one consideration that has received a lot of interest is the question of fairness and discrimination in these pricing models. With the rise of black-box, machine learning models in real estate, the question of discriminatory pricing, that is not necessarily intended but nonetheless happens due to the black-box nature of these algorithms, has been a major concern. Having full interpretability, thus, is important for fair and transparent pricing practices.

We randomly omitted 2000 samples from the dataset, which are left as test samples. To construct the prediction models for room prices, we consider the following three types of models:

- **Station-level model:** The dataset is grouped according to 438 nearest train stations and simple linear regression with 5 covariates (walking minutes, room areas, and three dummy variables for room arrangement) is applied to each grouped samples.

- **Ward-level model:** The dataset is grouped according to the 23 wards and an additive model with 6 continuous covariates and three dummy variables for room arrangement is applied to each grouped samples.

- **Full model:** An additive model with 6 continuous covariates, two-dimensional location information, and five dummy variables for room arrangement.

Since the sample size that can be used to estimate the models increase in the order of station-level model, ward-level model, and full model, we vary the model complexity (e.g. number of parameters) in the three types of models. We also note that the three models are fully interpretable. The above models provide the means and variances for each training sample, and we synthesize the predictions with BSPS by assuming normality for each prediction model. With an exponential kernel in $m = 5$ nearest neighbor Gaussian process for spatially varying model coefficients, we generated 7000 posterior samples after discarding 3000 samples as burn-in. For comparison, we applied spatial random forest (SRF) and spatial regression (SPR), as used in Section 4.2, to predict the room prices in the test data. Furthermore, we employed the extreme gradient boosting (XGB; Chen and Guestrin, 2016) with 1000 trees, where the optimal number of trees was selected via 5-fold

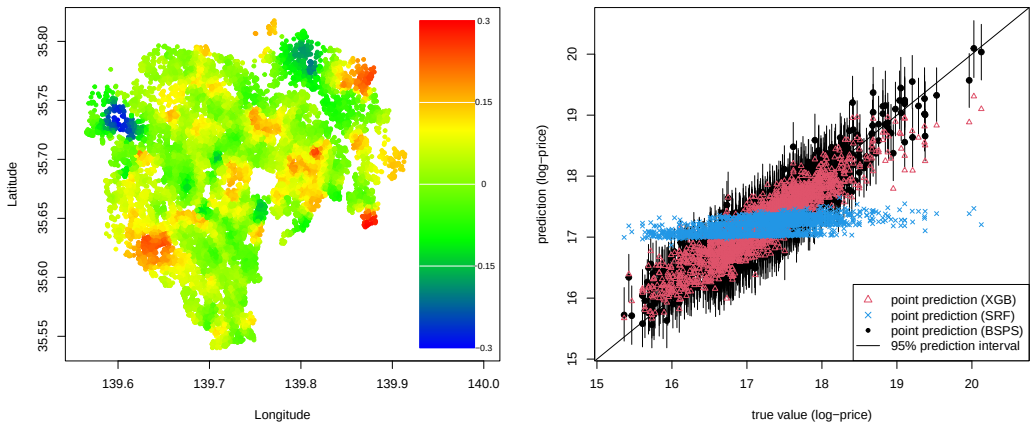

Figure 5: Left: Spatial plot of posterior means of the intercept term. Right: Point predictions of XGB (red circle), SRF (blue circle), and BSPS (black circle) with the 95% prediction intervals (vertical lines) based on the predictive distributions given by BSPS.

cross validation and learning and sub-sampling rates set to 0.01 and 0.1, respectively. We utilized the R package "xgboost" with the 6 covariates and the two-dimensional location variables as inputs.

The left panel in Figure 5 reports the spatial plot of $\beta_0(s_i)$, namely, the intercept term of BSPS. Since the intercept term captures the variability not captured by the model set, it effectively represents the model set uncertainty. Looking at the figure, we can see that the intercept is the largest in absolute value in certain regions. Each of these regions has different reasons for why the model set uncertainty is so high, some are due to new development skewing prices, some are due to heterogeneity in popular residential areas, and some are due to changes in disclosure rules. While the reason varies, the output of BSPS gives a clear and transparent indication for further inquiry.

We now consider comparing the predictive accuracy of each method for this application. The MSE values for predicting the test samples are

$$\text{BPS} : 0.240, \quad \text{XGB} : 0.257, \quad \text{SPR} : 0.268, \quad \text{SRF} : 0.579,$$

where, again, BSPS provides superior prediction accuracy. As with the previous applications, it should be noted that the resulting predictors made by BPS are interpretable, while XGB and SRF are not. We computed the 95% prediction intervals from the posterior distributions for the test samples. In the right panel in Figure 5, we report the point predictions of XGB, SRF, and BSPS, and the 95% prediction intervals for BPS. First, the relatively large MSE values of SRF come from the degeneracy of the point prediction, that is, the point prediction is much less variable than the true prices. This can be seen by the fact that the predictions are mostly horizontal, not deviating much from the mean. While XGB provides reasonable point prediction overall, XGB tends to under-predict the large true price, as with the ecological application. On the other hand, BSPS provides accurate point predictions and 95% prediction intervals with reasonable interval length regardless of the true price. The coverage proportion is 89.9%, which is well-calibrated for these tasks.

6 Concluding Remarks

Bayesian predictive synthesis provides a theoretically and conceptually sound framework to synthesize density prediction. Utilizing this framework, we develop a spatially varying synthesis method for the context of spatial data. With this new method, we can dynamically calibrate, learn, and update coefficients as the data changes across a spatial region. The simulations and real world applications demonstrate the efficacy of BSPS compared to conventional spatial models and modern machine learning techniques. Specifically, by dynamically synthesizing the predictive distribution from the models, BSPS can improve point and distributional predictions. Additionally, posterior inference on the full spatial region gives the decision maker information on how each model is related, and how their relationship changes across a region. In addition to the applications in this paper, our proposed framework can be applied to other fields, including, but not limited to, weather, GPS systems, and sports player tracking data. Further studies exploring different uses of BSPS, as well as specific developments catered towards a specific dataset, is of interest.

Regarding scalable computation algorithms for BSPS, it may be possible to use other types of scalable Gaussian processes, such as predictive process (Banerjee et al., 2008), meshed Gaussian process (Peruzzi et al., 2020), and fused Gaussian process (Ma and Kang, 2020). We leave the potential use of these techniques as future work. Apart from MCMC-based algorithms, the integrated nested Laplace approximation (Rue et al., 2009) may be an appealing strategy for fast computation. However, the latent factor spatially varying coefficient model (5) has $2(J+1)$ hyperparameters, which limits the use of the integrated nested Laplace approximation when J is not small (e.g. $J \geq 3$).

Finally, there are several ways to extend or apply the current BSPS approach. The first is to extend BSPS to spatio-temporal or multivariate data. This can potentially be done by using the recently developed techniques of graphical Gaussian processes (Dey et al., 2021; Peruzzi and Dunson, 2022). Moreover, BSPS can be used, not only for synthesizing multiple models, but also for saving computational cost under a large number of covariates. Specifically, if the number of covariates, say p, is very large, the standard spatially varying coefficient model (Gelfand et al., 2003) requires $p+1$ Gaussian processes for modeling spatially varying coefficients, which is computationally burdensome. On the other hand, it would be possible to first apply multiple, say J, regression models to get univariate spatial predictors, and then combine the multiple prediction models via BSPS. This reduces the number of Gaussian processes from $p+1$ to $J+1$, which makes the computation much less burdensome.

Acknowledgement

This work is partially supported by Japan Society for Promotion of Science (KAKENHI) grant numbers 21H00699.

References

Aastveit, K. A., F. Ravazzolo, and H. K. Van Dijk (2018). Combined density nowcasting in an uncertain economic environment. *Journal of Business & Economic Statistics 36*(1), 131–145.

Anselin, L. (1988). *Spatial econometrics: methods and models*, Volume 4. Springer Science & Business Media.

Banerjee, S., A. E. Gelfand, A. O. Finley, and H. Sang (2008). Gaussian predictive process models for large spatial data sets. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 70*(4), 825–848.

Bishop, C. M. et al. (1995). *Neural networks for pattern recognition*. Oxford university press.

Blei, D. M., A. Kucukelbir, and J. D. McAuliffe (2017). Variational inference: A review for statisticians. *Journal of the American statistical Association 112*(518), 859–877.

Breiman, L. (2001). Random forests. *Machine learning 45*(1), 5–32.

Brunsdon, C., S. Fotheringham, and M. Charlton (1998). Geographically weighted regression. *Journal of the Royal Statistical Society: Series D (The Statistician) 47*(3), 431–443.

Chen, T. and C. Guestrin (2016). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pp. 785–794.

Datta, A., S. Banerjee, A. O. Finley, and A. E. Gelfand (2016). Hierarchical nearest-neighbor gaussian process models for large geostatistical datasets. *Journal of the American Statistical Association 111*(514), 800–812.

Debarsy, N. and J. P. LeSage (2020). Bayesian model averaging for spatial autoregressive models based on convex combinations of different types of connectivity matrices. *Journal of Business & Economic Statistics*, 1–12.

Dey, D., A. Datta, and S. Banerjee (2021). Graphical gaussian process models for highly multivariate spatial data. *Biometrika (available online)*.

Diggle, P. J., J. A. Tawn, and R. A. Moyeed (1998). Model-based geostatistics. *Journal of the Royal Statistical Society: Series C (Applied Statistics) 47*(3), 299–350.

Finley, A. O., A. Datta, and S. Banerjee (2022). spNNGP R package for nearest neighbor Gaussian process models. *Journal of Statistical Software 103*(5), 1–40.

Friedman, J. H. (1991). Multivariate Adaptive Regression Splines. *The Annals of Statistics 19*(1), 1 – 67.

Gelfand, A. E., H.-J. Kim, C. Sirmans, and S. Banerjee (2003). Spatial modeling with spatially varying coefficient processes. *Journal of the American Statistical Association 98*(462), 387–396.

Genest, C. and M. J. Schervish (1985). Modelling expert judgements for Bayesian updating. *Annals of Statistics 13*, 1198–1212.

Genre, V., G. Kenny, A. Meyler, and A. Timmermann (2013). Combining expert forecasts: Can anything beat the simple average? *International Journal of Forecasting 29*, 108–121.

Geweke, J. and G. Amisano (2011). Optimal prediction pools. *Journal of Econometrics 164*(1), 130–141.

Greenaway-McGrevy, R. and K. Sorensen (2021). A spatial model averaging approach to measuring house prices. *Journal of Spatial Econometrics 2*(1), 1–32.

Hastie, T. and R. Tibshirani (1987). Generalized additive models: some applications. *Journal of the American Statistical Association 82*(398), 371–386.

Hoeting, J. A., D. Madigan, A. E. Raftery, and C. T. Volinsky (1999). Bayesian model averaging: a tutorial. *Statistical science*, 382–401.

Kim, M. and L. Wang (2021). Generalized spatially varying coefficient models. *Journal of Computational and Graphical Statistics 30*(1), 1–10.

Lany, N. K., P. L. Zarnetske, A. O. Finley, and D. G. McCullough (2020). Complementary strengths of spatially-explicit and multi-species distribution models. *Ecography 43*(3), 456–466.

LeSage, J. P. and O. Parent (2007). Bayesian model averaging for spatial econometric models. *Geographical Analysis 39*(3), 241–267.

Liao, J., G. Zou, and Y. Gao (2019). Spatial mallows model averaging for geostatistical models. *Canadian Journal of Statistics 47*(3), 336–351.

Ma, P. and E. L. Kang (2020). A fused gaussian process model for very large spatial data. *Journal of Computational and Graphical Statistics 29*(3), 479–489.

McAlinn, K., K. A. Aastveit, J. Nakajima, and M. West (2020). Multivariate bayesian predictive synthesis in macroeconomic forecasting. *Journal of the American Statistical Association 115*(531), 1092–1110.

McAlinn, K. and M. West (2019). Dynamic bayesian predictive synthesis in time series forecasting. *Journal of econometrics 210*(1), 155–169.

Peruzzi, M., S. Banerjee, and A. O. Finley (2020). Highly scalable bayesian geostatistical modeling via meshed gaussian processes on partitioned domains. *Journal of the American Statistical Association*, 1–14.

Peruzzi, M. and D. B. Dunson (2022). Spatial meshing for general bayesian multivariate models. *arXiv preprint arXiv:2201.10080*.

Polson, N. G., J. G. Scott, and J. S. Windle (2013). Bayesian inference for logistic models using polya-gamma latent variables. *Journal of the American Statistical Association 108*, 1339–1349.

Rue, H., S. Martino, and N. Chopin (2009). Approximate bayesian inference for latent gaussian models by using integrated nested laplace approximations. *Journal of the royal statistical society: Series B (statistical methodology) 71*(2), 319–392.

Saha, A., S. Basu, and A. Datta (2021). Random forests for spatially dependent data. *Journal of the American Statistical Association*, 1–46.

Sekulić, A., M. Kilibarda, G. Heuvelink, M. Nikolić, and B. Bajat (2020). Random forest spatial interpolation. *Remote Sensing 12*(10), 1687.

Wang, F. and M. M. Wall (2003). Generalized common spatial factor model. *Biostatistics 4*(4), 569–582.

West, M. (1992). Modelling agent forecast distributions. *Journal of the Royal Statistical Society (Series B: Methodological) 54*, 553–567.

West, M. and J. Crosse (1992). Modelling of probabilistic agent opinion. *Journal of the Royal Statistical Society (Series B: Methodological) 54*, 285–299.

Zhang, X. and J. Yu (2018). Spatial weights matrix selection and model averaging for spatial autoregressive models. *Journal of Econometrics 203*(1), 1–18.

Supplementary Material

This Supplementary Material provides proofs of Theorem 1 and Theorem 2 in Section S1 and Section S2, respectively, details of sampling algorithm under nearest neighbor Gaussian processes in Section S3 and derivation of the mean field variational Bayes algorithm (Algorithm 1 in the main document) in Section S4.

S1 Proof of Theorem 1 and feasible forms of best approximation

S1.1 Proof of Theorem 1

First, $y(s), \{f_j(s)\}_{j=1,\cdots,J}$, which are Gaussian processes, can be expanded by the system of eigenvalues and eigenfunctions $\{\lambda_k, e_k(\cdot)\}_{k\in\mathbb{N}}$ in the following way (details of $\{\lambda_k, e_k(\cdot)\}_{k\in\mathbb{N}}$ are described in Section S1.3):

$$\begin{aligned} y(\cdot) &= m_y(\cdot) + \sum_{k=1}^{\infty} \langle y - m_y, e_k\rangle e_k(\cdot), \quad \frac{\langle y - m_y, e_k\rangle}{\sqrt{\lambda_k}} \sim \mathcal{N}(0,1), \\ f_j(\cdot) &= m_j(\cdot) + \sum_{k=1}^{\infty} \langle f_j - m_j, e_k\rangle e_k(\cdot), \frac{\langle f_j - m_j, e_k\rangle}{\sqrt{\lambda_k}} \sim \mathcal{N}(0,1), \\ \langle e_k, e_l\rangle &= \begin{cases} \sqrt{\lambda_k}, & (k=l), \\ 0, & (k\neq l), \end{cases} \end{aligned} \tag{S1}$$

where $m_y(\cdot), m_j(\cdot)$ are the mean functions of $y(\cdot), f_j(\cdot)$ and $\langle\cdot,\cdot\rangle$ represents the inner product on $L^2(\mathcal{S})$. Note that given the eigenvector (eigenfunction), $e_k(\cdot)$, if we take the definition function, $1_{B(s_k)}(\cdot)$, at the neighbor, $B(s_k)$, of the observation point, s_k, $\langle f_j, e_k\rangle$ can be seen as the variable value at point, s_k, on the Gaussian process, f_j.

From (S1), the projection of the ($\mathbb{R}$-valued) Gaussian random variable, $\langle y - m_y, e_k\rangle$, onto $\bigcup_j \text{span}\{\langle f_j - m_j, e_k\rangle; k\in\mathbb{N}\}$ can be expressed using $\sum_{j=1}^{J}\sum_{l=1}^{\infty} c_{jk}^2 < \infty$, with the double subscript sequence, $\{c_{jk}\}_{1\leqq j\leqq J,\ k\in\mathbb{N}}$, as

$$\mathrm{E}\left[\langle y - m_y, e_k\rangle \middle| \bigcup_j \text{span}\{\langle f_j - m_j, e_k\rangle; k\in\mathbb{N}\}\right] = \sum_{j=1}^{J}\sum_{k=1}^{\infty} c_{jk}\langle f_j - m_j, e_k\rangle. \tag{S2}$$

Here, the sequence, $\{c_{jk}\}_{1\leqq j\leqq J,\ k\in\mathbb{N}}$, is not unique. This is because all the elements in $\{\langle f_j - m_j, e_k\rangle; k\in\mathbb{N}, 1\leqq j\leqq J\}$ are not necessarily independent, making is redundant as the base of the Gaussian space, $\mathfrak{G} = \bigcup_{j=1}^{J}\text{span}\{\langle f_j - m_j, e_k\rangle; k\in\mathbb{N}\}$.

The statement of 1 consider the best approximate model of the Gaussian process, $y(\cdot)$, given the realization of the agent of Gaussian processes for each point, $\{f_j(s_i)\}_{1\leqq i\leqq n+1}$ (instead of using the Gaussian process, $\{f_j(\cdot)\}_{j=1,\cdots,J}$ as the agents' predictive value). Since the value, $\{f_j(s_i)\}_{1\leqq i\leqq n+1}$, can be seen as $\{\langle f_j, e_i\rangle; 1\leqq i\leqq n+1\}, \left(e_i = 1_{B(s_i)}\right)$, the conditional expectation, $\mathrm{E}[y(\cdot)|\,\sigma\{\langle f_j, e_i\rangle; 1\leqq i\leqq n+1\}]$, conditional on the σ-field, $\sigma\{\langle f_j, e_i\rangle; 1\leqq i\leqq n+1\}$, generated from $\{\langle f_j, e_i\rangle; 1\leqq i\leqq n+1\}$, can be expressed (from Bogachev, 1998, Corollary 3.5.2) as

$$\begin{aligned} &\mathrm{E}[Y(\cdot)|\,\sigma\{\langle f_j, e_i\rangle; 1\leqq i\leqq n+1\}] \\ &\quad = m_y(\cdot) + \sum_{j=1}^{J}\sum_{i=1}^{n+1}\langle f_j - m_j, e_i\rangle e_i(\cdot) + \sum_{j=1}^{J}\underbrace{\mathrm{E}\left[\sum_{i=n+1}^{\infty}\langle f_j - m_j, e_i\rangle e_i(\cdot)\right]}_{=0}. \end{aligned} \tag{S3}$$

Therefore, $\mathrm{E}\left[\langle Y, e_{n+1}\rangle | \, \sigma\left\{\langle f_j, e_i\rangle \, ; 1 \leqq i \leqq n+1\right\}\right]$ is

$$\begin{aligned}
&\langle m_y, e_{n+1}\rangle + \sum_{j=1}^{J}\sum_{i=1}^{n+1} \langle f_j - m_j, e_i\rangle \langle e_i, e_{n+1}\rangle \\
&\qquad = \langle m_y, e_{n+1}\rangle + \sum_{j=1}^{J} \langle f_j - m_j, e_{n+1}\rangle \langle e_{n+1}, e_{n+1}\rangle ,
\end{aligned}$$

wher $\langle m_y, e_{n+1}\rangle$ is the average value of Y at location, s_{n+1}, and since s_{n+1} moves with e_{n+1}, $\langle m_y, e_{n+1}\rangle$ is a function of s. Similarly for $\langle f_j - m_j, e_{n+1}\rangle \langle e_{n+1}, e_{n+1}\rangle$, $\langle f_j - m_j, e_{n+1}\rangle$ is the value of $f_j - m_j$ at s_{n+1}, and because $\langle e_{n+1}, e_{n+1}\rangle$ is the value of some function at s_{n+1}, they are both functions of s. If we let

$$\beta_0(s) = \langle m_y, e(s)\rangle - \sum_{j=1}^{J} \langle m_j, e(s)\rangle , \qquad \beta_j(s) = \langle e(s), e(s)\rangle , \qquad f_j(s) = \langle f_j, e(s)\rangle ,$$

then, from (S3), the random variable, $y(s)$ at s on the Gaussian process, $y(\cdot)$, can be expressed using a spatially-varying coefficient model:

$$y(s) = \beta_0(s) + \sum_{j=1}^{J} \beta_j(s) f_j(s) + \varepsilon(s), \ \mathrm{E}\left[f(s)\varepsilon(s)\right] = 0, \tag{S4}$$

for $s \in \mathcal{S}$.

S1.2 On feasible form of best approximation

From (S2), the projection, $P_{\mathfrak{G}} y$, of the Gaussian process, $y(\cdot)$, to the space $\mathfrak{G}$ can be expressed as

$$P_{\mathfrak{G}} y(\cdot) = \sum_{j=1}^{J} \int_S K_j(\cdot, t) \, f_j(t) \, dt \tag{S5}$$

using the integral kernel, $K(\cdot, \cdot)$, which is not the same as the kernel function of the reproducing kernel Hilbert space (RKHS), κ_R. The best approximate model must be represented by this integral kernel; $K(\cdot, \cdot)$.

The estimation of the integral kernel, $K(\cdot, \cdot)$, is difficult. Therefore, we instead consider the best approximate model of the Gaussian process, $y(\cdot)$, given the realization of the agent Gaussian process for each point, $\{f_j(s_i)\}_{1 \leqq i \leqq n+1}$ (instead of using the Gaussian process, $\{f_j(\cdot)\}_{j=1,\cdots,J}$ as the agents' predictive value). The expression (S4) is the best approximate model of the random variable at s on $y(\cdot)$, given the realization of $\{f_j(s)\}_{j=1,\cdots,J}$ and not the best approximate model of the Gaussian process itself. For our proposed BSPS, realizations of the Gaussian process at each point is provided to the decision maker to construct the predictive value, thus making the above the implementable best approximate model. This best approximate model, in turn, is the best approximate synthesis function.

S1.3 Eigenvalues and eigenfunctions $\{\lambda_k, e_k(\cdot)\}_{k \in \mathbb{N}}$

Consider the task of predicting the data generating process (DGP), $\{y(s); s \in \mathcal{S}\}$, using the agents' predictive value, $\{f_j(s)\}_{j=1,\cdots,J}$. We assume that $\{y(s)\}$ and $\{f_j(s)\}_{j=1,\cdots,J}$ all follow Gaussian processes. Let H be a closed convex subspace of $C(\mathcal{S})$, a continuous real-valued functional space on $\mathcal{S}$. H is a separable Hilbert space, where $\langle \cdot, \cdot \rangle$ denotes

the inner product in H. Let $y(\cdot)$ denote the random variable taking value on H and μ_y the probability measure of y. Further, let $W(\cdot)$ be a H-valued Gaussian random variable with mean vector 0, taking value on H. Similarly, for the agents' predictive value, $\{f_j(\cdot)\}_{j=1,\cdots,J}$, let it be a random variable taking value in H, with probability measure, $\{\mu_{f_j}\}_{j=1,\cdots,J}$. Denote $\mu_{Y,f}$ as the joint probability measure of $\left(y(\cdot), \{f_j(\cdot)\}_{j=1,\cdots,J}\right)$. $\mu_{Y,f}$ is a probability measure on $H^{\otimes(J+1)}$, where the marginals for each random variable are $\mu_y, \{\mu_{f_j}\}_{j=1,\cdots,J}$. Note that $\mu_{Y,f}, \mu_y, \{\mu_{f_j}\}_{j=1,\cdots,J}$ are all Gaussian measures, thus the marginals, $\mu_y, \{\mu_{f_j}\}_{j=1,\cdots,J}$, and tensors, $\mu_y \otimes \mu_{f_j}$, are all Gaussian measures.

Denote m as the mean vector and R as the covariance operator of the Gaussian measure, $\mu_{Y,f}$, on $H^{\otimes(J+1)}$. If we set the positive definite kernel function, $\kappa_R : H^{\otimes(J+1)} \times H^{\otimes(J+1)} \to \mathbb{R}$, as

$$\kappa_R(x,y) = \langle Rx, y\rangle, \ x, y \in H^{\otimes(J+1)},$$

we can construct the RKHS, $\mathcal{H}_\kappa^{\otimes(J+1)}$, that corresponds to κ_R. Then, there exists a complete orthonormal system, $\{e_k; k \in \mathbb{N}\}$, for $\mathcal{H}_\kappa^{\otimes(J+1)}$, with corresponding eigenvalue, λ_k, and eigenvector, e_k, such that $Re_k = \lambda_k e_k$. From this, we have the expansion representation of a general Gaussian random variable, $X(\cdot)$, on $H^{\otimes(J+1)}$, as

$$X(\cdot) = m(\cdot) + \sum_{k=1}^{\infty} \sqrt{\lambda_k} \frac{\langle X - m, e_k\rangle}{\sqrt{\lambda_k}} e_k(\cdot), \text{ a.s. } \mu.$$

Here, $\{\lambda_k^{-1/2}\langle X, e_k\rangle; \ k \in \mathbb{N}\}$ are independent (for each k) standard Gaussian random variable sequences, namely, $\lambda_k^{-1/2}\langle X, e_k\rangle \sim \mathcal{N}(0,1)$. If we series expand the DGP, $y(\cdot)$, we have

$$y(\cdot) = m_y(\cdot) + \sum_{k=1}^{\infty} \langle Y - m_y, e_k\rangle e_k(\cdot),$$

which means that it is an element of a space of independent countably infinite Gaussian vectors, span $\{\langle y, e_k\rangle; k \in \mathbb{N}\}$. The agents' predictive value, $\{f_j(s)\}_{j=1,\cdots,J}$, is

$$f_j(\cdot) = m_j(\cdot) + \sum_{k=1}^{\infty} \langle f_j - m_j, e_k\rangle e_k(\cdot),$$

and the Gaussian space generated from the agents' predictive value is

$$\mathfrak{G} = \bigcup_{j=1}^{J} \text{span}\{\langle f_j - m_j, e_k\rangle; k \in \mathbb{N}\}.$$

S2 Proof of Theorem 2

To show exact minimaxity, we first define the transformation group that makes the statistical decision problem invariant under the KL risk. Here, the orthogonal group, $\mathcal{O}_{n+1}$, is the group of $(n+1) \times (n+1)$ orthogonal matrices, with $\mathbb{R}_+$ representing the positive region, $(0, \infty]$. The group, G, is

$$G = \mathbb{R}_+ \times \mathcal{O}_{n+1} \times \mathbb{R}^{n+1}, \quad c \in \mathbb{R}_+, \quad A \in \mathcal{O}_{n+1}, \quad F \in \mathbb{R}^{n+1},$$

where the operation g to the sample space of $(\boldsymbol{y}, y(s_{n+1}))$ is defined as

$$g\begin{bmatrix} \boldsymbol{y} \\ y(s_{n+1}) \end{bmatrix} = cA\begin{bmatrix} \boldsymbol{y} \\ y(s_{n+1}) \end{bmatrix} + F,$$

and the operation $\overline{g}$ to the parameter space, $\theta = \left(\boldsymbol{\mu}, \mu(s_{n+1}), C_w, \sigma^2\right)$, is defined as

$$\overline{g}\boldsymbol{\mu}^{n+1} = cA\boldsymbol{\mu}^{n+1} + F, \quad \overline{g}C^{n+1} = cAC^{n+1}A^\top, \quad \overline{g}\sigma^2 = c\sigma^2 A^\top A,$$

where $\boldsymbol{\mu}^{n+1} = (\boldsymbol{\mu}, \mu(s_{n+1}))$ and C^{n+1} is a $(n+1) \times (n+1)$ covariance matrix. The transformation, $\tilde{g}$, to the probability density, $q(y)$, is defined as $\tilde{g}q(y) = q(gy)$. The transformation group, $(g, \overline{g}, \tilde{g})$, operates transitively on the sample, $(\boldsymbol{y}, y(s_{n+1}))$, and the parameter space.

The statistical decision problem is invariant under the transformation group, $(g, \overline{g}, \tilde{g})$. Thus, for the sample, $y(s_{n+1})$, that follows a probability distribution, p^*_θ, it holds that $p^*_{\overline{g}\theta} = \tilde{g}p^*_\theta$, which entails that the loss is invariant under the KL loss,

$$\begin{aligned}\mathsf{KL}\left(p^*_{\overline{g}\theta} \,|\tilde{g}q\right) &= \int \log \frac{p^*_{\overline{g}\theta}\left(y(s_{n+1})\right)}{\tilde{g}q\left(y(s_{n+1}) \,|\boldsymbol{y}, \boldsymbol{f}(s_{n+1})\right)} p^*_{\overline{g}\theta}\left(y(s_{n+1})\right) dy(s_{n+1}) \\ &= \int \log \frac{p^*_\theta\left(g(s_{n+1})\right)}{q\left(g(s_{n+1}) \,|\boldsymbol{y}, \boldsymbol{f}(s_{n+1})\right)} p^*_\theta\left(g(s_{n+1})\right) dy(s_{n+1}) = \mathsf{KL}\left(p^*_\theta \,|q\right).\end{aligned}$$

The group, $(g, \overline{g}, \tilde{g})$, is an amenable group, which satisfies the Hunt-Stein condition (Bondar and Milnes, 1981). The conditions for minimaxity in Kiefer (1957) is thus satisfied. Therefore, the minimax solution for the given statistical decision problem exists in the solution of the invariant statistical decision problem:

$$\min_q \max_\theta \mathrm{E}_y\left[\mathsf{KL}\left(p^*_\theta \,|q\right)\right] = \min_{q:g\text{-invariant}} \max_\theta \mathrm{E}_y\left[\mathsf{KL}\left(p^*_\theta \,|q\right)\right],$$

where the minimum is taken over a class of g-invariant distributions.

From this argument, the best equivariant predictive distribution from the class of g-invariant distributions is the minimax solution out of all probability distributions. The best equivariant predictive distribution is $\overline{g}$-invariant, i.e., the Bayes decision based on the prior, ρ, that satisfies $\rho\left(\overline{g}\boldsymbol{\beta}^{n+1}\right) = \rho\left(\boldsymbol{\beta}^{n+1}\right)$, $\rho(\overline{g}\sigma) = \rho(\sigma)$ provides the best equivariant solution (Zidek, 1969). Under KL risk, the Bayesian predictive distribution under a $\overline{g}$-invariant prior is the best equivariant solution (Komaki, 2002). For the BSPS model with Gaussian processes for $\beta_1(s), \ldots, \beta_J(s)$, if we use the following $\overline{g}$-invariant prior distributions

$$\rho(\boldsymbol{\beta}_0) = 1_{[-b,a]^n}(\boldsymbol{\beta}_0), \quad \rho(\sigma) = \frac{1}{\sigma}, \quad \rho(\tau_j) = \frac{1}{\tau_j}, \quad j = 1, \cdots, J, \tag{S6}$$

for some positive constants, a and b. Here, $1_{[-b,a]^n}(\boldsymbol{\beta}_0)$ is an indicator function, where it is 1 when $\boldsymbol{\beta}_0$ is in the region, $[-b,a]^n$, and 0 otherwise. Therefore, the predictive distribution under the prior (S6), and all predictive distributions that dominate it, is a minimax solution.

S3 Sampling steps under nearest neighbor Gaussian process

The use of the m-nearest neighbor Gaussian process for $\beta_j(s)$ leads to a multivariate normal distribution with a sparse precision matrix for $\beta_j(s_1), \ldots, \beta_j(s_n)$, defined as

$$\pi(\beta_j(s_1), \ldots, \beta_j(s_n)) = \prod_{i=1}^n \phi(\beta_j(s_i); \boldsymbol{B}_j(s_i)\beta_j(N(s_i)), \tau_j \boldsymbol{F}_j(s_i)), \quad j = 0, \ldots, J$$

where

$$\begin{aligned}\boldsymbol{B}_j(s_i) &= C_j(s_i, N(s_i))C_j(N(s_i), N(s_i))^{-1}, \\ \boldsymbol{F}_j(s_i) &= C_j(s_i, s_i) - C_j(s_i, N(s_i))C_j(N(s_i), N(s_i))^{-1}C_j(N(s_i), s_i),\end{aligned}$$

and $N(s_i)$ denotes an index set of m-nearest neighbors of s_i. Here $C_j(\cdot,\cdot)$ is the same correlation function used in the original Gaussian process for $\beta_j(s)$.

The full conditional distributions of the latent factors, f_{ji}, and error variance, σ^2, are the same as the ones given in the main document. The full conditional distributions of the other parameters are given as follows:

- (Sampling of spatially varying weights) For $i = 1, \ldots, n$, the full conditional distribution of $(\beta_0(s_i), \ldots, \beta_J(s_i))$ is given by $N(A_i B_i, A_i)$, where

$$A_i = \left\{ \frac{f_i f_i^\top}{\sigma^2} + \mathrm{diag}(\gamma_{0i}, \ldots, \gamma_{Ji}) \right\}^{-1}, \quad \gamma_{ji} = \frac{1}{\tau_j F_j(s_i)} + \sum_{t; s_i \in N(t)} \frac{B_j(t; s_i)^2}{\tau_j F_j(t)},$$

$$B_i = \frac{f_i y_i}{\sigma^2} + (m_{0i}, \ldots, m_{Ji})^\top,$$

$$m_{ji} = \frac{B_j(s_i)^\top \beta_j(N(s_i))}{\tau_j F_j(s_i)} + \sum_{t; s_i \in N(t)} \frac{B(t; s_i)}{\tau_j F_j(t)} \Big\{ \beta_j(t) - \sum_{s \in N(t), s \neq s_i} B(t; s) \beta_j(s) \Big\}$$

  where $f_i = (1, f_{1i}, \ldots, f_{Ji})$ and $B_j(t; s)$ denotes the scalar coefficient for $\beta_j(s_i)$ among the element of the coefficient vector $B_j(t)$.

- (Sampling of τ_j) For $j = 0, \ldots, J$, the full conditional distribution of τ_j is

$$\mathrm{IG}\left(a_\tau + \frac{n}{2}, b_\tau + \frac{1}{2} \sum_{i=1}^{n} \frac{\{\beta_j(s_i) - B_j(s_i)\beta_j(N(s_i))\}^2}{F_j(s_i)} \right).$$

- (Sampling of g_j) For $j = 0, \ldots, J$, the full conditional distribution of g_j is proportional to

$$\prod_{i=1}^{n} \phi(\beta_j(s_i); B_j(s_i; \theta_j)\beta_j(N(s_i)), \tau_j F_j(s_i; g_j)), \quad g_j \in (\underline{g}, \overline{g}),$$

  where

$$B_j(s_i; g_j) = C_j(s_i, N(s_i); g_j) C_j(N(s_i), N(s_i); g_j)^{-1},$$
$$F_j(s_i; g_j) = C_j(s_i, s_i; g_j) - C_j(s_i, N(s_i); g_j) C_j(N(s_i), N(s_i); g_j)^{-1} C_j(N(s_i), s_i; g_j),$$

  and $C_j(\cdot, \cdot; g_j)$ is the correlation function with spatial range g_j.

S4 Derivation of variational Bayes algorithm

Remember that the mean filed variational Bayes (MFVB) approximates the posterior distributions through the form

$$q(\{f_{ji}\}, \{\beta_j\}, \{\tau_j\}, \{h_j\}, \sigma^2) = q(\sigma^2) \prod_{j=0}^{J} q(\beta_j) q(\tau_j) q(h_j) \prod_{i=1}^{n} q(f_{ji}).$$

It is known that the optimal form of the variational posterior is given by, for example, $q(\beta_j) \propto \exp(E_{-\beta_j}[\log p(y, \Theta)])$, where $\Theta = (\{f_{ji}\}, \{\beta_j\}, \{\tau_j\}, \{h_j\}, \sigma^2)$ and $E_{-\beta_j}$ denotes the expectation with respect to the marginal variational posterior of the parameters other

than β_j. From the forms of full conditional posterior distributions given in the main document, we can use the following distributions as optimal distributions:

$$q(\sigma^2) \sim \mathrm{IG}(\widetilde{a}_\sigma, \widetilde{b}_\sigma), \quad q(\tau_j) \sim \mathrm{IG}(\widetilde{a}_{\tau_j}, \widetilde{b}_{\tau_j}), \quad q(\beta_j) \sim N(\widetilde{\mu}_j, \widetilde{\Sigma}_j),$$
$$q(h_j) \sim \mathrm{D}(\widetilde{p}_{j1}, \ldots, \widetilde{p}_{jL}), \quad q(f_{ji}) \sim N(\widetilde{m}_{ji}, \widetilde{s}^2_{ji}),$$

where $\mathrm{D}(\widetilde{p}_{j1}, \ldots, \widetilde{p}_{jL})$ is a discrete distribution on $\{\eta_1, \ldots, \eta_L\}$ such that $P(h_j = \eta_\ell) = \widetilde{p}_{j\ell}$. The derivation of the updating steps of MFVB is given as follows.

- **(update of f_{ji})** It follows that

$$E_{-f_{ji}}[\log p(y, \Theta)] = (\text{const.}) - \frac{1}{2} f_{ij}^2 (A^f_{ji})^{-1} + f_{ji} B^f_{ji},$$

where

$$A^f_{ji} = \left(\frac{1}{b_{ji}} + E_q[\beta^2_{ji}] E_q\left[\frac{1}{\sigma^2}\right] \right)^{-1} = \left\{ \frac{1}{b_{ji}} + (\widetilde{\mu}^2_{ji} + \widetilde{\Sigma}_{jii}) \frac{\widetilde{a}_\sigma}{\widetilde{b}_\sigma} \right\}^{-1},$$

and

$$\begin{aligned} B^f_{ji} &= \frac{a_{ji}}{b_{ji}} + E_q[\beta_{ji}] E_q\left[\frac{1}{\sigma^2}\right] \left(y_i - E_q[\beta_{0i}] - \sum_{k \neq j} E_q[\beta_{ki}] E_q[f_{ki}] \right) \\ &= \frac{a_{ji}}{b_{ji}} + \widetilde{\mu}_{ji} \frac{\widetilde{a}_\sigma}{\widetilde{b}_\sigma} \left(y_i - \widetilde{\mu}_{0i} - \sum_{k \neq j} \widetilde{\mu}_{ki} \widetilde{m}_{ki} \right). \end{aligned}$$

Then, the parameters in the variational posterior of f_{ji} can be updated as $\widetilde{m}_{ji} = A^f_{ji} B^f_{ji}$ and $\widetilde{s}^2_{ji} = A^f_{ji}$.

- **(update of β_j)** It follows that

$$E_{-\beta_j}[\log p(y, \Theta)] = (\text{const.}) - \frac{1}{2} \beta^\top (A^{(\beta)}_j)^{-1} \beta + \beta^\top B^{(\beta)}_j,$$

where

$$\begin{aligned} A^{(\beta)}_j &= \left\{ E_q[\Omega_j] E_q\left[\frac{1}{\sigma^2}\right] + E_q[H(h_j)^{-1}] E_q\left[\frac{1}{\tau_j}\right] \right\}^{-1} = \left\{ \Omega^*_j \frac{\widetilde{a}_\sigma}{\widetilde{b}_\sigma} + \sum_{\ell=1}^{L} \widetilde{p}_{j\ell} H(\eta_\ell)^{-1} \frac{\widetilde{a}_{\tau_j}}{\widetilde{b}_{\tau_j}} \right\}^{-1} \\ B^{(\beta)}_j &= E_q\left[\frac{1}{\sigma^2}\right] E[F_j] \circ \left(y - E_q[\beta_0] - \sum_{k \neq j} E_q[\beta_k] \circ E_q[F_k] \right) \\ &= \frac{\widetilde{a}_\sigma}{\widetilde{b}_\sigma} \widetilde{m}_j \circ \left(y - \widetilde{\mu}_0 - \sum_{k \neq j} \widetilde{\mu}_k \circ \widetilde{m}_k \right), \end{aligned}$$

where $\Omega^* = \mathrm{diag}(\widetilde{m}^2_{j1} + \widetilde{s}^2_{j1}, \ldots, \widetilde{m}^2_{jn} + \widetilde{s}^2_{jn})$. Then, the parameters in the variational posterior of β_j can be updated as $\widetilde{\mu}_j = A^{(\beta)}_j B^{(\beta)}_j$ and $\widetilde{\Sigma}_j = A^{(\beta)}_j$.

- **(update of τ_j)** It follows that

$$E_{-\tau_j}[\log p(y, \Theta)] = (\text{const.}) - \left(\frac{n}{2} + a_\tau + 1 \right) \log \tau_j - \frac{1}{\tau_j} \left(b_\tau + \frac{1}{2} E_q\left[\beta_j^\top H(h_j)^{-1} \beta_j \right] \right),$$

noting that

$$E_q\left[\beta_j^\top H(h_j)^{-1}\beta_j\right] = \mathrm{tr}\left\{E_q[\beta_j\beta_j^\top]E_q[H(h_j)^{-1}]\right\} = \mathrm{tr}\left\{(\widetilde{\mu}_j\widetilde{\mu}_j^\top + \widetilde{\Sigma}_j)\sum_{\ell=1}^{L}\widetilde{p}_{j\ell}H(\eta_\ell)^{-1}\right\}.$$

Then, the parameters in the variational posterior of τ_j can be updated as

$$\widetilde{a}_{\tau_j} = a_\tau + \frac{n}{2}, \quad \widetilde{b}_{\tau_j} = b_\tau + \frac{1}{2}\mathrm{tr}\left\{(\widetilde{\mu}_j\widetilde{\mu}_j^\top + \widetilde{\Sigma}_j)\sum_{\ell=1}^{L}\widetilde{p}_{j\ell}H(\eta_\ell)^{-1}\right\}$$

- **(update of h_j)** It follows that

$$E_{-h_j}[\log p(y,\Theta)] = (\text{const.}) - \frac{1}{2}\log|H(h_j)| - \frac{1}{2}E_q\left[\frac{1}{\tau_j}\right]E_q[\beta_j^\top H(h_j)^{-1}\beta_j],$$

where $E_q[\beta_j^\top H(h_j)^{-1}\beta_j] = \mathrm{tr}\{(\widetilde{\mu}_j\widetilde{\mu}_j^\top + \widetilde{\Sigma}_j)H(h_j)^{-1}\}$. Then, the parameters in the variational posterior of h_j can be updated as

$$\widetilde{p}_{j\ell} = \frac{|H(\eta_\ell)|^{-1/2}\exp\left(-\widetilde{a}_{\tau_j}\mathrm{tr}\{(\widetilde{\mu}_j\widetilde{\mu}_j^\top + \widetilde{\Sigma}_j)H(\eta_\ell)^{-1}\}/2\widetilde{b}_{\tau_j}\right)}{\sum_{\ell'=1}^{L}|H(\eta_{\ell'})|^{-1/2}\exp\left(-\widetilde{a}_{\tau_j}\mathrm{tr}\{(\widetilde{\mu}_j\widetilde{\mu}_j^\top + \widetilde{\Sigma}_j)H(\eta_{\ell'})^{-1}\}/2\widetilde{b}_{\tau_j}\right)}$$

- **(update of σ^2)** It follows that

$$E_{-\sigma^2}[\log p(y,\Theta)] = (\text{const.}) - \frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}E_q\left[\left(y_i - \beta_{0i} - \sum_{j=1}^{J}\beta_{ji}f_{ji}\right)^2\right],$$

where

$$\begin{aligned} I_q(\sigma^2) &\equiv \sum_{i=1}^{n}E_q\left[\left(y_i - \beta_{0i} - \sum_{j=1}^{J}\beta_{ji}f_{ji}\right)^2\right] \\ &= \left(y - \widetilde{\mu}_0 - \sum_{j=1}^{J}\widetilde{\mu}_j\circ\widetilde{m}_j\right)^\top\left(y - \widetilde{\mu}_0 - \sum_{j=1}^{J}\widetilde{\mu}_j\circ\widetilde{m}_j\right) + \mathrm{tr}(\widetilde{\Sigma}_0) \\ &\quad + \sum_{j=1}^{J}\mathrm{tr}\left\{(\widetilde{\mu}_j\widetilde{\mu}_j^\top + \widetilde{\Sigma}_j)\circ(\widetilde{m}_j\widetilde{m}_j^\top + \widetilde{S}_j)\right\} - \sum_{i=1}^{n}\sum_{j=1}^{J}\widetilde{m}_{ji}^2\widetilde{\mu}_{ji}^2 \end{aligned}$$

with $\widetilde{S}_j = \mathrm{diag}(\widetilde{s}_{j1}^2,\ldots,\widetilde{s}_{jn}^2)$. Then, the parameters in the variational posterior of σ^2 can be updated as

$$\widetilde{a}_\sigma = a_\sigma + \frac{n}{2}, \quad \widetilde{b}_\sigma = b_\sigma + \frac{1}{2}I_q(\sigma^2)$$

References

Bogachev, V. I. (1998). *Gaussian measures.* Number 62. American Mathematical Society.

Bondar, J. V. and P. Milnes (1981). Amenability: A survey for statistical applications of Hunt-Stein and related conditions on groups. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete* 57(1), 103-128.

Kiefer, J. (1957). Invariance, minimax sequential estimation, and continuous time processes. *The Annals of Mathematical Statistics* 28(3), 573-601.

Komaki, F. (2002). Bayesian predictive distribution with right invariant priors. *Calcutta Statistical Association Bulletin* 52(1-4), 171–180.

Zidek, J. V. (1969). A representation of Bayes invariant procedures in terms of Haar measure. *Annals of the Institute of Statistical Mathematics* 21(1), 291–308.