Spatially-Varying Bayesian Predictive Synthesis for Flexible and Interpretable Spatial Prediction

Danielle Cabel\textsuperscript{1}, Shonosuke Sugasawa\textsuperscript{2}, Masahiro Kato\textsuperscript{3}, Kösaku Takanashi\textsuperscript{4}, and Kenichiro McAlinn\textsuperscript{1}

\textsuperscript{1}Fox School of Business, Temple University, \\textsuperscript{2}Center for Spatial Information Science, The University of Tokyo, \\textsuperscript{3}Cyberagent, Inc \\textsuperscript{4}Center for Advanced Intelligence Project, Riken

Abstract

Spatial data are characterized by their spatial dependence, which is often complex, non-linear, and difficult to 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 spatial Bayesian 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. 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. A finite sample theoretical guarantee is given for the predictive performance of our methodology, showing that the predictions are exact minimax. Through simulation examples and two real data applications, we demonstrate that our proposed spatial Bayesian predictive synthesis outperforms standard spatial models 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, with several methods being developed in response. This includes geographically weighted regressions (Brunsdon et al., 1998), simultaneous autoregressive models (Anselin 1988), and a class of hierarchical models based on latent spatial Gaussian process, such as 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). Beyond these conventional approaches, machine learning techniques, such as gradient boosting trees (e.g. Chen and Guestrin, 2016), (spatial) random forests (Breiman 2001; Sekulić et al., 2020; Saha et al., 2021), and neural networks (e.g. Bishop et al., 1995) could potentially offer more precise spatial predictions by incorporating spatial information as features. The main drawback of these machine learning approaches, however, is that the prediction mechanism is often a “black-box,” where the interpretation of the model structure is impossible. While several techniques to interpret black-box models have been proposed (e.g. Apley and Zhu 2020; Lundberg and Lee 2017), the conventional model-based approaches are much more preferable in terms of interpretation. Spatial prediction methods with accuracy comparable to machine learning methods, while maintaining interpretability at the level of conventional spatial models, are, thus, strongly desired. The aim of this paper is to construct such a methodology by introducing a new framework for synthesizing multiple spatial models.

There are several approaches in dealing with uncertainty arising from multiple candidate models (i.e. model uncertainty). One approach is model selection, where the “best” model is selected by, for example, AIC, BIC, cross validation, and so on. Alternatively, ensemble methods (including model averaging) have been used to mitigate model uncertainty by averaging over several candidate models. A simple approach is to take the arithmetic
mean over the models, which is surprisingly effective in many applications (Genre et al., 2013). For dealing with model uncertainty in spatial data, however, neither approaches provide much solace. 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 (a similar phenomenon can be seen in McAlinn et al. [2020] in the context of multivariate time-series forecasting). 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 has 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.

We contribute to this field by introducing a framework to deal with model uncertainty in spatial data. Our approach builds upon the framework of Bayesian predictive synthesis (BPS; McAlinn and West, 2019), which is a coherent Bayesian framework for synthesizing multiple sources of information. Utilizing this framework, we develop an extension for spatial data that explicitly takes into account the spatially dependent biases and dependencies among models. Specifically, our proposed methodology involves a latent factor spatially-varying coefficient model as a synthesis function within BPS. The predictions from the spatial models are treated as latent factors, utilizing a Gaussian process, allowing for spatially dependent biases and dependencies to be captured across regions. Spatial BPS, thus, effectively learns the spatially-varying coefficients that, in turn, improves predictive accuracy. In other words, spatial BPS learns that certain models do better or worse in different regions, learns their dependence across regions, and proportions their confidence in the models depending on the geographical location. 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). Further, theoretical support of our proposed strategy is presented. We show that spatial predictions produced by our method are exact minimax, which provides finite sample guarantees for its predictive performance. 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 analysis, such as neural networks, spatial random forest (Saha et al., 2021), and extreme gradient boosting trees (Chen and Guestrin, 2016), in terms of predictive accuracy and calibration. Moreover, by synthesizing only “white-box” prediction methods, such as generalized linear models and generalized additive models (Hastie and Tibshirani, 1987), that are fully interpretable, we maintain interpretability, while attaining predictive accuracy comparable to black-box machine learning methods.

The rest of the article will proceed as follows. Section 2 introduces spatial BPS under the general BPS framework. Full details of our proposed spatial BPS, using a latent factor spatially-varying coefficient model as the synthesis function, its MCMC computational strategy, and extensions to general responses are developed in Section 3. Section 4 develops alternative computational strategies for scalable inference. Section 5 gives the theoretical properties of the method. Simulation studies are presented in Section 6. Real world applications with occurrence of Tsuga canadensis and apartment prices in Tokyo are presented in Section 7. The paper concludes with additional comments and closing remarks in Section 8.
2 Spatial Bayesian Predictive Synthesis

Consider predicting a univariate outcome, \( y(s) \), at some unobserved site, \( s \in \mathcal{S} \). 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, \( h_j(f_j(s)) \), for \( j = 1:J \). These forecast densities represent the individual inferences from the models and the collection of these forms the information set, \( \mathcal{H}(s) = \{h_1(\cdot), \ldots, h_J(\cdot)\} \). 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.

McAlinn and West (2019)– extending earlier work on expert opinion analysis (Genest and Schervish, 1985; West and Crosse, 1992; West, 1992)– proposes the Bayesian predictive synthesis framework. The BPS framework provides a general and coherent way for Bayesian updating, given multiple predictive distributions. Specifically, the Bayesian posterior is given as,

\[
p(y(s)|\Phi(s), \mathcal{H}(s)) = \int \alpha(y(s)|f(s), \Phi(s)) \prod_{j=1}^{J} h_j(f_j(s)) df_j(s),
\]

where \( \Phi(s) \) represents the spatially-varying parameters defining the synthesis function, \( \alpha(y(s)|f(s), \Phi(s)) \), and \( f(s) = (f_1(s), \ldots, f_J(s)) \) is a vector of latent variables. Here, \( \alpha(y(s)|f(s), \Phi(s)) \) determines how the predictive distributions are synthesized and it includes a variety of existing combination methods, such as standard and advanced model averaging methods (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, \( p(y(s), \mathcal{H}(s)) \), and it does not restrict the functional form of the synthesis function, \( \alpha(y(s)|f(s), \Phi(s)) \). This allows \( \mathcal{D} \) to flexibly specify how they want the information to be synthesized. For example, if the interest is in predicting univariate time series, a dynamic linear model can be specified (McAlinn and West, 2019), or if the interest is in predicting multivariate time series, a dynamic seemingly unrelated regression
can be specified (McAlinn et al., 2020). Note that (1) is only a valid posterior if it satisfies the consistency condition. This condition states that, prior to observing $H(s)$, $D$ specifies their own prior predictive, $p(y(s))$, as well as their prior expectation of the model forecasts, $E[\prod_{j=1}^{J} h_j(f_j(s))]$. Then $p(y(s)) = \int \alpha(y(s)|f(s), \Phi(s)) \prod_{j=1}^{J} h_j(f_j(s)) df_j(s)$ must hold; meaning that the two priors that $D$ specifies must be consistent with each other.

3 Latent Factor Spatially-Varying Coefficient Models

The methodological developments at the core of this paper, work to adapt and extend the basic BPS framework to forecasting with spatial datasets. In particular, we develop an extension involving spatially-varying parameters to characterize and allow for model-specific biases, patterns of miscalibration, inter-dependencies among models, and differing forecast accuracy over different regions.

3.1 Setting and model

Consider a decision maker $D$, who is predicting a series of spatially-varying points $y(s)$, for $s \in S$. For each location, $D$ receives forecast densities from each of the $J$ models, forming the information set, $H(s) = \{h_1(f(s)), \ldots, h_J(f(s))\}$, for each $s \in S$. As $D$ accrues predictions from the models of the observed data, they gain more insight into the relationship among the models, their dependency characteristics, and forecast accuracy. The Bayesian model will involve parameters that define the BPS framework, for which information is updated along spatial regions. The BPS formulation in (1) gives the spatial extension of the BPS framework with the specific form of the synthesis function adopted as,

$$\alpha(y(s)|H(s), \Phi(s)) = \phi\left(y(s); \beta_0(s) + \sum_{j=1}^{J} \beta_j(s) f_j(s), \sigma^2(s)\right),$$

(2)

where $\phi(\cdot; a, b^2)$ denotes the univariate normal density with mean $a$ and variance $b^2$, and $\Phi(s) = (\beta_0(s), \beta_1(s), \ldots, \beta_J(s), \sigma^2(s))$ represents the spatially-varying parameters defining the synthesis pdf parameters. For simplicity, we assume $\sigma^2(s) = \sigma^2$. Here, $\beta_j(s)$ defines the spatially-varying coefficient for the $j$th model and $\beta_j(s)$ is assumed to be a smooth function of $s$. We employ a Gaussian process to estimate $\beta_j(s)$ at sampled locations, which enables us to obtain a predictive distribution of unknown synthesis coefficients at
unobserved locations.

The specification above defines the following latent factor spatially-varying coefficient model for \( y(s) \):

\[
y(s) = \beta_0(s) + \sum_{j=1}^{J} \beta_j(s)f_j(s) + \varepsilon(s), \quad \varepsilon(s) \sim N(0, \sigma^2), \tag{3}
\]

\[
\beta_j(s) \sim \text{GP}(\tau_j, g_j), \quad \text{independently for } j = 0, 1, \ldots, J,
\]

where \( \text{GP}(\tau, g) \) denotes a Gaussian process with bandwidth parameter, \( h \), scale parameter, \( \tau \), spatial range parameter, \( g \), and \( \varepsilon(s) \) denotes the residual variance in predicting \( y(s) \), given the set of model forecast distributions. 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. Based on the model in (3), \( D \) is able to perform coherent Bayesian predictions on \( y(s) \) at any non-sampled location \( s \), as long as the predictive density of \( f_j(\cdot) \) is available.

Suppose we observe samples at \( n \) locations, \( s_1, \ldots, s_n \in S \). 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

\[
y_i = \beta_{0i} + \sum_{j=1}^{J} \beta_{ji}f_{ji} + \varepsilon_i, \quad \varepsilon_i \sim N(0, \sigma^2), \quad i = 1, \ldots, n, \tag{4}
\]

\[
\beta_j \equiv (\beta_{j1}, \ldots, \beta_{jn})^\top \sim N(0, \tau_j G(g_j)), \quad j = 0, \ldots, J,
\]

where the \( (i, i') \)-element of \( G(g_j) \) 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. Instead of assuming that the prior mean of \( \beta_j \) is 0, it may be useful to use a non-zero prior mean, such as \( 1/J \), but the posterior is not sensitive to such choice as long as \( n \) is large. For the prior distributions of the unknown parameters, we use \( \sigma^2 \sim \text{IG}(a_\sigma, b_\sigma), \tau_j \sim \text{IG}(a_\tau, b_\tau), \) and \( g_j \sim U(g, g) \), independently for \( j = 1, \ldots, J \). We obtain the joint distribution

\[
\pi(\sigma^2) \prod_{j=0}^{J} \pi(\tau_j) \pi(g_j) \phi_n(\beta_j; 0, \tau_j G(g_j)) \times \prod_{j=1}^{J} \prod_{i=1}^{n} h_j(f_{ji}) \times \prod_{i=1}^{n} \phi(y_i; \beta_{0i} + \sum_{j=1}^{J} \beta_{ji}f_{ji}, \sigma^2)
\]

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

### 3.2 Posterior computation and prediction

At location $s$, the BPS analysis will include inferences on the latent factor states, $f_j(s)$, as well as the spatially-varying BPS 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 $\sigma^2$ is given as follows:

1. **Sampling of $f_{ji}$** Generate $f_{ji}$ from $N(A^{(f)}_{ji} B^{(f)}_{ji}, A^{(f)}_{ji})$, where
   $$A^{(f)}_{ji} = \left(\frac{\beta^2_{ji}}{\sigma^2} + \frac{1}{b_{ji}}\right)^{-1}, \quad B^{(f)}_{ji} = \frac{\beta_{ji}}{\sigma^2} \left(y_i - \beta_0 - \sum_{k \neq j} \beta_{ki} f_{ki}\right) + \frac{a_{ji}}{b_{ji}}$$

2. **Sampling of $\beta_j$** Generate $\beta_j$ from $N(A^{(\beta)}_j B^{(\beta)}_j, A^{(\beta)}_j)$, where
   $$A^{(\beta)}_j = \left\{\sigma^{-2} \Omega_j + \tau_j^{-1} G(g_j)^{-1}\right\}^{-1}, \quad B^{(\beta)}_j = \frac{1}{\sigma^2} f_j \circ \left(y - \beta_0 - \sum_{k \neq j} f_k \circ \beta_k\right),$$
   with $\Omega_j = \text{diag}(f^2_{j1}, \ldots, f^2_{jn})$ and $f_j = (f_{j1}, \ldots, f_{jn})$. Note that $\circ$ denotes the Hadamard product.

3. **Sampling of $\tau_j$** Generate $\tau_j$ from IG$(a_\tau + n/2, b_\tau + \beta_j^T G(g_j)^{-1} \beta_j / 2)$.

4. **Sampling of $g_j$** The full conditional of $g_j$ is proportional to
   $$|G(g_j)|^{-1/2} \exp\left(-\frac{1}{2\tau_j} \beta_j^T G(g_j)^{-1} \beta_j\right), \quad g_j \in (g, \overline{g}).$$
   A random-walk Metropolis-Hastings is used to sample from this distribution.

5. **Sampling of $\sigma^2$** Using the conditionally conjugate prior $\sigma^2 \sim \text{IG}(a_\sigma, b_\sigma)$, the full conditional is $\sigma^2 \sim \text{IG}(a_\sigma + n/2, b_\sigma + \sum_{i=1}^n (y_i - \beta_0 - \sum_{j=1}^J \beta_{ji} f_{ji})^2 / 2)$.

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 $t$ be a new location where we are interested in predicting $y(t)$, assuming that the predictive distributions of $f(t) = (f_1(t), \ldots, f_J(t))$, namely, predictive distributions of the $J$ models, are available. Then, the posterior predictive distribution of $y(t)$ is obtained as

$$p(y(t)|y, f(t)) = \int \phi(y(t); \beta_0(t) + \sum_{j=1}^J \beta_j(t)f_j(t), \sigma^2) \prod_{j=0}^J p(\beta_j(t)|\beta_j; \tau_j, g_j) d\beta_j(t)$$

$$\times \prod_{j=1}^J h_j(f_j(t)) df_j(t) \times \pi(\Theta|y) d\Theta,$$

where $\Theta$ is a collection of $\{f_{ji}, \{\beta_j\}, \{\tau_j\}, \{h_j\}$ and $\sigma^2$, $p(\beta_j(t)|\beta_j; \tau_j, g_j)$ is the conditional distribution of $\beta_j(t)$ given $\beta_j$, and $\pi(\Theta|y)$ is the posterior distribution of $\Theta$. Under the assumption of Gaussian process on $\beta_j(s)$, the conditional distribution of $\beta_j(t)$ is given by $N(G_t(g_j)^T G(g_j)^{-1} \beta_j, \{\tau_j - \tau_j G_t(g_j)^T G(g_j)^{-1} G_t(g_j)\}^{-1})$, where $G_t(g_j) = (C(||t - s_1||; g_j), \ldots, C(||t - s_n||; g_j))$. Sampling from the predictive distribution (5) can be easily carried out by using the posterior samples of $\Theta$. First, independently generate $f_j(t)$ from the predictive distribution of the $j$th model and generate $\beta_j(t)$ from its conditional distribution given $\Theta$. Then, we can generate $y(t)$ from $N(\beta_0(t) + \sum_{j=1}^J \beta_j(t)f_j(t), \sigma^2)$.

3.3 Spatial BPS under general types of response variables

The proposed spatial BPS framework (1) can also be applied to situations with general types of outcomes, using 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 7.1. The linear latent factor model (3) for continuous response can be modified as

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

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

9
To enhance the efficiency of posterior computation, we employ the following Pólya-gamma data augmentation [Polson et al., 2013]:

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

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

\[
A_j^{(\beta)} = \text{diag}(\omega_1 f_{j1}^2, \ldots, \omega_n f_{jn}^2) + \tau_j^{-1} G(g_j)^{-1},
\]

\[
B_j^{(\beta)} = F_j \circ \{ y^* - \omega \circ (\beta_0 + \sum_{k \neq j} f_k \circ \beta_k) \},
\]

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

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

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

4 Scalable Computation for Spatial BPS

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 spatial BPS. In this section, we consider two scalable alternatives to carry out spatial BPS.

4.1 Nearest-neighbor Gaussian process

The first approach is to employ a nearest-neighbor Gaussian process [Datta et al., 2016] for \( \beta_j(s) \), which uses a multivariate normal distribution with a sparse precision matrix for \( \beta_j(s_1), \ldots, \beta_j(s_n) \), defined as

\[
\pi(\beta_j(s_1), \ldots, \beta_j(s_n)) = \prod_{i=1}^n \phi(\beta_j(s_i); B_j(s_i)\beta_j(N(s_i)), \tau_j F_j(s_i)), \quad j = 0, \ldots, J
\]
where

\[ \mathbf{B}_j(s_i) = C_j(s_i, N(s_i))C_j(N(s_i), N(s_i))^{-1}, \]

\[ \mathbf{F}_j(s_i) = C_j(s_i, s_i) - C_j(s_i, N(s_i))C_j(N(s_i), N(s_i))^{-1}C_j(N(s_i), s_i), \]

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 \( \beta_j(s) \). Under the settings, the full conditional distribution of \( \beta_j(s_i) \) is a univariate normal distribution, where the computation of its mean and variance requires only \( m \times m \) matrix calculation at most, rather than \( n \times n \) matrix calculation required for the full Gaussian process. Hence, the computational cost at each iteration is reduced to \( O(\text{Jnm}^2) \), which is a drastic reduction from the original computation cost \( O(\text{Jn}^3) \), 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.

### 4.2 Variational Bayes approximation

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

\[ q(\{f_{ji}\}, \{\beta_j\}, \{\tau_j\}, \{g_j\}, \sigma^2) = q(\sigma^2) \prod_{j=0}^{J} q(\beta_j) q(\tau_j) q(g_j) \prod_{i=1}^{n} q(f_{ji}), \]

and each variational posterior can be iteratively updated by computing, for example, \( q(\beta_j) \propto \exp(E_{-\beta_j} [\log p(y, \Theta)]) \), where \( \Theta = (\{f_{ji}\}, \{\beta_j\}, \{\tau_j\}, \{g_j\}, \sigma^2) \), and \( E_{-\beta_j} \) denotes the expectation with respect to the marginal variational posterior of the parameters other than \( \beta_j \). From the forms of full conditional posterior distributions given in Section 3.2.
the following distributions can be used as variational distributions:

\[ q(f_{ji}) \sim N(\bar{m}_{ji}, \bar{s}_{ji}^2), \quad q(\beta_j) \sim N(\bar{\mu}_j, \bar{\Sigma}_j), \quad q(\tau_j) \sim IG(\bar{\alpha}_{\tau_j}, \bar{\beta}_{\tau_j}), \]

\[ q(g_j) \sim D(\bar{p}_{j1}, \ldots, \bar{p}_{jL}), \quad q(\sigma^2) \sim IG(\bar{a}_\sigma, \bar{b}_\sigma), \]

where \( D(\bar{p}_{j1}, \ldots, \bar{p}_{jL}) \) is a discrete distribution on \{\eta_1, \ldots, \eta_L\}, such that \( P(g_j = \eta_l) = \bar{p}_{jl} \).

The MFVB algorithm is described as follows:

**Algorithm 1.** Starting with \( \bar{m}_{ji}^{(0)}, \bar{s}_{ji}^{2(0)}, \bar{\mu}_j^{(0)}, \bar{\Sigma}_j^{(0)}, \bar{a}_{\tau_j}, \bar{b}_{\tau_j}, \bar{p}_{j1}^{(0)}, \bar{\alpha}_{\tau_j}^{(0)}, \bar{\beta}_{\tau_j}^{(0)}, \bar{\alpha}_{\sigma}, \bar{\beta}_{\sigma} \) and \( r = 0 \), repeat the following process until numerical convergence: for \( j = 1, \ldots, J \), update \( \bar{m}_{ji} \) and \( \bar{s}_{ji}^2 \) as

\[
\bar{s}_{ji}^{2(t+1)} \leftarrow \left\{ \frac{1}{b_{ji}} + (\bar{p}_{ji}^{(t)} + \bar{\Sigma}_{ji}^{(t)}) \frac{\bar{a}_{\sigma}^{(t)}}{b_{\sigma}^{(t)}} \right\}^{-1},
\]

\[
\bar{m}_{ji}^{(t+1)} \leftarrow \frac{a_{ji}}{b_{ji}} + \frac{\bar{p}_{ji}^{(t)} \bar{a}_{\sigma}^{(t)}}{b_{\sigma}^{(t)}} \left( y_i - \bar{\mu}_0 - \sum_{k<j} \bar{\mu}_k \bar{m}_k - \sum_{k>j} \bar{\mu}_k \bar{m}_k \right) / \bar{s}_{ji}^{2(t+1)}.
\]

For \( j = 1, \ldots, J \), update \( \bar{\mu}_j \) and \( \bar{\Sigma}_j \) as

\[
\bar{\Sigma}_j^{(t+1)} \leftarrow \left\{ \Omega_j^{(t+1)} \frac{\bar{a}_{\sigma}^{(t)}}{b_{\sigma}^{(t)}} + \sum_{\ell=1}^L \bar{p}_{j\ell} G(\eta_\ell) \frac{\bar{a}_{\tau_j}^{(t)}}{b_{\tau_j}^{(t)}} \right\}^{-1},
\]

\[
\bar{\mu}_j^{(t+1)} \leftarrow \left( \bar{\Sigma}_j^{(t+1)} \right)^{-1} \frac{\bar{a}_{\sigma}^{(t)}}{b_{\sigma}^{(t)}} \left( \bar{m}_j^{(t+1)} \circ \left( y_i - \bar{\mu}_0 - \sum_{k<j} \bar{\mu}_k \bar{m}_k - \sum_{k>j} \bar{\mu}_k \bar{m}_k \right) / \bar{s}_{ji}^{2(t+1)} \right),
\]

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

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

For \( j = 0, \ldots, J \), update \( \bar{p}_{j\ell} \) as

\[
\bar{p}_{j\ell}^{(t+1)} \leftarrow \frac{\|G(\eta_\ell)^{-1}\|^{-1/2} \exp \left( -\bar{a}_{\tau_j}^{(t+1)} \text{tr} \left( \bar{\mu}_j^{(t+1)} \bar{\tau}_j^{(t+1)} + \bar{\Sigma}_j^{(t+1)} \right) G(\eta_\ell)^{-1} / 2\bar{\beta}_{\tau_j}^{(t+1)} \right)}{\sum_{l'=1}^L \|G(\eta_{l'})\|^{-1/2} \exp \left( -\bar{a}_{\tau_j}^{(t+1)} \text{tr} \left( \bar{\mu}_j^{(t+1)} \bar{\tau}_j^{(t+1)} + \bar{\Sigma}_j^{(t+1)} \right) G(\eta_{l'})^{-1} / 2\bar{\beta}_{\tau_j}^{(t+1)} \right)}.
\]
Set $\tilde{a}_\sigma^{(t+1)} = a_\sigma + n/2$ and update $\tilde{b}_\sigma$ as

\[
\tilde{b}_\sigma^{(t+1)} \leftarrow \left( y - \tilde{\mu}_0^{(t+1)} - \sum_{j=1}^J \tilde{\mu}_j^{(t+1)} \circ \tilde{m}_j^{(t+1)} \right) + \text{tr} \tilde{\Sigma}_0^{(t+1)}
\]

\[
+ \sum_{j=1}^J \text{tr} \left\{ (\tilde{\mu}_j^{(t+1)})^\top \circ (\tilde{m}_j^{(t+1)})^\top + \tilde{\Sigma}_j^{(t+1)} + \tilde{S}_j^{(t+1)} \right\} - \sum_{i=1}^n \sum_{j=1}^J \tilde{m}_j^{(t+1)2}\tilde{\mu}_j^{(t+1)2}.
\]

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.

5 Theoretical Properties

We now discuss the theoretical properties of the predictions obtained from spatial BPS. Due to the characteristics of spatial data, where locations are finite and predictions are limited, we provide theoretical guarantees for predictive performances under finite sample. This is done by showing that spatial BPS is exact minimax under Kullback-Leibler (KL) loss. Consider the task of predicting the data generating process (DGP), $y(s)$, with the predictive values from the $J$ models (predictive distributions), $f(s) = (f_1(s), \ldots, f_J(s))$. Assume that both the DGP and predictive values have second moments. We define the DGP as

\[
y(s) = \mu(s) + W(s) + \varepsilon(s),
\]

unrelated to $f(s)$, where $\mu(s)$ is the average process, $\mu(s) = \mathbb{E}[y(s)]$, $W(s)$ is an isotropic and second order stationary process (without assuming Gaussian) with $\mathbb{E}[W(s)] = 0$ and covariance function $C_\omega(\cdot, \cdot)$, and $\varepsilon(s)$ is a white noise process with $\mathbb{E}[\varepsilon(s)] = 0$ and $\text{Var}[\varepsilon(s)] = \sigma^2$. Further, assume that $\mu(s)$ takes a value within a compact region, $[-b, a]$, that includes the origin. Let $s_1, \ldots, s_n$ be $n$ observed locations, and $t$ be an unobserved location. We consider discrete models for the $n+1$ locations, where the set of unknown parameters in the model is $\theta = (\mu(s_1), \ldots, \mu(s_n), \mu(t), C_\omega, \sigma^2)$. The joint distribution of $y = (y(s_1), \ldots, y(s_n))$ and $y(t)$ given $\theta$, induced from the DGP, is denoted by $p_\theta^*(y, y(t))$, and the marginal distribution of $y(t)$ is $p_\theta^*(y(t)) = \int_{\mathbb{R}^n} p_\theta^*(y, y(t))dy$. The
goal is to construct a predictive distribution, \( q(y(t)|y) \), that is minimax with regard to the KL risk against \( p_\theta^*(y(t)) \). Here, the KL loss is

\[
\text{KL}(p_\theta^*|q) = \int_y \log \frac{p_\theta^*(y(t))}{q(y(t)|y)} p_\theta^*(y(t)) dy(t),
\]

where the loss value depends on \( y \) and \( \theta \). The KL risk is the expectation of the loss, given by \( \mathbb{E}_y[\text{KL}(p_\theta^*|q)] \) with regard to the observation \( y \).

We consider the predictive distribution of \( y(t) \) obtained from spatial BPS without assuming a Gaussian process on \( \beta_0(s) \) in the spatially-varying model (4). In this case, the predictive distribution \( q(y(t)|y, f(t)) \) can be obtained in the same way as (5). We then show the following minimax property of the predictive distribution.

**Theorem 1.** Under a suitable class of priors on \( \beta_0, \beta_0(t), \{\tau_j\}, \{g_j\} \) and \( \sigma^2 \), the predictive distribution \( q(y(t)|y, f(t)) \) is exact minimax in terms of the KL risk \( \mathbb{E}_y[\text{KL}(p_\theta^*|q)] \).

The proof is in the Supplementary Material.

### 6 Simulation Studies

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

#### 6.1 Empirical behavior of spatial BPS

We first illustrate how our proposed spatial BPS 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: \( \text{Cov}(z_k(s_i), z_k(s_j)) = \exp(-\|s_i - s_j\|/\eta), k = 1,2, \) with \( \eta = 0.2 \). 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 \( r = 0.75 \), 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 model:

**Scenario 1:**

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

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, $\varepsilon(s_i)$ are mutually independent and $\varepsilon(s_i) \sim N(0, 1)$. 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 scenario 1, we apply a quadratic regression model, $y(s_i) = \beta_0 + \beta_1 x_1(s_i) + \beta_2 x_1(s_i)^2 + \beta_3 x_2(s_i) + \beta_4 x_2(s_i)^2 + \varepsilon_i$, 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 spatial BPS 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 Gaussian processes. We generate 1500 posterior samples after discarding the first 500 samples as burn-in.

We first evaluate the predictive performance in non-sampled locations. We generated 100 additional locations, as with $x_1(s_i)$, $x_2(s_i)$, and $y(s_i)$, according to the same DGP. Using the generated posterior samples in the proposed latent model, 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 spatial BPS, as well as predictors of the two quadratic regressions. We also applied extreme gradient boosting tree (Chen and Guestrin 2016), denoted by XGB, with two covariates and two-dimensional spatial information using the R package “xgboost” with 1000 trees, where the sub-sampling rate is 0.1 and the learning rate is 0.01. The MSE values are

$$BPS : 1.52, \quad QR1 : 16.0, \quad QR2 : 16.9, \quad XGB : 4.82.$$ 

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
which leads to QR1 having a large MSE. The same explanations can be given for QR2 and its MSE. It is reasonable that XGB outperforms both QR1 and QR2, but the proposed spatial BPS can outperform XGB. This is done by successfully synthesizing the two prediction models, QR1 and QR2.

To see how spatial BPS works in this example, we compute the ratio of two coefficients, $|\hat{\beta}_1|/(|\hat{\beta}_1| + |\hat{\beta}_2|)$, where $\hat{\beta}_1$ and $\hat{\beta}_2$ are posterior means of $\beta_1$ (weight for QR1) and $\beta_2$ (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. The model structures of QR1 and QR2 are easy to interpret, and the resulting predictor is also highly interpretable, while XGB is not.

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 0.96, which is practically equivalent to the nominal level, illustrating how well-calibrated spatial BPS is. For comparison, we also present the point prediction made by XGB in the right panel of Figure 1. XGB gives accurate predictions for samples whose true values are close to 0, but fails to capture the variability in the test data. Indeed, the prediction accuracy of XGB worsens as the true value moves away from 0. This is problematic for many applications, where prediction locations of interest are further away from the mean.

6.2 Performance comparison

Next, we compare the performance of our proposed method with other methods under three additional scenarios. Let $s \in [0, 1]^2$ be the spatial location generated from the uniform distribution on the region. First, we consider the following two scenarios, mimicking
the simulation settings of \cite{Saha2021}:

**Scenario 2:** \( y(s) = 10 \sin(\pi x) + w(s) + \varepsilon, \quad x \sim U(-1, 1), \quad \varepsilon \sim N(0, 5) \)

**Scenario 3:** \( y(s) = \frac{1}{6} \left( 10 \sin(\pi x_1 x_2) + 20(x_3 - 0.5)^2 + 10x_4 + 5x_5 \right) + w(s) + \varepsilon, \)

\[ x_1, \ldots, x_5 \sim U(0, 1), \quad \varepsilon \sim N(0, 0.75), \]

where \( \omega(s) \) is an unobserved spatial effect following a two-dimensional Gaussian process, \( \omega(s) \sim \text{GP}(0, C(\cdot, \cdot; \theta)) \) and \( C \) is a covariance function defined as \( C(s, s'; \theta) = \tau^2 \exp(-\|s - s'\|/\phi) \). Note that the mean function in Scenario 3 is the well-known Friedman function \cite{Friedman1991}. We set \((\tau^2, \phi) = (7, 0.4)\) and \((3, 0.2)\) in Scenarios 2 and 3, respectively.

Furthermore, we adopt the following scenario:

**Scenario 4:** \[ y(s) = \begin{cases} 
\beta_0(s) + \beta_1(s)x_1(s) + \beta_2(s)x_2(s) + \varepsilon, & s \in D_1, \\
x_1(s)^2 + x_2(s)^2 + \varepsilon, & s \in D_2, 
\end{cases} \]

where \( \varepsilon \sim N(0, 1), \beta_k(s) \sim \text{GP}(0, C(\cdot, \cdot; \theta_k)) \) with \( \tau_k = 1 \) and \( \phi_k = k + 1 \) for \( k = 0, 1, 2 \). The settings of \( D_1, D_2, x_1(s) \) and \( x_2(s) \) are the same as those in Section 6.1.

We generated 200 training samples and 50 test samples of \((y, x, s)\) from each DGP. In addition to XGB used in Section 6.1, we fitted the following four prediction models based on this simulation setting.
We used the R package “spgwr” to fit the GWR model (Brunsdon et al., 1998), where the optimal bandwidth is selected by cross-validation.

- GAM (generalized additive model): We fitted the GAM model (Hastie and Tibshirani, 1987) with covariates \( (x_i, s_i) \) by using the R package “gam” with default settings for tuning parameters.

- NN (Neural network): We applied a NN model that consists of five-layer perception with 32 nodes and the ReLU activation function. To train the NN model, we used the Adam optimizer in which the learning rate is fixed at 0.001, which is the default value of the open source “torch for R” (https://torch.mlverse.org/), and the batch size is fixed at 50.

- SRF (spatial random forest): We applied the recently proposed SRF (Saha et al., 2021) using the R package “RandomForestsGLS” with 10 trees.

It should be noted that prediction structures of both GWR and GAM models are easy to interpret, since their mean functions are relatively simple functions of \( x_i \) and \( s_i \). On the other hand, the three models, NN, XGB, and SRF, are so-called “black-box models” with little to no interpretability, despite their flexibility to capture the underlying structure.

We synthesize the prediction models via spatial BPS. In particular, we consider two sets of models to be synthesized, BPS1 (synthesis of GWR and GAM) and BPS2 (synthesis of GWR, GAM, NN, XGB, and SRF). In applying BPS, we used the means and variances of \( y(s) \) for the training and test sets, as done in Section 6.1. Note that BPS1 combines two “white-box” models and the resulting synthesized predictor is also interpretable. The motivation of using BPS1 is to see whether a combination of two simple models can perform as well as black-box prediction models. On the other hand, BPS2 combines all the prediction models, including the black-box models, and the performance of BPS2 will show the effectiveness of spatially-varying model synthesis. We applied spatial BPS with the exponential kernel in the Gaussian process for spatially-varying model coefficients.

We generated 1000 posterior samples of the model coefficients, as well as the unknown
parameters, after discarding the first 500 samples as burn-in, to obtain the posterior means of the test data. We also applied the variational Bayes (VB) algorithm described in Section 4.2 to efficiently compute approximate posterior distributions under BPS1 and BPS2.

We compute the mean squared error (MSE) of the test data and present boxplots of MSE values under 100 iterations in Figure 2. The associated summary statistics are presented in Table 1. Most notably, BPS1, which synthesizes two white-box models, performs almost as well as black-box models. This shows the strength of BPS; taking simple, interpretable models and synthesizing them to produce equivalent (or very close to) or superior performances to more complex, uninterpretable models. As an illustration of the robustness of BPS, it provides stable performances even when some of the underlying models perform poorly (GWR under Scenario 2 and NN under Scenario 1, 3, and 4). We note that BPS2, which synthesizes all the models, is shown to perform almost the same as the best model among the five models being synthesized. This shows that the spatially-varying model coefficients in BPS2 can be informative in highlighting which models are useful for prediction. The variational approximation of BPS performs in the same way as the BPS with MCMC (except for variational approximation for BPS1 under Scenario 2), which motivates the use of the approximation to speedily compute point predictions. In this simulation study, NN did not perform well, possibly due to the sample size being small compared to the number of parameters.

7 Real Data Applications

We consider two distinct real world applications to highlight the predictive performance of spatial BPS. 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 spatial BPS and compare different methods under distinctly different situations, to provide a more holistic assessment.
7.1 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).
Table 1: The mean of 25%, 50% and 75% quantile MSE values of five prediction methods and four types of BPS, repeated over 100 iterations. The smallest and second smallest values are highlighted in bold.

| Scenario | BPS1 | BPS2 | BPS1 | BPS2 | GWR | GAM | NN | XGB | SRF |
|----------|------|------|------|------|-----|-----|----|-----|-----|
| 25%      | 1.54 | 1.52 | 1.82 | 1.61 | 1.72 | 1.80 | 2.49 | 1.57 | 1.79 |
| 50%      | 1.90 | 1.95 | 2.21 | 1.95 | 2.12 | 2.21 | 3.49 | 1.98 | 2.28 |
| 75%      | 2.40 | 2.40 | 2.76 | 2.32 | 2.58 | 2.86 | 5.09 | 2.54 | 3.06 |
| 100%     | 6.84 | 6.99 |

There are several reasons why the prediction of 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, interpretability of the prediction is relevant for future conservation.

To investigate the predictive performance of spatial BPS 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):

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

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

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

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 BPS, we apply the logistic model \([6]\) 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 \( m = 5 \) 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 XGBoost \([7]\) with 1000 trees, where the optimal number of trees was selected via 5-fold 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. We also applied a linear logistic model with spatial random effects, where the spatial effects are modeled via the nearest-neighbor Gaussian process \([8]\), which was fitted using the R package “spNNGP.” We generated 7000 posterior samples after discarding the first 3000 samples. Spatial random forest \([9]\) 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. 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 figure shows the superiority of spatial BPS to all other methods, including both XGB and NNGP, 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 XGB and NNGP, but the AUC value of the synthesized prediction, through spatial BPS, is higher. This indicates the effectiveness of our proposed method in synthesizing simple models to outperform more complex methods, even if the performance of those models is relatively poor. More importantly, resulting predictors via spatial BPS are weighted combinations of the three models, whose prediction structures are fully interpretable, so the proposed predictor is also superior to XGB in terms of interpretation.

7.2 Apartment prices in Tokyo

Our second application is to apply spatial BPS 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 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, this application requires both accurate and in-
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 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 spatial BPS 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 XGB and SRF, as in 
Section 6.2 and NNGP, as used in Section 7.1 to predict the room prices in the test data.

The left panel in Figure 5 reports the spatial plot of $\beta_0(s_i)$, i.e., the intercept term of 
spatial BPS. 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 
spatial BPS 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

$$\begin{align*}
\text{BPS} : & \ 0.240, \\
\text{XGB} : & \ 0.257, \\
\text{NNGP} : & \ 0.268, \\
\text{SRF} : & \ 0.579,
\end{align*}$$

where, again, spatial BPS provides superior prediction accuracy. As with the previous 
applications, it should be noted that the resulting predictors made by BPS are inter-
pretable, 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 spatial BPS, 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, 
spatial BPS provides accurate point predictions and 95% prediction intervals with reason-
able interval length regardless of the true price. The coverage proportion is 89.9%, which 
is well-calibrated for these tasks.
8 Concluding Remarks

Bayesian predictive synthesis provides a theoretically and conceptually sound framework to synthesize density forecasts. Utilizing this framework, we develop a spatially-varying extension for the context of spatial data. With this new extension, 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 spatial BPS compared to conventional spatial models and modern machine learning techniques. Specifically, by dynamically synthesizing the predictive distribution from the models, spatial BPS can improve point and distributional predictions. Additionally, posterior inference on the full spatial region gives the decision maker the 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 spatial BPS, as well as specific developments catered towards a specific dataset, is of interest.

Regarding scalable computation algorithms for spatial BPS, 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 (4) 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 spatial BPS approach. The first is to extend spatial BPS 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, spatial BPS 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 spatial BPS. 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.

Supplementary Material

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

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

- 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

$$A_i = \left\{ \frac{f_i f_i^\top}{\sigma^2} + \text{diag}(\gamma_0, \ldots, \gamma_J) \right\}^{-1}, \quad \gamma_{ji} = 1 + \tau_j F_j(s_i) + \sum_{t:s_i \in N(t)} B_j(t; s_i)^2 \tau_j F_j(t),$$

$$B_i = \frac{f_i y_i}{\sigma^2} + (m_0, \ldots, m_J)^\top,$$

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

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 $B_j(t)$.

- For $j = 0, \ldots, J$, the full conditional distribution of $\tau_j$ is

$$\text{IG} \left( a_r + \frac{n}{2}, b_r + \frac{1}{2} \sum_{i=1}^n \left\{ \frac{\beta_j(s_i) - B_j(s_i) \beta_j(N(s_i))}{F_j(s_i)} \right\}^2 \right).$$

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

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

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 $g_j$. 

29
S2 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 \(\beta_j\). From the forms of full conditional posterior distributions given in the main document, we can use the following distributions as optimal distributions:

\[
q(\sigma^2) \sim IG(\tilde{a}_\sigma, \tilde{b}_\sigma), \quad q(\tau_j) \sim IG(\tilde{a}_{\tau_j}, \tilde{b}_{\tau_j}), \quad q(\beta_j) \sim N(\tilde{\mu}_j, \tilde{\Sigma}_j),
\]

\[
q(h_j) \sim D(\tilde{p}_{j1}, \ldots, \tilde{p}_{jL}), \quad q(f_{ji}) \sim N(\tilde{m}_{ji}, \tilde{s}_{ji}^2),
\]

where \(D(\tilde{p}_{j1}, \ldots, \tilde{p}_{jL})\) is a discrete distribution on \(\{\eta_1, \ldots, \eta_L\}\) such that \(P(h_j = \eta_\ell) = \tilde{p}_{j\ell}\).

The derivation of the updating steps of MFVB is given as follows.

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

\[
E_{-f_{ji}}[\log p(y, \Theta)] = (\text{const.}) - \frac{1}{2} f_{ji}^2 (A_{ji}^f)^{-1} + f_{ji} B_{ji}^f,
\]

where

\[
A_{ji}^f = \left( \frac{1}{b_{ji}} + E_q[\beta_{j1}^2]E_q\left[ \frac{1}{\sigma^2} \right] \right)^{-1} = \left\{ \frac{1}{b_{ji}} + (\tilde{\mu}_{ji}^2 + \tilde{\Sigma}_{ji}) \frac{\tilde{a}_\sigma}{b_{\sigma}} \right\}^{-1},
\]

and

\[
B_{ji}^f = \frac{a_{ji}}{b_{ji}} + E_q[\beta_{j1}]E_q\left[ \frac{1}{\sigma^2} \right] (y_i - E_q[\beta_0] - \sum_{k \neq j} E_q[\beta_k]E_q[f_{ki}])
\]

\[
= \frac{a_{ji}}{b_{ji}} + \tilde{\mu}_{ji}\frac{\tilde{a}_\sigma}{b_{\sigma}} \left( y_i - \tilde{\mu}_0 - \sum_{k \neq j} \tilde{\mu}_k \tilde{m}_{ki} \right).
\]
Then, the parameters in the variational posterior of $f_{ji}$ can be updated as $\tilde{m}_{ji} = A^f_{ji} B^f_{ji}$ and $\tilde{s}^2_{ji} = A^f_{ji}$.

- **(update of $\beta_j$)** It follows that

$$E_{-\beta_j} \left[ \log p(y, \Theta) \right] = \text{(const.)} - \frac{1}{2} \beta^T (A_j^{(\beta)})^{-1} \beta + \beta^T B_j^{(\beta)},$$

where

$$A_j^{(\beta)} = \left\{ E_q[\Omega_j] E_q \left[ \frac{1}{\sigma^2} \right] + E_q[H(h_j)^{-1}] E_q \left[ \frac{1}{r_j} \right] \right\}^{-1} = \left\{ \Omega^*_r \tilde{a}_{\tau_j} + \sum_{\ell=1}^L \tilde{p}_{\ell j} H(\eta_\ell)^{-1} \tilde{a}_{\tau_j} \right\}^{-1},$$

$$B_j^{(\beta)} = E_q \left[ \frac{1}{\sigma^2} \right] E[F_j] \circ \left( y - E_q[\beta_0] - \sum_{k \neq j} E_q[\beta_k] \circ E_q[F_k] \right) = \frac{a_{\tau_j}}{b_{\tau_j}} \tilde{m}_j \circ \left( y - \tilde{\mu}_0 - \sum_{k \neq j} \tilde{\mu}_k \circ \tilde{m}_k \right),$$

where $\Omega^*_r = \text{diag}(\tilde{m}^2_{j1} + \tilde{s}^2_{j1}, \ldots, \tilde{m}^2_{jn} + \tilde{s}^2_{jn})$. Then, the parameters in the variational posterior of $\beta_j$ can be updated as $\tilde{\mu}_j = A_j^{(\beta)} B_j^{(\beta)}$ and $\tilde{\Sigma}_j = A_j^{(\beta)}$.

- **(update of $\tau_j$)** It follows that

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

noting that

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

Then, the parameters in the variational posterior of $\tau_j$ can be updated as

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

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

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

31
where $E_q[\beta_j^\top H(h_j)^{-1}\beta_j] = \text{tr}\{(\tilde{\mu}_j^\top \tilde{\Sigma}_j)H(h_j)^{-1}\}$. Then, the parameters in the variational posterior of $h_j$ can be updated as

$$
\tilde{p}_j\ell = \frac{|H(\eta_{\ell})|^{-1/2} \exp \left(-\tilde{\alpha}_j \text{tr}\{(\tilde{\mu}_j^\top + \tilde{\Sigma}_j)H(\eta_{\ell})^{-1}\}/2\tilde{b}_j\right)}{\sum_{\ell'=1}^L |H(\eta_{\ell'})|^{-1/2} \exp \left(-\tilde{\alpha}_j \text{tr}\{(\tilde{\mu}_j^\top + \tilde{\Sigma}_j)H(\eta_{\ell'})^{-1}\}/2\tilde{b}_j\right)}
$$

- (update of $\sigma^2$) It follows that

$$
E_{-\sigma^2}[\log p(y, \Theta)] = (\text{const.}) - \frac{n}{2} \log \sigma^2 - \frac{1}{2\sigma^2} \sum_{i=1}^n E_q\left[\left(y_i - \beta_0 - \sum_{j=1}^J \beta_{ji} f_{ji}\right)^2\right],
$$

where

$$
I_q(\sigma^2) = \sum_{i=1}^n E_q\left[\left(y_i - \beta_0 - \sum_{j=1}^J \beta_{ji} f_{ji}\right)^2\right] = \left(y - \bar{\mu}_0 - \sum_{j=1}^J \tilde{\mu}_j \circ \tilde{m}_j\right)^\top \left(y - \bar{\mu}_0 - \sum_{j=1}^J \tilde{\mu}_j \circ \tilde{m}_j + \text{tr}(\tilde{\Sigma}_0)\right) + \sum_{j=1}^J \text{tr}\{(\tilde{\mu}_j^\top + \tilde{\Sigma}_j) \circ (\tilde{m}_j^\top + \tilde{S}_j)\} - \sum_{i=1}^n \sum_{j=1}^J \tilde{m}_{ji}^2 \tilde{\mu}_{ji}^2
$$

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

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

**S3 Proof of Theorem 1**

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, $O_{n+1}$, is the group of $(n+1) \times (n+1)$ orthogonal matrices, with $\mathbb{R}_+$ representing the positive region, $(0, \infty]$. The group, $G$, is

$$
G = \mathbb{R}_+ \times O_{n+1} \times \mathbb{R}^{n+1}, \quad c \in \mathbb{R}_+, \quad A \in O_{n+1}, \quad F \in \mathbb{R}^{n+1},
$$
where the operation $g$ to the sample space of $(y, y(t))$ is defined as
\[
g \begin{bmatrix} y \\ y(t) \end{bmatrix} = cA \begin{bmatrix} y \\ y(t) \end{bmatrix} + F,
\]
and the operation $\gamma$ to the parameter space, $\theta = (\mu, \mu(t), C_w, \sigma^2)$, is defined as
\[
\gamma \mu^{n+1} = cA \mu^{n+1} + F, \quad \gamma C^{n+1} = cAC^{n+1}A^\top, \quad \gamma \sigma^2 = c\sigma^2 A^\top A,
\]
where $\mu^{n+1} = (\mu, \mu(t))$ and $C^{n+1}$ is a $(n+1) \times (n+1)$ covariance matrix. The transformation, $g$, to the probability density, $q(y)$, is defined as $gq(y) = q(gy)$. The transformation group, $(g, \gamma, \tilde{g})$, operates transitively on the sample, $(y, y(t))$, and the parameter space.

The statistical decision problem is invariant under the transformation group, $(g, \gamma, \tilde{g})$. Thus, for the sample, $y(t)$, that follows a probability distribution, $p^\ast_{\theta}$, it holds that $p^\ast_{\gamma \theta} = \tilde{g} p^\ast_{\theta}$, which entails that the loss is invariant under the KL loss,
\[
\text{KL} (p^\ast_{\gamma \theta} \mid \tilde{g} q) = \int \log \frac{p^\ast_{\gamma \theta} (y(t))}{gq (y(t) \mid y, f(t))} p^\ast_{\gamma \theta} (y(t)) \, dy(t)
\]
\[
= \int \log \frac{p^\ast_{\theta} (gy(t))}{q(gy(t) \mid y, f(t))} p^\ast_{\theta} (gy(t)) \, dy(t) = \text{KL} (p^\ast_{\theta} \mid q).
\]

The group, $(g, \gamma, \tilde{g})$, is an amenable group (Bondar and Milnes, 1981, p.114), which satisfies the Hunt-Stein condition (Bondar and Milnes, 1981, p.110). 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:
\[
\min_{q} \max_{\theta} \mathbb{E} \left[ \text{KL} (p^\ast_{\theta} \mid q) \right] = \min_{q \text{-} g \text{-} invariant} \max_{\theta} \mathbb{E} \left[ \text{KL} (p^\ast_{\theta} \mid q) \right].
\]

From this, 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 $\gamma$-invariant, i.e., the Bayes decision based on the prior, $\rho$, that satisfies $\rho (\gamma \beta^{n+1}) = \rho (\beta^{n+1})$, $\rho (\gamma \sigma) = \rho (\sigma)$ provides the best equivariant solution (Zidek, 1969). Under KL loss, the Bayesian predictive distribution under a $\gamma$-invariant prior is the best equivariant solution (Komaki, 2002). For the spatial BPS model with Gaussian
processes for $\beta_1(s), \ldots, \beta_J(s)$, if we use the prior distributions

$$
\rho (\beta_0) = 1_{[-b,a]^n} (\beta_0), \quad \rho (\sigma) = \frac{1}{\sigma}, \quad \rho (\tau_j) = \frac{1}{\tau_j}, \quad j = 1, \ldots, J,
$$

(S1)

$\rho$ is $g$-invariant. Here, $1_{[-b,a]^n} (\beta_0)$ is an indicator function, where it is 1 when $\beta_0$ is in the region, $[-b,a]^n$, and 0 otherwise. Therefore, the predictive distribution under the prior (S1), and all predictive distributions that dominate it, is a minimax solution.

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.

Apley, D. W. and J. Zhu (2020). Visualizing the effects of predictor variables in black box supervised learning models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 82(4), 1059–1086.

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.

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.

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.

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.

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

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

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

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.

Lundberg, S. M. and S.-I. Lee (2017). A unified approach to interpreting model predictions. *Advances in neural information processing systems* 30.

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.

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.