Journal of Data Science 20 (4), 533–544
October 2022
DOI: 10.6339/22-JDS1073
Statistical Data Science

Scalable Predictions for Spatial Probit Linear Mixed Models Using Nearest Neighbor Gaussian Processes

Arkajyoti Saha[1], Abhirup Datta[2], and Sudipto Banerjee[3,*]

[1]*Department of Statistics, University of Washington, Seattle, WA, USA*
[2]*Department of Biostatistics, Johns Hopkins University, Baltimore, MD, USA*
[3]*UCLA Department of Biostatistics, 650 Charles E. Young Drive South, University of California Los Angeles, CA 90095-1772, USA*

Abstract

Spatial probit generalized linear mixed models (spGLMM) with a linear fixed effect and a spatial random effect, endowed with a Gaussian Process prior, are widely used for analysis of binary spatial data. However, the canonical Bayesian implementation of this hierarchical mixed model can involve protracted Markov Chain Monte Carlo sampling. Alternate approaches have been proposed that circumvent this by directly representing the marginal likelihood from spGLMM in terms of multivariate normal cummulative distribution functions (cdf). We present a direct and fast rendition of this latter approach for predictions from a spatial probit linear mixed model. We show that the covariance matrix of the cdf characterizing the marginal cdf of binary spatial data from spGLMM is amenable to approximation using Nearest Neighbor Gaussian Processes (NNGP). This facilitates a scalable prediction algorithm for spGLMM using NNGP that only involves sparse or small matrix computations and can be deployed in an embarrassingly parallel manner. We demonstrate the accuracy and scalability of the algorithm via numerous simulation experiments and an analysis of species presence-absence data.

Keywords *binary data; generalized linear mixed models; spatial, Gaussian processes*

1 Introduction

Spatial analysis of non-Gaussian outcomes is widely prevalent in diverse fields within the natural and environmental sciences (De Oliveira et al., 1997; Diggle et al., 1998; Zhang et al., 2022). In several instances we measure variables that are not naturally modeled as Gaussian and may not even be continuous. For example, the outcome of interest might be a spatially referenced binary variable measuring whether the rainfall at a given spatial location was above a specified threshold or not. In ecological and forestry applications, data are collected at spatial locations indicating whether a particular species is present or absent at that location. First proposed in Heagerty and Lele (1998), De Oliveira (2000), there is, by now, a considerable literature on the modeling and analysis of spatial non-Gaussian, and binary in particular, data for which a comprehensive review is beyond the scope of this article. We will focus on settings where the outcome can reasonably be treated as point-referenced and modeled using a stochastic process embedded within a hierarchical model. Spatial referencing for non-Gaussian data may sometimes be aggregated into rates or counts over larger regions, but high-resolution spatial data compiled

*Corresponding author. Email: sudipto@ucla.edu.

Received August 16, 2022; Accepted October 6, 2022

over areas that are sufficiently small are often analyzed using process-based methods for point-referenced data (see, e.g. Diggle et al., 1998).

Following Diggle et al. (1998), we can introduce spatially dependent stochastic processes for non-Gaussian data as follows. Let $Y(s)$ be the outcome of interest at location s, where the probability law for $Y(s)$ cannot be reasonably assumed to be Gaussian. Therefore, we endow $Y(s)$ with a probability law corresponding to a member of the exponential family of densities,

$$f(Y(s) \mid \beta, w(s), \gamma) = h(y(s), \gamma) \times \exp\{\gamma\,[y(s)\eta(s) - \psi(\eta(s))]\}, \tag{1}$$

where γ is a dispersion parameter, $g(\eta(s)) = x(s)'\beta + w(s)$ for specified real-valued link function $g(\cdot)$, $w(s)$ is a spatial random effect at an arbitrary location s assumed to be realized from a Gaussian process, $\psi(\cdot)$ is a known real-valued function, and $h(\cdot,\cdot)$ is a non-negative real-valued function of the random variable that may depend on the dispersion parameter (see, e.g., Lee and Nelder, 1996). The link function $g(\cdot)$ is a key component of such models and connects the support of $Y(s)$ to the real line. Probit and logistic link functions are customary choices for binary data.

Our intended contribution in this manuscript is to offer a scalable inferential framework for point-referenced binary spatial data, where we account for a linear fixed effect and a spatial random field. To accommodate spatial dependence while maintaining computational efficiency, we deviate from the traditional Bayesian implementations of the spatial probit model (Albert and Chib, 1993; Berrett and Calder, 2016) that uses the conditional latent variable representation (1) of the model to sample both w's and the hyper-parameters in a Markov Chain Monte Carlo (MCMC) algorithm. We eschew Markov chain Monte Carlo (MCMC) sampling by exploiting some analytic expressions for marginal likelihood from (1) using a probit link function as proposed in Cao et al. (2022). Avoiding MCMC offers the significant added advantage of obviating some of the challenges of poor mixing of posterior sampling high-dimensional chains resulting from weak identifiability of model hyperparameters. Our approach closely aligns with the recently proposed MCMC-free approach in Cao et al. (2022). A key innovation over this work is the use of Nearest Neighbor Gaussian Process (NNGP) (Datta et al., 2016b) to approximate the conditional prediction probabilities. This reduces the computational complexity over the algorithm of Cao et al. (2022) both theoretically and empirically. We demonstrate that our approach delivers fast and exact spatial predictive inference and scales to very large data sets using a Nearest Neighbor Gaussian Process (NNGP) approximation for the covariance function corresponding to the cumulative distribution function (cdf).

The rest of this paper is organized as follows. In Section 2, we describe the method, and discuss its advantages in terms of computational overhead. In Section 3, we demonstrate the utility of the proposed approach through simulation studies and application in invasive species data. We conclude with a discussion in Section 4.

2 Method

Let $Y(s_i)$ denote the binary outcome and $X(s_i)$ denote the set of covariates at location i, for $i = 1, \ldots, n$. A probit model for the data is given by

$$Y(s_i) = I(Z_i \leqslant X(s_i)'\beta + w(s_i)) \text{ where } Z_i \overset{iid}{\sim} N(0,1) \perp\!\!\!\perp (X(s_i), w(s_i)). \tag{2}$$

Here $w(s_i)$ are the spatial random effects, customarily modeled as a Gaussian Process (GP), to account for spatial correlation in the response. Modeling $w(\cdot) \sim GP(0, C(\cdot,\cdot|\theta))$ we have

$w = (w(s_1), \ldots, w(s_n))' \sim N(0, C(\theta))$ where $C(\theta) = (C(s_i, s_j|\theta))$, and θ denotes the spatial parameters.

Letting $\Phi(\cdot)$ denote the cdf of a standard normal distribution, the joint likelihood of the data is given by

$$\prod_{i=1}^{n} \Phi(X(s_i)'\beta + w(s_i))^{Y(s_i)}(1 - \Phi(X(s_i)'\beta + w(s_i)))^{(1-Y(s_i))} \times N(w|0, C(\theta)), \tag{3}$$

where $N(w|0, C(\theta))$ denotes the density of w, given by the pdf of $N(0, C(\theta))$. Denote $Y = (Y(s_1), \ldots, Y(s_n))'$, $X = (X(s_1), \ldots, X(s_n))'$ and $\Phi_n(m, \Sigma)$ to be the cdf of a multivariate (n-dimenional) $N(0, \Sigma)$ distribution evaluated at m. As derived in Cao et al. (2022), the marginal likelihood of Y is given by

$$\begin{aligned}
p(Y) &= \int \prod_{i=1}^{n} \Phi(X(s_i)'\beta + w(s_i))^{Y(s_i)}(1 - \Phi(X(s_i)'\beta + w(s_i)))^{(1-Y(s_i))} \times N(w|0, C(\theta))dw \\
&= \int \prod_{i=1}^{n} \Phi((2Y(s_i) - 1)(X(s_i)'\beta + w(s_i))) \times N(w|0, C(\theta))dw \\
&= \int \Phi_n(DX\beta + Dw, I_n) \times N(w|0, C(\theta))dw, \\
&\qquad\qquad \text{where } D = \text{diag}(2Y(s_1) - 1, \ldots, 2Y(s_n) - 1) \\
&= \Phi_n(DX\beta, I_n + DC(\theta)D) \text{ by Lemma 7.1 of Azzalini and Capitanio (2014).}
\end{aligned}$$

Hence the marginal likelihood of Y is given by a multivariate normal cdf

$$p(Y) = \Phi_n(m, \Sigma) \text{ where } m = DX\beta, \Sigma = I_n + DC(\theta)D. \tag{4}$$

Following the separation of variables trick proposed in Genz (1992), such multivariate cdf can be evaluated as

$$\Phi_n(m, \Sigma) = E_u \prod_{i=1}^{n} u_i I(0 \leqslant u_i \leqslant e_i) = E_u \prod_{i=1}^{n} e_i \text{ where } u = (u_1, \ldots, u_n)' \text{ with } u_i \overset{iid}{\sim} U[0, 1] \tag{5}$$

and the upper-limit vector $e = (e_1, \ldots, e_n)'$ is given by

$$e_1 = \Phi(m_1/l_{11}), \; e_i = \Phi\left(\left[m_i - \sum_{j=1}^{i-1} l_{ij}\Phi^{-1}(u_j e_j)\right]/l_{ii}\right), \; \forall i \geqslant 2. \tag{6}$$

Here $L = (l_{ij})$ is the lower-triangular Cholesky factor of Σ, i.e., $\Sigma = LL'$.

2.1 Nearest Neighbor GP Cholesky Factors

Evaluating the marginal probabilities in (4) requires computing the Cholesky factor L of $\Sigma = I_n + DC(\theta)D$ for generating the quantities in (6). Cholesky factorization typically needs $O(n^3)$ floating point operations or *flops*. Cao et al. (2022) reduced the complexity to $O(n^{5/2})$ by using a tile-low-rank approximation of Σ. In this manuscript, we propose an improved $O(n^2)$-complexity algorithm using Cholesky factors from Nearest neighbor Gaussian Process (NNGP, Datta et al., 2016a) covariance matrices. NNGP constructs a valid joint distribution for w on a set of locations

$\{s_1, \ldots, s_n\}$ by sequentially specifying $w(s_i) \mid w(s_1), \ldots, w(s_{i-1})$ only using GP spatial correlations between $w(s_i)$ and $w(s)$ for m-nearest neighbors s of s_i among $s_1, \ldots, s_{i-1}$. This specification was proposed as a scheme for GP likelihood approximation originally by Vecchia (1988) and was shown to correspond to a multivariate Gaussian distribution with a valid covariance matrix in Datta et al. (2016b).

The key rationale for motivation and success of NNGP as an excellent surrogate for full GP is that if the full GP covariance function monotonically decreases with distance, the m-nearest neighbors constitute the set of m locations of among $s_1, \ldots, s_{i-1}$ which has the highest correlation with s_i. Thus NNGP approximations are used either on the latent spatial random effects w with covariance matrix C, or, for Gaussian responses, directly on the response vector after marginalizing out w which has the covariance matrix of the form $C + \tau^2 I$, τ^2 denoting the unstructured (non-spatial) variance. If the covariance function $C(\cdot, \cdot \mid \theta)$ monotonically decreases with distance then so do the entries of both the matrices C and $C + \tau^2 I_n$ (Finley et al., 2019), and replacing either of these covariances with their respective NNGP analogs is justified and works well.

For spatial probit linear mixed model, we propose approximating the covariance matrix $\Sigma = I_n + DCD$ with a NNGP. The justification is that the off-diagonal entries of $\Sigma = (\sigma_{ij})$ satisfy

$$|\sigma_{ij}| = |(2Y(s_i) - 1)C(s_i, s_j|\theta)(2Y(s_j) - 1)| = |C(s_i, s_j|\theta)|$$

as $Y(s_i)$'s are binary. Hence, the absolute values of the covariances of Σ still decrease with distances and the principles of using NNGP hold.

Let $\widetilde{\Sigma}$ denotes the NNGP covariance matrix corresponding to Σ. It is well-known (Datta et al., 2016b) that $\widetilde{\Sigma}$ has the following properties:

$$\widetilde{\Sigma}^{-1} = (I - A)' F^{-1} (I - A)$$

where A is a sparse strictly lower-triangular matrix and $F = \text{diag}(f_1, \ldots, f_n)$ is diagonal, both of which can be evaluated in $O(n)$ flops. The lower-triangular Cholesky factor $\widetilde{L}$ of $\widetilde{\Sigma}$ is then given by $\widetilde{L} = (\tilde{l}_{ij}) = (I - A)^{-1} F^{1/2}$. Letting $\tilde{l}_{\cdot j}$ denote the jth column of $\widetilde{L}$, and η_j the n-dimensional vector with 1 at the jth position and 0's elsewhere, one can solve for $\tilde{l}_{\cdot j}$ as

$$\tilde{l}_{\cdot j} = (I - A)^{-1} f_j^{\frac{1}{2}} \eta_j \iff (I - A)\tilde{l}_{\cdot j} = f_j^{\frac{1}{2}} \eta_j \iff \tilde{l}_{\cdot j} = \text{trsolve}(I - A, f_j^{\frac{1}{2}} \eta_j), \tag{7}$$

where `trsolve` denotes solution of a triangular linear system. As A is strictly lower triangular with at-most $O(1)$ entries per row, the linear system in (7) can be solved in $O(n)$ flops (see Saha and Datta, 2018a; Datta, 2021, for the algorithm). Repeating this for $j = 1, \ldots, n$, the entire Cholesky factor $\widetilde{L}$ can be obtained in $O(n^2)$ flops or in $O(n^2/K)$ flops if parallelized over K computing cores as the solves for $\tilde{l}_{\cdot j}$ for different j's can proceed in an embarrassingly parallel manner.

2.2 Predictions in Probit Model

For prediction at a new location s_{n+1}, let $Y^* = (Y(s_1), \ldots, Y(s_n), Y(s_{n+1}))'$. Define the quantities X^*, D^* and C^* analogous to X, D and C respectively but for the $n + 1$ data-points. Let $m^* = D^* X^* \beta$ and $\widetilde{\Sigma}^*$ denote the NNGP matrix corresponding to $\Sigma^* = I_{n+1} + D^* C^* D^{*'}$. Then we have

$$\widehat{Y(s_{n+1})} = p(Y(s_{n+1}) = 1|Y) = \frac{\Phi_{n+1}(m^*, \widetilde{\Sigma}^*)}{\Phi_n(m, \widetilde{\Sigma})}. \tag{8}$$

Evaluating the denominator using (6) requires the Cholesky factor $\widetilde{L}$ of $\widetilde{\Sigma}$ which can be computed using $O(n^2)$ flops via (7). For the numerator, one needs to repeat the procedure using the Cholesky factor $\widetilde{L}^*$ of $\widetilde{\Sigma}^*$. However, another advantage of NNGP for this task is that, having computed $\widetilde{L}$, one can compute $\widetilde{L}^*$ in only $O(n)$ additional flops. To see this, let A^* and F^* respectively denote the $n+1$ dimensional analogs of A and F. Then we have

$$F^* = \text{diag}(F, f_{n+1}),\ A^* = \begin{pmatrix} A & 0 \\ a'_{n+1} & 0 \end{pmatrix} \text{ and } \widetilde{L}^* = \begin{pmatrix} \widetilde{L} & 0 \\ a'_{n+1}\widetilde{L} & f_{n+1}^{1/2} \end{pmatrix}. \tag{9}$$

Hence, the only added computations for $\widetilde{L}^*$ are computing the nearest-neighbor kriging weight-vector a_{n+1} and the nearest-neighbor kriging variance f_{n+1}, and then computing $v' = a'_{n+1}\widetilde{L}$. Obtaining both a_{n+1} and f_{n+1} are done in $O(1)$ flops as they only involve m-dimensional vectors and matrices, m being the number of nearest neighbors. To get v, we solve the sparse lower-triangular system $\widetilde{L}^{-1}v = a_{n+1}$ for v. This is akin to the linear system in (7) and can be done in $O(n)$ additional flops due to sparsity of $\widetilde{L}^{-1} = (I - A)F^{-1/2}$.

Having computed $\widetilde{L}$ and $\widetilde{L}^* = (\tilde{l}^*_{ij})$, we will evaluate the ratio in (8) via (5) and (6). Let $u_{n+1} \sim U[0,1] \perp\!\!\!\perp]u$ and $u^{*'} = (u', u_{n+1})$. Then, for the denominator, we first calculate the e_i's defined in (6) by replacing l_{ij} with $\tilde{l}_{ij}$. Subsequently, noting from (9) that $\widetilde{L}$ and $\widetilde{L}^*$ agree on the top $n \times n$ block, these e_i' can also be reused for evaluating the numerator. One only needs to compute an additional

$$e_{n+1} = \Phi\left(\left[m_{n+1} - \sum_{j=1}^{n} \tilde{l}^*_{n+1,j}\Phi^{-1}(u_j e_j)\right] / \tilde{l}^*_{n+1,n+1}\right). \tag{10}$$

Now we can compute the prediction $\widehat{Y}(s_{n+1})$ from (8) as

$$\widehat{Y(s_{n+1})} = \frac{\Phi_{n+1}(m^*, \widetilde{\Sigma}^*)}{\Phi_n(m, \widetilde{\Sigma})} = \frac{E_{u^*}\prod_{i=1}^{n+1} e_i}{E_u \prod_{i=1}^{n} e_i} \approx \frac{\frac{1}{R}\sum_{r=1}^{R}\prod_{i=1}^{n+1} e_i^{(r)}}{\frac{1}{R}\sum_{r=1}^{R}\prod_{i=1}^{n} e_i^{(r)}}. \tag{11}$$

Here the last approximation reflects the practice where the expectation is replaced by Monte-Carlo expectation using samples $u_i^{(r)}$ from $U[0,1]$ and generating the $e_i^{(r)}$ for $r = 1, \ldots, R$. The entire evaluation requires $O(n^2)$ flops – an improvement over the total $O(n^{5/2})$ flops ($O(n^{3/2})$ flops for each Monte Carlo sampling and $O(n^{5/2})$ flops for the Cholesky factor) algorithm proposed in Cao et al. (2022).

In practice, the parameters β and θ are unknown. They can be evaluated by cross-validation as outlined in Cao et al. (2022) using the mean square error (MSE) loss. Other loss functions like the mis-classification loss or the Kullback-Leibler divergence (KLD)-loss can also be considered that accounts for the binary nature of the response.

3 Illustrations

In this section, we illustrate the predictive performance and the computational scalability of the probit-NNGP model proposed in Section 2.1 in large simulation experiments and real world geospatial data. The implementation makes use of code snippets from the existing R package `BRISC` (Saha and Datta, 2018b), and heavily leverages Intel Math Kernel Library's threaded `BLAS` and `LAPACK` routines. In this section, we will be using 15 nearest neighbors (m in Section 2.1) and a location ordering scheme based on sum of coordinates for the NNGP based conditional probabilities derived in Section 2.

3.1 Simulation Experiment

For evaluating the performance of the proposed approach, we closely follow the simulation setup in Cao et al. (2022). We restrict ourselves to the scenario where $\beta = 0$ and simulate data on an equispaced $g \times g$ grid on a unit square. We simulate binary response Y at these $n = g^2$ locations as follows:

$$Y(s_i) \sim \text{Bernoulli}\left(\Phi\left(w\left(s_i\right)\right)\right); \quad \left(w(s_1), w(s_2), \ldots, w(s_n)\right) \sim N(0, C),$$

where C is the covariance matrix corresponding to exponential covariance kernel with spatial variance σ^2 and spatial decay ϕ, i.e.

$$C_{i,j} = \sigma^2 \exp\left(-\phi \left\| s_i - s_j \right\|\right).$$

To demonstrate the performance of probit-NNGP in a wide spectrum of sample sizes, we use the following choices of $g \in \{15, 25, 50, 100\}$. We first simulate the probabilities and the corresponding binary response for $g = 100$. For all of the other choices of g, we obtain the corresponding data by subsetting the full data over the suitable equally spaced subgrid. We set the spatial parameters as follows: $\sigma^2 = 1$, $\phi = \sqrt{30}$. To assess the predictive performance of probit-NNGP, we consider two sets of 100 out-of-sample locations in Cao et al. (2022):

1. 100 random locations from the unit square
2. a grid of 100 locations, non overlapping with the training dataset.

For each choice of g, we replicate the process 100 times. Instead of using the true parameter values for prediction purposes, we first estimate the parameters and next use the estimates for prediction. We estimate the parameters by optimizing the likelihood in Section 2. Performing a global optimization with respect to the spatial parameters can lead to significant computational overhead. In Cao et al. (2022), it was demonstrated that prediction performances are robust with respect to minor variation in parameter choices. We use a grid search based optimization over $\{\sigma^2, \phi\}$, by evaluating the likelihood (5) on a grid of values in the joint parameter space. We use a 10×10 grid on $\left[\sqrt{\frac{1}{2}}, \sqrt{\frac{3}{2}}\right] \times \left[\sqrt{15}, \sqrt{45}\right]$. We use mean squared error between the estimated predictive probabilities at the out-of-sample locations and the true ones.

In order to compare the accuracy of the proposed approach with a state-of-the-art solution for this problem, we also consider the minimax tilting method based evaluation of truncated normal (TN) distribution (Botev, 2017), implemented in `R` package `TruncatedNormal` (Botev and Belzile, 2021). The authors proposed a novel minimax tilting method to simulate i.i.d. observations from a multivariate Gaussian distribution. This provides with an efficient estimator for otherwise intractable cumulative distribution function of multivariate Gaussian distribution. Cao et al. (2022) developed a scalable probit model prediction based on Genz (1992) separation of variable algorithm. They achieve scalability through tile-low-rank (TLR) representation of the covariance matrix, to approximate the required cholesky decomposition in (6). This method is implemented with `R` package `tlrmvnmvt` (Cao et al., 2020, 2022) and the repository https://github.com/danieledurante/PredProbitGP. As far as the proposed approach (TLR) in Cao et al. (2022) is concerned, TN was used as the gold standard for accuracy whereas TLR vastly outperformed TN in computational overhead. Hence we compare the accuracy and runtime of probit-NNGP with that of both TN and TLR throughout the article. For implementing TN and TLR, we follow the model parameter choices in the Tutorial in the aforementioned repository. The packages to implement probit-NNGP, TN and TLR are implemented in `R`, with the base code written in `C/C++/RCPP`.

Table 1: MSE of predicted out-of-sample probabilities for probit-NNGP, TLR and TN, for different choices of n. Empty cells indicate the corresponding configurations to have a runtime exceeding our 24 hours computational budget.

Methods	$n = 15^2$	$n = 25^2$	$n = 50^2$	$n = 100^2$
(a) Out-of-sample random locations				
probit-NNGP	0.037	0.028	0.019	0.013
TLR	0.036	0.028	0.020	0.014
TN	0.037	0.028	–	–
(b) Out-of-sample grid locations				
probit-NNGP	0.031	0.026	0.018	0.013
TLR	0.030	0.025	0.020	0.014
TN	0.031	0.024	–	–

We set a constraint of 24 hours on the computational budget (for estimation and prediction together) and report the results for TN and TLR for values of g, where the whole prediction task for one replicate does not exceed the 24 hours. We note that both TN and TLR rely on parameter estimation by optimization of the probit likelihood. To be consistent in our empirical comparison, we use the same grid-search strategy and grid points as probit-NNGP for this estimation part.

Tables 1a and 1b demonstrate that the accuracy of the proposed approach is comparable to that of the TN for both sets of out-of-sample locations. Note that for some of the larger sample sizes, the TN method could not be completed within the total allotted time. In all the settings for which all the methods were able to be completed, the predictive performances of all the methods are nearly identical.

Next, we consider the computational overhead of the methods under consideration. The total runtimes of the concerned methods can be decomposed into two components, the first component being performing a grid search to optimize the (log)likelihood with respect to the spatial parameters. The second component consists of performing prediction on K out-of-sample locations with the parameter estimates obtained in the first step. Any NNGP based method incurs a one-time cost of determining the set of ordered nearest-neighbors for each location, which can be significantly time consuming for large datasets. Here we note that once we obtain the nearest-neighbor set, it can be used in the grid search and the subsequent prediction process. As far as prediction in out-of-sample data is concerned, unlike TN or TLR, probit-NNGP performs predictions on all locations together. Recall that, since the estimation process in probit-NNGP returns the $\widetilde{L}$ in (7), the $\widetilde{L}^*$ in (9) can be obtained in $O(n)$ computation. The reported runtime for TN for prediction at an out-of-sample location in Cao et al. (2022) includes the runtime needed to evaluate the denominator in (8) (i.e. our reported runtime in Table 2a). This shows that our reported total runtime for TN is consistent with that of Cao et al. (2022) for comparable sample sizes. The TLR runtimes in Table 2b also are consistent with the reported timings in Cao et al. (2022). Here we note that unlike TN, TLR predicts both the numerator and denominator of (8) simultaneously, hence given values of spatial parameters, the total runtime for prediction at one location only involves the timing reported in Table 2b.

The runtimes reveal the superior capability of the probit-NNGP model for scaling to settings involving massive geospatial data. For each runtime, we see that the probit-NNGP is orders of

Table 2: Runtime in seconds for probit-NNGP, TLR and TN for different choices of n. Empty cells indicate the corresponding configurations to have a runtime exceeding our 24 hours computational budget.

Methods	$n = 15^2$	$n = 25^2$	$n = 50^2$	$n = 100^2$
(a) Grid search for **one** parameter combination				
probit-NNGP	0.065	0.5	9	166
TL	0.57	2.9	28	187
TN	3.5	22	–	–
(b) Prediction at **one** out-of-sample location following estimation				
probit-NNGP	< 0.01	< 0.01	< 0.01	0.025
TLR	1.2	5.8	40	271
TN	3.5	22	–	–

Table 3: Runtime in seconds to predict at K locations with known parameters.

Methods	$n = 15^2$	$n = 25^2$	$n = 50^2$	$n = 100^2$
probit-NNGP	$0.065 + 0.01K$	$0.5 + 0.01K$	$9 + 0.01K$	$166 + 0.025K$
TL	$1.2K$	$5.8K$	$40K$	$271K$
TN	$3.5(K + 1)$	$22(K + 1)$	–	–

magnitude faster than TN and TLR. For the largest sample size of $100^2 = 10000$, NNGP completes estimation and prediction in less than a total of 3 minutes. As we have seen from Tables 1a and 1b, this expedition in computational speed comes at no discernible compromise in terms of predictive accuracy.

In case the parameters are known, i.e. we do not need to estimate the parameters. The time to predict at K locations in this scenario are given in Table 3.

3.2 Invasive Species Data Analysis

We demonstrate the utility of the probit-NNGP model in analysis of binary spatial data. The dataset under consideration informs about presence or absence of an invasive plant species *Celastrus orbiculatus* in the state of Connecticut, USA. The data set consists of 603 locations; the response is a presence–absence binary indicator (0 for absence) for *Celastrus orbiculatus*. We refer the readers to Banerjee and Gelfand (2006) for details on the data. The data set contains covariates, but consistent with the theme of the present article, we focus on assessing prediction performance only using the location information and hence do not use the covariates. In order to compare the predictive performances of the concerned methods, we first scale the 603 location coordinates to $[0, 1] \times [0, 1]$ unit square and divide the data into 500 training and 103 out-of-sample data. For all the models, we model the dependence structure in the response variable with an exponential covariance. We first obtain a crude estimate of the spatial parameters $\{\sigma^2, \phi\}$, by evaluating the likelihood (5) on a 25×10 equispaced grid on $[0, \sqrt{200}] \times [0, 10]$. For TLR, the Cholesky decomposition failed on the grid search, so we only compare the performances of probit-NNGP and TN in this data. Subsequent to estimating the spatial parameters, we obtain the predictive probabilities at the 103 held-out locations corresponding to probit-NNGP and TN.

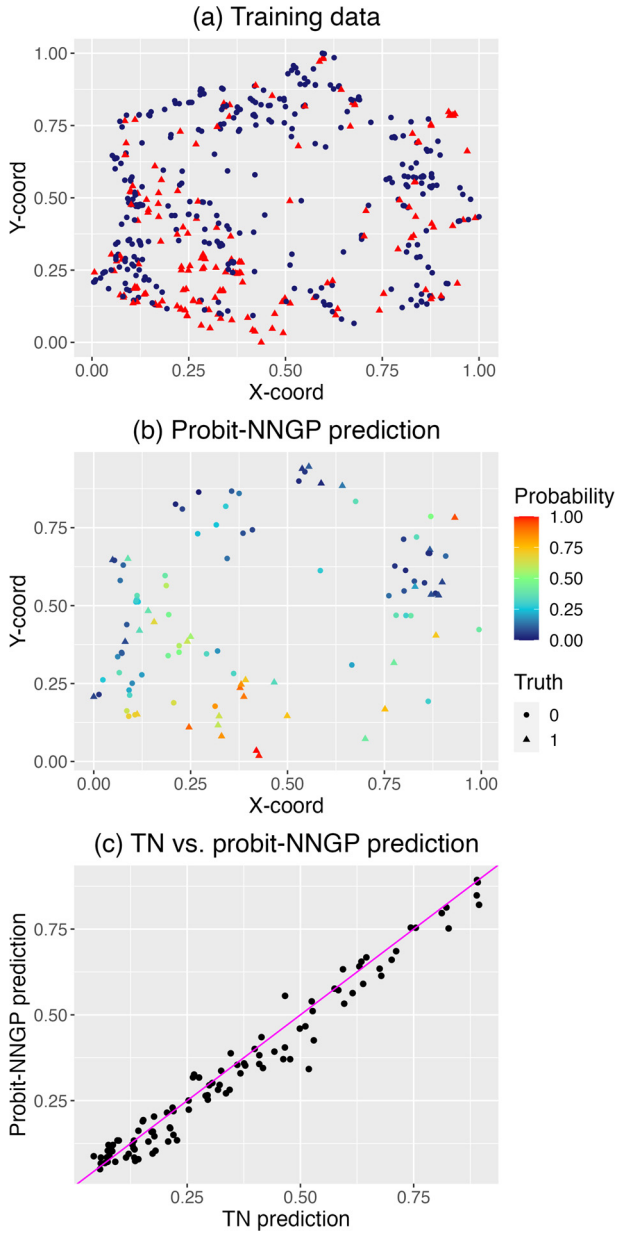

Figure 1: Invasive species data analysis with probit-NNGP. (a) shows the 500 training data with red denoting the presence and blue denoting the absence of *Celastrus orbiculatus*. (b) Shows the predicted probabilities of presence of *Celastrus orbiculatus*, along with the training data. (c) Shows that scatter plot of the predictions from the TN and the probit-NNGP models, demonstrating the similarity of the two.

Fig. 1 plots the training data and the predicted probabilities from probit-NNGP for the test data. The predictions are quite similar in nature with the mean square difference between the two being 0.002. This is also corroborated from the scatterplot in Panel (c) of Fig. 1 revealing the close alignment of the two sets of predictions. Unlike in the case of simulated data, we do not have access to the true probabilities at the held-out locations. Hence we measure the out-

Table 4: AUC and total (estimation + prediction) runtime in seconds for probit-NNGP and TN in *Celastrus orbiculatus* data.

Methods	AUC	Time
probit-NNGP	0.700	75
TN	0.662	4977

of-sample predictive performance via the area under the ROC curve (AUC) following Cao et al. (2022). Table 4 shows that the AUC for both the methods are similar. As the covariates effects for this data turned out to be statistically significant in Banerjee and Gelfand (2006), accounting for covariate effects may have lead to a better overall AUC for both the methods but would require optimization or grid-search over a higher dimensional space to obtain estimates of the regression coefficients. We also show the total runtime of probit-NNGP and TN in Table 4, which shows that probit-NNGP significantly outperforms TN in terms of computational overhead, even with the smaller sample size. Here, we note that the timing in Table 4 may seem higher than Table 2, but indeed are consistent with the ones reported in Table 2 due to the use of a denser grid in this scenario and Table 4 reporting the total runtime for all predictions as opposed to the run-time for one prediction location reported in Table 2a. For example, for TN, estimating the likelihood in (5) takes 14.1 seconds. We need to perform this operation 250 times for the grid search and 103 times for prediction at new locations. This gives us the reported time in Table 4.

4 Discussion

We proposed a modification of the spatial probit model prediction algorithm of Cao et al. (2022) using Nearest Neighbor Gaussian Process (NNGP) approximation. NNGP is a natural candidate for this problem as the covariance matrices involved in the probit marginal cdf's still respect monotonicity with respect to distance – the central rational for dimension-reduction using a few nearest neighbors, and that the approach relies on Cholesky factors which are conveniently obtained for NNGP. The NNGP-based algorithm proposed here theoretically reduces the computational complexity, which is reflected in considerably improved run times in the data analysis over competing methods. We also learn that this expedition in terms of computational efficiency does not sacrifice prediction accuracy.

As suggested by one of the reviewers of the paper, we also considered using Monte Carlo (MC) simulation to estimate $p(Y)$ directly using (4). The approach has a theoretical computational complexity of $\mathcal{O}(n)$, as both the approximate inverse Cholesky computation and one MC sample cost $\mathcal{O}(n)$. This method requires a high number of MC simulations to produce a stable estimate, as shown in the supplementary material. This makes this approach infeasible even for moderately high values of n.

One limitation of probit-NNGP is that, the cross-validation strategy only works if the number of covariates is small which limits the dimensionality of β. All the approaches (probit-NNGP, TN, TLR) perform prediction for a given set of parameters. In Cao et al. (2022) and in the present article, we use this for parameter estimation by maximising the likelihood through grid search. The overall computational time depends on the number of points in the parameters space. With zero mean, the dimension of the parameter space is equal to the number of spatial parameters, i.e. 2 for exponential covariance model. This requires a grid search over $O(K^2)$

points to search the parameter space, assuming K evaluation points for each parameter. If we want to account for unknown linear effect of d covariates, the dimension of the parameter space would be $d + 2$ (d parameters denoting the effect of d covariates, i.e. β in (2) of the article). In order to get a good coverage of the parameters space, we have to perform grid search over $O(K^{d+2})$ points, which will exponentially increase the overall cost of parameter estimation. We note that if we know the effect of the covariates, i.e. β is known or estimated apriori, the computational cost will not increase, as the unknown parameter space remains unchanged compared to the $\beta = 0$ scenario. Future research will explore alternative estimation strategies for moderate or large number of covariates. Future developments may also comprise generalizing the proposed methodology for spatial multinomial probit models that can replace spatially-varying multinomial logistic regression models to predict forest type groups across large forested landscapes (Finley et al., 2009).

Supplementary Material

This supplementary material contains discussion on why is it infeasible to directly use a Monte Carlo sampling to estimate $p(Y)$ in (4), evaluation of the algorithms under consideration with respect to misclassification error, and details of the code and data used in the article.

Funding

Abhirup Datta was partially supported by National Institute of Environmental Health Sciences (NIEHS) grant R01 ES033739 and by National Science Foundation (NSF) Division of Mathematical Sciences grant DMS-1915803. Sudipto Banerjee was partially supported by the National Science Foundation (NSF) from grants NSF/DMS 1916349 and NSF/IIS 1562303, and by the National Institute of Environmental Health Sciences (NIEHS) from grants R01ES030210 and 5R01ES027027.

References

Albert JH, Chib S (1993). Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association*, 88(422): 669–679.

Azzalini A, Capitanio A (2014). *The Skew-Normal and Related Families*, volume 3. Cambridge University Press.

Banerjee S, Gelfand AE (2006). Bayesian wombling: Curvilinear gradient assessment under spatial process models. *Journal of the American Statistical Association*, 101(476): 1487–1501.

Berrett C, Calder CA (2016). Bayesian spatial binary classification. *Spatial Statistics*, 16: 72–102.

Botev ZI (2017). The normal law under linear restrictions: simulation and estimation via minimax tilting. *Journal of the Royal Statistical Society, Series B, Statistical Methodology*, 79(1): 125–148.

Botev Z, Belzile L (2021). TruncatedNormal: Truncated Multivariate Normal and Student Distributions. R package version 2.2.2.

Cao J, Durante D, Genton MG (2022). Scalable computation of predictive probabilities in probit models with gaussian process priors. *Journal of Computational and Graphical Statistics*, 1–12. https://doi.org/10.1080/10618600.2022.2036614.

Cao J, Genton MG, Keyes DE, Turkiyyah GM (2022). tlrmvnmvt: Computing high-dimensional multivariate normal and student-t probabilities with low-rank methods in r. *Journal of Statistical Software*, 101: 1–25.

Cao J, Genton M, Keyes D, Turkiyyah G (2020). tlrmvnmvt: Low-Rank Methods for MVN and MVT Probabilities. R package version 1.1.0.

Datta A (2021). Nearest-neighbor sparse cholesky matrices in spatial statistics. *Wiley Interdisciplinary Reviews: Computational Statistics*, 14(5): e1574.

Datta A, Banerjee S, Finley AO, Gelfand AE (2016a). Hierarchical Nearest-Neighbor Gaussian Process Models for Large Geostatistical Datasets. *Journal of the American Statistical Association*, 111(514): 800–812.

Datta A, Banerjee S, Finley AO, Gelfand AE (2016b). On nearest-neighbor gaussian process models for massive spatial data. *Wiley Interdisciplinary Reviews: Computational Statistics*, 8(5): 162–171.

De Oliveira V (2000). Bayesian prediction of clipped gaussian random fields. *Computational Statistics & Data Analysis*, 34(3): 299–314.

De Oliveira V, Kedem B, Short DA (1997). Bayesian prediction of transformed gaussian random fields. *Journal of the American Statistical Association*, 92(440): 1422–1433.

Diggle PJ, Tawn JA, Moyeed RA (1998). Model-based geostatistics. *Journal of the Royal Statistical Society. Series C. Applied Statistics*, 47(3): 299–350.

Finley AO, Banerjee S, McRoberts RE (2009). Hierarchical spatial models for predicting tree species assemblages across large domains. *Annals of Applied Statistics*, 3(3): 1052–1079.

Finley AO, Datta A, Cook BD, Morton DC, Andersen HE, Banerjee S (2019). Efficient algorithms for bayesian nearest neighbor gaussian processes. *Journal of Computational and Graphical Statistics*, 28(2): 401–414.

Genz A (1992). Numerical computation of multivariate normal probabilities. *Journal of Computational and Graphical Statistics*, 1(2): 141–149.

Heagerty PJ, Lele SR (1998). A composite likelihood approach to binary spatial data. *Journal of the American Statistical Association*, 93(443): 1099–1111.

Lee Y, Nelder JA (1996). Hierarchical generalized linear models. *Journal of the Royal Statistical Society, Series B, Methodological*, 58(4): 619–656.

Saha A, Datta A (2018a). Brisc: bootstrap for rapid inference on spatial covariances. *Stat*, 7(1): e184.

Saha A, Datta A (2018b). BRISC: Fast Inference for Large Spatial Datasets using BRISC. R package version 0.1.0.

Vecchia AV (1988). Estimation and model identification for continuous spatial processes. *Journal of the Royal Statistical Society, Series B, Methodological*, 50(2): 297–312.

Zhang Z, Arellano-Valle RB, Genton MG, Huser R (2022). Tractable bayes of skew-elliptical link models for correlated binary data. *Biometrics*. https://doi.org/10.1111/biom.13731.