

Conjugate Bayesian Regression Models for Massive Geostatistical Data Sets



Sudipto Banerjee

Abstract Geographic Information Systems and related technologies are routinely used to construct massive amounts of spatially oriented data. This, in turn, has generated substantial interest among statisticians for modelling and analysing large spatial datasets. Scalable spatial process models have been found especially attractive due to their richness and flexibility and, particularly so in the Bayesian paradigm, due to their presence in hierarchical model settings. A substantial amount of research articles focus upon innovative theory and more complex model development, but limited attention has been accorded to approaches for easily implementable scalable hierarchical models for the practising scientist or spatial analyst. This article outlines how point-referenced spatial process models can be cast within the framework of conjugate Bayesian linear regression that can rapidly deliver inference on spatial processes. The approach directly samples from the exact joint posterior distribution of regression parameters, the latent process and the predictive random variables, and can be easily implemented on statistical programming environments such as R.

1 Introduction

The modelling and analysis for spatial and spatial-temporal data have witnessed an explosion of interest stemming from computerized Geographic Information Systems (GIS) and accompanying technologies. Bayesian hierarchical spatiotemporal process models have become widely deployed statistical tools for researchers to better understand the complex nature of spatial and temporal variability; see, for example, the books by Schabenberger and Gotway (2004), Gelfand et al. (2010), Cressie and Wikle (2011), and Banerjee et al. (2014) for a variety of methods and applications. Technological advances in diverse scientific disciplines have produced massive spatially and temporally indexed databases on a variety of health outcomes

S. Banerjee (✉)

UCLA Department of Biostatistics, University of California, Los Angeles, CA, USA

e-mail: sudipto@ucla.edu

© Springer Nature Switzerland AG 2020

A. Bekker et al. (eds.), *Computational and Methodological Statistics and Biostatistics*, Emerging Topics in Statistics and Biostatistics,

https://doi.org/10.1007/978-3-030-42196-0_10

255

and risk factors that are accessible to public health researchers, administrators, and policy-makers. This “data deluge” poses new challenges and critical barriers in data analysis for the next generation of biostatisticians and spatial data analysts.

Several methods and approaches, both classical and Bayesian, have been developed and evaluated to address the needs of spatial analysts encountering massive spatial datasets and increasingly complex scientific questions. There is already a substantial literature on modelling and analysing massive spatial datasets and a comprehensive review is beyond the scope of this article; see, e.g., Banerjee (2017) for a focused review on a couple of popular Bayesian approaches and Heaton et al. (2019) for a comparative evaluation for contemporary statistical methods for large spatial data. Here, I will provide develop how elementary conjugate Bayesian linear regression models can be exploited to provide a quick Bayesian analysis of massive spatial datasets. These approaches can be described as model-based solutions for very large spatial datasets that can be executed on modest computing environments.

2 Bayesian Modelling for Point-Referenced Data

Point-referenced spatial data are referenced by locations with coordinates (latitude-longitude, Easting-Northing etc.) and are customarily modelled using a random field. This random field is an uncountable set of random spatial surfaces, say $\{w(\ell) : \ell \in \mathcal{L}\}$, defined over a domain of interest \mathcal{L} . This uncountable set is modelled using a stochastic process which ensures the existence of a well-defined probability law for any finite collection of random variables from the underlying random field. Furthermore, the process models the spatial association among the random variables as a function of the locations, typically of the distance between pairs of locations.

For example, in spatial modelling \mathcal{L} is often assumed to be a subset of points in the Euclidean space \mathfrak{R}^d (usually $d = 2$ or 3) or, perhaps, a set of geographic coordinates over a sphere or ellipsoid. Such processes are specified with a *covariance function* $K_\theta(\ell, \ell')$ that gives the covariance between $w(\ell)$ and $w(\ell')$ for any two points ℓ and ℓ' in \mathcal{L} . For any finite collection $\mathcal{U} = \{\ell_1, \ell_2, \dots, \ell_n\}$ in \mathcal{L} , the covariance matrix for $w(\ell_i)$'s over \mathcal{U} is the $n \times n$ matrix K_θ whose (i, j) -th entry is the covariance $K_\theta(\ell_i, \ell_j)$. Covariance functions cannot be any function and need to ensure positive-definiteness of the resulting covariance matrix for any finite sample of locations in the domain. A rich literature exists on characterizations for covariance functions, their different properties and their impact on subsequent inference; see, e.g., any of the aforementioned books on spatial statistics. For any two finite subsets \mathcal{A} and \mathcal{B} of \mathcal{L} , we will let $K_\theta(\mathcal{A}, \mathcal{B})$ be the matrix whose (i, j) -th entry is the covariance function $K_\theta(\cdot, \cdot)$ evaluated between the i -th location in \mathcal{A} and the j -th location in \mathcal{B} . In particular, we denote the $n \times n$ spatial covariance matrix $K_\theta(\mathcal{L}, \mathcal{L})$ simply by K_θ .

A geostatistical setting customarily assumes a response or dependent variable $y(\ell)$ observed at a generic point ℓ along with a $p \times 1$ vector of spatially referenced

predictors $x(\ell)$. Model-based geostatistical data analysis customarily envisions a spatial regression model,

$$y(\ell) = x^\top(\ell)\beta + w(\ell) + \epsilon(\ell) , \tag{1}$$

where β is the $p \times 1$ vector of slopes, and the residual from the regression is the sum of a spatial process, $w(\ell) \sim GP(0, K_\theta(\cdot, \cdot))$ capturing spatial dependence, and an independent process, $\epsilon(\ell)$, modelling measurement error or fine scale variation attributed to disturbances at distances smaller than the minimum observed inter-site distance. A Bayesian spatial model can now be constructed from (1) as

$$p(\theta, w, \beta, \tau | y) \propto p(\theta, \beta, \tau) \times N(w | 0, K_\theta) \times N(y | X\beta + w, D_\tau) , \tag{2}$$

where $y = (y(\ell_1), y(\ell_2), \dots, y(\ell_n))^\top$ is the $n \times 1$ vector of observed outcomes, X is the $n \times p$ matrix (we assume $p < n$) of regressors with i -th row $x^\top(\ell_i)$ and the noise covariance matrix $D(\tau)$ represents measurement error or micro-scale variation and depends upon a set of variance parameters τ . A common specification is $D_\tau = \tau^2 I_n$, where τ^2 is called the ‘‘nugget.’’ The hierarchy is completed by assigning prior distributions to β , θ and τ .

The primary computational bottleneck emerges from the size of K_θ in computing (2). Since θ is unknown, each iteration of the model fitting algorithm will involve decomposing or factorising K_θ , which typically requires $\sim n^3$ floating point operations (flops). Memory requirements are of the order $\sim n^2$. These become prohibitive for large values of n when K_θ has no exploitable structure. For Gaussian likelihoods, one can integrate out the random effects w from (2) and work with the posterior

$$p(\theta, \beta, \tau | y) \propto p(\theta, \beta, \tau) \times N(y | X\beta, K_\theta + D_\tau) , \tag{3}$$

This reduces the parameter space to $\{\tau^2, \theta, \beta\}$, but one still needs to work with $K_\theta + D_\tau$, which is still $n \times n$. These settings are referred to as ‘‘big- n ’’ or ‘‘high-dimensional’’ problems in geostatistics and are widely encountered in environmental sciences today.

3 Conjugate Bayesian Linear Geostatistical Model

A conjugate Bayesian linear regression model is written as

$$y | \beta, \sigma^2 \sim N(X\beta, \sigma^2 V_y) ; \quad \beta | \sigma^2 \sim N(\beta | \mu_\beta, \sigma^2 V_\beta) ; \quad \sigma^2 \sim IG(a_\sigma, b_\sigma) , \tag{4}$$

where y is an $n \times 1$ vector of observations of the dependent variable, X is an $n \times p$ matrix (assumed to be of rank p) of independent variables (covariates or predictors)

and its first column is usually taken to be the intercept, V_y is a fixed (i.e., known) $n \times n$ positive definite matrix, μ_β , V_β , a_σ and b_σ are assumed to be fixed hyper-parameters specifying the prior distributions on the regression slopes β and the scale σ^2 . This model is easily tractable and the posterior distribution is

$$p(\beta, \sigma^2 | y) = \underbrace{IG(\sigma^2 | a_\sigma^*, b_\sigma^*)}_{p(\sigma^2 | y)} \times \underbrace{N(\beta | Mm, \sigma^2 M)}_{p(\beta | \sigma^2, y)}, \tag{5}$$

where $a_\sigma^* = a_\sigma + n/2$, $b_\sigma^* = b_\sigma + (1/2) \left\{ \mu_\beta^\top V_\beta^{-1} \mu_\beta + y^\top V_y^{-1} y - m^\top Mm \right\}$, $M^{-1} = V_\beta^{-1} + X^\top V_y^{-1} X$ and $m = V_\beta^{-1} \mu_\beta + X^\top V_y^{-1} y$. Sampling from the joint posterior distribution of $\{\beta, \sigma^2\}$ is achieved by first sampling $\sigma^2 \sim IG(a_\sigma^*, b_\sigma^*)$ and then sampling $\beta \sim N(Mm, \sigma^2 M)$ for each sampled σ^2 . This yields marginal posterior samples from $p(\beta | y)$, which is a non-central multivariate t distribution but we do not need to work with its complicated density function. See Gelman et al. (2013) for further details on the conjugate Bayesian linear regression model and sampling from its posterior.

We will adapt (4) to accommodate (2) or (3). Let us first consider (3) with the customary specification $D_\tau = \tau^2 I$ and let $K_\theta = \sigma^2 R(\phi)$, where $R(\phi)$ is a correlation matrix whose entries are given by a correlation function $\rho(\phi; \ell_i, \ell_j)$. Thus, $\theta = \{\sigma^2, \phi\}$, where σ^2 is the spatial variance component and ϕ is a spatial decay parameter controlling the rate at which the spatial correlation decays with separation between points. A simple example is $\rho(\phi; \ell_i, \ell_j) = \exp(-\phi \|\ell_i - \ell_j\|)$, although much richer choices are available (Banerjee et al. 2014, see, e.g., Ch 3 in). Therefore, we can write $K_\theta = \sigma^2 V_y$, where $V_y = R(\phi) + \delta^2 I$ and $\delta^2 = \tau^2 / \sigma^2$ is the ratio between the ‘‘noise’’ variance and ‘‘spatial’’ variance. If we assume that ϕ and δ^2 are fixed and that the prior on $\{\beta, \sigma^2\}$ are as in (4), then we have reduced (3) to (4) and direct sampling from its posterior is easily achieved as described below (5). We will return to the issue of fixing $\{\phi, \delta^2\}$ shortly.

Let us turn to accommodating (2) within (4), which would include directly sampling the spatial random effects w from their marginal posterior $p(w | y)$. Here, it is instructive to write the joint distribution of y and w in (2) as a linear model,

$$\underbrace{\begin{bmatrix} y \\ \mu_\beta \\ 0 \end{bmatrix}}_{y_*} = \underbrace{\begin{bmatrix} X & I_n \\ I_p & O \\ O & I_n \end{bmatrix}}_{X_*} \underbrace{\begin{bmatrix} \beta \\ w \end{bmatrix}}_{\gamma} + \underbrace{\begin{bmatrix} \eta_1 \\ \eta_2 \\ \eta_3 \end{bmatrix}}_{\eta}, \tag{6}$$

where $\eta \sim N(0, \sigma^2 V_{y_*})$ and $V_{y_*} = \begin{bmatrix} \delta^2 I_n & O & O \\ O & V_\beta & O \\ O & O & R(\phi) \end{bmatrix}$. If we assume that δ^2 and ϕ are fixed at known values, then V_{y_*} is fixed. Under this parametrisation, we have a conjugate Bayesian linear regression model $y_* = X_* \gamma + \eta$, where γ has a flat prior

and $\sigma^2 \sim IG(a_\sigma, b_\sigma)$. Thus,

$$p(\gamma, \sigma^2 | y) = \underbrace{IG(\sigma^2 | a_\sigma^*, b_\sigma^*)}_{p(\sigma^2 | y)} \times \underbrace{N(\beta | M_* m_*, \sigma^2 M_*)}_{p(\gamma | \sigma^2, y)}, \tag{7}$$

where $a_\sigma^* = a_\sigma + (2n + p)/2$, $b_\sigma^* = b_\sigma + (1/2) \{y_*^\top V_{y_*}^{-1} y_* - m_*^\top M_* m_*\}$, $M_*^{-1} = X_*^\top V_{y_*}^{-1} X_*$ and $m_* = X_*^\top V_{y_*}^{-1} y_*$. Note that the posterior mean of γ is given by $\hat{\gamma} = Mm = \left(X_*^\top V_{y_*}^{-1} X_*\right)^{-1} X_*^\top V_{y_*}^{-1} y_*$, which is the generalized least squares estimate obtained from the augmented linear system in (6). Sampling from the posterior proceeds analogous to that described below (5).

From the preceding account we see that fixing the spatial range decay parameter ϕ and the noise-to-spatial variance ratio δ^2 casts the Bayesian geostatistical model into a conjugate framework that will allow inference on $\{\beta, w, \sigma^2\}$. Note that multiplying the posterior samples of σ^2 by the fixed quantity δ^2 fetches us the posterior samples of τ^2 . Therefore, the uncertainty quantification is entirely lost only for the spatial range parameter ϕ and partially for one of the variance components due to fixing their ratio. This, however, provides the computational advantage that inference can be carried out without resorting to expensive iterative algorithms such as Markov chain Monte Carlo that require several iterations before sampling from the posterior distribution. This computational benefit becomes especially relevant when handling massive spatial data. Furthermore, fixing the values of δ^2 and ϕ is not entirely unreasonable given that the identifiability of these parameters from the data are known to be problematic and thwarts posterior learning in any case. Nevertheless, the inference will depend upon these fixed parameters so we discuss a practical approach to fix ϕ and δ^2 at reasonable values.

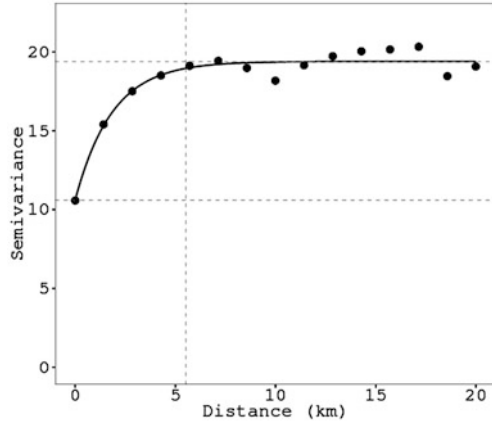
One simple approach to setting values for ϕ and δ^2 is by conducting some simple spatial exploratory data analysis using the ‘‘variogram’’. The variogram for a zero-centred spatial process $w(\ell)$ is defined as

$$E[w(\ell + h) - w(\ell)]^2 = \text{var}\{w(\ell + h) - w(\ell)\} = 2\gamma(h) , \tag{8}$$

which is meaningful only if the above expression depends solely on h and, whereupon, $\gamma(h)$ is called the ‘‘semivariogram’’. If the process $w(\ell)$ is weakly stationary in the sense that the covariance between $w(\ell)$ and $w(\ell')$ is a function only of the separation $h = \ell' - \ell$, then a simple calculation reveals that $\gamma(h) = K_\theta(0) - K_\theta(h)$, where $K_\theta(\ell, \ell') = K_\theta(\ell' - \ell) = K(h)$. The variogram is usually computed for the observations $y(\ell)$ or for the residuals from a linear model to ascertain the presence of spatial structure underlying the data after adjusting for explanatory variables.

Several practical algorithms exist for empirically calculating the variogram (or semivariogram) from observations by approximating (8) using finite sample moments. Many of these methods for variograms are now offered in user-friendly R packages hosted by the Comprehensive R Archive Network (CRAN) (<https://cran.r-project.org/>).

Fig. 1 Variogram of the residuals from non-spatial regression indicates strong spatial pattern



r-project.org). As one example, Finley et al. (2019) investigate the impact of tree cover and occurrence of forest fires on forest height. They first fit an ordinary linear regression of the form $y_{FH} = \beta_0 + \beta_1 x_{\text{tree}} + \beta_2 x_{\text{fire}} + \epsilon$ and then compute a variogram for the residuals from the ordinary linear regression.

Figure 1 depicts the variogram, which helps glean from three process parameters. The lower horizontal line represents the “nugget” or the micro-scale variation captured by the measurement error variance component τ^2 . The top horizontal line represents the “sill” (or ceiling) which is the total variation captured by $\sigma^2 + \tau^2$. Therefore, the difference between the two horizontal lines is called the “partial sill” and is captured by σ^2 . Finally, the vertical line represents the distance beyond which the variogram flattens or the covariance tends to zero. One can provide “eye-ball” estimates for these quantities and, in particular, fix the values of ϕ and $\delta^2 = \tau^2/\sigma^2$. Fixing these values from the variogram yields the desired highly accessible conjugate framework and the models can be estimated without resorting to Markov chain Monte Carlo (MCMC) as described earlier.

4 Bayesian Modelling for Massive Spatial Data

Conjugate models can be estimated by sampling directly from their joint posterior density and, therefore, completely obviates problems associated with MCMC convergence. This is a major computational benefit. However, the challenges in analysing massive spatial data do not quite end here. When the number of spatial locations providing measurements are in the order of millions as in Finley et al. (2019), then the matrices K_θ , V_y or V_{y^*} that we encountered earlier in different model parametrisations will be too massive to be efficiently loaded on to the machine’s CPU, let alone be computed with. This precludes efficient likelihood computations and has led several researchers to propose models specifically adapted

for spatial analysis. We briefly present adaptations of (6) using two different classes of models for massive spatial data: (1) low-rank process models and (2) nearest-neighbour Gaussian process models.

In low rank models, the spatial process is approximated as $w(\ell) \approx b_\theta^\top(\ell)z$, where $b_\theta(\ell)$ is an $r \times 1$ vector of r basis functions, each evaluated at ℓ , and z is an $r \times 1$ vector of coefficients. This means that the $n \times 1$ spatial effect w in (2) is replaced by $B_\theta z$, where B_θ is the $n \times r$ matrix whose i -th row is $b_\theta^\top(\ell_i)$. Dimension reduction is achieved by fixing r to be much smaller than n so that we only deal with r random effects instead of n . The framework in (6) can be easily adapted to this situation as below:

$$\underbrace{\begin{bmatrix} y \\ \mu_\beta \\ 0 \end{bmatrix}}_{y_*} = \underbrace{\begin{bmatrix} X & B_\theta \\ I_p & O \\ O & I_r \end{bmatrix}}_{X_*} \underbrace{\begin{bmatrix} \beta \\ z \end{bmatrix}}_{\gamma} + \underbrace{\begin{bmatrix} \eta_1 \\ \eta_2 \\ \eta_3 \end{bmatrix}}_{\eta}, \tag{9}$$

where $\eta \sim N(0, \sigma^2 V_{y_*})$ and $V_{y_*} = \begin{bmatrix} \delta^2 I_n & O & O \\ O & V_\beta & O \\ O & O & V_z \end{bmatrix}$ is $(n + p + r) \times (n + p + r)$

and fixed, and V_z is now $r \times r$ instead of the $n \times n$ matrix $R(\phi)$ in (6). Benefits accrue in terms of storage and the number of floating point operations (flops) when conducting the exact conjugate Bayesian analysis for this model. Note that the marginal density $p(y_* | \gamma, \theta, \tau)$ corresponds to the linear model $y_* = X_* \hat{\gamma} + \eta$, where $\hat{\gamma}$ is the generalized least square estimate of γ obtained by solving the linear system $X_*^\top V_{y_*}^{-1} X_* \gamma = X_*^\top V_{y_*}^{-1} y_*$. Computational benefits accrue from the block diagonal structure of V_{y_*} . To be precise, let $V_z^{1/2}$ and $V_\beta^{1/2}$ be matrix square roots of V_z and V_β , respectively. For example, $V_\beta^{1/2}$ and $V_z^{1/2}$ can be the triangular (upper or lower) Cholesky factor of the $r \times r$ matrices V_β and V_z , respectively. Then, the corresponding Cholesky factor of V_{y_*} is given by the block diagonal matrix

$$V_{y_*}^{1/2} = \begin{bmatrix} \delta I_n & O & O \\ O & V_\beta^{1/2} & O \\ O & O & V_z^{1/2} \end{bmatrix}. \text{ Once we obtain the square root } V_{y_*}^{1/2}, \text{ we can make}$$

the transformations $\tilde{y}_* = V_{y_*}^{-1/2} y_*$ and $\tilde{X}_* = V_{y_*}^{-1/2} X_*$, where $V_{y_*}^{-1/2}$ is cheaply obtained from $V_{y_*}^{1/2}$ because it inverts only a triangular matrix. Now the posterior mean $\hat{\gamma}$ can be obtained using ordinary least squares from the model $\tilde{y}_* = \tilde{X}_* \hat{\gamma} + e_*$, where $e_* \sim N(0, I_{n+p+r})$. Banerjee (2017) provides a more detailed discussion on hierarchical low-rank models, biases they induce and how bias-adjustments and improvements can be made.

Low-rank models continue to be popular choices for analysing spatial data. The cost for fitting low-rank models typically decrease from $O(n^3)$ to $O(nr^2 + r^3) \approx O(nr^2)$ flops since $n \gg r$. However, when n is large, empirical investigations suggest that r must be fairly large to adequately approximate the original process and

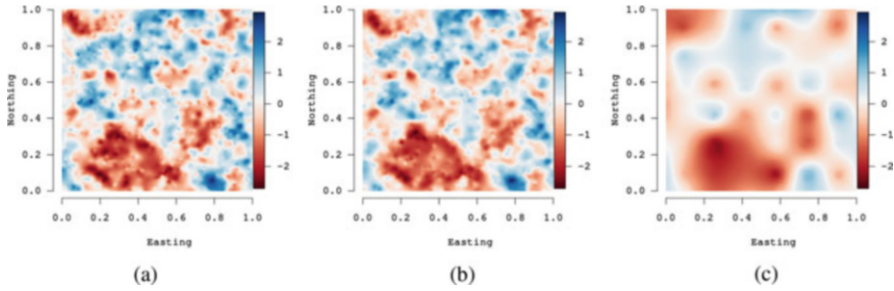


Fig. 2 Comparing estimates of a simulated random field using a full-rank Gaussian Process (Full GP) and a Gaussian Predictive process (PPGP) with 64 knots. The oversmoothing by the low-rank model is evident. (a) True w . (b) Full GP. (c) PPGP 64 knots

the nr^2 flops become exorbitant. Furthermore, low-rank models can perform poorly depending upon the smoothness of the underlying process or when neighbouring observations are strongly correlated and the spatial signal dominates the noise.

As an example, consider part of the simulation experiment presented in Datta et al. (2016a), where a spatial random field was generated over a unit square using a Gaussian process with fixed spatial process parameters over a set of 2500 locations. We then fit a full Gaussian process model and a particular low-rank model called the predictive process model (Banerjee et al. 2008) with 64 knots. Figure 2 presents the results. While the estimated random field from the full Gaussian process is almost indistinguishable from the true random field, the surface obtained from the predictive process with 64 locations substantially oversmooths. This oversmoothing can be mitigated by using a larger number of knots, but this adds to the computational burden.

Figure 2 serves to reinforce findings that low-rank models may be limited in their ability to produce accurate representation of the underlying process at massive scales. They will need a considerably larger number of basis functions to capture the features of the process and will require substantial computational resources for emulating results from a full GP. As the demands for analysing large spatial datasets increase from the order of $\sim 10^4$ to $\sim 10^6$ locations, low-rank models may struggle to deliver acceptable inference. In this regard, enhancements such as the multi-resolution predictive process approximations (Katzfuss 2017) are highly promising.

An alternative to low-rank models is to develop full rank models that can exploit sparsity. Here, too, there are different options. One approach draws on the concept of sparse precision matrices. There are numerous specifications but the one that is effective, scalable and easy to compute is based upon modelling the Cholesky decomposition of the precision matrix of w in a sparse manner.

Writing $N(w | 0, \sigma^2 R_\phi)$ as $p(w_1) \prod_{i=2}^n p(w_i | w_1, w_2, \dots, w_{i-1})$ is equivalent to the following set of linear models,

$$w_1 = 0 + \eta_1 \quad \text{and} \quad w_i = a_{i1}w_1 + a_{i2}w_2 + \dots + a_{i,i-1}w_{i-1} + \eta_i \quad \text{for } i = 2, \dots, n,$$

or, more compactly, simply $w = Aw + \eta$, where A is $n \times n$ strictly lower-triangular with elements $a_{ij} = 0$ whenever $j \geq i$ and $\eta \sim N(0, D)$ and D is diagonal with diagonal entries $d_{11} = \text{var}\{w_1\}$ and $d_{ii} = \text{var}\{w_i | w_j : j < i\}$ for $i = 2, \dots, n$. From the structure of A it is evident that $I - A$ is unit lower-triangular, hence nonsingular, and $R_\phi = (I - A)^{-1} D (I - A)^{-\top}$.

We now introduce sparsity in $R_\phi^{-1} = (I - A)^\top D (I - A)$ by letting $a_{ij} = 0$ whenever $j \geq i$ (since A is strictly lower-triangular) and also whenever ℓ_j is not among the m nearest neighbours of ℓ_i , where m is fixed by the user to be a small number. It turns out that a very effective approximation emerges by recognising that the lower-triangular elements of A are precisely the coefficients of a linear combination of $w(\ell_j)$'s equating to the conditional expectation $E[w(\ell_i) | \{w(\ell_j) : j < i\}]$. Thus, the $m \times 1$ vector \tilde{a}_i of non-zero entries in the i -th row of A are obtained by solving the $m \times m$ linear system $\tilde{R}_{\phi, N_i, N_i} \tilde{a}_i = R_{\phi, N_i, i}$, where $\tilde{R}_{\phi, N_i, N_i}$ is the $m \times m$ principal submatrix extracted from R_ϕ corresponding to the m neighbours of i (indexed by elements of a neighbour set N_i) and $R_{\phi, N_i, i}$ is the $m \times 1$ vector extracted by choosing the m indices in N_i from the i -th column of R_ϕ . Once \tilde{a}_i is obtained, the i -th diagonal entry of D is obtained as $d_{ii} = R_\phi[i, i] - \tilde{a}_i^\top R_{\phi, N_i, i}$. These computations need to be carried out for each $i = 2, \dots, n$ (note that for $i = 1$, $d_{11} = \sigma^2$ and $a_{11} = 0$), but m can be kept very small (say 5 or 10 even if $n 10^7$) so that the expense is $O(nm^3)$ and still feasible. The details can be found in Banerjee (2017). This notion is familiar in Gaussian Graphical models and have been used by Vecchia (1988) and, more recently, by Datta et al. (2016a) and Finley et al. (2019) to tackle massive amounts of spatial locations.

The framework in (6) now assumes the form

$$\underbrace{\begin{bmatrix} y \\ \mu_\beta \\ 0 \end{bmatrix}}_{y_*} = \underbrace{\begin{bmatrix} X & I_n \\ I_p & O \\ O & D^{-1/2}(I - A) \end{bmatrix}}_{X_*} \underbrace{\begin{bmatrix} \beta \\ w \end{bmatrix}}_{\gamma} + \underbrace{\begin{bmatrix} \eta_1 \\ \eta_2 \\ \eta_3 \end{bmatrix}}_{\eta}, \tag{10}$$

where $\eta \sim N(0, \sigma^2 V_{y_*})$ and $V_{y_*} = \begin{bmatrix} \delta^2 I_n & O & O \\ O & V_\beta & O \\ O & O & I_n \end{bmatrix}$ is $(2n + p) \times (2n + p)$ and

fixed with much greater sparsity. While this approach can also be subsumed into the framework of (6), its efficient implementation on standard computing architectures needs careful consideration and involves solving a large linear system with $(n + p) \times (n + p)$ coefficient matrix $X_*^\top X_*$. This matrix is large, but is sparse because of sparsity in $(I - A)^\top D^{-1}(I - A)$. Since $(I - A)$ has at most $m + 1$ nonzero entries in each row, an upper bound of nonzero entries in $(I - A)$ is $n(m + 1)$ and, therefore, the upper bound in $(I - A)^\top D^{-1}(I - A)$ is $n(m + 1)^2$. This sparsity can be exploited by sparse linear solvers such as conjugate gradient methods that can be implemented on modest computing environments.

Sampling from the joint posterior distribution $p(\gamma, \sigma^2 | y_*)$ is achieved in the following manner. First, the least-squares estimate $\hat{\gamma}$ is obtained using a sparse least-square solver using a preconditioned conjugate gradient algorithm. Subsequently, σ^2 is sampled from its marginal posterior density $IG(a_*, b_*)$, where $a_* = a_\sigma + n/2$ and $b_* = b_\sigma + (1/2)(y_* - X_* \text{gamma})^\top (y_* - X_* \hat{\gamma})$, and then for each sampled σ^2 , γ is sampled from $N\left(\hat{\gamma}, \sigma^2 \left(X_*^\top V_{y_*}^{-1} X_*\right)^{-1}\right)$. Details on such implementations can be found in a recent article by.

5 Spatial Prediction

Let $\tilde{\mathcal{L}} = \{\tilde{\ell}_1, \tilde{\ell}_2, \dots, \tilde{\ell}_{\tilde{n}}\}$ be a set of \tilde{n} locations where we wish to predict the outcome $y(\ell)$. Let \tilde{Y} be an $\tilde{n} \times 1$ vector with i -th element $\tilde{Y}(\tilde{\ell}_i)$ and let \tilde{w} be the $\tilde{n} \times 1$ vector with elements $w(\tilde{\ell}_i)$. The predictive model augments $p(\theta, w, \beta, \tau, y)$ to

$$p(\theta, \tau, \beta, w, y, \tilde{w}, \tilde{Y}) = p(\theta, \tau, \beta) \times p(w | \theta) \times p(y | \beta, w, \tau) \\ \times p(\tilde{w} | w, \theta) \times p(\tilde{Y} | \beta, \tilde{w}, \tau). \quad (11)$$

The factorisation in (11) implies that \tilde{Y} and w are conditionally independent of each other given \tilde{w} and β . Predictive inference for spatial data evaluates the posterior predictive distribution $p(\tilde{Y}, \tilde{w} | y)$. This is the joint posterior distribution for the outcomes and the spatial effects at locations in $\tilde{\mathcal{L}}$. This distribution is easily derived from (11) as

$$p(\tilde{Y}, \tilde{w}, \beta, w, \theta, \tau | y) \propto p(\beta, w, \theta, \tau | y) \times p(\tilde{w} | w, \theta) \times p(\tilde{Y} | \beta, \tilde{w}, \tau). \quad (12)$$

Sampling from (12) is achieved by first sampling $\{\beta, w, \theta, \tau\}$ from the posterior distribution $p(\beta, w, \theta, \tau | y)$. For each drawn sample, we make one draw of the $\tilde{n} \times 1$ vector \tilde{w} from $p(\tilde{w} | w, \theta)$ and then, using this sampled \tilde{w} , we make one draw of \tilde{Y} from $p(\tilde{Y} | \beta, \tilde{w}, \tau)$. The resulting samples of \tilde{w} and \tilde{Y} will be draws from the desired posterior predictive distribution $p(\tilde{w}, \tilde{Y} | y)$. This delivers inference on both the latent spatial random effect \tilde{w} and the outcome \tilde{Y} at arbitrary locations since \mathcal{L} can be any finite collection of samples. Summarizing these distributions by computing their sample means, standard errors, and the 2.5-th and 97.5-th quantiles (to produce a 95% credible interval) yields point estimates with associated uncertainty quantification.

It is instructive to see how the entire inference for Gaussian outcomes can be cast into an augmented linear regression model. The predictive model for \tilde{Y} can be written as a spatial regression

$$\tilde{Y} = \tilde{X}\beta + \tilde{w} + \tilde{\epsilon}; \quad \tilde{w} = Cw + \omega, \tag{13}$$

where \tilde{X} is the $\tilde{n} \times p$ matrix of predictors observed at locations in $\tilde{\mathcal{L}}$ and $\tilde{\epsilon} \sim N(0, \tilde{D}_\tau)$, where $\tilde{\epsilon}$ is the $\tilde{n} \times 1$ vector with elements $\epsilon(\tilde{\ell}_i)$. The second equation in (13) expresses the relationship between the spatial effects \tilde{w} across the unobserved locations in $\tilde{\mathcal{L}}$ and the spatial effects across the observed locations in \mathcal{L} . Since there is one underlying random field over the entire domain, the covariance function for the random field specifies the $\tilde{n} \times n$ coefficient matrix C . In particular, if $w \sim N(0, K_\theta)$, then $C = K_\theta(\tilde{\mathcal{L}}, \mathcal{L})K_\theta^{-1}$ and $\omega \sim N(0, F_\theta)$, where $F_\theta = K_\theta(\tilde{\mathcal{L}}, \tilde{\mathcal{L}}) - K_\theta(\tilde{\mathcal{L}}, \mathcal{L})K_\theta^{-1}K_\theta(\mathcal{L}, \tilde{\mathcal{L}})$. The model for the data and the predictions is combined into

$$\underbrace{\begin{bmatrix} y \\ \mu_\beta \\ 0 \\ 0 \\ 0 \end{bmatrix}}_{y_*} = \underbrace{\begin{bmatrix} X & I_n & O & O \\ I_p & O & O & O \\ O & C & -I_{\tilde{n}} & O \\ \tilde{X} & O & I_{\tilde{n}} & -I_{\tilde{n}} \end{bmatrix}}_{X_*} \underbrace{\begin{bmatrix} \beta \\ w \\ \tilde{w} \\ \tilde{Y} \end{bmatrix}}_{\gamma} + \underbrace{\begin{bmatrix} \eta_1 \\ \eta_2 \\ \eta_3 \\ \eta_4 \\ \eta_5 \end{bmatrix}}_{\eta}, \quad \text{where} \tag{14}$$

$$\eta \sim N \left(0, \begin{bmatrix} D_\tau & O & O & O & O \\ O & V_\beta & O & O & O \\ O & O & K_\theta & O & O \\ O & O & O & F_\theta & O \\ O & O & O & O & \tilde{D}_\tau \end{bmatrix} \right)$$

If locations where predictions are sought are fixed by study design, then fitting (14) using the Bayesian conjugate framework can be beneficial. On the other hand, one can first estimate $\{\beta, w, \sigma^2\}$ and store samples from their posterior distribution. Then, for any arbitrary set of points in $\tilde{\mathcal{L}}$, for each stored sample of the parameters we draw one sample of $\tilde{w} \sim N(Cw, F_\theta)$ followed by one draw of $\tilde{Y} \sim N(\tilde{X}\beta + \tilde{w}, \tilde{D}_\tau)$. The resulting $\{\tilde{w}, \tilde{Y}\}$ will be the desired posterior predictive samples for the latent spatial process and the unobserved outcomes.

6 An Example

We present a synopsis of the analysis by Zhang et al. (2019) of a spatial dataset from NASA comprising sea surface temperature observations over 2,827,252 spatial locations of which approximately 90% (2,544,527) were used for model fitting and the rest were withheld for cross-validatory predictive assessment. Details of the dataset can be found in <http://modis-atmos.gsfc.nasa.gov/index.html> and details on the analysis can be found in Zhang et al. (2019). The salient feature of the analysis

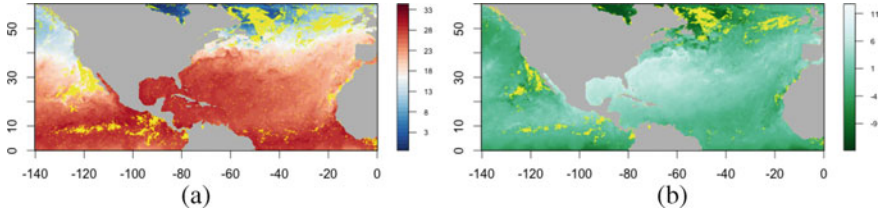


Fig. 3 Posterior predictive maps of sea-surface temperature and latent spatial effects. The land is colored by gray, locations in the ocean without observations are colored by yellow. **(a)** Posterior predictive map of sea-surface temperature. **(b)** Posterior predictive map of latent spatial effects

is that a conjugate Bayesian framework for the NNGP model as in (10) was able to deliver full inference including the estimation of the spatial latent effects in about 2387 s. Sampling from the posterior distribution was achieved using direct sampling as described below (10). Since this algorithm is fast and exact, it was run over a grid of values of $\{\delta^2, \phi\}$. For each such value, a posterior predictive assessment over the cross-validatory hold-out set was carried out and the value of $\{\delta^2, \phi\}$ producing the least root mean square prediction error (RMSPE) was selected as optimal inputs. Figure 3 presents the posterior predictive maps of (a) the response and (b) the latent spatial effects from the conjugate model.

7 Concluding Remarks

This short article has demonstrated how the familiar theory of conjugate Bayesian linear regression models can be adapted to spatial models and used effectively to analyse massive spatial datasets without requiring MCMC algorithms. The article has attempted to provide some insight into constructing highly scalable Bayesian hierarchical models for very large spatial datasets using low-rank and sparsity-inducing processes. Such models are increasingly being employed to answer complex scientific questions and analyse massive spatiotemporal datasets in the natural and environmental sciences. Exploratory data analysis tools such as the variogram can be used to fix the spatial decay parameter and the ratio between the spatial and non-spatial variance components. An alternative is a cross-validatory approach, where a grid of values of the process parameters is used and a fast and exact conjugate Bayesian analysis is performed for each of the values on the grid. The inference from the optimal value of the process parameters based upon RMSPE over hold-out locations is then presented. While these approaches may produce slightly shrunk credible and prediction intervals due to the effect of fixing a parameter, the effect is seen to be moderate in practical spatial analysis and the approach could form a useful tool for quick spatial analysis within the Bayesian paradigm for massive spatial datasets.

References

- Banerjee, S. (2017). High-dimensional Bayesian geostatistics. *Bayesian Analysis*, 12, 583–614.
- Banerjee, S., Carlin, B. P., & Gelfand, A. E. (2014). *Hierarchical modeling and analysis for spatial data*. Boca Raton, FL: CRC Press.
- Banerjee, S., Gelfand, A.E., Finley, A.O., & Sang, H. (2008). Gaussian predictive process models for large spatial datasets. *Journal of the Royal Statistical Society, Series B*, 70, 825–848.
- Cressie, N., & Wikle, C. K. (2011). *Statistics for spatio-temporal data*. Hoboken, NJ: Wiley.
- Datta, A., Banerjee, S., Finley, A. O., & Gelfand, A. E. (2016a). Hierarchical nearest-neighbor gaussian process models for large geostatistical datasets. *Journal of the American Statistical Association*, 111, 800–812. <http://dx.doi.org/10.1080/01621459.2015.1044091>
- Finley, A. O., Datta, A., Cook, B. C., Morton, D. C., Andersen, H. E., & Banerjee, S. (2019). Efficient algorithms for Bayesian nearest neighbor gaussian processes. *Journal of Computational and Graphical Statistics*, 28(2), 401–414.
- Gelfand, A. E., Diggle, P., Guttorp, P., & Fuentes, M. (2010). *Handbook of spatial statistics*. Boca Raton, FL: CRC Press.
- Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). *Bayesian data analysis. Chapman & Hall/CRC Texts in Statistical Science* (3rd edn.). Boca Raton, FL: Chapman & Hall/CRC.
- Heaton, M., Datta, A., Finley, A., Furrer, R., Guinness, J., Guhaniyogi, R., et al. (2019). Methods for analyzing large spatial data: A review and comparison. *Journal of Agricultural, Biological and Environmental Statistics*, 24(3), 398–425. <https://doi.org/10.1007/s13253-018-00348-w>
- Katzfuss, M. (2017). A multi-resolution approximation for massive spatial datasets. *Journal of the American Statistical Association*, 112, 201–214. <http://dx.doi.org/10.1080/01621459.2015.1123632>
- Schabenberger, O., & Gotway, C. A. (2004). *Statistical methods for spatial data analysis* (1st edn.). Boca Raton, FL: Chapman and Hall/CRC Press.
- Vecchia, A. V. (1988). Estimation and model identification for continuous spatial processes. *Journal of the Royal Statistical society, Series B*, 50, 297–312.
- Zhang, L., Datta, A., & Banerjee, S. (2019). Practical Bayesian modeling and inference for massive spatial datasets on modest computing environments. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, 12(3), 197–209. <https://doi.org/10.1002/sam.11413>