Spatial Factor Models for High-Dimensional and Large Spatial Data: An Application in Forest Variable Mapping

Daniel Taylor-Rodriguez[1], Andrew O. Finley[2,*], Abhirup Datta[3], Chad Babcock[4], Hans-Erik Andersen[5], Bruce D. Cook[6], Douglas C. Morton[6], Sudipto Banerjee[7]

[1]*Portland State University,* [2]*Michigan State University,* [3]*Johns Hopkins University,* [4]*University of Washington,* [5]*United States Forest Service,* [6]*National Aeronautics and Space Administration,* [7]*University of California Los Angeles,* * *Corresponding*

Supplementary Material

The supplementary materials include (1) background information on NNGPs and spatial factor models, (2) the sampling algorithm for the SF-NNGP, and (3) additional simulation results.

S1 Dimension reduction with factor models

For a general l-variate problem, where the dependence is specified through a Gaussian Process $\mathbf{w}^{\star}(\mathbf{s}) \sim \mathrm{GP}_l(\mathbf{0}, \mathcal{C}(\cdot \,|\, \boldsymbol{\theta}))$, the specification of $\mathcal{C}(\cdot|\boldsymbol{\theta})$ is complicated by the fact that it must be a nonnegative definite function, and it must meet the symmetry constraint $\mathcal{C}(\mathbf{h} \,|\, \boldsymbol{\theta}) = \mathcal{C}(-\mathbf{h} \,|\, \boldsymbol{\theta})$ (see Ver Hoef

and Barry, 1998; Chiles and Delfiner, 2009; Genton and Kleiber, 2015). In addition to the difficulty in specifying a suitable covariance function, the size of the LiDAR signals in our application makes modeling directly this joint high-dimensional spatial component a computationally daunting task. As such, we take advantage of the gains achieved using the spatial factor model structure, which reduces the computational burden in two ways. First, it dramatically reduces the dimensionality of the stochastic processes used, and second, it assumes that the multivariate stochastic processes considered are composed of independent univariate processes.

Under the SFM structure, the spatial dependence is introduced by defining the spatial process as $\mathbf{w}^{\star}(\mathbf{s}) = \mathbf{\Lambda}\mathbf{w}(\mathbf{s}) \sim \text{GP}(\mathbf{0}, \mathcal{H}(\cdot \,|\, \boldsymbol{\phi}))$, where $\mathbf{\Lambda}$ is a factor loadings matrix (commonly tall and skinny) and $\mathbf{w}(\mathbf{s})$ is a small-dimensional vector of independent spatial GP's, providing the non-separable multivariate cross-covariance function given by

$$\begin{aligned} \mathcal{H}(\mathbf{h} \,|\, \boldsymbol{\phi}) &= \text{cov}(\mathbf{\Lambda}\, \mathbf{w}(\mathbf{s}), \mathbf{\Lambda}\, \mathbf{w}(\mathbf{s}+\mathbf{h})) \\ &= \sum_{k=1}^{q} \rho_k(\mathbf{h}, \phi_k) \boldsymbol{\lambda}_k \boldsymbol{\lambda}_k' \;=\; \sum_{k=1}^{q} \mathcal{C}_k(\mathbf{h}, \phi_k) T_k, \end{aligned} \tag{S1.1}$$

for locations $\mathbf{s}, \mathbf{s}+\mathbf{h} \in \mathcal{D}$. Here, $\mathcal{C}_k(\mathbf{h}|\phi_k)$'s are univariate parametric correlation functions, and $\boldsymbol{\lambda}_k$ is the kth column of $\mathbf{\Lambda}$, which also corresponds to the eigenvector associated to the *only* positive eigenvalue of the rank one matrix T_k. This cross-covariance matrix is induced by q-variate ($q \leq l$) spatial fac-

tors $\mathbf{w}(\mathbf{s})$ with *independent* components $w_k(\mathbf{s}) \sim \mathrm{GP}(0, \mathcal{C}_k(\cdot \,|\, \phi_k))$. Hence, $\mathrm{var}(w_k(\mathbf{s})) = 1$, $\mathrm{cov}(w_k(\mathbf{s}), w_r(\mathbf{r})) = 0$ for $k \neq r$, and $\mathrm{cov}(w_k(\mathbf{s}), w_k(\mathbf{s}+\mathbf{h})) = \mathcal{C}_k(\mathbf{h} \,|\, \phi_k)$.

Additional constraints are required for factor models to be identifiable (Anderson, 2003). Nevertheless, conveniently with spatial factor models only two groups of orthogonal transformations lead to non-identifiability issues, as shown in Ren and Banerjee (2013) . The first of them is produced by an orthogonal matrix $\mathbf{P}_H$ resulting from the product of Householder reflectors, which is diagonal with 1's and -1's. The non-identifiability comes from the fact that $\mathbf{P}'_H \mathcal{H}(\mathbf{h} \,|\, \boldsymbol{\phi}) \mathbf{P}'_H = \mathcal{H}(\mathbf{h} \,|\, \boldsymbol{\phi})$. The second type are permutation matrices $\mathbf{P}_P$, given that $\boldsymbol{\Lambda}\mathbf{P}_P\mathbf{P}'_P\mathbf{w}(\mathbf{s}) \stackrel{d}{=} \boldsymbol{\Lambda}\mathbf{w}(\mathbf{s})$, where "$\stackrel{d}{=}$" represents equality in distribution. Both of these situations can be avoided, either through the conventional approach of making the upper triangle of the loadings matrix equal to 0 and its diagonal elements all equal to 1; or as in Ren and Banerjee (2013), by fixing the sign of one element in each column of $\boldsymbol{\Lambda}$, while enforcing an ordering constraint on the univariate correlation functions.

S2 Nearest Neighbor Gaussian Processes

In spite of the dimension reduction achieved with the factor model structure, given the formidable number of locations considered, even the factor model representation is prohibitive with dense Gaussian processes. Under a Bayesian approach, $q_w + q_v$ covariance matrices, each of dimension $n \times n$, have to be estimated and inverted at each iteration of the sampling algorithm. In view of this, we resort to the sparse approximation provided by the NNGP approach.

The Nearest Neighbor Gaussian Process approach belongs to the class of sparsity inducing methods that introduce zeros in the precision matrix to impose conditional independence, exploiting the graphical structure available for points distributed across space and/or time. The idea underlying this method is to derive a sparse approximation of a parent GP, which is a proper GP itself. The NNGP has been shown to provide an accurate and computationally efficient approximation to the dense parent GPs (see for example Datta et al., 2016b,c,a; Finley et al., 2017).

To elaborate, consider a univariate spatial Gaussian Process $w(\mathbf{s}) \sim \mathrm{GP}(0, \mathcal{C}(\cdot \,|\, \boldsymbol{\phi}))$ for $\mathbf{s} \in \mathcal{D} \subset \mathbb{R}^d$. Recall that when observed at a finite collection of locations $\mathcal{T} = \{\mathbf{s}_1, \ldots, \mathbf{s}_n\}$, the process constrained to these locations is such that $\mathbf{w} = (w(\mathbf{s}_1), \ldots, w(\mathbf{s}_n))' \sim \mathrm{N}_n(\mathbf{0}, \mathbf{C})$, with $\mathbf{C} =$

$\left(\left(\mathcal{C}(\|\mathbf{s}_i - \mathbf{s}_j\|; \boldsymbol{\phi})\right)\right)$. Alternatively, this joint density can be decomposed into the product of conditionals

$$p\left(w_1\right) p\left(w_2|w_1\right) \ldots p\left(w_i|w_j : 1 \leq j < i\right) \ldots p\left(w_n|w_j : 1 \leq j < n\right),$$

where $w_i = w(\mathbf{s}_i)$. This representation and the multivariate normality of $\mathbf{w}$ imply the linear model given by $w_1 = \eta_1$ and $w_i = \sum_{j=1}^{i-1} b_{ij} w_j + \eta_i$, with $\eta_i \sim N(0, \tau_i)$, where $\tau_1 = \text{var}(w_1)$ and $\tau_i = \text{var}(w_i|w_j : 1 \leq j < i)$.

If locations are suitably ordered, a good approximation can be obtained by replacing the conditioning set $\{w_j : 1 \leq j < i\}$, for $i = 2, \ldots, n$, by a subset $N(i)$ that contains a reduced number of nearest neighbors. When considering neighborhoods of sizes up to m, the sparse approximation to the dense linear model becomes

$$w_i \;\; = \;\; \sum_{\mathbf{s}_j \in N(i)} b_{ij} w_j + \xi_i \;\; = \;\; \mathbf{b}_i' \mathbf{w}_{N(i)} + \xi_i, \tag{S2.1}$$

where $N(i)$ contains the $m_i = \min\{m, i-1\}$ nearest neighbors within the conditioning set $\{w_j : 1 \leq j < i\}$, and $\mathbf{w}_{N(i)} = \{w_j : \mathbf{s}_j \in N(i)\}$. Denote by $\mathbf{C}_{U,V}$ represent the submatrix of $\mathbf{C}$ indexed by the rows corresponding to locations in set U and by columns indexed by locations in V, and $\mathbf{C}_U$ be the square matrix with rows and columns indexed by locations indexed by U. Using this notation, we have that $\mathbf{b}_i' = \mathbf{C}_{i,N(i)} \mathbf{C}_{N(i)}^{-1}$, i.e., the kriging weights conditioned on the neighbor set. Additionally, the last

term on the right hand side $\xi_i \sim N(0, F_i)$, where $F_i = \text{var}(w_i|\mathbf{w}_{N(i)}) = \mathbf{C}_i - \mathbf{C}_{i,N(i)}\mathbf{C}_{N(i)}^{-1}\mathbf{C}_{N(i),i}$. Hence, both $\mathbf{b}_i$ and F_i are entirely characterized by the covariance function $\mathcal{C}(\cdot,\cdot\,|\,\boldsymbol{\phi})$ from the parent process. Note that the dense and the sparse process share the same equations for locations $i = 1, 2, \ldots, m+1$. This implies that the covariance among the first $m+1$ locations under the NNGP is the same as that of the parent process.

In vector form, the sparse model can be written as $\mathbf{w} = \mathbf{B}\mathbf{w} + \boldsymbol{\xi}$. Here, $\boldsymbol{\xi} \sim \text{N}_n(\mathbf{0}, \mathbf{F})$, where $\mathbf{F} = \text{diag}\{F_i : i = 1, \ldots, n\}$, and $\mathbf{B}$ is the lower triangular matrix with zeros along the diagonal, and at most m nonzero values in each row. The nonzero values in the ith row of $\mathbf{B}$ are located in columns $\{j : \mathbf{s}_j \in N(i)\}$. Hence, this representation implies that

$$\mathbf{w} = (\text{I} - \mathbf{B})^{-1}\boldsymbol{\xi} \sim \text{N}_n(\mathbf{0}, \tilde{\mathbf{C}}), \tag{S2.2}$$

where $\tilde{\mathbf{C}} = (\text{I} - \mathbf{B})^{-1}\mathbf{F}(\text{I} - \mathbf{B})^{-T}$ (with $\tilde{\mathbf{C}}^{-1}$ sparse), which provides a good approximation to the original covariance matrix $\mathbf{C}$. For more details on the construction and appealing features of the NNGP methodology, we direct the reader to the meticulous construction presented in Datta et al. (2016a).

S3 Sampling algorithm

To begin with, given that we have a two-step process, where $\mathbf{w}(\mathbf{s})$ is assumed to exclusively capture the spatial patterns in the $\mathbf{z}(\mathbf{s})$, the full conditional for $\mathbf{w}(\mathbf{s}_i)$ is proportional to

$$\mathrm{N}_{q_w}\left(\mathbf{w}(\mathbf{s}_i)|\mathbf{B}_i^{(w)}\mathbf{w}_{N(i)},\mathbf{F}_i^{(w)}\right)\prod_{\mathbf{s}_j\in\mathcal{P}(i)}\mathrm{N}_{q_w}\left(\mathbf{w}(\mathbf{s}_j)\,|\,\mathbf{B}_j^{(w)}\mathbf{w}_{N(j)},\mathbf{F}_j^{(w)}\right)\times \\ \mathrm{N}_l\left(\mathbf{z}(\mathbf{s}_i)\,|\,\mathbf{X}_z(\mathbf{s}_i)'\boldsymbol{\beta}_z+\boldsymbol{\Lambda}_z\mathbf{w}(\mathbf{s}_i),\boldsymbol{\Psi}_z\right), \quad \text{(S3.1)}$$

where $\mathcal{P}(i)=\{\mathbf{s}_j\in\mathcal{T}:\mathbf{s}_i\in N(j)\}$, is the set of locations to which $\mathbf{s}_i$ is a neighbor. To simplify (S3.1), let $\mathbf{s}_{j_d}$ be the dth neighbor of $\mathbf{s}_j\in\mathcal{D}$ (for $1\leq d\leq m$). The columns in $\mathbf{B}_j^{(w)}$ indexed by the set

$$\mathcal{I}_{N(j),\mathbf{s}_{j_d}}=\{(d-1)q+1,\ldots,(d-1)q+q\}$$

relate $\mathbf{s}_j$ and $\mathbf{s}_{j_d}$. Denote by $B_{j,j_d}^{(w)}$ the $q\times q$ matrix containing the columns indexed by $\mathcal{I}_{N(j),\mathbf{s}_{j_d}}$ in $\mathbf{B}_j^{(w)}$. From the component of the expression above corresponding to locations $\mathbf{s}_j\in\mathcal{P}(i)$, we may rewrite the quadratic form within the exponential function in the normal density, as

$$\left(\mathbf{w}(\mathbf{s}_j)-\sum_{j_d\in N(j)}B_{j,j_d}^{(w)}\mathbf{w}(\mathbf{s}_{j_d})\right)'\mathbf{F}_j^{-1}\left(\mathbf{w}(\mathbf{s}_j)-\sum_{j_d\in N(j)}B_{j,j_d}^{(w)}\mathbf{w}(\mathbf{s}_{j_d})\right)\;= \\ \left(B_{j,i}^{(w)}\mathbf{w}(\mathbf{s}_i)-\boldsymbol{\chi}_{j,i}^{(w)}\right)'(\mathbf{F}_j^{(w)})^{-1}\left(B_{j,i}^{(w)}\mathbf{w}(\mathbf{s}_i)-\boldsymbol{\chi}_{j,i}^{(w)}\right),$$

where $\boldsymbol{\chi}_{j,i}^{(w)}=\mathbf{w}(\mathbf{s}_j)-\sum_{j_d\neq i}B_{j,j_d}^{(w)}\mathbf{w}(\mathbf{s}_{j_d})$.

Making use of this notation, the full conditional for $\mathbf{w}(\mathbf{s}_i)$ for $\mathbf{s}_i \in \mathcal{T}$, is $\mathrm{N}_{q_w}(\mathbf{w}(\mathbf{s}_i)|\boldsymbol{\Sigma}_i^{(w)}\boldsymbol{\mu}_i^{(w)}, \boldsymbol{\Sigma}_i^{(w)})$, with

$$\begin{aligned}
\boldsymbol{\mu}_i^{(w)} &= \left((\mathbf{F}_i^{(w)})^{-1}\mathbf{B}_i^{(w)}\mathbf{w}_{N(i)} + \sum_{\mathbf{s}_j \in \mathcal{P}(i)} (B_{j,i}^{(w)})'(\mathbf{F}_j^{(w)})^{-1}\boldsymbol{\chi}_{j,i}^{(w)} + \boldsymbol{\Lambda}_z'\boldsymbol{\Psi}_z^{-1}(\mathbf{z}(\mathbf{s}_i) - \mathbf{X}_z(\mathbf{s}_i)'\boldsymbol{\beta}_z) \right), \text{ and} \\
\boldsymbol{\Sigma}_i^{(w)} &= \left((\mathbf{F}_i^{(w)})^{-1} + \sum_{\mathbf{s}_j \in \mathcal{P}(i)} (B_{j,i}^{(w)})'(\mathbf{F}_j^{w})^{-1}B_{j,i}^{(w)} + \boldsymbol{\Lambda}_z'\boldsymbol{\Psi}_z^{-1}\boldsymbol{\Lambda}_z \right)^{-1}.
\end{aligned}$$

Similarly, the full conditional for $\mathbf{v}(\mathbf{s}_i)$ is $\mathrm{N}_{q_v}(\mathbf{v}(\mathbf{s}_i)|\boldsymbol{\Sigma}_i^{(v)}\boldsymbol{\mu}_i^{(v)}, \boldsymbol{\Sigma}_i^{(v)})$, where

$$\begin{aligned}
\boldsymbol{\mu}_i^{(v)} &= \left((\mathbf{F}_i^{(v)})^{-1}\mathbf{B}_i^{(v)}\mathbf{v}_{N(i)} + \sum_{\mathbf{s}_j \in \mathcal{P}(i)} (B_{j,i}^{(v)})'(\mathbf{F}_j^{(v)})^{-1}\boldsymbol{\chi}_{j,i}^{(v)} + \boldsymbol{\Gamma}'\boldsymbol{\Psi}_y^{-1}(\mathbf{y}(\mathbf{s}_i) - \mathbf{X}_y(\mathbf{s}_i)'\boldsymbol{\beta}_y - \boldsymbol{\Lambda}_y\mathbf{w}(\mathbf{s}_i)) \right), \text{ and} \\
\boldsymbol{\Sigma}_i^{(v)} &= \left((\mathbf{F}_i^{(v)})^{-1} + \sum_{\mathbf{s}_j \in \mathcal{P}(i)} (B_{j,i}^{(v)})'(\mathbf{F}_j^{(v)})^{-1}B_{j,i}^{(v)} + \boldsymbol{\Gamma}'\boldsymbol{\Psi}_y^{-1}\boldsymbol{\Gamma} \right)^{-1}.
\end{aligned}$$

To obtain the full conditional density for $\boldsymbol{\beta}_z$ and $\boldsymbol{\beta}_y$, let $\mathbf{W}_{\mathcal{T}_z}$ be the $n_z \times q_w$ matrix with rows given by $\mathbf{w}(\mathbf{s})$ for $\mathbf{s} \in \mathcal{T}_z$. Define analogously the $n_y \times q_w$ and $n_y \times q_v$ matrices $\mathbf{W}_{\mathcal{T}_y}$ and $\mathbf{V}_{\mathcal{T}_y}$, respectively. Represent $\boldsymbol{\Lambda}_z = (\boldsymbol{\lambda}_1^z : \cdots : \boldsymbol{\lambda}_{h_z}^z)'$, $\boldsymbol{\Lambda}_y = (\boldsymbol{\lambda}_1^y : \cdots : \boldsymbol{\lambda}_{h_y}^y)'$, and $\boldsymbol{\Gamma} = (\boldsymbol{\gamma}_1 : \cdots : \boldsymbol{\gamma}_{h_y})'$. Additionally, for $j = 1, \ldots, h_z$ and $k = 1, \ldots, h_y$, define $\mathbf{z}_j = (z_j(\mathbf{s}) : \mathbf{s} \in \mathcal{T}_z)'$, and $\mathbf{y}_k = (y_k(\mathbf{r}) : \mathbf{r} \in \mathcal{T}_y)'$. Also denote the $n_z \times p_{z,j}$ matrix of predictors for the jth outcome in $\mathbf{z}(\cdot)$ by $\mathbf{X}_j^z = (\mathbf{x}_j^z(\mathbf{s}) : \mathbf{s} \in \mathcal{T}_z)'$ for $j = 1, \ldots, h_z$. Similarly, let $\mathbf{X}_k^y = (\mathbf{x}_k^y(\mathbf{r}_1) : \ldots : \mathbf{x}_k^y(\mathbf{r}_{n_y}))'$ for $k = 1, \ldots, h_y$. Thus, assuming flat priors for $\boldsymbol{\beta}_z$ and $\boldsymbol{\beta}_y$ the full conditionals are

$$\boldsymbol{\beta}_z \,|\, \cdot \;\sim\; \prod_{j=1}^{h_z} \mathrm{N}_{p_{z,j}}(\boldsymbol{\beta}_j^z \,|\, \Omega_j^z\boldsymbol{\mu}_j^z, \Omega_j^z),$$

where $\boldsymbol{\mu}_j^z = \left(\frac{1}{\psi_j^z}(\mathbf{X}_j^z)'(\mathbf{z}_j - \mathbf{W}_{\mathcal{T}_z}\boldsymbol{\lambda}_j^z)\right)$ and $\Omega_j^z = \psi_j^z\left((\mathbf{X}_j^z)'\mathbf{X}_j^z\right)^{-1}$, and

$$\boldsymbol{\beta}_y \,|\, \cdot \;\sim\; \prod_{k=1}^{h_y} \mathrm{N}_{p_{y,k}}(\boldsymbol{\beta}_k^y \,|\, \Omega_k\boldsymbol{\mu}_k, \Omega_k),$$

where $\boldsymbol{\mu}_k^y = \left(\frac{1}{\psi_k^y}(\mathbf{X}_k^y)'(\mathbf{y}_k - \mathbf{W}_{\mathcal{T}_y}\boldsymbol{\lambda}_k^y - \mathbf{V}_{\mathcal{T}_y}\boldsymbol{\gamma}_k)\right)$ and $\Omega_k^y = \psi_k^y\left((\mathbf{X}_k^y)'\mathbf{X}_h\right)^{-1}$.

Given the identifiability restrictions imposed on $\boldsymbol{\Lambda}_z$, let $q_j = \min\{j-1, q_w\}$ for $2 \leq j \leq h_z$, and denote by $\tilde{\boldsymbol{\lambda}}_j^z = (\lambda_{j1}^z, \ldots, \lambda_{jq_j}^z)'$ the vector of unrestricted elements in the jth row of $\boldsymbol{\Lambda}_z$. Define $\mathbf{W}_{1:j} = (\dot{\mathbf{w}}_1 \cdots \dot{\mathbf{w}}_j)$ (i.e., the matrix with the first j columns of $\mathbf{W}_{\mathcal{T}_z}$). Using the definitions above, the full conditional density for $\boldsymbol{\Lambda}_z$ can be represented as $\prod_{j=2}^{h_z} \mathrm{N}_{q_j}\left(\tilde{\boldsymbol{\lambda}}_j \,|\, \Omega_{\lambda_j^z}\boldsymbol{\mu}_{\boldsymbol{\lambda}_j^z}, \Omega_{\lambda_j^z}\right)$, where

$$\boldsymbol{\mu}_{\tilde{\boldsymbol{\lambda}}_j^z} \;=\; \begin{cases} \frac{1}{\psi_j^z}\mathbf{W}'_{1:(j-1)}(\mathbf{z}_j - \mathbf{X}_j^z\boldsymbol{\beta}_j^z - \dot{\mathbf{w}}_j) & \text{if} \quad 2 \leq j \leq q_w \\ \frac{1}{\psi_j^z}\mathbf{W}'_{\mathcal{T}_z}(\mathbf{z}_j - \mathbf{X}_j^z\boldsymbol{\beta}_j^z) & \text{if} \quad j > q_w \end{cases}, \text{ and}$$

$$\Omega_{\tilde{\boldsymbol{\lambda}}_j^z} \;=\; \begin{cases} \left(\frac{1}{\psi_j^z}\mathbf{W}'_{1:(j-1)}\mathbf{W}_{1:(j-1)} + \mathrm{I}_{j-1}\right)^{-1} & \text{if} \quad 2 \leq j \leq q_w \\ \left(\frac{1}{\psi_j^z}\mathbf{W}'_{\mathcal{T}_z}\mathbf{W}_{\mathcal{T}_z} + \mathrm{I}_{q_w}\right)^{-1} & \text{if} \quad j > q_w \end{cases}$$

The elements in $\boldsymbol{\Lambda}_y$ are all unrestricted; hence, the full conditional posterior for $\boldsymbol{\Lambda}_y$ corresponds to $\prod_{k=1}^{h_y} \mathrm{N}_{q_w}\left(\boldsymbol{\lambda}_k^y \,|\, \Omega_{\lambda_k^y}\boldsymbol{\mu}_{\boldsymbol{\lambda}_k^y}, \Omega_{\lambda_k^y}\right)$, with

$$\boldsymbol{\mu}_{\boldsymbol{\lambda}_k^y} \;=\; \frac{1}{\psi_k^y}\mathbf{W}'_{\mathcal{T}_y}(\mathbf{y}_k - \mathbf{X}_k^y\boldsymbol{\beta}_k^y - \mathbf{V}_{\mathcal{T}_y}\boldsymbol{\gamma}_k), \quad \text{and} \quad \Omega_{\boldsymbol{\lambda}_k^y} \;=\; \psi_k^y\left(\mathbf{W}'_{\mathcal{T}_y}\mathbf{W}_{\mathcal{T}_y} + \mathrm{I}_{q_w}\right)^{-1}$$

The sampler for $\boldsymbol{\Gamma}$ has a similar form to that of $\boldsymbol{\Lambda}_z$, with the upper triangular elements equal to zero; however, given that we make no dimension reduction for the forest outcomes, the diagonal elements are only constrained

to be positive (instead of setting them to one). First, for $2 \leq k \leq h_y$, let $q_k = \min\{k-1, q_v\}$, denote $\tilde{\boldsymbol{\gamma}}_k = (\gamma_{k1}, \ldots, \gamma_{kq_k})'$, and let $\mathbf{V}_{1:k} = (\dot{\mathbf{v}}_1 \cdots \dot{\mathbf{v}}_k)$ denote the matrix with the first k columns of $\mathbf{V}_{\mathcal{T}_y}$. Then, the full conditional posterior density for $\boldsymbol{\Gamma}$ is $\prod_{k=2}^{h_y} \mathrm{N}_{q_k}\left(\tilde{\boldsymbol{\gamma}}_k \,|\, \Omega_{\gamma_k}\boldsymbol{\mu}_{\gamma_k}, \Omega_{\gamma_k}\right)$, with

$$\boldsymbol{\mu}_{\tilde{\gamma}_k} = \begin{cases} \frac{1}{\psi_k^y}\mathbf{V}'_{1:(k-1)}(\mathbf{y}_k - \mathbf{X}_k^y\boldsymbol{\beta}_k^y - \mathbf{W}_{\mathcal{T}_y}\boldsymbol{\lambda}_k^y - \gamma_{kk}\dot{\mathbf{v}}_k) & \text{if} \quad 2 \leq k \leq q_v \\ \frac{1}{\psi_k^y}\mathbf{V}'_{\mathcal{T}_y}(\mathbf{y}_k - \mathbf{X}_k^y\boldsymbol{\beta}_k^y - \mathbf{W}_{\mathcal{T}_y}\boldsymbol{\lambda}_k^y) & \text{if} \quad k > q_v \end{cases}, \text{ and}$$

$$\Omega_{\tilde{\gamma}_k} = \begin{cases} \left(\frac{1}{\psi_k^y}\mathbf{V}'_{1:(k-1)}\mathbf{V}_{1:(k-1)} + \mathrm{I}_{k-1}\right)^{-1} & \text{if} \quad 2 \leq k \leq q_v \\ \left(\frac{1}{\psi_k^y}\mathbf{V}'_{\mathcal{T}_y}\mathbf{V}_{\mathcal{T}_y} + \mathrm{I}_{q_v}\right)^{-1} & \text{if} \quad k > q_v \end{cases}$$

As mentioned before, the diagonal elements of $\boldsymbol{\Gamma}$ are assumed to be positive. Hence to sample the kth diagonal element in $\boldsymbol{\Gamma}$ we assume a prior $\propto \mathrm{I}_{\{\gamma_{kk}>0\}}$. This yields a truncated normal distribution, obtained from a normal with mean $\mu_{\gamma_{kk}}$ and variance $\xi_{\gamma_{kk}}$ and truncated to be greater than zero, where

$$\begin{aligned} \mu_{\gamma_{kk}} &= (\dot{\mathbf{v}}_r\dot{\mathbf{v}}_k)^{-1}\dot{\mathbf{v}}_r(\dot{\mathbf{y}}_k - \mathbf{X}_k^y\boldsymbol{\beta}_k^y - \mathbf{W}_{\mathcal{T}_y}\boldsymbol{\lambda}_k^y - \mathbf{V}_{1:(k-1)}\tilde{\boldsymbol{\gamma}}_k), \\ \xi_{\gamma_{kk}} &= \frac{\psi_k^y}{\dot{\mathbf{v}}_r\dot{\mathbf{v}}_k}. \end{aligned}$$

Given that the half-t distribution is a mixture of two Inverse-Gamma distributions, the full conditionals for $\boldsymbol{\psi}_z$ and $\boldsymbol{\psi}_y$ (the vectors of diagonal elements of $\boldsymbol{\Psi}_z$ and $\boldsymbol{\Psi}_y$, respectively) are conjugate with their corresponding

likelihoods. Sampling them amounts to drawing from

$$\boldsymbol{\psi}_z|\cdot \;\sim\; \prod_{j=1}^{h_z} \mathcal{IG}\left(\psi_j^z \,\middle|\, \frac{\nu+n_z}{2}, \frac{\nu}{a_j^z} + \frac{1}{2}\sum_{\mathbf{s}_i\in\mathcal{T}_z} (z_j(\mathbf{s}_i) - \mathbf{x}_j^z(\mathbf{s}_i)'\boldsymbol{\beta}_j^z - (\boldsymbol{\lambda}_j^z)'\mathbf{w}(\mathbf{s}_i))^2\right),$$

with hyperparameters $\mathbf{a}_z = (a_1^z, \ldots, a_{h_z}^z)$ sampled from

$$\mathbf{a}_z|\boldsymbol{\psi}_z \;\sim\; \prod_{j=1}^{h_z} \mathcal{IG}\left(a_j^z \,\middle|\, \frac{1}{2}, \frac{\nu}{\psi_j^z} + \frac{1}{A^2}\right)$$

$$\boldsymbol{\psi}_y|\cdot \;\sim\; \prod_{k=1}^{h_y} \mathcal{IG}\left(\psi_k^y \,\middle|\, \frac{\nu+n_y}{2}, \frac{\nu}{a_k^y} + \frac{1}{2}\sum_{\mathbf{s}_i\in\mathcal{T}_y} (y_k(\mathbf{s}_i) - \mathbf{x}_k^y(\mathbf{s}_i)'\boldsymbol{\beta}_k^y - (\boldsymbol{\lambda}_k^y)'\mathbf{w}(\mathbf{s}_i) - (\boldsymbol{\gamma}_k)'\mathbf{v}(\mathbf{s}_i))^2\right),$$

with hyperparameters $\mathbf{a}_y = (a_1^y, \ldots, a_{h_y}^y)$ sampled from

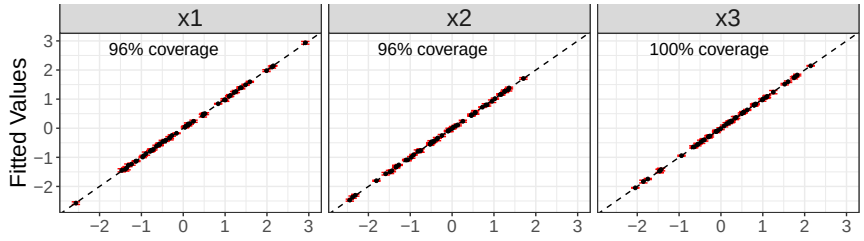
$$\mathbf{a}_y|\boldsymbol{\psi}_y \;\sim\; \prod_{k=1}^{h_y} \mathcal{IG}\left(a_k^y \,\middle|\, \frac{1}{2}, \frac{\nu}{\psi_k^y} + \frac{1}{A^2}\right)$$

Lastly, we may use a Metropolis within Gibbs steps to sample $\boldsymbol{\phi}_w$ and $\boldsymbol{\phi}_v$, with target densities proportional to

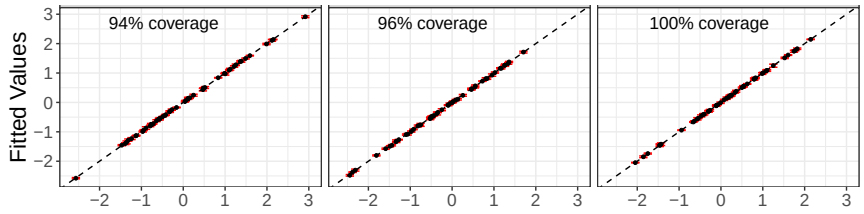
$$\begin{aligned}
\pi(\boldsymbol{\phi}_w)\mathrm{N}_{nq_w}(\mathbf{w}_{\mathcal{T}_z}|\mathbf{0}, \tilde{\mathbf{C}}^{(w)}) &= \prod_{k=1}^{q_w} \pi(\phi_{w,k})\mathrm{N}_n\left(\dot{\mathbf{w}}_k|\mathbf{0}, \tilde{\mathbf{C}}_k^{(w)}(\phi_{w,k})\right), \text{ and} \\
\pi(\boldsymbol{\phi}_v)\mathrm{N}_{nq_v}(\mathbf{v}_{\mathcal{T}_y}|\mathbf{0}, \tilde{\mathbf{C}}^{(v)}) &= \prod_{r=1}^{q_v} \pi(\phi_{v,r})\mathrm{N}_n\left(\dot{\mathbf{v}}_r|\mathbf{0}, \tilde{\mathbf{C}}_r^{(v)}(\phi_{v,r})\right),
\end{aligned}$$

which can both be sampled using a random walk Metropolis.

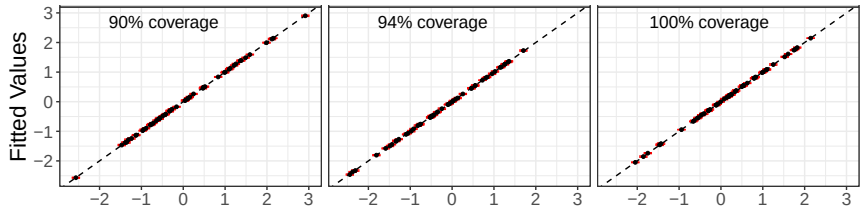
S4 Additional results from the simulation exercise



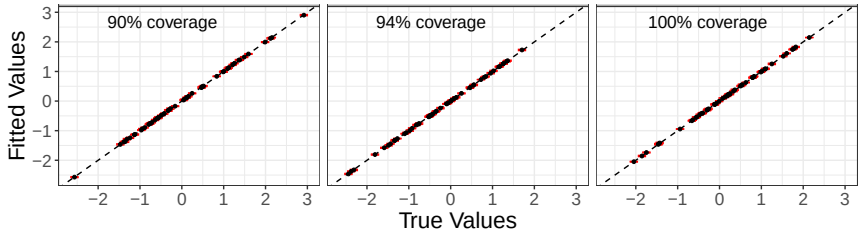
(a) $q_w = 3$



(b) $q_w = 5$

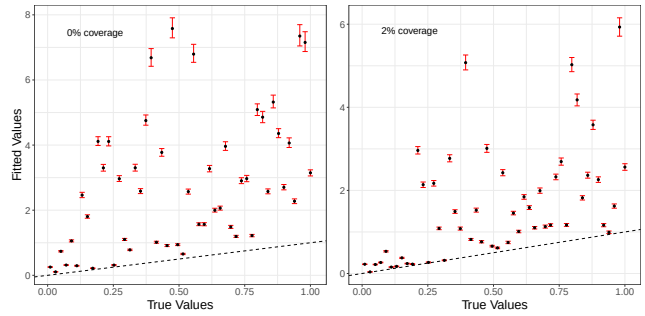


(c) $q_w = 8$

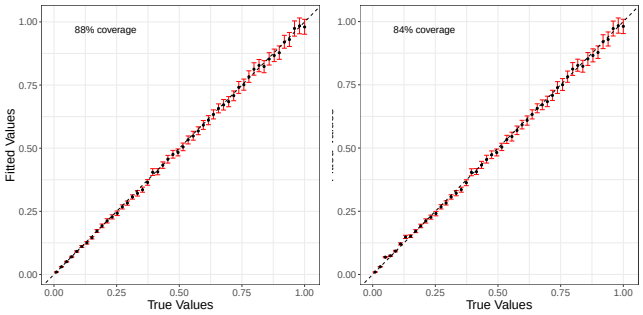


(d) $q_w = 10$

Figure 1: Posterior median and 95% credible set vs true values of $\boldsymbol{\beta}_z$ for $q_w \in \{3, 5, 8, 10\}$. The rows vary by the number of spatial factors used in model fitting, the columns vary by predictor.



(a) $q_w = 3$ (b) $q_w = 5$



(c) $q_w = 8$ (d) $q_w = 10$

Figure 2: Posterior median and 95% credible set vs true values of $\boldsymbol{\psi}_z$.

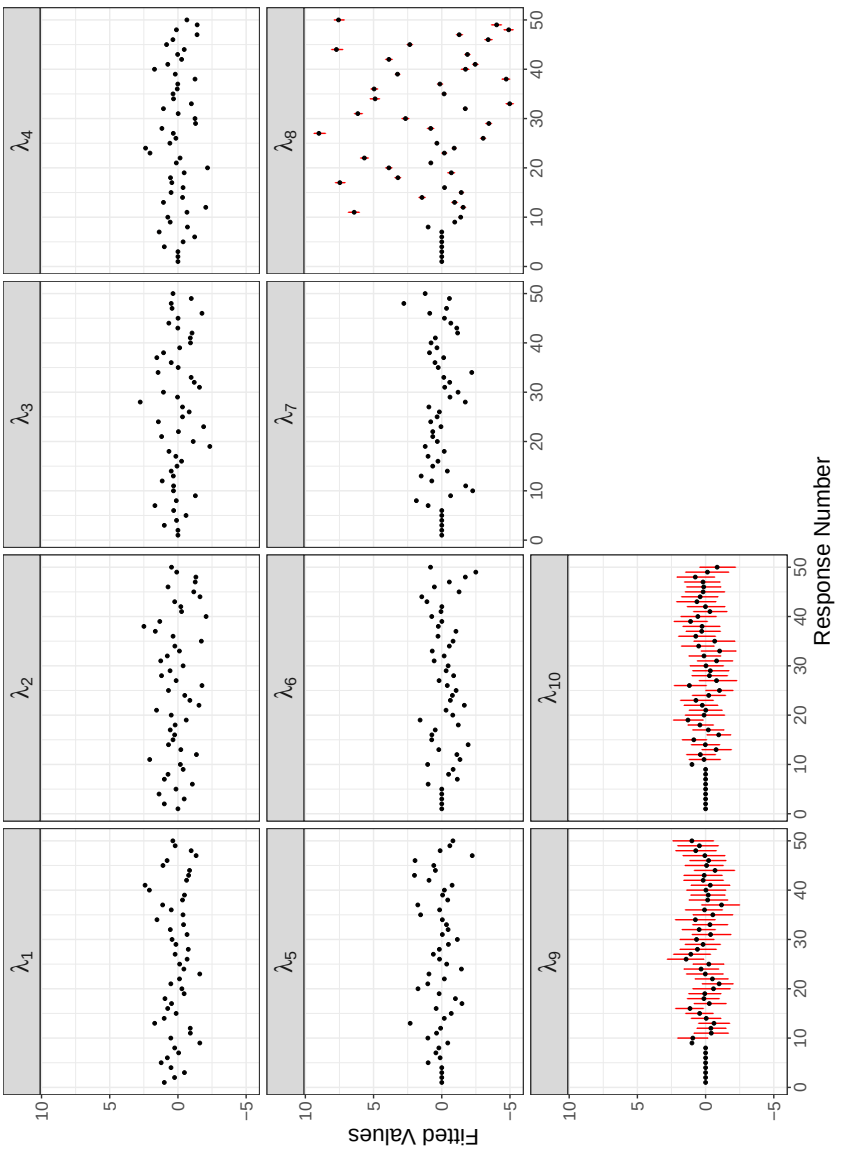


Figure 3: Posterior median and 95% credible set for the factor loadings matrix with $q_w = 10$. Each panel displays a column of the estimated $\mathbf{\Lambda}_z$ matrix.

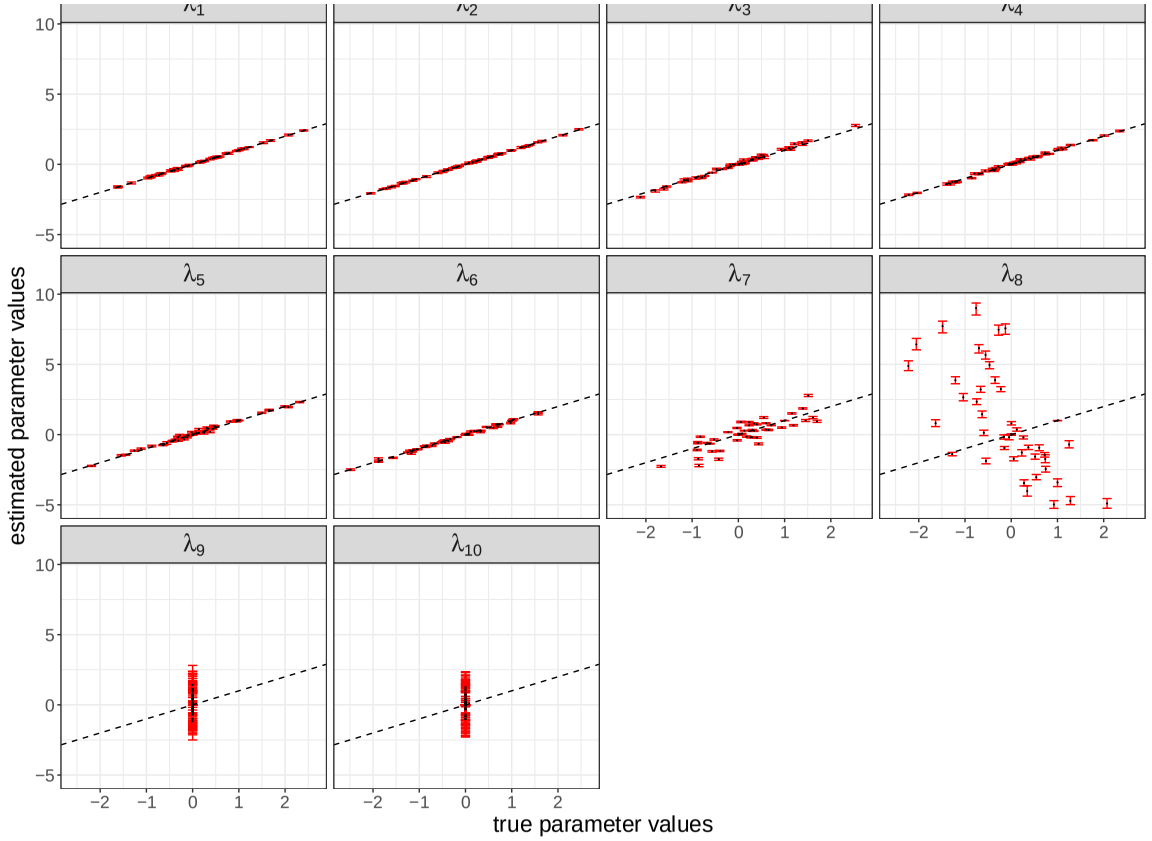
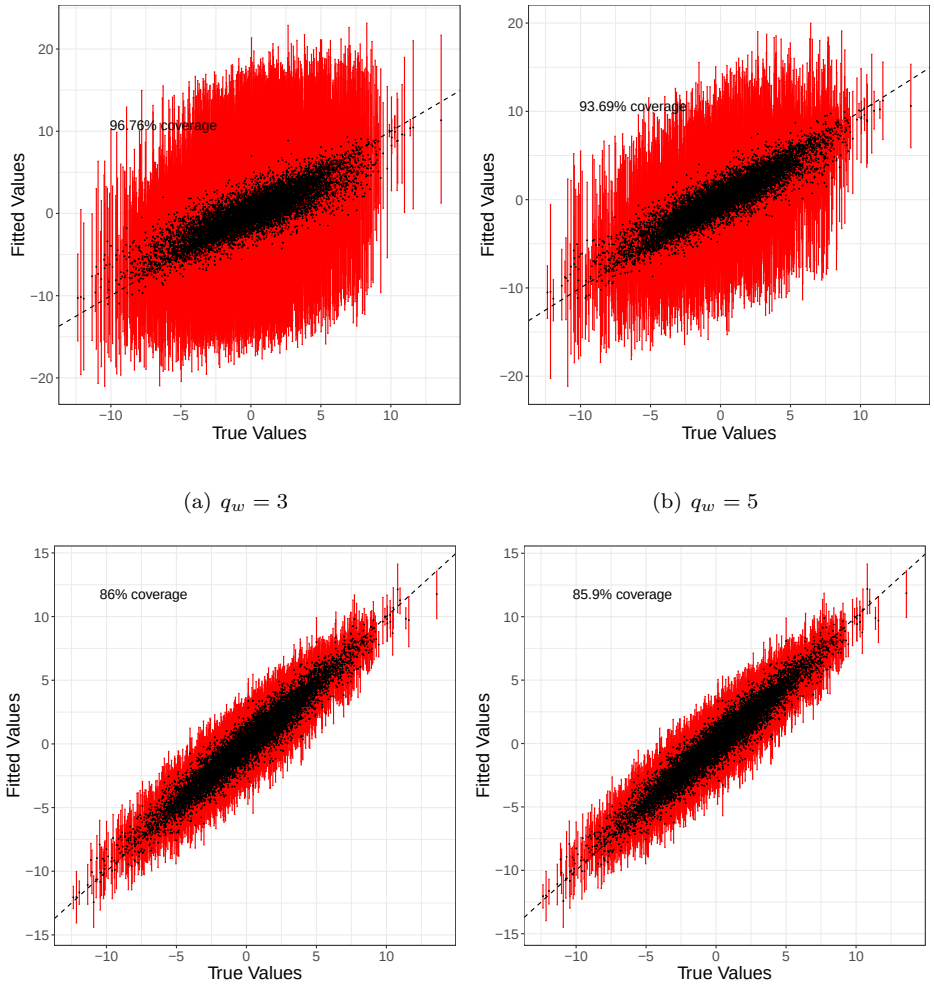


Figure 4: Fitted vs true factor loadings (95% credible sets and medians) for the model with $q_w = 10$. Each panel represents a column in $\mathbf{\Lambda}_z$. The true parameter values for columns $\boldsymbol{\lambda}_9$ and $\boldsymbol{\lambda}_{10}$ are set to zero since these are not part of the true model.



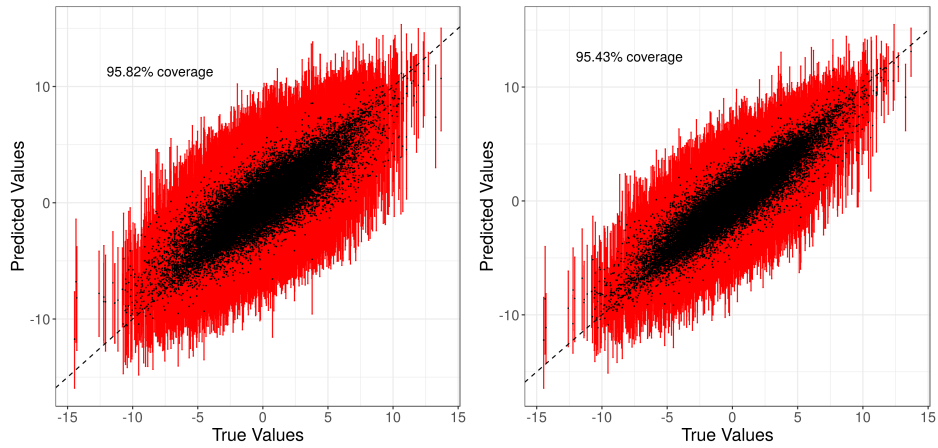
(a) $q_w = 3$

(b) $q_w = 5$

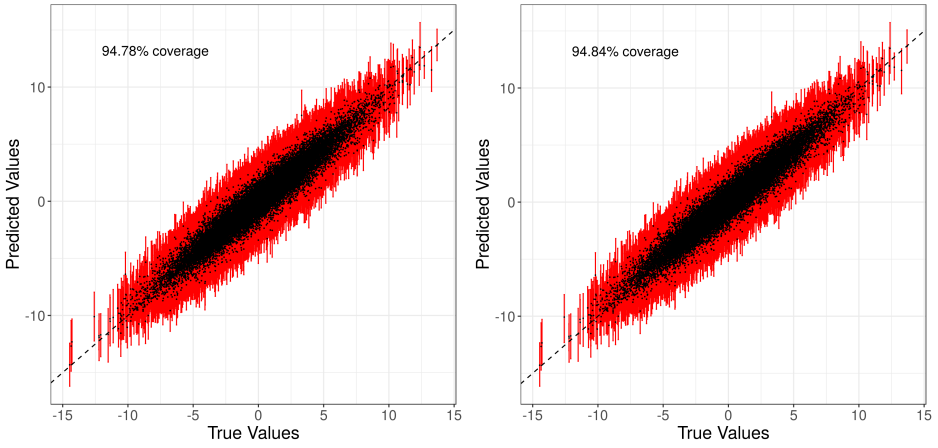
(c) $q_w = 8$

(d) $q_w = 10$

Figure 5: Imputed vs true values for missing outcomes from the simulation exercise.

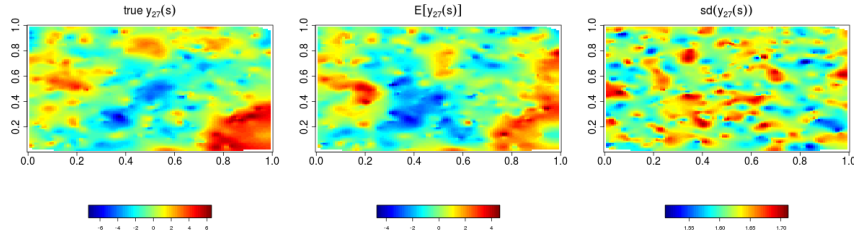


(a) $q_w = 3$ (b) $q_w = 5$

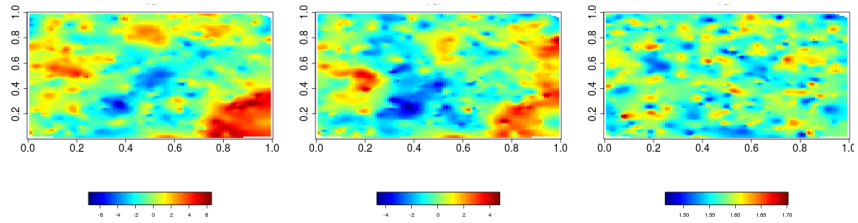


(c) $q_w = 8$ (d) $q_w = 10$

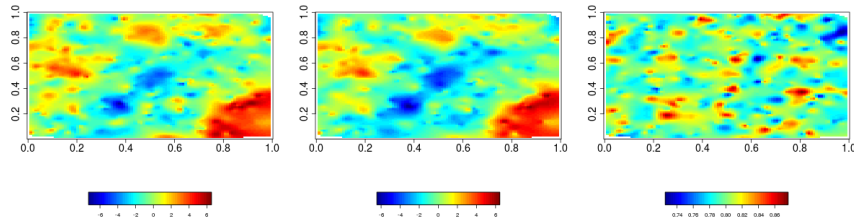
Figure 6: Predicted vs true values for 500 out-of-sample locations from the simulation exercise.



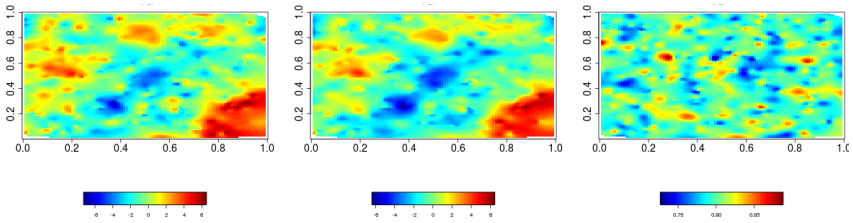
(a) $q_w = 3$



(b) $q_w = 5$



(c) $q_w = 8$



(d) $q_w = 10$

Figure 7: True, predicted and prediction uncertainty maps for $y_{27}(\cdot)$ at 500 out-of-sample locations from the simulation exercise.

Bibliography

Anderson, T. (2003). *An Introduction to Multivariate Statistical Analysis. 3rd edition*. Wiley Series in Probability and Statistics, Hoboken, NJ.

Chiles, J.-P. and Delfiner, P. (2009). *Geostatistics: modeling spatial uncertainty*, volume 497. John Wiley & Sons.

Datta, A., Banerjee, S., Finley, A., Hamm, N. A., and Schaap, M. (2016a). Non-Separable Dynamic Nearest-Neighbor Gaussian Process Models for Large Spatio-Temporal Data With an Application to Particulate Matter Analysis. *Annals of Applied Statistics Statistics*, 44(2):629–659.

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

Finley, A. O., Datta, A., Cook, B. C., Morton, D. C., Andersen, H. E., and Banerjee, S. (2017). Efficient algorithms for bayesian nearest neigh-

bor gaussian processes. *ArXiv e-prints*. https://arxiv.org/abs/1702.00434.

Genton, M. G. and Kleiber, W. (2015). Cross-Covariance Functions for Multivariate Geostatistics. *Statistical Science*, 30(2):147–163.

Ren, Q. and Banerjee, S. (2013). Hierarchical Factor Models for Large Spatially Misaligned Data: A Low-Rank Predictive Process Approach. *Biometrics*, 69(1):19–30.

Ver Hoef, J. and Barry, R. (1998). Modeling crossvariograms for cokriging and multivariable spatial prediction. *Journal of Statistical Planning and Inference*, 69:275–294.