SUPPLEMENTARY MATERIAL
Fast Bayesian inference of Block Nearest Neighbor Gaussian models for large data

Zaida C. Quiroz*

Department of Sciences, Pontificia Universidad Católica del Perú
Marcos O. Prates
Department of Statistics, Universidade Federal de Minas Gerais
Dipak K. Dey
Department of Statistics, University of Connecticut
and
Håvard Rue
Computer, Electrical and Mathematical Science and Engineering Division,
King Abdullah University of Science and Technology

A. Proofs of main results

Proof of Proposition 1. If $\pi(\boldsymbol{w_S})$ is a valid multivariate joint density, $\pi(\boldsymbol{w_{b_k}}|\boldsymbol{w_{N(b_k)}})$ is also proper, and we have that $\int \pi(\boldsymbol{w_{b_k}}|\boldsymbol{w_{N(b_k)}})d\boldsymbol{w_{b_k}} = 1, \forall k = 1, \ldots, M$. From the definitions of $\boldsymbol{G}$ and $\boldsymbol{G^b}$ there exists a set of nodes in block $\Delta(b_1)$, $\boldsymbol{s_{\Delta(b_1)}} \in \boldsymbol{G}$, such that "the last node" from a DAG $\boldsymbol{G^b}$ belongs to $\boldsymbol{s_{\Delta(b_1)}}$. Then the nodes in $\boldsymbol{s_{\Delta(b_1)}}$ do not have any directed edge originating from them. As a consequence, any node in block $\Delta(b_1)$ can not belong to the set of nodes of any other block. So the term in (2) where all locations of $\boldsymbol{\Delta(b_1)}$ appear is $\pi(\boldsymbol{w_{\Delta(b_1)}}|\boldsymbol{w_{N(\Delta(b_1))}})$. From Fubini's theorem, we can interchange the

*E-mail: zquiroz@pucp.edu.pe

product and integral, thus

$$\begin{aligned}\int \pi(\boldsymbol{w_S})d\boldsymbol{w_S} &= \int\cdots\int \prod_{i=1}^{M} \pi(\boldsymbol{w_{\Delta(b_i)}}|\boldsymbol{w_{N(\Delta(b_i))}})d\boldsymbol{w_{\Delta(i)}} \\ &= \int\cdots\int \prod_{i=2}^{M} \pi(\boldsymbol{w_{\Delta(b_i)}}|\boldsymbol{w_{N(\Delta(b_i))}})d\boldsymbol{w_{\Delta(i)}}.\end{aligned}$$

Then, removing every node of $\boldsymbol{\Delta(b_1)}$ from $\boldsymbol{G}$ and $\boldsymbol{G^b}$, we have the chain graph $\boldsymbol{G'}$ and DAG $\boldsymbol{G'_b}$, respectively. There exists another set of nodes $\boldsymbol{s_{\Delta(b_2)}}$ in $\boldsymbol{G'}$, such that "the last node" from a DAG $\boldsymbol{G'_b}$ belongs to $\boldsymbol{s_{\Delta(b_2)}}$. Then the nodes $\boldsymbol{s_{\Delta(b_2)}}$ do not have any directed edge originating from them. As consequence, any node in block $\boldsymbol{\Delta(b_2)}$ can not belong to the set of nodes of any other block. So the term in (2) where all locations of $\boldsymbol{\Delta(b_2)}$ appear is $\pi(\boldsymbol{w_{\Delta(b_2)}}|\boldsymbol{w_{N(\Delta(b_2))}})$. Applying the Fubini's theorem again,

$$\int \pi(\boldsymbol{w_S})d\boldsymbol{w_S} = \int\cdots\int \prod_{i=3}^{M} \pi(\boldsymbol{w_{\Delta(b_i)}}|\boldsymbol{w_{N(\Delta(b_i))}})d\boldsymbol{w_{\Delta(i)}}.$$

In a similar way, we find $\boldsymbol{s_{\Delta(b_3)}}, \ldots, \boldsymbol{s_{\Delta(M)}}$, such that,
$\int \pi(\boldsymbol{w_S})d\boldsymbol{w_S} = \int \prod_{i=1}^{M} \pi(\boldsymbol{w_{\Delta(b_i)}}|\boldsymbol{w_{N(\Delta(b_i))}})d\boldsymbol{w_{\Delta(i)}} = 1.$ □

Proof of Lemma 1. We need to prove that the finite dimensional distributions in (6) are consistent with a stochastic process. The Kolmogorov consistency conditions are checked as follows:

Symmetry under permutation: Let $\Delta_1, \ldots, \Delta_n$ be any permutation of $1, \ldots, n$, note that $\boldsymbol{S}$ is fixed, then it is clear that $\widetilde{\pi}(w(v_1), \ldots, w(v_n)) = \widetilde{\pi}(w(v_{\Delta_1}), \ldots, w(v_{\Delta_n}))$ if and only if the same holds for the distribution of $u_i|\boldsymbol{N(u_i)}$. Since $\boldsymbol{w_U}|\boldsymbol{w_S}$ follows a l-multivariate normal distribution, then the symmetry condition is satisfied by $\pi(\boldsymbol{w_U}|\boldsymbol{w_S})$, and it holds that the next condition $\widetilde{\pi}(w(u_1), \ldots, w(u_l)|\boldsymbol{w_S}) = \widetilde{\pi}(w(u_{\Delta_1}), \ldots, w(u_{\Delta_l})|\boldsymbol{w_S})$ is necessary and sufficient to prove the symmetry condition of $\widetilde{\pi}(\boldsymbol{w_V})$. To prove this we define the next pdfs,

$$\begin{aligned}\widetilde{\pi}(w(u_1), \ldots, w(u_l)|\boldsymbol{w_S}) &= |2\pi \boldsymbol{F_U}|^{-1/2} \exp\left\{-\frac{1}{2}(\boldsymbol{w_U} - \boldsymbol{B_U w_S})^T \boldsymbol{F_U}^{-1}(\boldsymbol{w_U} - \boldsymbol{B_U w_S})\right\} \\ &= |2\pi \boldsymbol{F_U}|^{-1/2} \exp\left\{\boldsymbol{Q(w_U)}\right\},\end{aligned}$$

and

$$\begin{aligned}\widetilde{p}(w(u_{\Delta_1}),\ldots,w(u_{\Delta_l})|\boldsymbol{w_S}) &= |2\pi\boldsymbol{\Sigma'}|^{-1/2}\exp\left\{-\frac{1}{2}(\boldsymbol{w_{U\Delta}}-\boldsymbol{m'})^T\boldsymbol{\Sigma'}^{-1}(\boldsymbol{w_{U\Delta}}-\boldsymbol{m'})\right\}\\ &= |2\pi\boldsymbol{\Sigma'}|^{-1/2}\exp\left\{\boldsymbol{Q(w_{U\Delta})}\right\}.\end{aligned}$$

Following Abrahamsen (1997), we also define a permutation matrix $\boldsymbol{P}$ such that $(\Delta_1,\ldots,\Delta_l)^T = \boldsymbol{P}(1,\ldots,l)^T$. Then $\boldsymbol{P}w_U = \boldsymbol{P}(w(u_1),\ldots,w(u_l))^T = (w(u_{\Delta_1}),\ldots,w(u_{\Delta_l}))^T = \boldsymbol{w_{U\Delta}}$. And the mean and covariance matrix of $\boldsymbol{w_U\Delta}|\boldsymbol{w_S}$ are $\boldsymbol{m'} = \boldsymbol{PB_Uw_S}$ and $\boldsymbol{\Sigma'} = \boldsymbol{PF_UP'}$. Since $\boldsymbol{P}^{-1} = \boldsymbol{P}^T$ it follows that $|\boldsymbol{P}| = \pm 1$ which implies that $|\boldsymbol{\Sigma'}| = |\boldsymbol{F_U}|$. Using this we have,

$$\begin{aligned}\boldsymbol{Q(w_{U\Delta})} &= (\boldsymbol{Pw_U}-\boldsymbol{m'})^T\boldsymbol{\Sigma'}^{-1}(\boldsymbol{Pw_U}-\boldsymbol{m'})\\ &= (\boldsymbol{Pw_U}-\boldsymbol{PB_Uw_S})^T(\boldsymbol{PF_UP'})^{-1}(\boldsymbol{Pw_U}-\boldsymbol{PB_Uw_S})\\ &= (\boldsymbol{w_U}-\boldsymbol{B_Uw_S})^T\boldsymbol{P}^T(\boldsymbol{P}^T\boldsymbol{F_U}^{-1}\boldsymbol{P}^T)\boldsymbol{P}(\boldsymbol{w_U}-\boldsymbol{B_Uw_S})\\ &= (\boldsymbol{w_U}-\boldsymbol{B_Uw_S})^T\boldsymbol{P}^T\boldsymbol{\Sigma'}^{-1}\boldsymbol{P}(\boldsymbol{w_U}-\boldsymbol{B_Uw_S})\\ &= (\boldsymbol{w_U}-\boldsymbol{B_Uw_S})^T\boldsymbol{F_U}^{-1}(\boldsymbol{w_U}-\boldsymbol{B_Uw_S}) = \boldsymbol{Q(w_U)}.\end{aligned}$$

Since both $|\boldsymbol{F_U}|$ and $\boldsymbol{Q(w_U)}$ are invariant under permutations, $\widetilde{\pi}(w(u_1),\ldots,w(u_l)|\boldsymbol{w_S}) = \widetilde{\pi}(w(u_{\Delta_1}),\ldots,w(u_{\Delta_l})|\boldsymbol{w_S})$ and hence the symmetry condition is satisfied.

Dimensional consistency: We also assume that $\boldsymbol{S}$ is fixed, so, this proof does not differ from the one found in Datta et al. (2016) although $\widetilde{\pi}(\boldsymbol{w_S})$ has a different definition. Let $\boldsymbol{V_1} = \boldsymbol{V}\cup\{v_0\}$ then $\boldsymbol{V_1} = \boldsymbol{S'}\cup\{v_0\}\cup\boldsymbol{U}$. We need to verify $\widetilde{\pi}(\boldsymbol{w_V}) = \int\widetilde{\pi}(\boldsymbol{w_{V_1}})d(w(v_0))$. So, we have two cases:

Case 1: If $v_0\in\boldsymbol{S}$. By definition $\widetilde{\pi}(\boldsymbol{w_{V_1}}) = \int\widetilde{\pi}(\boldsymbol{w_{V_1|S}}|\boldsymbol{w_S})\widetilde{p}(\boldsymbol{w_S})\prod_{si\in\boldsymbol{S}|\boldsymbol{V_1}}d(w_{s_i})$, where $V_1|\boldsymbol{S}$ denotes V_1 without $\boldsymbol{S}$ and $\boldsymbol{S}|V_1$ denotes $\boldsymbol{S}$ without V_1. Then

$$\int\widetilde{\pi}(\boldsymbol{w_{V_1}})d(w(v_0)) = \int\widetilde{\pi}(\boldsymbol{w_{V_1|S}}|\boldsymbol{w_S})\widetilde{\pi}(\boldsymbol{w_S})\prod_{si\in\boldsymbol{S}|\boldsymbol{V_1}}d(w(s_i))d(w(v_0)).$$

If $v_0\in\boldsymbol{S}$, and $\boldsymbol{V} = \boldsymbol{S'}\cup\boldsymbol{U}$ then $v_0\in(\boldsymbol{S'})^c$, and $\prod_{si\in\boldsymbol{S}|\boldsymbol{V_1}}d(w(s_i)d(w(v_0)) = \prod_{si\in(\boldsymbol{S'})^c}d(w(s_i)$, and

$$\int\widetilde{\pi}(w_{\boldsymbol{V_1}})d(w(v_0)) = \int\widetilde{\pi}(\boldsymbol{w_{V_1|S}}|\boldsymbol{w_S})\widetilde{\pi}(\boldsymbol{w_S})\prod_{si\in(\boldsymbol{S'})^c}d(w_{s_i}).$$

Also, $\boldsymbol{V_1}|\boldsymbol{S} = \boldsymbol{U}$ since $v_0 \in \boldsymbol{S}$, then

$$\int \widetilde{\pi}(\boldsymbol{w_{V_1}}) d(w(v_0)) = \int \widetilde{\pi}(\boldsymbol{w_U}|\boldsymbol{w_S})\widetilde{\pi}(\boldsymbol{w_S}) \prod_{si \in (\boldsymbol{S'})^c} d(w_{s_i}) = \widetilde{\pi}(\boldsymbol{w_V}).$$

Case 2: If $v_0 \notin \boldsymbol{S}$, then $\boldsymbol{V_1}|\boldsymbol{S} = \boldsymbol{U} \cup \{v_0\}$, $\widetilde{\pi}(\boldsymbol{w_{V_1|S}}|\boldsymbol{w_S}) = \widetilde{\pi}(\boldsymbol{w_{U|S}}|\boldsymbol{w_S})\widetilde{\pi}(w(v_0)|\boldsymbol{w_S})$ and $\boldsymbol{S}|\boldsymbol{V_1} = (\boldsymbol{S'})^c$. Now,

$$\begin{aligned} \widetilde{\pi}(\boldsymbol{w_{V_1}}) &= \int \widetilde{\pi}(\boldsymbol{w_{V_1|S}}|\boldsymbol{w_S})\widetilde{\pi}(\boldsymbol{w_S}) \prod_{si \in \boldsymbol{S}|\boldsymbol{V_1}} d(w_{s_i}) \\ &= \int \widetilde{\pi}(\boldsymbol{w_U}|\boldsymbol{w_S})\widetilde{\pi}(w(v_0)|\boldsymbol{w_S})\widetilde{\pi}(\boldsymbol{w_S}) \prod_{si \in (\boldsymbol{S'})^c} d(w_{s_i}). \end{aligned}$$

Hence,

$$\begin{aligned} \int \widetilde{\pi}(\boldsymbol{w_{V_1}}) d(w(v_0)) &= \int \widetilde{\pi}(\boldsymbol{w_U}|\boldsymbol{w_S})\widetilde{\pi}(w(v_0)|\boldsymbol{w_S})\widetilde{\pi}(\boldsymbol{w_S}) \prod_{si \in (\boldsymbol{S'})^c} d(w_{s_i}) d(w(v_0)) \\ &= \int \widetilde{\pi}(\boldsymbol{w_S})\widetilde{\pi}(\boldsymbol{w_U}|\boldsymbol{w_S})[\widetilde{\pi}(w(v_0)|\boldsymbol{w_S}) d(w(v_0))] \prod_{si \in (\boldsymbol{S'})^c} d(w_{s_i}), \end{aligned}$$

where $\int \widetilde{\pi}(w(v_0)|\boldsymbol{w_S}) d(w(v_0)) = 1$, since $w(v_0)$ does not appear in any other term. Finally,

$$\int \widetilde{\pi}(\boldsymbol{w_{V_1}}) d(w(v_0)) = \int \widetilde{\pi}(\boldsymbol{w_S})\widetilde{\pi}(\boldsymbol{w_U}|\boldsymbol{w_S}) \prod_{si \in (\boldsymbol{S'})^c} d(w_{s_i}) = \widetilde{\pi}(\boldsymbol{w_V}).$$

□

Proof of Theorem 1. To verify that $\widetilde{\pi}(\boldsymbol{w_V})$ is the pdf of the finite dimensional distribution of a Gaussian process, we only need to prove that $\widetilde{\pi}(\boldsymbol{w_V})$ is the pdf of a multivariate normal distribution. Since $\boldsymbol{w_U}|\boldsymbol{w_S}$ follows a l-multivariate normal distribution and $\boldsymbol{w_S}$ follows a n-multivariate normal distribution, the product of these densities is also a multivariate normal distribution.

Let $\widetilde{C}_{m,n}$ be an element of $\widetilde{\boldsymbol{C}}_{\boldsymbol{S}}$. The cross-covariance is computed for the next possible cases:

Case 1: If $v_1 \in \boldsymbol{S}$ and $v_2 \in \boldsymbol{S}$, that is, $v_1 = s_i$ and $v_2 = s_j$, then $\text{cov}(w(v_1), w(v_2)|\theta)) = \widetilde{C}_{s_i,s_j}$.

Case 2: If $v_1 \in U$ and $v_2 \in S$, we may suppose also that $v_2 \in b_l$. Using the law of total covariance,

$$\text{cov}(w(v_1), w(v_2)|\theta)) = \text{E}(\text{cov}(w(v_1), w(v_2)|w_S)|\theta) + \text{cov}(\text{E}(w(v_1)|w_S), \text{E}(w(v_2)|w_S)|\theta).$$

From our definition $w(v_1)|\boldsymbol{w_S} \perp w(b_l)|\boldsymbol{w_S}$ and $v_2 \in b_l$, then we have that $w(v_1)|\boldsymbol{w_S} \perp w(v_2)|\boldsymbol{w_S}$ and $\text{cov}(w(v_1)|\boldsymbol{w_S}, w(v_2)|\boldsymbol{w_S}) = 0$. Further, $\text{E}(w(v_1)|\boldsymbol{w_S}) = \boldsymbol{B_{v_1} w_{N(v_1)}}$ and using the next property, $\text{E}(g(X)|X) = g(X)$, $\text{E}(w(v_2)|\boldsymbol{w_S}) = w(v_2)$. It follows that,

$$\text{cov}(w(v_1), w(v_2)|\theta)) = \text{E}(0|\theta) + \text{cov}(\boldsymbol{B_{v_1} w_{N(v_1)}}, w(v_2)|\theta) = \boldsymbol{B_{v_1} \widetilde{C}_{N(v_1), w(v_2)}} = \boldsymbol{B_{v_1} \widetilde{C}_{N(v_1), w(s_j)}}.$$

Case 3: If $v_1 \in U$ and $v_2 \in U$. This part of the proof follows from (Datta et al., 2016). We have $\text{E}(w(v_1)|\boldsymbol{w_S}) = \boldsymbol{B_{v_1} w_{N(v_1)}}$ and $\text{E}(w(v_2)|\boldsymbol{w_S}) = \boldsymbol{B_{v_2} w_{N(v_2)}}$. Then,

$$\begin{aligned} \text{cov}(\text{E}(w(v_1)|\boldsymbol{w_S}), \text{E}(w(v_2)|\boldsymbol{w_S})|\theta) &= \text{cov}(\boldsymbol{B_{v_1} w_{N(v_1)}}, \boldsymbol{B_{v_2} w_{N(v_2)}}) \\ &= \boldsymbol{B_{v_1}} \text{cov}(\boldsymbol{w_{N(v_1)}}, \boldsymbol{w_{N(v_2)}}) \boldsymbol{B_{v_2}^T}. \end{aligned}$$

Observe that if $v_1 \neq v_2$, then $w(v_1)|\boldsymbol{w_S} \perp w(v_2)|\boldsymbol{w_S}$ and $\text{cov}(w(v_1), w(v_2)|\boldsymbol{w_S}) = 0$. Conversely, if $v_1 = v_2$ now $\text{cov}(w(v_1), w(v_2)|\boldsymbol{w_S}) = \text{var}(w(v_1)|\boldsymbol{w_S}) = F_{v_1}$. Then, $\text{cov}(w(v_1), w(v_2)|\boldsymbol{w_S}) = \delta(v_1 = v_2) F_{v_1}$, and $\text{E}(\delta(v_1 = v_2) F_{v_1}|\theta) = \delta(v_1 = v_2) F_{v_1}$.
Hence,

$$\text{cov}(w(v_1), w(v_2)|\theta)) = \delta(v_1 = v_2) F_{v_1} + \boldsymbol{B_{v_1} \widetilde{C}_{N(v_1), N(v_2)} B_{v_2}}^T.$$

□

B. Bayesian inference through full-MCMC and collpased MCMC

The LGM is also a hierarchical model, thus, in a Bayesian framework, inference can also be performed using simulation-based techniques, such as MCMC methods. Here we describe in detail how to achieve Bayesian inference for a Gaussian geostatistical model using the block-NNGP. This approach can also be extended for non-Gaussian families.

Let $\mathbf{Y} = (\mathbf{Y}(\mathbf{s_1}), \ldots, \mathbf{Y}(\mathbf{s_n}))^{\mathbf{T}}$ be a realization of a spatial process defined for all $s_i \in \boldsymbol{D} \subset \Re^2$, $i = 1, \ldots, n$. The basic geostatistical Gaussian model is of the form

$$Y(s_i) = \boldsymbol{X}^T(s_i)\boldsymbol{\beta} + w(s_i) + \epsilon(s_i), \tag{1}$$

where $\boldsymbol{\beta}$ is a coefficient vector (or regression parameter), $\boldsymbol{X}^T(s_i)$ is a vector of covariates, $w(s_i)$ is a spatial structured random effect, thus $\boldsymbol{w} = (w(s_1), \ldots, w(s_n))^T$ captures the spatial association, and $\epsilon(s_i) \sim N(0, \tau^2)$ models the measurement error.

We assigned priors to β, w, τ, and hyperparameters. The usual Gaussian process prior for $\boldsymbol{w} \sim N(0, \boldsymbol{C}(\theta_1))$, where $\boldsymbol{C}(.)$ is some specific covariance function which depends on $\theta_1 = (\phi, \sigma^2)$. Instead of this prior we assume that $\boldsymbol{w} \sim$ block-NNGP$(0, \widetilde{\boldsymbol{C}})$. We assumed $\boldsymbol{\beta} \sim N(\boldsymbol{\mu_\beta}, \boldsymbol{V_\beta})$ and $\boldsymbol{\theta} = (\phi, \sigma^2, \tau^2) \sim \Delta(\boldsymbol{\theta})$.

The joint posterior distribution for the model in (1) is given by

$$\Delta(\boldsymbol{\theta}, \boldsymbol{\beta}, \boldsymbol{w}|\boldsymbol{y}) \propto \Delta(\boldsymbol{\theta}) \times \Delta_G(\boldsymbol{\beta}|\boldsymbol{\mu_\beta}, \boldsymbol{\Sigma_\beta}) \times \Delta_G(\boldsymbol{w}|0, \widetilde{\boldsymbol{C}}) \times \Delta_G(\boldsymbol{y}|\boldsymbol{X\beta} + \boldsymbol{w}, \boldsymbol{D}), \qquad (2)$$

where $\Delta_G(|.,.)$ denotes the Gaussian density, and $\boldsymbol{D}$ is a diagonal matrix with entries τ^2. The parameters $\boldsymbol{\theta}, \boldsymbol{\beta}, \boldsymbol{w}$ are updated in a Gibbs sampler within Metropolis random-walk step (full-MCMC) through the following algorithm:

(i) Block updating of $\boldsymbol{\theta}$ through Metropolis Random walk. The target log-density is

$$\log(\Delta(\boldsymbol{\theta}^\star|\boldsymbol{y}, \boldsymbol{w}, \boldsymbol{\beta})) \propto \log \Delta(\boldsymbol{\theta}) - \frac{1}{2}\log|\boldsymbol{D}| - \frac{1}{2}\log|\widetilde{\boldsymbol{C}}| - \frac{1}{2}(\boldsymbol{y} - \boldsymbol{X\beta} - \boldsymbol{w})^T \boldsymbol{D}^{-1}(\boldsymbol{y} - \boldsymbol{X\beta} - \boldsymbol{w}) - \frac{1}{2}\boldsymbol{w}^T \widetilde{\boldsymbol{Q}} \boldsymbol{w},$$

where $\widetilde{\boldsymbol{Q}} = \widetilde{\boldsymbol{C}}^{-1}$;

(ii) Gibbs sampler that updates $\boldsymbol{\beta}$ from the full conditional $\boldsymbol{\beta}|\boldsymbol{y}, \boldsymbol{w}, \boldsymbol{\theta} \sim N(\boldsymbol{Bb}, \boldsymbol{B})$, where $\boldsymbol{B} = (\boldsymbol{\Sigma_\beta}^{-1} + \boldsymbol{X}^T \boldsymbol{D}^{-1} \boldsymbol{X})^{-1}$ and $\boldsymbol{b} = \boldsymbol{\Sigma_\beta}^{-1} \boldsymbol{\mu_\beta} + X^T \boldsymbol{D}^{-1} \boldsymbol{y} - X^T \boldsymbol{D}^{-1} \boldsymbol{w}$;

(iii) Gibbs sampler that updates $\boldsymbol{w}$ from the full conditional $\boldsymbol{w}|\boldsymbol{y}, \boldsymbol{\beta}, \boldsymbol{\theta}^\star \sim N(\boldsymbol{Ff}, \boldsymbol{F})$, where $\boldsymbol{F} = (\widetilde{\boldsymbol{Q}} + \boldsymbol{D}^{-1})^{-1}$ and $\boldsymbol{f} = \boldsymbol{D}^{-1}(\boldsymbol{y} - \boldsymbol{X\beta})$. Repeat step i).

In general it is fast to compute $\widetilde{\boldsymbol{Q}}$ and $\log|\widetilde{\boldsymbol{C}}|$ using properties of block matrices and Cholesky decomposition. Nevertheless, although $\widetilde{\boldsymbol{Q}}$ is an sparse precision matrix and $(\widetilde{\boldsymbol{Q}} + \boldsymbol{D}^{-1})$ has the same sparsity, the inverse of this last expression is not sparse, therefore subsequent computations are performed using this huge dense matrix, and the cost to sample $\boldsymbol{\theta}, \boldsymbol{\beta}, \boldsymbol{w}$ through this approach is still too costly.

An alternative method for simulation from the conditional posterior $\Delta(\boldsymbol{\theta}, \boldsymbol{\beta}, \boldsymbol{w}|\boldsymbol{y})$ is to use the collapsed MCMC sampling (Liu, 1994). The collapsed MCMC for the Gaussian block-NNGP model follows the next steps:

(i) Block updating of $\boldsymbol{\theta}$ through a Metropolis random walk. The target log-density is

$$\log(\Delta(\boldsymbol{\theta}|\boldsymbol{y})) \propto \log \Delta(\boldsymbol{\theta}) - \frac{1}{2}\log|\boldsymbol{\Sigma}_{y|\beta,\theta}| - \frac{1}{2}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}-\boldsymbol{w})^T\boldsymbol{\Sigma}^{-1}_{y|\beta,\theta}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}-\boldsymbol{w}),$$

where $\boldsymbol{\Sigma}_{\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{\theta}} = \widetilde{\boldsymbol{C}} + \boldsymbol{D}$ and $\boldsymbol{\Sigma}^{-1}_{\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{\theta}} = \boldsymbol{D}^{-1} - \boldsymbol{D}^{-1}(\widetilde{\boldsymbol{Q}} + \boldsymbol{D}^{-1})^{-1}\boldsymbol{D}^{-1}$.
(ii) Gibbs sampler that updates $\boldsymbol{\beta}$ from $\boldsymbol{\beta}|\boldsymbol{y} \sim N(\boldsymbol{B}\boldsymbol{b}, \boldsymbol{B})$, where $\boldsymbol{B} = (\boldsymbol{\Sigma}^{-1}_{\beta} + \boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{\Sigma}_{y|\beta,\theta}\boldsymbol{X})^{-1}$ and $\boldsymbol{b} = \boldsymbol{\Sigma}^{-1}_{\beta}\boldsymbol{\mu}_{\beta} + \boldsymbol{X}^T\boldsymbol{\Sigma}^{-1}_{y|\beta,\theta}\boldsymbol{y}$. Repeat step i) and ii) until convergence;
(iii) Post-MCMC sampling: Use all the posterior samples of $\boldsymbol{\theta}$ and $\boldsymbol{\beta}$ to estimate $\boldsymbol{w}$ from $\boldsymbol{w}|\boldsymbol{\beta},\boldsymbol{\theta},\boldsymbol{y} \sim N(\boldsymbol{F}\boldsymbol{f}, \boldsymbol{F})$, where $\boldsymbol{F} = (\widetilde{\boldsymbol{Q}} + \boldsymbol{D}^{-1})^{-1}$ and $\boldsymbol{f} = \boldsymbol{D}^{-1}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})$.

We recall that this scheme also has the same dense matrix $\boldsymbol{F} = (\widetilde{\boldsymbol{Q}} + \boldsymbol{D}^{-1})^{-1}$, however the main advantage of the composite MCMC approach is to draw samples of $\boldsymbol{\theta}$ and $\boldsymbol{\beta}$, and then use these samples to recover $\boldsymbol{w}$. Further, Finley et al. (2019) argued that the blocking and sampling schemes of the composite sampling result in good convergence properties.

C. Supplementary simulation results

In this section more results on simulations are presented. Figure S1 and Figure S2 show the criteria assessment and time requirements for scenarios SIM I ($\phi = 12$) and SIM II ($\phi = 6$). In general, NNGP models with $nb = 4$ or $nb = 6$ neighbor blocks show a better performance in terms of computational cost and goodness of fit. Almost all the models run in less than an hour, and in the best scenarios, they run in less than 1000 seconds, showing the great advantage of running the block-NNGP models through INLA. It is observed that the computing times requirements for the block-NNGP models decreases as the number of neighbor blocks increases. While for the NNGP models, the computational cost increases as the number of neighbor increases. The LPML for all block-NNGP models with $nb = 4$ or $nb = 6$ neighbor blocks did not significantly change. A similar pattern is presented for NNGP-models with $nb = 20$ or higher. The WAIC for block-NNGP models with regular blocks, $nb = 4$ or $nb = 6$ neighbor blocks are quite similar, while for irregular blocks, it tend to increase its value when the number of neighbor blocks increases.

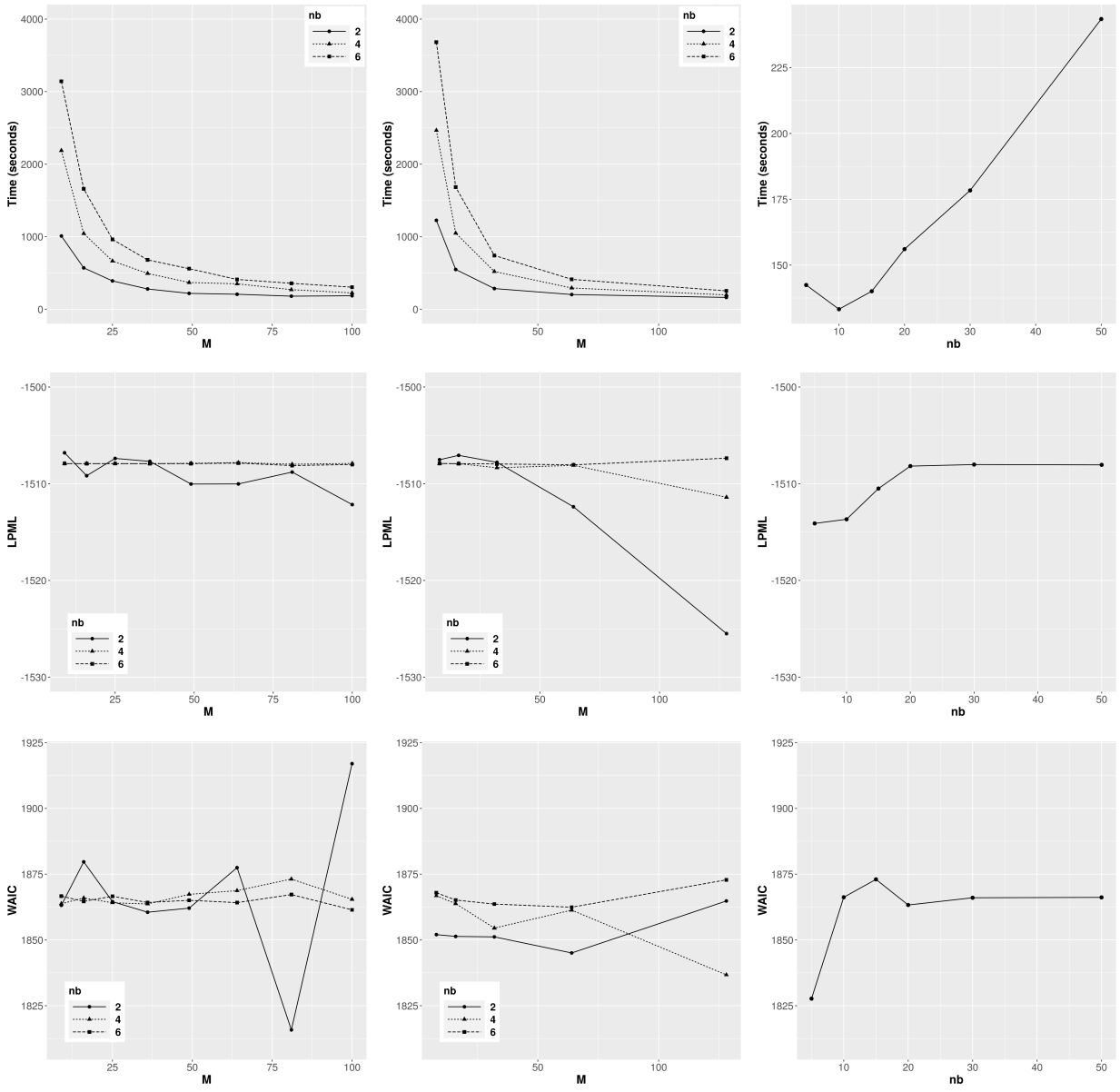

Figure S1: SIM I ($\phi = 12$). INLA results. Criteria assessment: Running times (first row), LPML (second row) and WAIC (third row), under block-NNGP models using regular blocks (left column), irregular blocks (middle column) and NNGP models (right column).

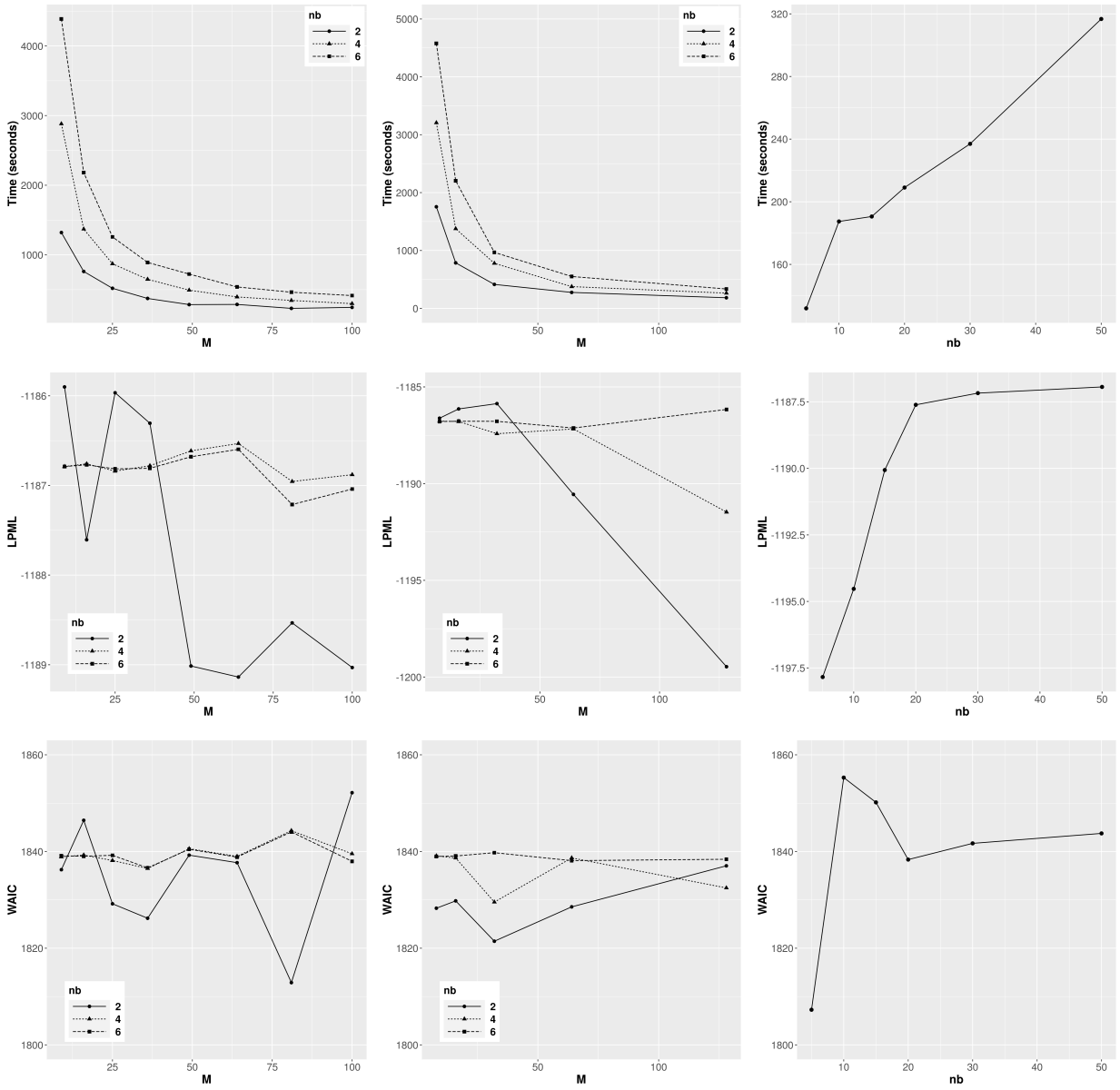

Figure S2: SIM II ($\phi = 6$). INLA results. Criteria assessment: Running times (first row), LPML (second row) and WAIC (third row), under block-NNGP models using regular blocks (left column), irregular blocks (middle column) and NNGP models (right column).

The posterior mean estimation of w_S for scenarios SIM I ($\phi = 12$) and SIM II ($\phi = 6$) are displayed in Figure S3 and Figure S4, respectively. These results confirm that the NNGP and block-NNGP models show a very good performance when the range is not too large and, as the range increases, the larger the credible intervals of the spatial effects.

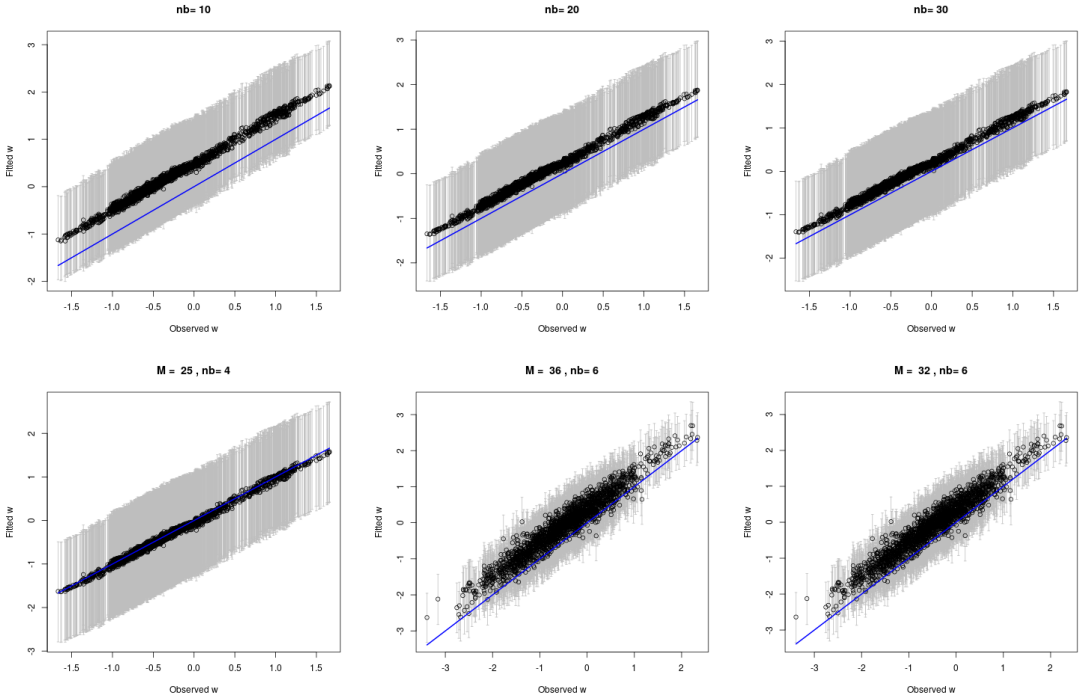

Figure S3: SIM I ($\phi = 12$). INLA results. Mean posterior estimates of spatial effects for different NNGP (upper panel) and block-NNGP models (lower panel).

Parameter estimates for specific models under SIM III ($\phi = 3$) are provided in Table S1. In general, the posterior mean estimates for all the models fitted are close to the true parameter values which are included within credible intervals, but the posterior mean estimates of β_0 for the NNGP are far away from the true values.

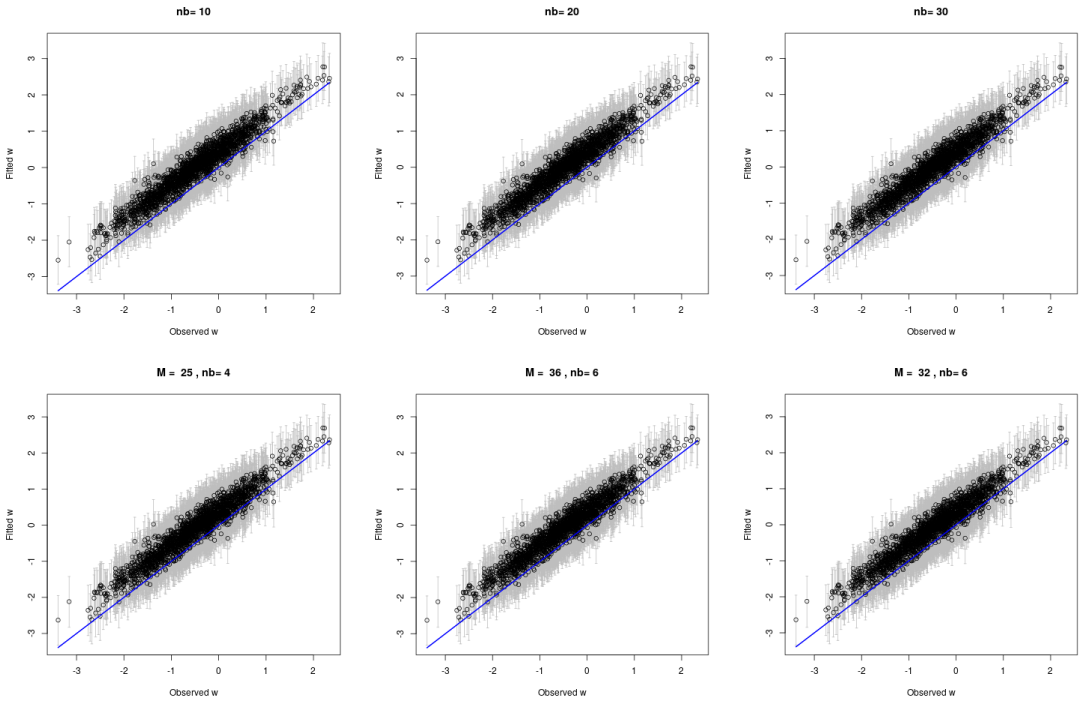

Figure S4: SIM II ($\phi = 6$). INLA results. Mean posterior estimates of spatial effects for different NNGP (upper panel) and block-NNGP models (lower panel).

Figure S5 displays the posterior mean estimates of the spatial random effects interpolated over the domain for SIM III ($\phi = 3$). The block-NNGP models show better approximations than the NNGP models. We can observe that the NNGP model with 10 neighbors did not approximate well the spatial process, it overestimate the spatial effects, specially in the south region. Indeed, the NNGP has proven to be successful in capturing local/small-scale variation of spatial processes, however, it might have one disadvantage: inaccuracy in representing global/large scale dependencies. This might happen because the NNGP built the DAG based on observations, adversely, the block-NNGP built a chain graph based on blocks of observations, which captures both small and large dependence. The previous result for large r is magnified when the spatial process is smoother.

Table S1: SIM III ($\phi = 3$). INLA results. Summary of posterior mean parameter estimates, parameter posterior summary credible intervals (2.5, 97.5) and criteria assessment.

		NNGP (10)	**NNGP (20)**	(R)**M=49 nb=4**	(R)**M=64 nb=4**	(I)**M=32 nb=4**	(I)**M=64 nb=4**
β_0	1	0.283 (-0.653,1.053)	0.412 (-0.500,1.203)	0.594 (-0.321,1.501)	0.585 (-0.300,1.459)	0.592 (-0.237,1.504)	0.622 (-0.249,1.523)
β_1	5	4.994 (4.977,5.011)	4.994 (4.977,5.011)	4.994 (4.977,5.011)	4.994 (4.977,5.011)	4.994 (4.977,5.011)	4.994 (4.977,5.011)
σ^2	1	1.081 (0.845,1.250)	1.044 (0.813,1.208)	1.053 (0.829,1.216)	1.045 (0.815,1.209)	1.019 (0.775,1.183)	1.036 (0.804,1.200)
ϕ	3	3.168 (2.473,3.717)	3.278 (2.560,3.851)	3.240 (2.559,3.780)	3.286 (2.578,3.852)	3.437 (2.652,4.084)	3.322 (2.597,3.907)
τ^2	0.1	0.092 (0.088,0.096)	0.092 (0.088,0.096)	0.092 (0.088,0.095)	0.092 (0.088,0.095)	0.091 (0.087,0.095)	0.092 (0.088,0.095)
Time (sec)		240.515	**171.927**	496.271	359.056	687.980	335.161

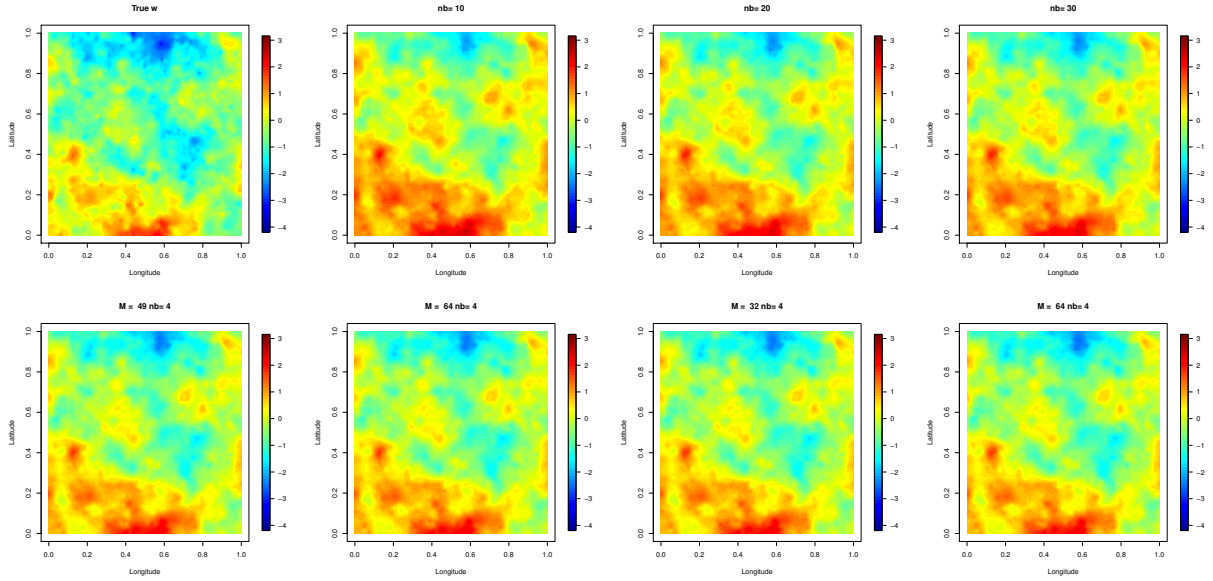

Figure S5: SIM III ($\phi = 3$). True spatial random effects w, and their posterior mean estimates for NNGP models (upper panel) with $nb = 10, 20, 30$ neighbors, and different block-NNGP models (lower panel) with regular blocks (R) and irregular blocks (I), using INLA.

The root of mean square prediction error (RMSP) for all the fitted models, under SIM III, is shown in Figure S6. For block-NNGP models, as expected the prediction metric increases when the number of blocks is increased, being lower for a higher number of neighbor blocks. This result is a little bit more evident for irregular blocks. For NNGP models, the RMSP is higher when the number of neighbors is small.

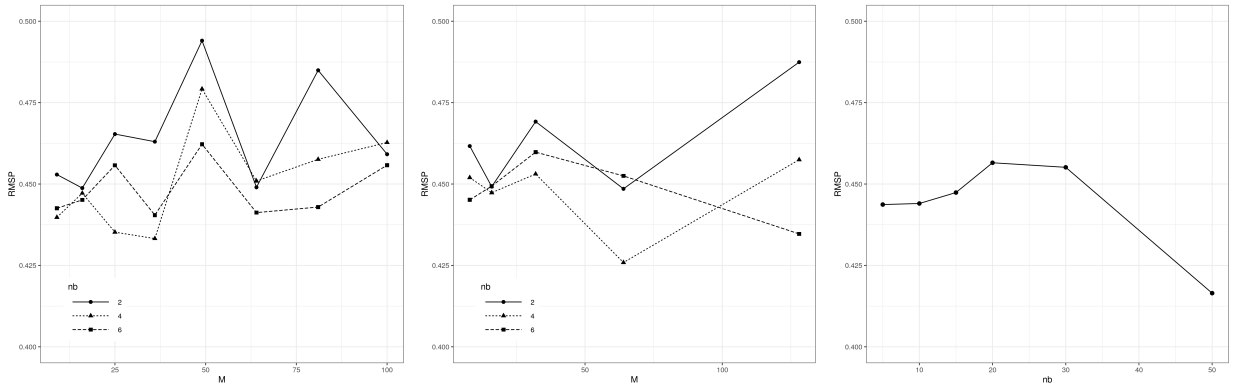

Figure S6: SIM III ($\phi = 3$). INLA results. Root mean square prediction error under block-NNGP models using regular blocks (left column), irregular blocks (middle column) and NNGP models (right column).

We also simulate data from the model in Eq. (9) but setting $\nu = 1.5$. We computed the approximate posterior marginals for the fixed effects (Figure S7). The block-NNGP models give more accurate results than the NNGP models. For block-NNGP models, the posterior marginals for β_0 and β_1 with different number of blocks are overlaid. A different result is obtained for NNGP models, where the posterior marginals for β_0 with different number of neighbors dramatically change.

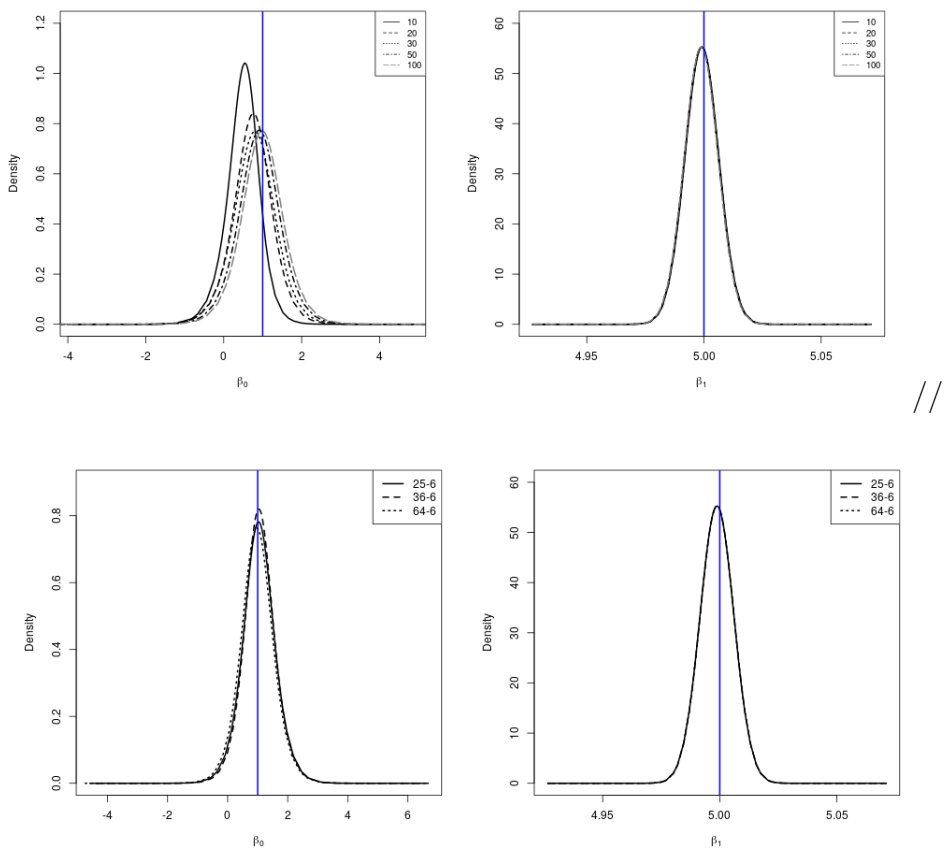

Figure S7: INLA results for simulation of GP with Matérn covariance function ($\nu = 1.5$, $\phi = 3.5$, $\sigma^2 = 1$ and $\tau^2 = 0.1$). Posterior marginal densities of regression coefficient effects for NNGP models (upper panel)with $nb = 10, 20, 30, 50, 100$ neighbors and block-NNGP models (lower panel) with $M = 25, 36, 64$ regular blocks and $nb = 6$ neighbor blocks. The solid vertical blue lines represent the true parameter values.

Fig. S8 shows the posterior mean spatial effect estimates compared to the simulated ones for block-NNGP models and NNGP models. This result shows the inaccuracy of NNGP models when the number of neighbors is small.

Finally, we also computed the theoretical Matérn covariance function (black line) and the empirical covariance function for the NNGP models and block-NNGP models (blue dots) in Figure S9 for this scenario. In general, the match between the theorical and empirical covariance is better for block-NNGP models, specially when the number of blocks is smaller. We further note that for NNGP models when the number of neighbors is

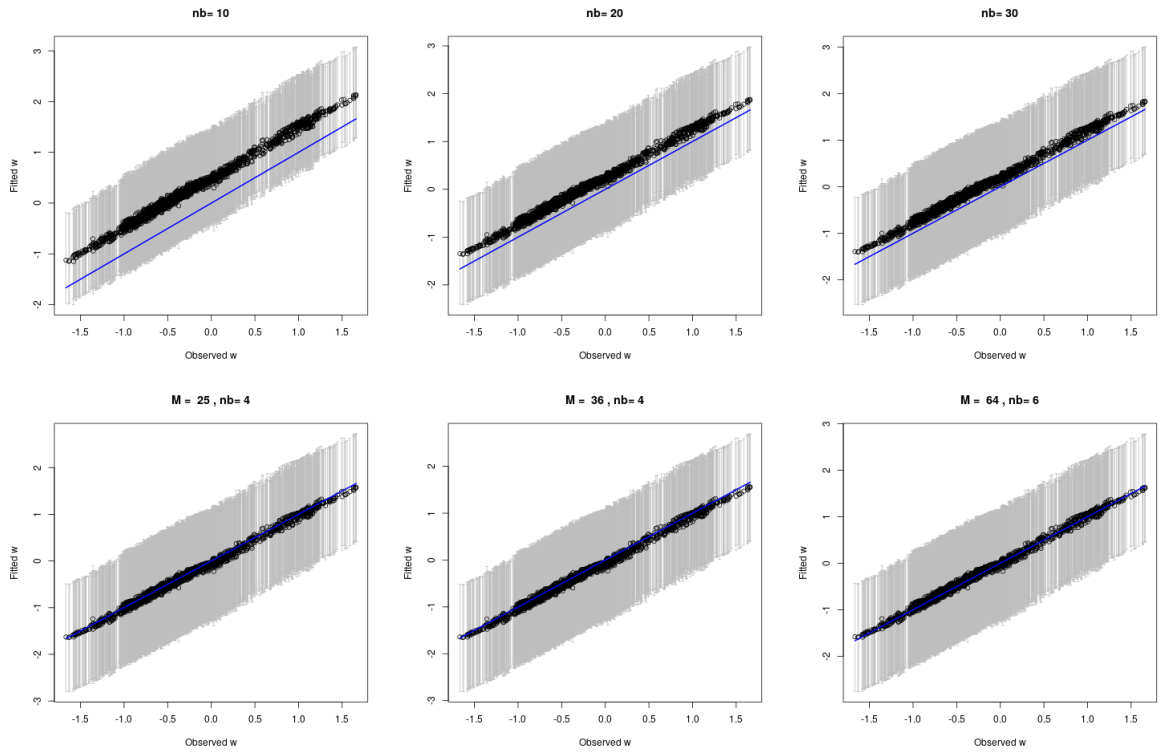

Figure S8: INLA results for simulation of GP with Matérn covariance function($\nu = 1.5$, $\phi = 3.5$, $\sigma^2 = 1$ and $\tau^2 = 0.1$). Mean posterior estimates of spatial effects for different NNGP (upper panel) and block-NNGP models using regular blocks (lower panel).

small and the range is large, the theoretical covariance function is far away from the true covariance function.

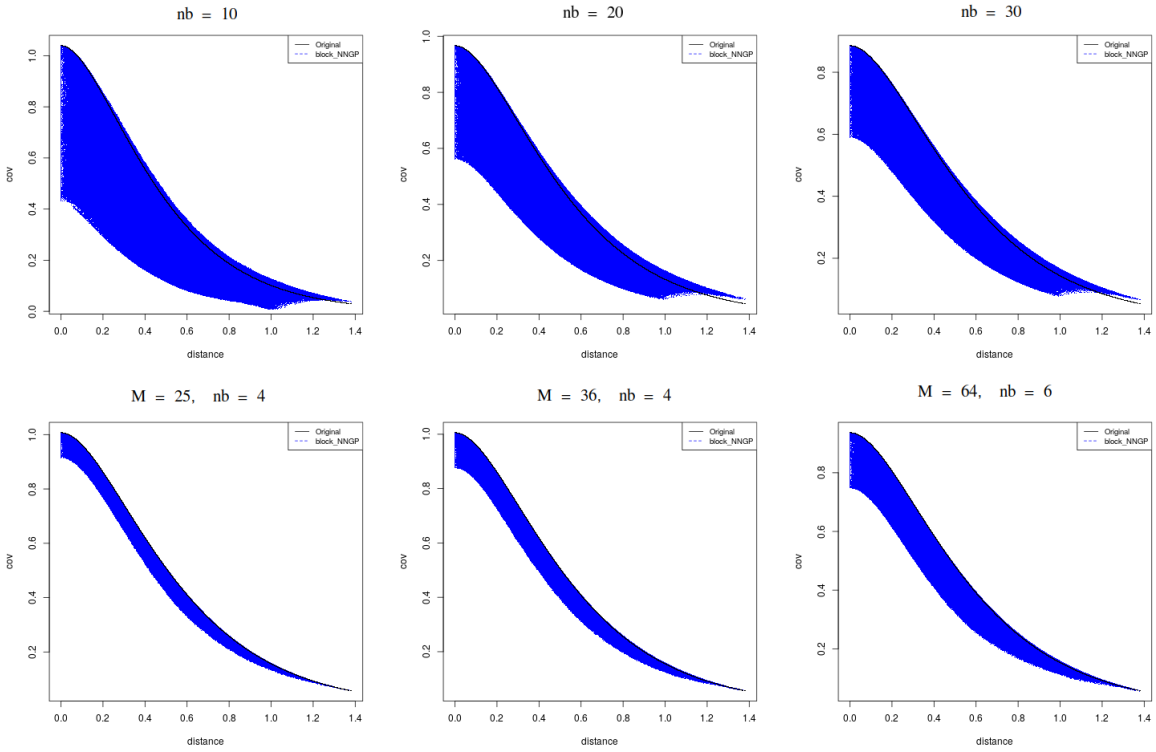

Figure S9: True Matérn covariance function ($\nu = 1.5$, $\phi = 3.5$, $\sigma^2 = 1$ and $\tau^2 = 0.1$) of GP against distance (black lines) and empirical approximated covariance of block-NNGP (Regular blocks) against distance (blue dots) for different NNGP (upper panel) and block-NNGP models (lower panel).

D. Supplementary application results

Fig. S10 shows an example of DAGs, built using one location for each block, for the mining and precipitation data used in applications.

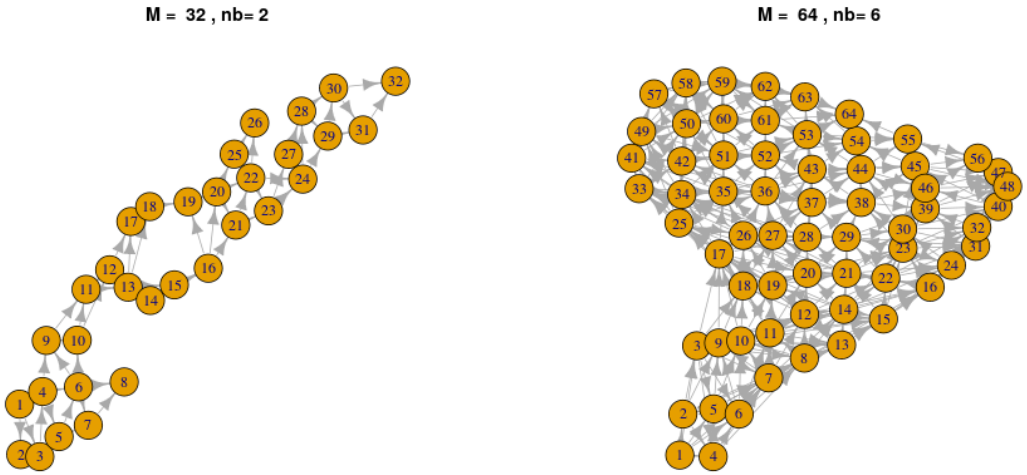

Figure S10: DAG of blocks for mining data (left) using $M = 32$ blocks and $nb = 2$ neighbor blocks. DAG of blocks for precipitation data (right) using $M = 64$ blocks and $nb = 6$ neighbor blocks.

Figure S11 shows maps of interpolated posterior mean estimates of joint-frequency data. We see little difference between these models.

Table S2 presents the selection criteria of the fitted models. This result shows that it is quite difficult to choose the number of neighbors, because there is not a clear pattern that the more neighbors, the better the model. While the best block-NNGP model is the one with $M = 64$ and $nb = 4$ neighbor blocks, followed by the model with $M = 64$ and $nb = 6$ neighbor blocks. Overall, the results are more stable for less blocks and more neighbor blocks. Table S2 also presents the total running time for each model.

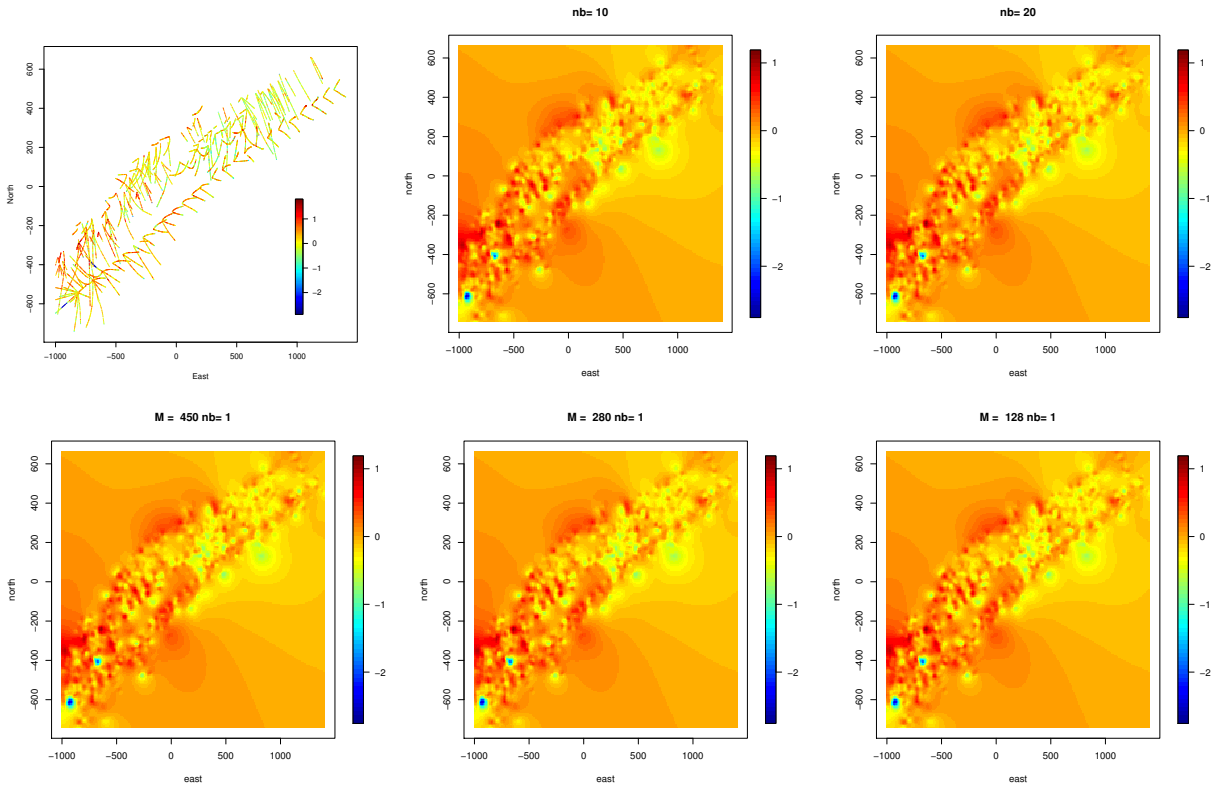

Figure S11: Original joint-frequency data (left upper plot). INLA results. Mean posterior of joint-frequency data using NNGP models with $nb = 10, 20$ neighbors (upper panel). Mean posterior of joint-frequency data using block-NNGP models with regular blocks (M=450 and M=280) and irregular blocks (M=128), and nb=1 neighbor block (lower panel).

Table S2: Precipitation data. INLA results. Criteria assessment and time requirements.

	M	nb	LPML	WAIC	RSME	RSMP	time (sec)
NNGP		10	-18407.420	-11705.390	0.0052	0.523	1535.664
		20	-18863.340	-11750.800	0.0051	0.467	2588.953
		30	-19167.690	-11666.300	0.0051	0.626	1356.085
		50	-19228.520	-11680.250	0.0051	0.561	1745.064
		100	-18875.250	-11783.170	0.0050	0.508	5019.320
block-NNGP (I)	64	2	-19533.230	-11642.630	0.0050	0.598	7191.660
	64	4	-19141.680	-11736.000	0.0050	0.680	13587.350
	64	6	-19436.940	-11708.210	0.0050	0.645	19992.250
	128	2	-19661.130	-11595.540	0.0049	0.645	2740.382
	128	4	-19483.990	-11625.660	0.0051	0.595	5434.684
	128	6	-19381.560	-11680.610	0.0050	0.742	5092.628

References

Abrahamsen, P. (1997). A review of gaussian random fields and correlation functions. Technical report.

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

Finley, A. O., Datta, A., Cook, B. D., Morton, D. C., Andersen, H. E., and Banerjee, S. (2019). Efficient algorithms for bayesian nearest neighbor gaussian processes. *Journal of Computational and Graphical Statistics*, 28(2):401–414. PMID: 31543693.

Liu, J. S. (1994). The collapsed gibbs sampler in bayesian computations with applications to a gene regulation problem. *Journal of the American Statistical Association*, 89(427):958–966.