Statistica Sinica Preprint N S-2018-0005	
Title	Spatial Factor Mod for High-Dimen and arge Spatial Data: An Appli Mapping
Manuscript ID	SS-2018-0005
URL	http://www.stat.sinica.edu.tw/stati
DOI	10.5705/ss.202018.0005
Complete List of Authors	Daniel Taylor-R iguez Andrew O. Finle Abhirup tta had ck Ha Erin en ce Cook s C. Morton and ud o Banerjee
Correspond Auth	Andrew O. Finley
	finleya@msu.edu
tice: Accept ver subject to English editing.	

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0005

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 author*

Abstract: Gathering information about forest variables is an expensive and arduous activity. Therefore, directly collecting ground data is used to produce high-resolution maps over large spatial domains. Next-generation collection initiatives for remotely sensed light detection and ranging (LiDAR) data are specifically aimed at producing complete-coverage maps over large spatial domains. Given that LiDAR data and forest characteristics are often strongly correlated, it is possible to use the former to model, predict, and map forest variables over regions of interest. This is challenging with high-dimensional ($\sim 10^2$) spatially dependent LiDAR outcomes observed at a large number of locations ($\sim 10^5 - 10^6$). With this in mind, we propose a spatial factor nearest neighbor Gaussian process (SF-NNGP) model, which we implement in a two-stage approach that connects the spatial structure found in LiDAR signals with forest variables. We provide a simulation experiment that demonstrates the inferential and predictive performance of the SF-NNGP, and use the two-stage modeling strategy to generate complete-coverage maps of the forest variables, with associated uncertainty, over a large region of boreal forests in

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0005

interior Alaska.

Key words and phrases: LiDAR data, forest outcomes, nearest neighbor Gaussian processes, spatial prediction.

1. Introduction

Strong relationships between remotely sensed [illegible]tection and rangi[illegible] DAR) data and forest variables have been do[illegible]umented in the [illegible] ner et al., 2009; Babcock et al., 2013; Næsset, 2014). [illegible]rested settings, LiDAR data provide a high-dimensional signal that ch[illegible]cterizes the vertical structure of the forest canopy at point-refe[illegible]ations. Traditionally, LiDAR data acquisition campa[illegible]s have sought complete-coverage at a high spatial resolution over relatively s[illegible]al domains, resulting in a fine grid of point-referenced LiDAR signals. [illegible]ch settings, the link between the LiDAR data and the fo[illegible]easurements on sparsely sampled forest inventory plots has [illegible]loited to create high-resolution complete-coverage predic[illegible] forest variables. Commonly, this link is established b[illegible] first [illegible]g the relevant features of the high-dimensional LiDA[illegible] [illegible]mension-reduction step (Babcock et al., 2015; Junttila and [illegible]e, [illegible]). Then the LiDAR features are used as predictors in a regression [illegible] to explain the variability in the spatially coinciding forest variable[illegible]es. Lastly, the model is applied to predict the forest outcomes at all locations across the domain where LiDAR signals have been observed.

Considerably more ambitious next-generation LiDAR collection initia-

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0005

tives, such as ICESAT-2 (ICESat-2, 2015), Global Ecosystem Dynamics Investigation LiDAR (GEDI) (GEDI, 2014), and NASA Goddard's LiDAR, Hyper-Spectral, and Thermal imager (G-LiHT) (G-LiHT, 2016), seek to quantify and map forest variables over vast spatial extents. To fulfill their goals in a cost-effective manner, these data-gathering programs do not collect LiDAR data over the entire domain. Instead, they sparsely sample locations across the domain extent and over forest inventory plots (i.e., where forest variables have been measured). While [illegible] coverage high-resolution maps of forest outcomes remains the primary intended use of these data, there is also interest in creating maps of LiDAR [illegible] over non-sampled locations and assessing the spatial dependence within and among LiDAR signals.

Our motivating application focuses on forest variable prediction and mapping in the boreal forests of interior Alaska using sparsely sampled LiDAR and forest variable measurements. Within these regions, acquiring complete-coverage LiDAR data is prohibitive from a cost perspective (Andersen et al., 2011; Bolton et al., 2013; Nelson et al., 2012). Because generating complete-coverage maps of forest variables (and perhaps LiDAR signals) is still the goal, the [illegible] sampled LiDAR must be leveraged to inform the forest variable predictions. One attractive solution is to move the LiDAR predictor variables to the left-hand side of the regression and then to model them jointly with the forest outcomes. When the number of LiDAR and forest variables is small, such joint models are possible via linear models of coregionalization; for example, see, Babcock et al. (2017) and Finley et al.

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0005

(2014a). Alternatively, if the LiDAR signal is high-dimensional, but observed at a small number of locations, reduced-rank models can be employed. For example, Banerjee et al. (2008), Ren and Banerjee (2013), and Finley et al. (2017) applied a reduced-rank *predictive process* modeling strategy to analyze similar high-dimensional data. However, such approaches cannot scale to data sets with tens of thousands of locations and can yield poor predictive performance (Stein, 2014).

Models able to handle high-dimensional signals observed over a large number of locations and capable of estimating within and among location dependence structures are needed. Recent modeling developments reviewed in Heaton et al. (2017) and Banerjee (2017) highlight several options for robust and practical approximation of univariate Gaussian process (GP) models. A subset of these models can be easily extended to accommodate relatively small multivariate responses (e.g., spectral process) for example, see (Datta et al., 2016a). Nevertheless, for our particular application, we require an approach that can cope with both high-dimensional LiDAR measurements, ∼50 outcomes at a location, and up to the large collection of observed locations.

The nearest-neighbor Gaussian process (NNGP) developed in Datta et al. (2016a), Datta et al. (2016b), and Datta et al. (2016c) can be used with a large number of locations, because its scalability is not mediated by the number of observed locations, but rather by the size of the nearest neighbor set considered—a quality that yields minimal storage and computational requirements. These models belong to the class of methods that induce sparsity on the spatial precision matrix, and exploit the natural representation

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0005

of sparsity provided by graphical models (Lauritzen, 1996; Murphy, 2012) to build a sparse GP that accurately approximates the original dense GP.

To tackle the high-dimensional LiDAR data set, we develop a Bayesian NNGP spatial factor model (SFM), referred to as the SF-NNGP. Following Christensen and Amemiya (2002), Hogan and Tchernis (2004), and Ren and Banerjee (2013), the SFM structure enables approximating the cross-dependence between multivariate (spatially dependent) outcomes through a low-dimensional set of spatial factors, alleviating the computational bottleneck of working directly with high-dimensional outcomes. The SF-NNGP allows us to model and map the LiDAR signals on both observed and unobserved locations, and, conditioning on the LiDAR spatial signatures, we can similarly map the forest variables over the entire spatial domain of interest. Furthermore, using a Bayesian approach for model fitting enables us to equip the derived estimates and predictions with associated measures of uncertainty, an essential requirement of many high-profile inferential studies. Our methods are fully implemented in C++, using BLAS (Blackford et al., 2001; Zhang, 2016) to leverage efficient multiprocessor matrix operations and openMP (Dagum and Menon, 1998) to improve key aspects of the algorithm through parallelization.

The structure of the remainder of this paper is as follows. Section 2 introduces the Wanza Creek data set. In Section 3, we formulate the proposed hierarchical Bayesian modeling strategy. Section 4 presents an analysis of a synthetic data set to validate the performance of the SF-NNGP model. Using the available LiDAR and forest inventory data, in Section 5, we develop and validate a predictive model for the forest variables. We close by providing

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0005

insights, recommendations, and directions for future in Section 6.

2. Data Description

The Bonanza Creek Experimental Forest (BCEF) is a Long-Term Ecological Research (LTER) site consisting of vegetation and landforms typical of interior Alaska. The BCEF is 21,000 ha and includes a section of Tanana River floodplain along the southeastern borders (Bonanza Creek LTER, 2016). Figure 1 shows the location and [illegible] F data detailed in this section.

Forest variables were collected on 197 plots in 2[illegible] the USDA Forest Service Forest Inventory and Analysis Program protocol (Bechtold and Patterson, 2005). We consider three forest variables commonly used by forest professionals to make management decisions: above-ground biomass (AGB); tree density (TD); and basal area (BA). The AGB for individual trees was estimated using the Component Ratio Method described in Woodall et al. (2015). The TD for a plot is expressed in thousands of trees per hectare. The BA for a plot is the sum of the individual trees' cross-sectional areas in m^2 at breast height and is scaled to a per hectare basis.

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0005

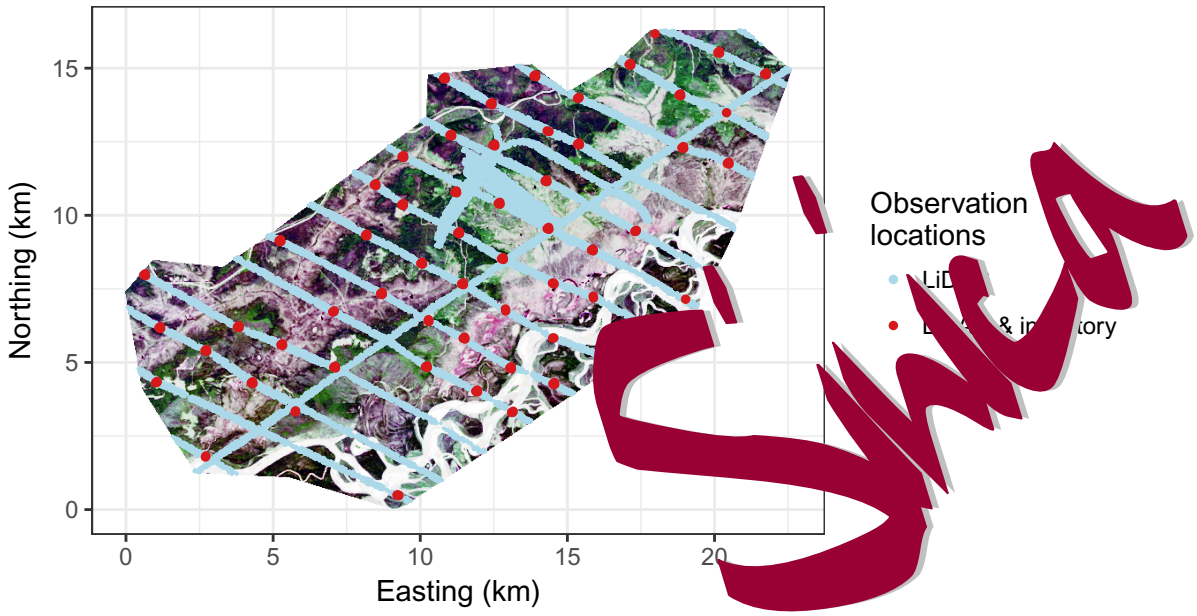


Figure 1: Bonanza Creek Experimental Forest extent with color enhanced Landsat image and locations where the LiDAR signals were measured (*LiDAR* in the legend) and locations where both LiDAR signals and forest variables were measured (*LiDAR & inventory* in the legend).

In the summer of 2012, LiDAR data were collected using a flight-line strip sampling approach by NASA Goddard's G-LiHT sensor (Cook et al., 2013), which is a portable multisensor system that accurately characterizes complex terrain and the vertical distribution of canopy elements (Jakubowski et al., 2013; Wu et al., 2013). Point cloud information was summarized to a 15×15 m grid cell size to approximate field plot areas. Over each grid cell, psuedo-waveforms were generated by calculating the LiDAR return count densities for .5 m height bins between 0 and 28.5 m (i.e., 57 LiDAR outcomes per location). The LiDAR return count density for height bin l is defined as the number of returns in height bin l divided by the total number of

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0005

LiDAR returns over the grid cell. Identical LiDAR psuedo-waveforms were obtained using point clouds extracted over each field plot. G-LiHT data for the study area are available online at `https://gliht.gsfc.nasa.gov`. For this analysis, 50,197 LiDAR observations were used for model-fitting.

A Landsat 8 top of atmosphere (TOA) reflectance product was prepared for the BCEF area for June of 2015. The June 2015 image was preferred to the June 2014 image owing to the excessive cloud cover in the latter image. A tasseled cap transformation was applied to the Landsat 8 TOA reflectance bands to obtain *brightness*, *greenness*, and *wetness* tasseled cap indices (Baig et al., 2014). These indices are used as covariates in the subsequent analysis.

Further details on the data set and our analysis are provided in Section 5.

3. Modeling Strategy

Our goal is to model and predict uncertainty-equipped predictions of forest variables using information contained in LiDAR signals. Consider a LiDAR signal, $\mathbf{z}(\cdot)$, observed at a finite collection of locations, $\mathcal{T}_z = \{\mathbf{s}_1, \ldots, \mathbf{s}_{n_z}\}$, and a set of forest outcomes, $\mathbf{y}(\cdot)$, observed at locations in the set $\mathcal{T}_y = \{\mathbf{r}_1, \ldots, \mathbf{r}_{n_y}\} \subset \mathcal{D}$. Furthermore, let $\mathcal{T}_\emptyset = \{\mathbf{t}_1, \ldots, \mathbf{t}_{n_\emptyset}\}$ denote a set of locations where neither LiDAR signals nor forest outcomes are available, but where predictions are of interest. Thus, the set of locations where both LiDAR and forest outcomes are mapped corresponds to $\mathcal{T} = (\mathcal{T}_z \cup \mathcal{T}_\emptyset)$, with $\mathcal{T} \subset \mathcal{D} \subset \mathbb{R}^2$, where $\mathcal{D}$ is the spatial domain of interest. Note that although

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0005

$\mathbf{z}(\cdot)$ and $\mathbf{y}(\cdot)$ are "observed" at locations in $\mathcal{T}_z$ and $\mathcal{T}_y$, respectively, we allow for missing values that are to be imputed in these sets. We make this distinction because locations where imputation is performed are part of the model fitting, whereas for locations in $\mathcal{T}_\emptyset$, predictions are drawn *ex post facto* from the posterior predictive distribution; see Section 3.4.

The LiDAR signals are high-dimensional vectors of measurements (h_z whereas the forest outcomes are relatively small-dimensional ($h_y << h_z$), assumed to have support on $\mathbb{R}^{h_y}$. The LiDAR signals are strongly dependent on each other; LiDAR signals vary with the composition of a forest, and, as a plethora of examples in the literature have demonstrated (Ene et al., 2018; Finley et al., 2014b; Nelson et al., 2017), variability in forest outcome variables can be partially explained by LiDAR characteristics.

3.1 Linking the LiDAR and Forest Inventory Data

We seek to connect the forest outcomes and LiDAR signals as a two-step process. First, we formulate a generative model to extract the spatial signature from the LiDAR data at locations in $\mathcal{T}_z$, which can also be used to interpolate LiDAR signals at $\mathcal{T}_\emptyset$. Along with other spatially referenced predictors, the LiDAR spatial signatures for locations in $\mathcal{T}_y$ are used as predictors to build the model for the forest outcomes. Moreover, a component that captures the spatial variation exclusive to the forest outcomes can also be specified, if

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0005

required. For $\mathbf{s} \in \mathcal{D}$, this two-stage model is given by

$$\text{Stage 1: } \mathbf{z}(\mathbf{s}) = \mathbf{X}_z(\mathbf{s})'\boldsymbol{\beta}_z + \mathbf{w}^\star(\mathbf{s}) + \boldsymbol{\varepsilon}_z(\mathbf{s}), \tag{3.1}$$

$$\text{Stage 2: } \mathbf{y}(\mathbf{s}) = \mathbf{X}_y(\mathbf{s})'\boldsymbol{\beta}_y + \boldsymbol{\Upsilon}\mathbf{w}^\star(\mathbf{s}) + \mathbf{v}^\star(\mathbf{s}) + \boldsymbol{\varepsilon}_y(\mathbf{s}). \tag{3.2}$$

Note that the influence of $\mathbf{z}(\mathbf{s})$ over $\mathbf{y}(\mathbf{s})$ in (3.2) is exerted solely through its spatial component, $\mathbf{w}^\star(\mathbf{s})$. There are several arguments in favor of this approach, as opposed to substituting $\mathbf{z}(\mathbf{s})$ or $\boldsymbol{\mu}_z(\mathbf{s}) = \mathbf{X}_z(\mathbf{s})'\boldsymbol{\beta}_z$ and $\mathbf{w}^\star(\mathbf{s})$ as covariates directly into (3.2). Among these, and most important for our setting, $\mathbf{z}(\mathbf{s})$, $\boldsymbol{\mu}_z(\mathbf{s})$ and $\mathbf{w}^\star(\mathbf{s})$ are all high-dimensional objects, so using $\mathbf{w}^\star(\mathbf{s})$ reduces the dimensionality of the problem by casting it under the factor model structure, as shown in Section 3.2. In addition, the elements within $\mathbf{z}(\mathbf{s})$ are strongly correlated, hence multicollinearity issues would arise if it was included directly in (3.2).

In (3.1) and (3.2), the terms $\mathbf{X}_z(\mathbf{s})'\boldsymbol{\beta}_z$ and $\mathbf{X}_y(\mathbf{s})'\boldsymbol{\beta}_y$ capture large-scale variation. For $\kappa \in \{z, y\}$, $\mathbf{X}_\kappa(\mathbf{s})'$ represents a fixed $h_\kappa \times p_\kappa$ block-diagonal matrix of spatially referenced predictors, where $p_\kappa = \sum_{j=1}^{h_\kappa} p_{\kappa,j}$, having as its jth diagonal block the length-$p_{\kappa,j}$ vector $\mathbf{x}_j^\kappa(\mathbf{s})'$. The length-$p_\kappa$ vector $\boldsymbol{\beta}_\kappa$ corresponds to the regression coefficients associated with $\mathbf{X}_\kappa(\mathbf{s})'$. The vectors $\mathbf{w}^\star(\mathbf{s})$ and $\mathbf{v}^\star(\mathbf{s})$ are h_z- and h_y-dimensional zero-centered stochastic processes over $\mathcal{D}$, respectively. The process $\mathbf{w}^\star(\mathbf{s})$ captures the spatial variation of $\mathbf{z}(\mathbf{s})$, and $\mathbf{v}^\star(\mathbf{s})$ synthesizes additional spatial variation in the forest outcomes. The $h_y \times h_z$ matrix $\boldsymbol{\Upsilon}$ connects the spatial information extracted from the LiDAR model into the forest outcomes model. The vectors $\boldsymbol{\varepsilon}_z(\mathbf{s}) \sim \mathrm{N}_{h_z}(\mathbf{0}, \boldsymbol{\Psi}_z)$ and

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0005

$\boldsymbol{\varepsilon}_y(\mathbf{s}) \sim \mathrm{N}_{h_y}(\mathbf{0}, \boldsymbol{\Psi}_y)$ represent uncorrelated random errors (i.e., $\boldsymbol{\Psi}_z$ and $\boldsymbol{\Psi}_y$ are diagonal) at finer scales.

Implementing this modeling strategy directly is challenging owing to the high-dimensionality of the LiDAR signals ($h_z \sim 50$) and the massive number of spatially dependent observations ($n \sim 10^5$). Thus, it is impossible to attempt using common computing resources. In the following section we formulate a viable alternative to models (3.1) and (3.2).

3.2 The Spatial Factor NNGP Model

To make models (3.1) and (3.2) tractable with limited computing power, we combine a dimension-reduction approach and a sparsity-inducing technique. In particular, we introduce the SF-NNGP model which brings together the SFM structure (Schmidt and Gelfand, 2003; Lopes et al., 2008; Zhang, 2007; Ren and Banerjee, 2013) with NNGPs (Datta et al., 2016b,c,a).

While the SFM structure enables the analysis of high-dimensional response vectors by using linear combinations of a relatively small number of independent stochastic processes, NNGPs make it possible to fit spatial process models when the number of spatial observations is particularly large. NNGPs approximate the *parent* (dense) GP using the natural representation of sparsity provided by graphical models (Lauritzen, 1996; Murphy, 2012), by assuming conditional independence—where conditioning is on the nearest neighbors—with locations outside of the neighbor set. The result is a proper (but sparse) GP that accurately approximates the original dense GP. In contrast to other sparsity-inducing approaches, NNGPs allow for interpolation

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0005

at unobserved locations and can be used to make full inference on model parameters, including the latent processes. Combining the SFM structure with NNGPs provides a methodology capable of coping simultaneously with high-dimensional response vectors and a large number of spatially dependent observations.

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

$$\begin{aligned}\mathcal{H}(\mathbf{h}\,|\,\boldsymbol{\phi}) &= \mathrm{cov}(\mathbf{\Lambda}\,\mathbf{w}(\mathbf{s}), \mathbf{\Lambda}\,\mathbf{w}(\mathbf{s}+\mathbf{h})) \\ &= \sum_{k=1}^{q_w} \mathcal{C}_k(\mathbf{h}\,|\,\phi_k)\boldsymbol{\lambda}_k\boldsymbol{\lambda}_k', \end{aligned} \tag{3.3}$$

for locations $\mathbf{s}, \mathbf{s}+\mathbf{h} \in \mathcal{D}$, where $\mathcal{C}_k(\cdot\,|\,\phi_k)$ denotes a univariate parametric correlation function, and $\boldsymbol{\lambda}_k$ is the kth column of $\mathbf{\Lambda}$. This cross-covariance matrix is induced by q-variate ($q \leq l$) spatial factors $\mathbf{w}(\mathbf{s})$ with *independent* components $w_k(\cdot) \sim \mathrm{GP}(0, \mathcal{C}_k(\cdot\,|\,\phi_k))$.

As such models (3.1) and (3.2) can be reformulated as SF-NNGPs by [illegible] the spatial processes $\mathbf{w}^\star(\mathbf{s})$ and $\mathbf{v}^\star(\mathbf{s})$ as

$$\mathbf{w}^\star(\mathbf{s}) = \mathbf{\Lambda}_z\mathbf{w}(\mathbf{s}) \quad \text{and} \quad \mathbf{v}^\star(\mathbf{s}) = \mathbf{\Gamma}\,\mathbf{v}(\mathbf{s}), \tag{3.4}$$

where the matrices $\mathbf{\Lambda}_z = ((\lambda_{hk}^{(z)}))_{h_z \times q_w}$ and $\mathbf{\Gamma} = ((\gamma_{lr}))_{h_y \times q_v}$ correspond to the

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0005

factor loadings matrices, and the new spatial factors for $\mathbf{s} \in \mathcal{D}$ are given by

$$\begin{aligned} \mathbf{w}(\mathbf{s}) &\sim \prod_{k=1}^{q_w} \text{NNGP}\left(0, \tilde{\mathcal{C}}(\cdot \,|\, \phi_k^w)\right), \text{ and} \\ \mathbf{v}(\mathbf{s}) &\sim \prod_{r=1}^{q_v} \text{NNGP}\left(0, \tilde{\mathcal{C}}(\cdot \,|\, \phi_r^v)\right). \end{aligned}$$

The expressions $\text{NNGP}\left(0, \tilde{\mathcal{C}}(\cdot \,|\, \phi_k^w)\right)$ and $\text{NNGP}\left(0, \tilde{\mathcal{C}}(\cdot \,|\, \phi_r^v)\right)$ denote [illegible] derived from the parent processes $\text{GP}(0, \mathcal{C}(\cdot \,|\, \phi_k^w))$ and $\text{GP}(0, \mathcal{C}(\cdot \,|\, \phi_r^v))$, respectively. Here, $\mathcal{C}(\cdot \,|\, \phi)$ represents the spatial correlation function with spatial decay parameter ϕ. The factor model representation in (3.2) leads to a significant reduction in the dimensionality of the problem because the spatial factors $\mathbf{w}(\mathbf{s}) = (w_k(\mathbf{s}) : 1 \leq k \leq q_w)$ and $\mathbf{v}(\mathbf{s}) = (v_r(\mathbf{s}) : 1 \leq r \leq q_v)$ have dimensions $q_w << h_z$ and $q_v \leq h_y$, respectively.

Combining these elements, and letting $\mathbf{\Lambda}_y = \mathbf{\Upsilon}\mathbf{\Lambda}_z = ((\lambda_{lk}^{(y)}))_{h_y \times q_w}$, a computationally viable version of (3.2) is

$$\text{Stage 1: } \mathbf{y}(\mathbf{s}) = \mathbf{X}_z(\mathbf{s})'\boldsymbol{\beta}_z + \mathbf{\Lambda}_z \mathbf{w}(\mathbf{s}) + \boldsymbol{\varepsilon}_z(\mathbf{s}) \tag{3.5}$$

$$\text{Stage 2: } \mathbf{y}(\mathbf{s}) = \mathbf{X}_y(\mathbf{s})'\boldsymbol{\beta}_y + \mathbf{\Lambda}_y \mathbf{w}(\mathbf{s}) + \mathbf{\Gamma}\mathbf{v}(\mathbf{s}) + \boldsymbol{\varepsilon}_y(\mathbf{s}). \tag{3.6}$$

[illegible] additional constraints are required for factor models to be identifiable (Anderson, 2003). Identifiability for SFMs can be achieved either by [illegible] upper triangle of the loadings matrix equal to zero and its diagonal elements all equal to one (Geweke and Zhou, 1996; Lopes and West, 2004; Aguilar and West, 2010), or, as in Ren and Banerjee (2013), by fixing

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0005

the sign of one element in each column of the factor loadings matrix, while enforcing an ordering constraint among the spatial decay parameters of the univariate correlation functions. We choose to ensure rotation and scale identifiability by using the former approach.

With the SFM structure in place, introducing the NNGP reduces the expensive ($\sim n_z^3 q_w$ and $\sim n_y^3 q_v$) calculation required to invert the dense covariance matrices from the parent GPs by $n_z q_w$ and $n_y q_v$ parallelizable, each of order m^3. Here, m is the number of neighbors for the NNGP, with $m << n_y \leq n_z$. In simulations, Datta et al. (2016a) found that, in most cases, $10 \leq m \leq 20$ provides an excellent approximation to the parent process; thus, the number of operations required is nearly linear in n.

For completeness, additional details on SFMs, NNGPs, and the sampling algorithm are included in the online supplement. For a more thorough treatment of SFM's, refer to Fruehwirth-Schnatter (2013) and Genton and Kleiber (2015), and for NNGPs, refer to Datta et al. (2016c).

3.3 Prior Specification and Hierarchical Formulation

Importantly, models (3.5) and (3.6) are fitted separately such that $\mathbf{w}(\mathbf{s})$ exclusively captures the spatial signal present in the LiDAR signals. However, using point estimates for $\mathbf{w}(\mathbf{s})$ (e.g., the posterior means) in (3.6) disregards the uncertainty present in the LiDAR spatial signal. Thus, to propagate this uncertainty through the forest outcome predictions, at each iteration of the Markov Chain Monte Carlo (MCMC) algorithm for $\mathbf{y}(\mathbf{s})$, we draw a sample for $\mathbf{w}(\mathbf{s})$ ($\mathbf{s} \in \mathcal{T}_y$) MCMC samples obtained when fitting model (3.5).

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0005

As mentioned in the previous section, the stochastic processes that capture the spatial structure are assumed to follow NNGPs. Given that an NNGP is a proper GP, at a finite collection of locations, the NNGPs induce zero-centered multivariate normal priors, with covariance matrices given by $\tilde{\mathbf{C}}^{(w)}$ and $\tilde{\mathbf{C}}^{(v)}$, respectively. Additionally, we use suitably noninformative priors for all other parameters, thus providing [illegible] sampling strategy.

In particular, we assume that $\boldsymbol{\beta}$ is either flat or conjugate [illegible] matrices $\boldsymbol{\Gamma}$ and $\boldsymbol{\Lambda}_z$ are constrained as described [illegible] below the diagonal assumed to be standard normal. All elements in $\mathbf{A}$ are also assumed to follow a standard normal distribution. The [illegible] entries in $\boldsymbol{\Psi}_z$ and $\boldsymbol{\Psi}_y$ are assigned half-t priors. Lastly, we assume uniform priors for the elements of the spatial decay vectors $\boldsymbol{\phi}_w = (\phi_{w,k} : 1 \leq k \leq q_w)$ and $\boldsymbol{\phi}_v = (\phi_{v,r} : 1 \leq r \leq q_v)$ in the interval $(-\log 0.05/\zeta_{\max}, -\log 0.01/\zeta_{\min})$, where $\zeta_{\min}$ and $\zeta_{\max}$ are the minimum and maximum distances, respectively, across all locations. Given that $\boldsymbol{\phi}_z$ and $\boldsymbol{\phi}_y$ are not conjugate with their corresponding likelihood, they are sampled using random walk Metropolis steps.

The joint posterior densities for the first and second stages of the algo-

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0005

rithm are proportional to

Stage 1:

$$\pi(\boldsymbol{\phi}_w)\,\mathrm{N}_{n_z q_w}(\mathbf{w}_{\mathcal{T}_z}|\mathbf{0},\tilde{\mathbf{C}}^{(w)})\left(\prod_{k=1}^{q_w}\prod_{j>k}^{h_z}\mathrm{N}(\lambda_{jk}^{(z)}|0,\right)$$
$$\times\,\pi(\boldsymbol{\beta}_z)\left(\prod_{j=1}^{h_z}\mathcal{IG}(\psi_j^z|\nu/2,\nu/\;)\mathcal{G}(a_{z,j}|1/2,1/A^2)\right)$$
$$\times\left(\prod_{\mathbf{s}_i\in\mathcal{T}_z}\mathrm{N}_{h_z}(\mathbf{z}(\mathbf{s}_i)|\mathbf{X}_z(\mathbf{s}_i)\;\boldsymbol{\Lambda}_z\mathbf{w}(\mathbf{s}_i),\boldsymbol{\Psi}_z)\right) \tag{3.7}$$

Stage 2:

$$\pi(\boldsymbol{\phi}_v)\,\mathrm{N}_{n_y q_v}(\;_{\mathcal{T}_y}|\mathbf{0},\tilde{\mathbf{C}}^{(v)})\left(\prod_{k=1}^{q_w}\prod_{j}(\lambda_{jk}^{(y)}|0,1)\right)\left(\prod_{r=1}^{q_v}\prod_{j>r}^{h_y}\mathrm{N}(\gamma_{jr}|0,1)\right)$$
$$\times\,\pi(\boldsymbol{\beta}_y)\left(\prod_{j}\mathcal{IG}(\;^{y}|2,\nu/\;_{,j})\mathcal{G}(a_{y,j}|1/2,1/A^2)\right)$$
$$\times\left(\prod_{\mathbf{s}_i}\mathrm{N}\;(\mathbf{s}_i)'\mathbf{X}_y(\mathbf{s}_i)'\boldsymbol{\beta}_y+\boldsymbol{\Lambda}_y\mathbf{w}(\mathbf{s}_i)+\boldsymbol{\Gamma}\mathbf{v}(\mathbf{s}_i),\boldsymbol{\Psi}_y)\right), \tag{3.8}$$

where $\mathbf{w}_{\mathcal{T}_z}=(\mathbf{w}(\mathbf{s}_i)'\;\mathbf{s}_i\in\;)'$ and $\mathbf{v}_{\mathcal{T}}=(\mathbf{v}(\mathbf{s}_i)':\mathbf{s}_i\in\mathcal{T}_y)'$, such that

$$\begin{aligned}\mathrm{N}_{n_w}(\mathbf{w}\;\;\tilde{\mathbf{C}}^{(w)}) &= \prod_{\mathbf{s}_i\in\mathcal{T}_z}\mathrm{N}_{q_w}(\mathbf{w}(\mathbf{s}_i)\,|\,\mathbf{B}_i^{(w)}\mathbf{w}_{N(i)},\mathbf{F}_i^{(w)}),\ \text{and}\\ \mathrm{N}_{n_w}\;_{\mathcal{T}}|\mathbf{0},\tilde{\mathbf{C}}^{(v)}) &= \prod_{\mathbf{s}_i\in\mathcal{T}_y}\mathrm{N}_{q_v}(\mathbf{v}(\mathbf{s}_i)\,|\,\mathbf{B}_i^{(v)}\mathbf{v}_{N(i)},\mathbf{F}_i^{(v)}).\end{aligned} \tag{3.9}$$

The expressions on the right-hand side of (3.9) result from the construction of the NNGP (see online supplement). For an m-neighbor NNGP, let

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0005

$m_i = \min\{m, i-1\}$ denote the number of neighbors for location $\mathbf{s}_i$. The index set $N(i)$ for location $\mathbf{s}_i \in \mathcal{T}_z$ contains its m_i nearest neighbors; thus, $\mathbf{w}_{N(i)}$ corresponds to the vector $(\mathbf{w}(\mathbf{s}_j)' : \mathbf{s}_j \in N(i) \subset \mathcal{T}_z)'$. The neighbor set for $\mathbf{v}(\mathbf{s}_i)$ is defined analogously. Letting $u \in \{w, v\}$, $\mathbf{B}_i^{(u)}$ denotes a $q_u \times m_i q_u$ block matrix, with the $q_u \times q_u$ diagonal blocks containing the kriging weights for the q_u spatial factors for each neighbor. In addition, $\mathbf{F}_i^{(u)}$ corresponds to the $q_u \times q_u$ diagonal matrix with the variances for the q_u spatial factors conditioned on the neighbor set $N(i)$ (see Section S1 in the online supplement for details on $\mathbf{B}_i^{(u)}$ and $\mathbf{F}_i^{(u)}$). Lastly, the parameters $\{a_{y,j}\}_{j=1}^{h_y}$ and $\{a_{z,k}\}_{k=1}^{h_z}$ complete the hierarchical representation of the half-t prior distributions assigned for ψ_j^y and ψ_k^z, respectively, and the hyperparameter A is simply chosen to be some large value (say, 100).

Owing to prior conjugacy, the full conditional densities for all parameters except $\boldsymbol{\phi}_w$ and $\boldsymbol{\phi}_v$ can be sampled using simple Gibbs steps. Further details on the sampling algorithm are referred to the online supplement.

3.4 Imputation and Prediction

As mentioned previously, LiDAR signals are collected over the large spatial region $\mathcal{T}_z$, whereas forest outcome observations are confined to the smaller subset of locations $\mathcal{T}_y$. Additionally, there are relevant out-of-sample locations where neither LiDAR nor forest outcomes are observed, $\mathcal{T}_\emptyset$. Finally, there are some locations within the corresponding reference sets $\mathcal{T}_z$ and $\mathcal{T}_y$ that have some or all missing outcomes. It is thus essential for this modeling effort to provide the means to accurately impute the missing values in $\mathcal{T}_z$ or

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0005

$\mathcal{T}_y$. This enables us to generate LiDAR predictions in $\mathcal{T}_\emptyset$ and forest outcome predictions within $\mathcal{T}_\emptyset \cup (\mathcal{T}_z \setminus \mathcal{T}_y)$. Given the NNGP formulation, both the imputation and the out-of-sample prediction are remarkably inexpensive.

Imputation is straightforward. Let $\mathbf{s}_\bullet \in \mathcal{T}_z$ be a location where $\mathbf{z}(\mathbf{s}_\bullet)$ is missing. Then, $\mathbf{z}(\mathbf{s}_\bullet)$ is drawn as part of the sampling algorithm from $\mathrm{N}_{h_z}(\mathbf{X}_z(\mathbf{s}_\bullet)'\boldsymbol{\beta}_z + \boldsymbol{\Lambda}_z\mathbf{w}(\mathbf{s}_\bullet), \boldsymbol{\Psi}_z)$, where $\mathbf{w}(\mathbf{s}_\bullet)$ is sampled from the full conditional posterior density in Equation (S3.1) of the online supplement. For a missing value $\mathbf{y}(\mathbf{s}_\bullet)$, where $\mathbf{s}_\bullet \in \mathcal{T}_y$, the procedure uses the full conditional posterior for $\mathbf{v}(\mathbf{s}_\bullet)$ and the likelihood for $\mathbf{y}(\mathbf{s}_\bullet)$.

The procedure to predict a new LiDAR observation at $\mathbf{s}_\circ \in \mathcal{T}_\emptyset$, begins by sampling the spatial factor $\mathbf{w}(\mathbf{s}_\circ)$ from $\mathrm{N}_{q_w}(\mathbf{B}_\circ^{(w)}\mathbf{w}_{N(\mathbf{s}_\circ)}, \mathbf{F}_\circ^{(w)})$, with $\mathbf{B}_\circ^{(w)}$ and $\mathbf{F}_\circ^{(w)}$ defined as before. Note that the nearest neighbor set $N(\mathbf{s}_\circ)$ is assumed to be in $\mathcal{T}_z$. Then, we draw $\mathbf{z}(\mathbf{s}_\circ) \,|\, \mathbf{z}_{\mathcal{T}_z}$ from $\mathrm{N}_{h_z}(\mathbf{X}_z(\mathbf{s}_\circ)'\boldsymbol{\beta}_z + \boldsymbol{\Lambda}_z\mathbf{w}(\mathbf{s}_\circ), \boldsymbol{\Psi}_z)$. This is done by conditioning on the posterior samples of $\{\boldsymbol{\beta}_z, \boldsymbol{\Lambda}_z, \boldsymbol{\Psi}_z, \boldsymbol{\phi}_w\}$ obtained from the fitting algorithm.

To predict the forest outcomes $\mathbf{y}(\mathbf{s}_\circ)$ at $\mathbf{s}_\circ \in \mathcal{T}_\emptyset \cup (\mathcal{T}_z \setminus \mathcal{T}_y)$, we first generate samples of $\mathbf{v}(\mathbf{s}_\circ) \sim \mathrm{N}_{q_v}(\mathbf{B}_\circ^{(v)}\mathbf{v}_{N(\mathbf{s}_\circ)}, \mathbf{F}_\circ^{(v)})$. Given that $\mathbf{y}(\mathbf{s}_\circ)$ depends on $\mathbf{w}(\mathbf{s}_\circ)$, we combine these posterior draws of $\{\boldsymbol{\beta}_y, \boldsymbol{\Lambda}_y, \boldsymbol{\Gamma}, \boldsymbol{\Psi}_y, \boldsymbol{\phi}_v\}$ with those of $\mathbf{w}(\mathbf{s}_\circ)$, obtained when predicting $\mathbf{z}(\mathbf{s}_\circ)$, and draw predicted values for $\mathbf{y}(\mathbf{s}_\circ) \,|\, \mathbf{y}_{\mathcal{T}_y}$ from $\mathrm{N}(\mathbf{X}_y(\mathbf{s}_\circ)'\boldsymbol{\beta}_y + \boldsymbol{\Lambda}_y\mathbf{w}(\mathbf{s}_\circ) + \boldsymbol{\Gamma}\mathbf{v}(\mathbf{s}_\circ), \boldsymbol{\Psi}_y)$.

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0005

4. Simulation: Recovering Low-dimensional Structure

In the following simulation exercise we focus exclusively on the high-dimensional component (i.e., the first stage) of the model described above. The simulation below was devised to illustrate the ability of our approach to recover the true low-dimensional structure when the data are generated from a low-dimensional SFM with dense spatial factors.

We generate a synthetic data set for $h_z = 50$ outcomes in $n = 10{,}000$ locations from the spatial factor model $\mathbf{z}(\mathbf{s}) = \mathbf{X}_z(\mathbf{s})'\tilde{\boldsymbol{\beta}}_z + \boldsymbol{\Lambda}_z \tilde{\mathbf{w}}(\mathbf{s}) + \tilde{\boldsymbol{\varepsilon}}(\mathbf{s})$. Here, $\mathbf{X}_z(\mathbf{s})'$ is a 50×150 block-diagonal matrix of predictors, and $\tilde{\boldsymbol{\beta}}$ is the vector of regression coefficients, both defined as before. We consider the same three predictors for all outcomes. The spatial factors $\tilde{\mathbf{w}}(\mathbf{s}) \sim \prod_{k=1}^{8} \mathrm{GP}(0, \mathcal{C}(\cdot \mid \tilde{\phi}_k^z))$, where $\mathcal{C}(\cdot \mid \tilde{\phi}_k^z)$ is an exponential correlation function with decay parameter $\tilde{\phi}_k^z$. Additionally, for identifiability we assume that the 50×8 factor loadings matrix $\boldsymbol{\Lambda}_z$ has zeros in the upper triangle and ones along the diagonal. Finally, $\tilde{\boldsymbol{\varepsilon}}_z \sim \mathrm{N}_{h_z}\left(\mathbf{0}, \tilde{\boldsymbol{\Psi}}_z\right)$ with $\tilde{\boldsymbol{\Psi}}_z = \mathrm{diag}(\tilde{\psi}_k^z : k = 1, \ldots, 8)$.

We assess the ability of (3.5) to recover the model parameters from the true data generating process, impute missing outcomes, and predict at out-of-sample locations. The SF-NNGP model was fitted for $q_w \in \{3, 5, 8, 10\}$ spatial factors and $m = 10$ neighbors. Of the $10{,}000$ locations, we assume all 50 outcomes to be missing in 200 locations chosen at random. These outcomes are to be imputed. Additionally, we use $n_0 = 500$ locations for out-of-sample prediction and model validation.

The first result worth highlighting is the gains in computational efficiency

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0005

provided by the SF-NNGP. For this simulation exercise—a relatively computationally challenging problem—fitting the largest model considered (i.e., $q_w = 10$) with 50,000 MCMC iterations on a Linux server with an Intel i7 processor (two eight-core) and 16 GB of memory, the runtime was 4.88 hours. As shown below, the proposed approach is able to recover the true model parameters, accurately impute missing data, and generate precise predictions, all with suitable uncertainty estimates.

For all values of q_w, the SF-NNGP accurately recovered the regression coefficients $\tilde{\boldsymbol{\beta}}_z$ for all predictors and responses (Figure 1 in the online supplement). In contrast, the quality of the estimates for the variance components $\tilde{\psi}_k^z$'s was compromised when q_w was lower than the true number of spatial factors. This behavior is expected: for lower values of q_w, the ψ_k^z's attempt to compensate for the additional signal that the spatial component with too few spatial factors is unable to capture (Figure 2 in the supplementary material). For $q_w = 8$ and $q_w = 10$, the coverage for $\tilde{\boldsymbol{\psi}}_z$ was 88% and 84%, respectively, with all close to $\tilde{\psi}_k^z$ with tight 95% credible sets.

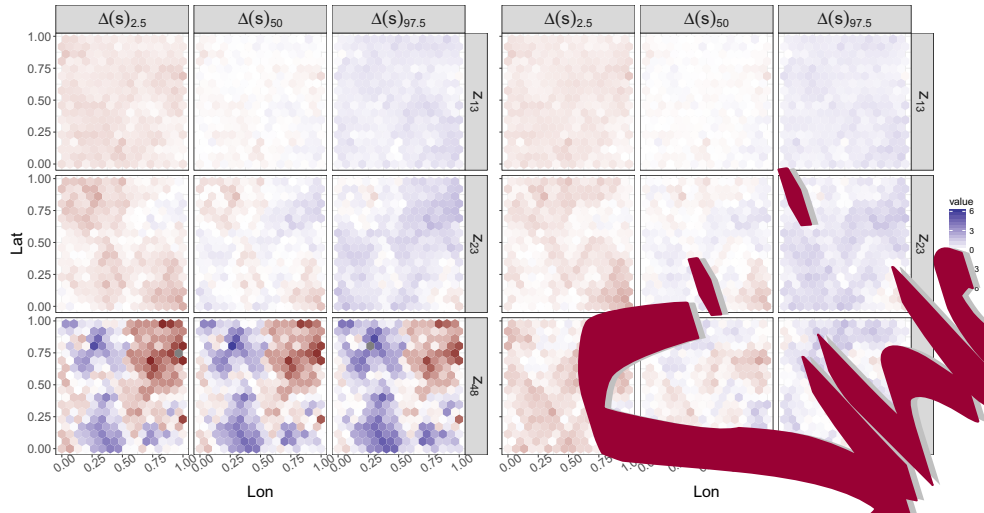
When $q_w \neq 8$, the dimensions of the fitted $\boldsymbol{\Lambda}_z$, $\boldsymbol{\phi}_w$, and $\mathbf{w}(\mathbf{s})$ do not match those of their analogs in the true model. Therefore, to assess the quality of the fit for the spatial signal for all values of q_w considered, we instead compare the fitted spatial component $\mathbf{w}^\star(\mathbf{s}) = \boldsymbol{\Lambda}_z \mathbf{w}(\mathbf{s})$, for $\mathbf{s} \in \mathcal{T}_z$, to that of the true model, given by $\tilde{\mathbf{w}}^\star(\mathbf{s}) = \tilde{\boldsymbol{\Lambda}}_z \tilde{\mathbf{w}}(\mathbf{s})$.

At locations in $\mathcal{T}_z$, we calculate $\Delta(\mathbf{s}) = \mathbf{w}^\star(\mathbf{s}) - \tilde{\mathbf{w}}^\star(\mathbf{s})$ (fitted minus true spatial signal) for each MCMC draw of the parameters. For all $\mathbf{s} \in \mathcal{T}_z$, we obtained the median and 95% credible set for $\Delta(\mathbf{s})$. To facilitate visu-

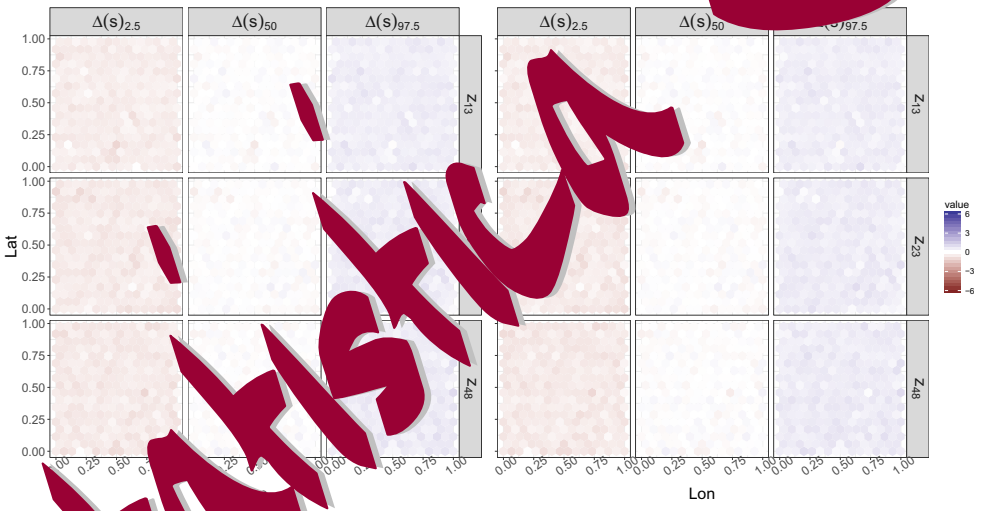
Statistica Sinica: Preprint
doi:10.5705/ss.202018.0005

alization, in Figure 2, we show the results for only three responses, selected at random, from the 50 considered. The columns of each panel map the quantiles 2.5, 50, and 97.5 for $\Delta(\mathbf{s})$, with three locations (13, 23, and 48) plotted by row. The fitted spatial signal when $q_w \in \{3, 5\}$ only partially recovers the true signal, with coverages of 26.13% and 42.06%, respectively, for $q_w = 3$ and $q_w = 5$. When $q_w \in \{8, 10\}$, the recovery of the spatial signal is extremely accurate: over all responses, the coverage is 94.78% with $q_w = 8$, and 94.18% with $q_w = 10$.

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0005



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



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

Fig. … minus true spatial signal, $\Delta(\mathbf{s}) = \mathbf{w}^\star(\mathbf{s}) - \tilde{\mathbf{w}}^\star(\mathbf{s})$, for locations $\mathbf{s}_{13}, \mathbf{s}_{23}, \mathbf{s}_{48}$. … m left to right, the columns in each panel show percentiles 2.5, 50, … for $\Delta(\mathbf{s})$, respectively.

In addition to the previous results, it is also encouraging to find that when the dimension of the SF-NNGP model matches that of the true model,

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0005

both the factor loadings ($\tilde{\mathbf{\Lambda}}_z$) and the spatial decay parameters ($\tilde{\boldsymbol{\phi}}_z$) from the true spatial process can be recovered accurately (Figures 3 and 4).

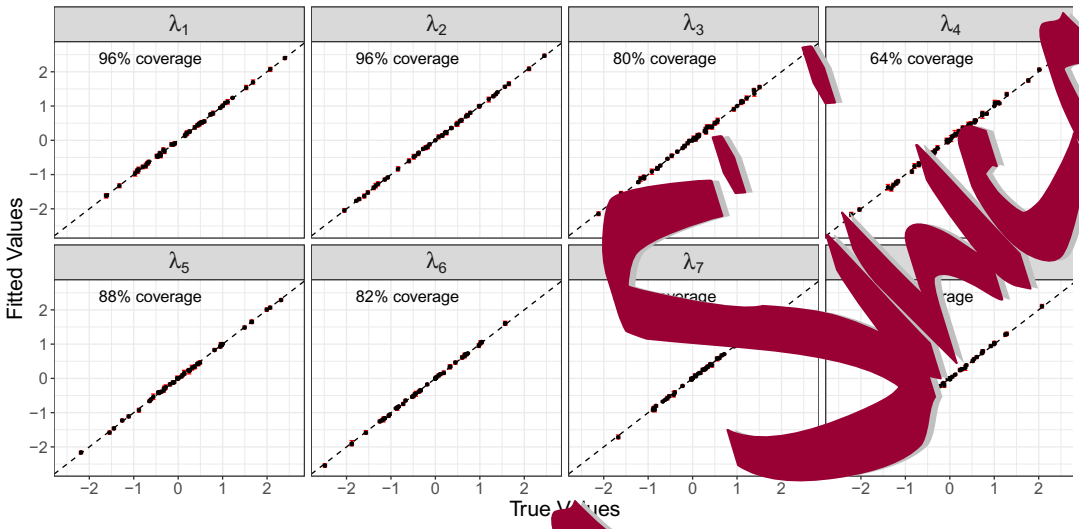


Figure 3: Fitted vs. true factor loadings and matrix parameters (95% credible sets and medians) for $q_w = 8$.

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0005

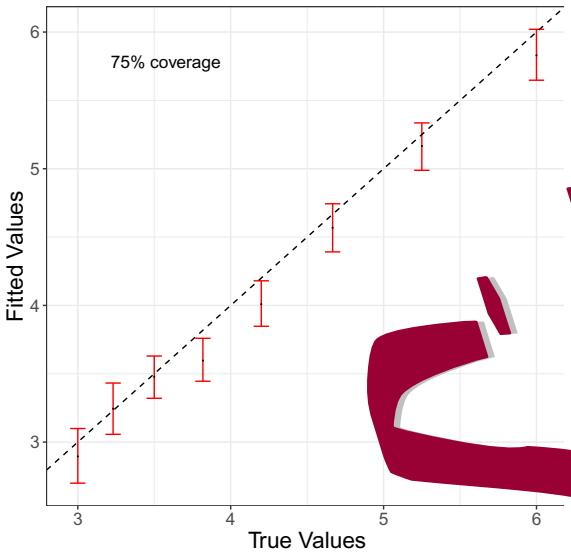


Figure 4: Fitted vs. true spatial decay parameters (95% credible sets and medians) for $q_w = 8$.

Model performance in terms of the accuracy of imputation and prediction improves drastically as the number of factors approaches that of the true model; see Figures 5 and 6 in the online Supplement.

Table 1 compares q_w SVCNNGP using different measures of out-of-sample prediction error magnitude. In particular, the continuous rank probability score (CRPS) (Equation (21) in Gneiting and Raftery, 2007) and the root mean squared prediction error (RMSPE) (Yeniay and Goktas, 2002) with $q_w = 8$. The coverage of the 95% credible intervals of the prediction was close to the nominal value for all q_w; however, the width of the interval rapidly decreases as q_w approaches the true number of spatial factors.

Both the fitted values for the spatial signals and the out-of-sample pre-

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0005

Table 1: Out-of-sample prediction comparison across models with different numbers of spatial factors.

q_w	CRSP	RMSPE	95% Coverage	95% CI Width
3	0.85	1.61	95.82	6.14
5	0.67	1.28	95.43	4.79
8	0.45	0.83	94.78	3.10
10	0.45	0.83	94.84	3.10

dictions with $q_w = 8$ and $q_w = 10$ are practically [illegible] each other. Furthermore, the model with $q_w = 8$ accurately recovers all [illegible] the true factor loadings (Figure 3). Interestingly, with $q_w = 10$ [illegible] spection of the estimates for columns 1 through 6 in [illegible] indicates that this model accurately estimates the corresponding true pa[illegible]alues (see Figure 4 in the online supplement). However, in t[illegible] sam[illegible]del, the estimated parameter values in columns 7 and 8 [illegible] epartures from their true values. Furthermore, the 95% cre[illegible]e[illegible] for all unconstrained elements in the 9th and 10th colum[illegible] of [illegible] zero (see Figure 3 in the online supplement). These results provi[illegible] guid[illegible]ce on the number of factors q_w to use. Because there [illegible] no [illegible] the model with $q_w = 10$ over that with $q_w = 8$ in terms [illegible] predictive accuracy or parameter fit, the results favor the more parsimonious [illegible]odel of the two.

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0005

5. Modeling LiDAR Signals and Forest Structure

Our focus in the subsequent analysis is to assess and interpret the utility of SF-NNGP spatial factors to explain the variability in the three forest outcomes defined in Section 2, measured on the BCEF. Following the two-stage model developed in Section 3.2, we fit (3.5) using $q \in \{1, 2, 3, 4, 6, 8\}$ spatial factors and $m = 10$ neighbors to the [illegible] LiDAR data comprising n_z=50,197 signals, each of length h_z=57. The model mean [illegible] only an intercept. The specification for the priors follows [illegible] with the support for elements in $\boldsymbol{\phi}_w$ adjusted to match the BCEF spatial extent.

The n_y=197 locations with h_y=3 forest outcomes were used in the second-stage model (3.6). To more clearly interpret the spatial factors' ability to explain the variability in forest outcomes, we decided to avoid potential issues with spatial confounding (Hanks et al., 2015) and set $\mathbf{v}(\mathbf{s})$ to zero. In practice, however, if our main objective is to maximize predictive performance, then this residual spatial random effect should likely be included in the model. In addition to the latent factors, the second-stage model was informed by the three Landsat-based cap predictor variables defined in Section 2, which, along with an intercept, were included in $\mathbf{X}_y(\mathbf{s})$. Importantly, these predictor variables are available across the entire BCEF; hence, given the predicted values of the spatial factors at unobserved locations, we can create complete coverage forest outcome maps.

Posterior inference for all candidate models was based on three chains of 50,000 post-burn-in MCMC samples. Chains converged by 20,000 MCMC

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0005

iterations. Using the same computer configuration detailed in Section 4, the total runtime for the most demanding model, $q_w = 8$, was ∼36 hours.

The eight candidate models, specified by q_w, were assessed based on their ability to inform the forest outcome predictions. This was done by fitting each of the first-stage models, then fitting their corresponding second-stage models using data from 99 of the 197 available locations in $\mathcal{T}_y$. The three forest outcomes were then predicted for the remaining 98 out-of-sample locations. The scoring rules and other summaries of the predictive distributions for the 98 out-of-sample locations are presented in Table 2.

Increasing the number of spatial factors improves the CRPS and RMSPE for each forest outcome shown in Table 2. Exploratory analysis showed that the gains in predictive performance were negligible beyond $q_w = 4$ for AGB and $q_w = 5$ for TD and BA. Given that the q_w=5 model generally yielded the "best" predictions, it was selected for exposition below.

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0005

Table 2: Cross-validation prediction summary for forest outcomes given increasing number of spatial factors q_w. Bold values identify lowest CRPS and RMSPE.

	q_w	CRSP	RMSPE	95% Coverage	95% CI Width
AGB	1	26.21	51.37	91.88	161.24
	2	26.36	52.02	92.39	162.14
	3	23.64	46.95	9	155.71
	4	**23.53**	**46.93**	9	15
	5	24	47.54	9	157.
	6	24.47	47.8	94.	
	7	24.75	47.84	95.43	17
	8	24.76	48.02	96.45	182.12
TD	1	1017.7	1980.62	92.39	
	2	1006.02	1957.54	93.4	5944.81
	3	1007.72	1954.87	93.4	6068.29
	4	997. 2	1955.2		6040.06
	5	**989.31**	**1930.76**	94. 2	6182.2
	6	998.3	1944.2	2	6223.73
	7	1005.2	196 8	5.43	6450.5
	8	1004.36	1	96.95	6503.17
BA	1	5.53	29	91.88	36.34
	2	4	10.	94.42	36.85
			54	93.91	35.16
	4	1	52	93.4	36.21
		16	**9.58**	93.4	36.51
		5	9.59	96.45	38.62
			9.73	95.43	38.34
		5.27	9.72	94.42	37.93

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0005

Table 3: Elements of $\mathbf{\Lambda}_y$ median and 95% credible intervals for the $q_w = 5$ model. Bold entries indicate where the 95% credible interval excludes zero.

Parameter	50% (2.5%, 97.5%)
$\lambda^{(y)}_{\mathsf{AGB},1}$	**-6.65 (-8.89, -4.23)**
$\lambda^{(y)}_{\mathsf{AGB},2}$	27.20 (-14.11, 65.1
$\lambda^{(y)}_{\mathsf{AGB},3}$	**-278.29 (-324 2.28)**
$\lambda^{(y)}_{\mathsf{AGB},4}$	-46.15 (-16 6, 75.91)
$\lambda^{(y)}_{\mathsf{AGB},5}$	**-308.81 (-524 90.45)**
$\lambda^{(y)}_{\mathsf{TD},1}$	-1.77 (-21.35, 17.60)
$\lambda^{(y)}_{\mathsf{TD},2}$	**-357.49 (-718.82, -7.86)**
$\lambda^{(y)}_{\mathsf{TD},3}$	269.03 (-137.51, 667.
$\lambda^{(y)}_{\mathsf{TD},4}$	**-1777.21 (-2696.67, -708.08)**
$\lambda^{(y)}_{\mathsf{TD},5}$	**2457.52 (18, 4337.97)**
$\lambda^{(y)}_{\mathsf{BA},1}$	**-2.93 (75)**
$\lambda^{(y)}_{\mathsf{BA},2}$	07 (15.79)
$\lambda^{(y)}_{\mathsf{BA},3}$	**(9.79, -76.24)**
$\lambda^{(y)}_{\mathsf{BA},4}$	**72. 20.60, -23.00)**
$\lambda^{(y)}_{\mathsf{BA},5}$	0.55 (-177.44, 20.51)

Table 3 provides estimates for the second-stage model's spatial factor regression coefficients, that is, the elements in $\mathbf{\Lambda}_y$. These results show that [illegible] spatial factors explain a substantial portion of the variability in the forest outcomes. It is, however, difficult to interpret the different $\lambda^{(y)}$ [illegible] sense of what characteristic of $\mathbf{z}(\mathbf{s})$ the spatial factors are capturing. When considered with the estimates in Table 3, Figure 5 provides a biological interpretation of the spatial factors. Specifically, each panel in

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0005

Figure 5 represents a spatial factor. The 50 lines in each panel are observed LiDAR signals, with the lines corresponding to the 25 largest (lighter colored lines) and 25 smallest (darker lines) estimated spatial factor values.

There are some general biological relationships between the forest canopy structure and AGB, TD, and BA. A very low maximum canopy height is indicative of a young regenerating forest (e.g., growth after a fire), which would be characterized by low AGB, high TD, and low BA. The trees of trees in a forest have a high canopy height, high AGB, low TD, and high BA (i.e., a few large-diameter mature trees dominate the area). When the forest is characterized by trees of many different heights (i.e., tree crowns in several vertical strata), then we might expect moderate/high AGB, moderate TD, and moderate/high BA. Some of these expected relationships are observed when comparing Table 3 and Figure 5. For example, the top left panel in Figure 5 differentiates well regenerating forests and all other forest structures, that is, the darker lines show a spike of energy returned at or near ground level versus the lighter lines which show the majority of the energy is returned at or above a height of several meters. Hence, we have negative regression coefficients $\lambda^{(y)}_{AGB,1}$ and $\lambda^{(y)}_{BA,1}$ in Table 3. The LiDAR signals shown in the top right panel in Figure 5 differentiate between young and old single-cohort forests (i.e., all trees were regenerated around the same time and there is little vertical variation in canopy height); hence, we have negative $\lambda^{(y)}_{AGB,3}$ and $\lambda^{(y)}_{BA,3}$ in Table 3. The top middle and bottom left panels in Figure 5 generally separate the signal for mature 20+ and ∼20 meter canopy heights (lighter lines), respectively, from the lower stature ∼10 meter canopy height

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0005

forest (darker lines). Consistent with the biological expectation, the negative $\lambda_{TD,2}^{(y)}$ and $\lambda_{TD,4}^{(y)}$ suggest that forests associated with red LiDAR signals have higher tree density relative to the older taller forests.

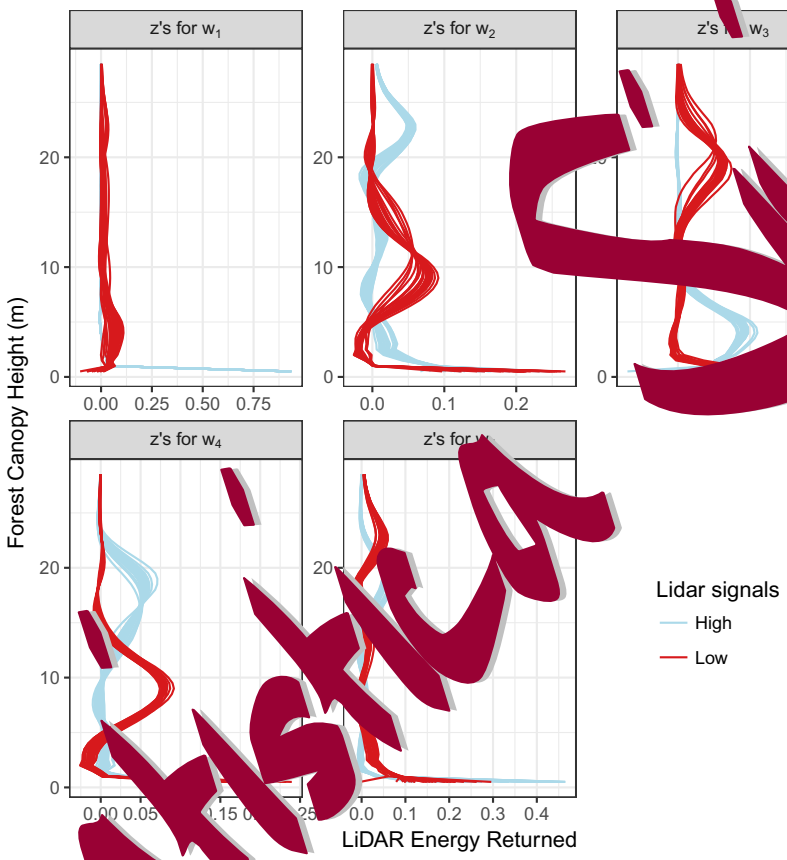


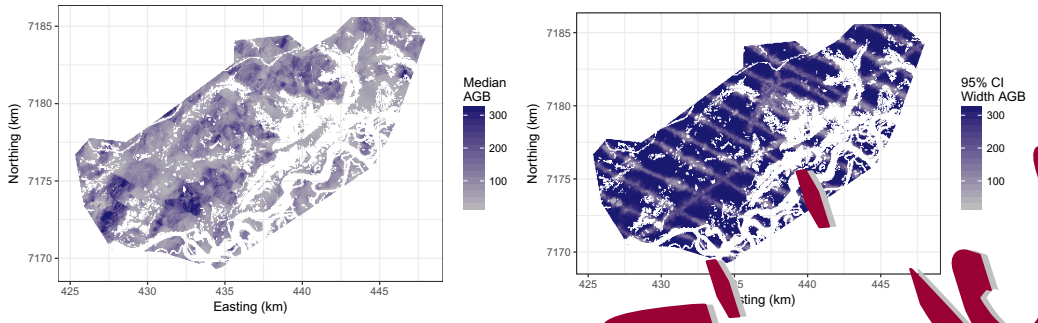
Figure 3: Scatterplots of LiDAR signals with the 25 largest (*High* in the legend) and 25 smallest (*Low* in the legend) values of $\mathbf{w}(\mathbf{s})$ from the $q_w = 5$ model.

As detailed in Section 1, complete-coverage maps of the forest outcomes with associated uncertainty estimates are important data products that can be delivered by the proposed two-stage model. Following Section 3.4 and using the full data set depicted in Figure 1, we predicted the forest outcomes

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0005

on a 30×30 m grid over the BCEF. Figure 6 provides the median and 95% credible interval width maps for each outcome. Nonforested areas are omitted (white regions on the maps). The posterior predictive point estimates match well with the distribution of the forest outcomes across the BCEF and are clearly informed by the LiDAR factors, which are capturing key forest structure characteristics. Most importantly, the prediction uncertainty maps, displayed in the right column of Figure 6, accurately reflect the lack of information for prediction units that are far from where LiDAR data are available, that is, we achieve more precise posterior predictive distributions along and adjacent to locations where LiDAR data are available. Far from the LiDAR flight lines, prediction is only informed by the Landsat 8 tassel cap predictor variables which in this study explained very little variability in the forest outcomes.

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0005



(a) AGB (Mg ha^{-1}) (b) 95% CI Width (Mg ha

(c) TD Median (1000 trees ha^{-1}) (d) 95% CI Width (1000 trees ha^{-1})

(e) BA Median (m^2 ha^{-1}) (f) BA 95% CI Width (m^2 ha^{-1})

Figure 6: $q_w = 5$ posterior predictive distribution median and 95% CI width for AGB, TD, and BA forest variables over Bonanza Creek Experi-

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0005

6. Concluding Remarks

We formulated an approach to model high-dimensional spatial data over a large set of locations, and developed an efficient implementation in C++. The SF-NNGP enables the analysis of multivariate spatially referenced data sets that, due to their magnitude, could not be rigorously explored before. It does so by combining the ability of SFMs to compress the signal in high-dimensional structures into a few dimensions with the computational scalability of NNGPs.

The algorithm was used to exploit the information from the high-dimensional LiDAR signals to jointly model and generate LiDAR-based maps of multiple forest variables. Importantly, the proposed two-stage model provides a viable approach to producing spatially continuous maps from sparsely sampled LiDAR and forest measurements. Furthermore, the model delivers spatially explicit uncertainty quantification that captures the irregular distribution of information across the domain of interest. Such frameworks will become increasingly important as emerging LiDAR systems, such as GEDI, come online in the near future. These approaches can also be extended to help guide LiDAR mission data acquisition to minimize prediction uncertainty.

Importantly, when fitting a spatial factor model, one must choose the number of factors q_w to use in the model; there are different strategies to address this issue. Here, we consider out-of-sample evaluation metrics for different choices of q_w and select the one where the curves flatten out. This is a pragmatic solution, similar in spirit to cross-validation approaches commonly

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0005

used to tune hyper-parameters in richly parametrized models. Like any other cross-validation approach, this leads to additional computation, but parallel computing opens the possibility of conducting simultaneous MCMC runs for different values of q_w. As shown, both in the simulation experiment and in the BCEF data analysis, this heuristic provides sufficiently good results. Other automated rank selection schemes are available in the literature, such as those proposed in Lopes and West (2004) and in Ren and Banerjee, however, these drastically increase the computational burden of already computationally costly problem.

In future research, we would like to explore analogous models for spatio-temporal data. For this type of data, it is necessary to posit a strategy to select the neighbors in the spatio-temporal domain, following the discussion presented in Datta et al. (2016a).

Although our method represents substantial improvement in terms of scalability over existing approaches, further efforts are required to scale multivariate spatial methods to truly massive data sets. For instance, the ultimate goal for forest biomass mapping assisted by sampled LiDAR in interior Alaska is a complete-coverage map of the entire domain (e.g., 46 million ha), which could easily require models capable of assimilating LiDAR signals in more than locations.

Supplementary Materials

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

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0005

(3) additional simulation results.

Acknowledgments

The research presented in this study was partially supported by NASA's Arctic-Boreal Vulnerability Experiment (ABoVE) and Carbon Monitoring System (CMS) programs. Additional support was provided by the United States Forest Service Pacific Northwest Research Station. Finley was supported by National Science Foundation (NSF) DMS-1513481, EF-1137309, and EF-1241874, and Finley and Taylor-Rodriguez were supported by EF-1253225. Banerjee was supported by NSF DMS-1513654, NSF IIS-1562303, and NIH/NIEHS 1R01ES027027-01.

References

Aguilar, O. and West, M. (2000). Bayesian Dynamic Factor Models and Portfolio Allocation. *Journal of Business & Economic Statistics*, 18:338–357.

Andersen, H.-E., Strunk, J. and Temesgen, H. (2011). Using airborne light detection and ranging as a sampling tool for estimating forest biomass resources in the upper Tanana Valley of interior Alaska. *Western Journal of Applied Forestry*, 26(4):157–164.

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

Asner, G., Hughes, R., Varga, T., Knapp, D., and Kennedy-Bowdoin, T. (2009). Environmental and biotic controls over aboveground biomass throughout a tropical rain forest. *Ecosystems*, 12(2):261–278.

Babcock, C., Finley, A. O., Andersen, H.-E., Pattison, R., Cook, B. D., Morton, D. C., Alonzo,

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0005

M., Nelson, R., Gregoire, T., Ene, L., Gobakken, T., and Næsset, E. (2017). Geostatistical estimation of forest biomass in interior alaska combining landsat-derived tree cover, sampled airborne lidar and field observations. *ArXiv e-prints*. `https://arxiv.org/pdf/1705.03534.pdf`.

Babcock, C., Finley, A. O., Bradford, J. B., Kolka, R., Birdsey, R., and Ryan, M. G. (2015). Lidar based prediction of forest biomass using hierarchical models with spatially varying coefficients. *Remote Sensing of Environment*, 169:113–127.

Babcock, C., Matney, J., Finley, A., Weiskittel, A., and Cook, B. (2018). Bayesian multivariate spatial regression models for predicting individual tree structure variables using lidar data. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 11(1, SI):6–14.

Baig, M. H. A., Zhang, L., Shuai, T., and Tong, Q. (2014). Derivation of a tasselled cap transformation based on landsat 8 at-satellite reflectance. *Remote Sensing Letters*, 5(5):423–431.

Banerjee, S. (2017). High-dimensional bayesian geostatistics. *Bayesian Anal.*, 12(2):583–614.

Banerjee, S., Gelfand, A. E., , A. [illegible] (2008). Gaussian predictive process models for large spatial data sets. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, [illegible]

Bechtold, W. A. and Patterson, P. L. (2005). *The Enhanced Forest Inventory and Analysis Program - National Sampling Design and Estimation Procedures*. US Department of Agriculture Forest Service, Southern Research Station Asheville, North Carolina.

Blackford, L. S., Demmel, J., Dongarra, J., Duff, I., Hammarling, S., Henry, G., Heroux, M., Kaufman, L., Lumsdaine, A., Petitet, A., Pozo, R., Remington, K., and Whaley, R. C. (2001). An updated set of basic linear algebra subprograms (blas). *ACM Transactions on Mathematical Software*, 28:135–151.

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0005

Bolton, D. K., Coops, N. C., and Wulder, M. A. (2013). Measuring forest structure along productivity gradients in the Canadian boreal with small-footprint lidar. *Environmental monitoring and assessment*, 185(8):6617–6634.

Bonanza Creek LTER (2016). Bonanza Creek Experimental Forest. `http://www.lter.uaf.edu/research/study-sites-bcef`. Accessed: 12-16-2017.

Christensen, W. F. and Amemiya, Y. (2002). Latent variables of multivariate spatial. *Journal of the American Statistical Association*, 97(457):302–317.

Cook, B., Corp, L., Nelson, R., Middleton, E., Morton, D., Ranson, K., Ly, V., and Montesano, P. (2013). NASA Goddard's lidar, hyperspectral and thermal (G-LiHT) airborne imager. *Remote Sensing*, 5(8):4045–4066.

Dagum, L. and Menon, R. (1998). Openmp: an industry standard api for shared-memory programming. *Computational Science & Engineering*, (1):46–55.

Datta, A., Banerjee, S., Finley, A., Hamm, A., and Schlather, 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.

Ene, L. T., Gobakken, T., Andersen, H.-E., Nsset, E., Cook, B. D., Morton, D. C., Babcock,

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0005

C., and Nelson, R. (2018). Large-area hybrid estimation of aboveground biomass in interior alaska using airborne laser scanning data. *Remote Sensing of Environment*, 204(Supplement C):741 – 755.

Finley, A. O., Banerjee, S., and Cook, B. D. (2014a). Bayesian hierarchical models for spatially misaligned data in R. *Methods in Ecology and Evolution*, 5(6):514–523.

Finley, A. O., Banerjee, S., Ek, A. R., and McRobert[illegible]008). Bayesian mu[illegible] process modeling for prediction of forest attributes. *Journal of Agricultu[illegible]* *Environmental Statistics*, 13(1):60.

Finley, A. O., Banerjee, S., Weiskittel, A. R., Babcock, C., and Cook, B. D. (201[illegible] Dynamic spatial regression models for space-varying forest stand tables. [illegible]s, 25(8):596–609.

Finley, A. O., Banerjee, S., Zhou, Y., Cook, B. D., [illegible], C. (2017). Joint hierarchical models for sparsely sampled high-d[illegible]nal Li[illegible]nd forest variables. *Remote Sensing of Environm[illegible]nt*, 190:149–161.

G-LiHT (2016). Goddard's lid[illegible], per[illegible]tral and thermal (G-LiHT) imager. `http://www.gliht.gsfc.na[illegible]` [illegible]1-2017.

GEDI (2014). [illegible]bal [illegible]stem[illegible]ynamics investigation lidar. `http://science.nasa.gov/mis[illegible]e/[illegible]d`: 8-11-2017.

[illegible] K[illegible]r, W. (2015). Cross-Covariance Functions for Multivariate Geostatistics. *Statistical S[illegible]nce*, 30(2):147–163.

G[illegible]ou, G. (1996). Measuring the pricing error of the arbitrage pricing theory.

Gneiting, T. and Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378.

Hanks, E. M., Schliep, E. M., Hooten, M. B., and Hoeting, J. A. (2015). Restricted spatial regression in practice: geostatistical models, confounding, and robustness under model misspecification. *Environmetrics*, 26(4):243–254.

Heaton, M. J., Datta, A., Finley, A. O., Furrer, R., Guhaniyogi, R., Gerber, F., Gramacy, R. B., Hammerling, D., Katzfuss, M., Lindgren, F., Nychka, D. W., Sun, F., and Zammit-Mangion, A. (2017). Methods for analyzing large spatial data: A review and comparison. *arXiv e-prints*. `https://arxiv.org/abs/1710.05013`.

Hogan, J. W. and Tchernis, R. (2004). Bayesian factor analysis for spatially correlated data, with application to summarizing area-level material deprivation from census data. *Journal of the American Statistical Association*, 99(466):314–324.

ICESat-2 (2015). Ice, cloud, and land elevation satellite-2. `http://icesat.gsfc.nasa.gov/icesat2`. Accessed: 8-11-2017.

Jakubowski, M. K., Guo, Q., and Kelly, M. (2013). Tradeoffs between lidar pulse density and forest measurement accuracy. *Remote Sensing of Environment*, 130(Supplement C):245 – 253.

Junttila, V. and Laa[illegible] (2015). Bayesian principal component regression model with spatial effects for forest inventory variables under small field sample size. *Remote Sensing of Environment*, 171(Supplement C):45 – 57.

[illegible] (1996). *Graphical Models*. Clarendon Press, Oxford, United Kingdom.

Lopes, H. F. and West, M. (2004). Bayesian model assessment in factor analysis. *Statistica Sinica*,

Murphy, K. (2012). *Machine Learning: A probabilistic perspective*. The MIT Press, Cambridge, MA.

Næsset, E. (2011). Estimating above-ground biomass in young forests with airborne laser scanning. *International Journal of Remote Sensing*, 32(2):473–501.

Nelson, R., Gobakken, T., Næsset, E., Gregoire, T., Ståhl, G., Holm, S., and Flewelling, J. (2012). Lidar sampling – using an airborne profiler to estimate forest biomass in Hedmark County, Norway. *Remote Sensing of Environment*, 123:563–578.

Nelson, R., Margolis, H., Montesano, P., Sun, G., Cook, L., Andersen, H.-E., B., Pellat, F. P., Fickel, T., Kauffman, J., and Pri y, S. (2017). Lidar- f aboveground biomass in the continental us and mexic satellite observations. *Remote Sensing of Environment*, 188(Supplement C):127 – 140.

Ren, Q. and Banerjee, S. (2013). Hierarchical Factor Models for Larg ligned Data: A Low-Rank Predictive Process Approach. *Bi trics*, 69(1):19–30.

Schmidt, A. M. and Gelfand, A. E. (2003). A bayesian o ation approach for multivariate pollutant data. *Journal of Geophys al esearch spheres*, 108(D24).

Stein, M. L. (20). Limitations o a mations for covariance matrices of spatial data. *Spatial Statistic*, 8:1

White, J. C., Wulder, Vastaranta, M., Coops Nicholas, C., Cook, B. D., Pitt, D., and W M.). A st practices guide for generating forest inventory attributes fro ing data using an area-based approach. *The Forestry Chronicle*, 723.

Woodall, C. W. ulston, J. W., Domke, G. M., Walters, B. F., Wear, D. N., Smith, J. E., E., Clough, B. J., Cohen, W. B., Griffith, D. M., et al. (2015). The US forest carbon accounting framework: Stocks and stock change, 1990-2016.

Yeniay, O. and Goktas, A. (2002). A comparison of partial least squares regression with other

Statistica Sinica: Preprint
doi:10.5705/ss.202018.0005

prediction methods. *Hacettepe Journal of Mathematics and Statistics*, 31(99):99–101.

Zhang, H. (2007). Maximum-likelihood estimation for multivariate spatial linear coregionalization models. *Environmetrics*, 18(2):125–139.

Zhang, X. (2016). An optimized blas library based on gotoblas2. `https://github.com/xianyi/OpenBLAS/`. Accessed 2015-06-01.

Department of Mathematics & Statistics, Portland State University, Portland

E-mail: dantayrod@pdx.edu

Department of Forestry, Michigan State University, East Lansing, MI

E-mail: finleya@msu.edu

Department of Biostatistics, Johns Hopkins University, Baltimore, MA

E-mail: abhidatta@jhu.edu

School of Environmental and Forest Sciences, University of Washington, Seattle, WA

E-mail: babcoc7@uw.edu

USDA Forest Service Pacific Northwest Research Station, Seattle, WA

E-mail: handersen@fs.fed.us

Biospheric Sciences Laboratory, NASA Goddard Space Flight Center, Greenbelt, MD

E-mail:

Biospheric Sciences Laboratory, NASA Goddard Space Flight Center, Greenbelt, MD

E-mail: douglas.morton@nasa.gov

Department of Statistics, University of California Los Angeles, Los Angeles, CA

E-mail: sudipto@ucla.edu