Supplementary materials for this article are available at https:// doi.org/ 10.1007/ s13253-023-00548-z.

A Variance Partitioning Multi-level Model for Forest Inventory Data with a Fixed Plot Design

Isa MARQUES, Paul F. V. WIEMANN, and Thomas KNEIB

Forest inventories are often carried out with a particular design, consisting of a multi-level structure of observation plots spread over a larger domain and a fixed plot design of exact observation locations within these plots. Consequently, the resulting data are collected intensively within plots of equal size but with much less intensity at larger spatial scales. The resulting data are likely to be spatially correlated both within and between plots, with spatial effects extending over two different areas. However, a Gaussian process model with a standard covariance structure is generally unable to capture dependence at both fine and coarse scales of variation as well as for their interaction. In this paper, we develop a computationally feasible multi-level spatial model that accounts for dependence at multiple scales. We use a data-driven approach to determine the weight of each spatial process in the model to partition the variability of the measurements. We use simulated and German small tree inventory data to evaluate the model's performance.

Supplementary material to this paper is provided online.

Key Words: Bayesian inference; Forestry; Markov chain Monte Carlo simulations; Multi-level modeling; Spatial statistics.

1. INTRODUCTION

Designs for collecting spatially oriented data in agricultural, biological, or environmental research often entail multi-level structures where data are collected at very different intensities in different parts of the domain. As an example that will serve as the application of interest later on in this paper, consider the design illustrated in Fig. 1 that arose from a large-scale project on biodiversity research in Germany (BIOKLIM Project).[1] As a part of this project, information on forest cover of European blueberry in the Bavarian Forest National Park in Germany was collected for a number of plots distributed over a large spatial range

Isa Marques and Paul F. V. Wiemann have contributed equally to this work.

I. Marques (✉) · P. F. V. Wiemann · T. Kneib, University of Göttingen, Chair of Statistics, Humboldtallee 3, 37073 Göttingen, Germany (E-mail: *imarques@uni-goettingen.de*). P. F. V. Wiemann, Department of Statistics, Texas A&M University, College Station, TX, USA.

Journal of Agricultural, Biological, and Environmental Statistics, Volume 28, Number 4, Pages 706–725
https://doi.org/10.1007/s13253-023-00548-z

[1] The dataset is available from the corresponding author on reasonable request https://keep.eu/project-ext/26176/.

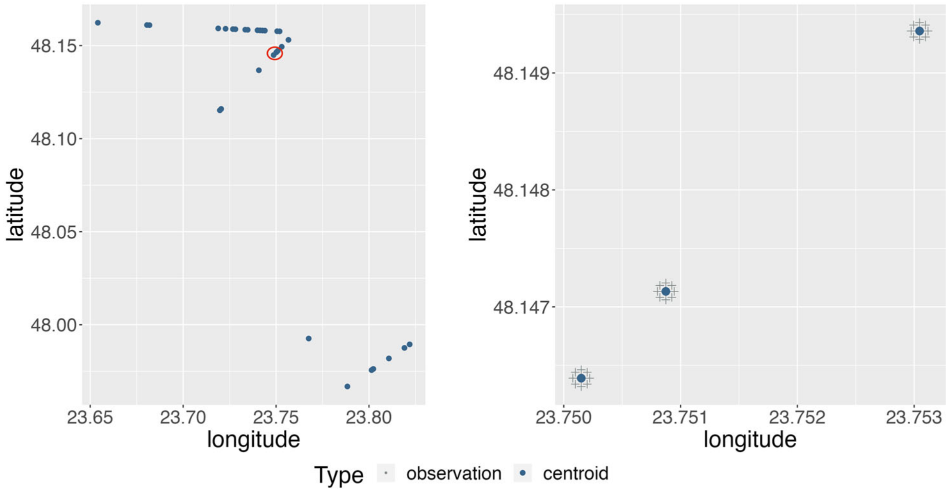

Figure 1. Locations of 30 identically sized plots distributed over a large spatial range (left panel) with identical distribution of observation locations within the plots (zoom in for the area represented by the red circle on the left in the right panel). The plots are located along straight transects following the altitude gradient. The data collected included more plots, but we restricted the dataset to plots of European blueberries.

(left panel) along altitude gradients where within each plot intensive data collection takes place on locations that are distributed in exactly the same setup around the centroid location of the plot (magnified in the right panel). As a consequence, the actual data are organized according to a specific multi-level structure representing two different spatial scales with large distances between the plots and much smaller distances between the actual locations of observations within the plots. Similar designs are very common, especially in forestry, where forest inventories are probably the most accurate source of data, but are sparsely collected due to the high financial costs involved (see, for example, Bässler et al. 2010; Junttila et al. 2013; Finley et al. 2009, 2011).

Considering the analysis of forest cover as a dependent variable within a regression scenario, it is most likely not sufficient to relate this to measured environmental factors. It is therefore common practice to add a spatially correlated effect, e.g., based on a Gaussian process, to account for unexplained spatial dependence in the data. However, due to the specific design of the data collection, standard covariance functions are unlikely to be flexible enough to represent the multi-level structure. More specifically, they cannot simultaneously capture spatial variation between and within plots, as spatial variation occurs at completely different spatial scales. Previous work on multi-resolution models relates to our objective of modeling data available at different spatial scales (Nychka et al. 2015; Katzfuss 2017), but we can cast our problem into a more straightforward multi-level model framework directly addressing the data structure of the forest inventories.

The goal of this paper is to develop efficient Bayesian inference with Markov Chain Monte Carlo (MCMC) for spatial regression models acknowledging the different spatial scales arising from the fixed plot design. More specifically, we aim to (1) adequately account for unobserved spatial variation at different scales, (2) allow for interactions between effects at different scales, (3) obtain appropriate uncertainty estimates for the regression effects in the model, (4) obtain predictions at new locations within the observed plots as well as for new plots, and (5) use efficient ways of handling Gaussian process models to make inference tractable.

In our data, there are two different spatial scales to consider: the coarser plot level, where only plot centroids are considered, and the finer within-plot level, which corresponds to the area around the centroids (see again Fig. 1), where the circles (i.e., plots) are assumed to be replicates of each other. This can easily be extended to include more scales, for example if the plots themselves are organized into clusters. We allow any two scales to interact by using Kronecker products of the dependence structures on the two scales. This follows ideas developed in Knorr-Held (2000) and Franco-Villoria et al. (2022) for interactions in space-time models. However, here we extend the concept to the case of two spatial effects at different scales, i.e., a space–space interaction.

For inference, we follow a Bayesian approach based on MCMC simulations. To improve the computational efficiency of the MCMC sampler, we exploit the techniques developed in Stegle et al. (2011), which have rarely been used in the spatial statistics literature. This technique allows for efficient inference in matrix-variate Gaussian models with i.i.d. observation noise by rotation of the data prior to evaluating the multivariate normal likelihood. The resulting (marginalized) likelihood has a diagonal covariance that is easier to factorize than a dense covariance. Indeed, although one can explore the Kronecker product for computational efficiency in spatial models, this is generally not possible in models that additionally include i.i.d. observation noise in the marginalized multivariate normal likelihood. Thus, this technique has benefits that extend beyond our space–space interactions to any other interactions (such as space-time), and speed-up inference with MCMC in any matrix-variate Gaussian model with i.i.d. observation noise.

Finally, our model incorporates a data-driven variance partitioning approach to determine the contribution of each spatial structure (within plot, between plots, interaction) and nugget to the model, thus avoiding the need to postulate the presence or absence of an effect a priori. This also helps to stabilize inference in situations where certain effects are absent, and improves the interpretability of the model.

The paper is organized as follows: In Sect. 2, we introduce our novel model more formally. In Sect. 3, details on inference are provided, while in Sect. 4, we explain how predictions are obtained. A simulation study is provided in Sect. 5. Finally, in Sect. 6 we consider a German inventory of European blueberries that exemplifies the usefulness of allowing for interaction between effects on different spatial scales.

2. MODEL STRUCTURE

2.1. FIXED PLOT DESIGNS WITH DIFFERENT SCALES

We consider regression data collected on a spatial domain $\mathcal{S} \subset \mathbb{R}^2$. Within $\mathcal{S}$, data are only available at m equally sized areas/plots, $\mathcal{S}_i \subset \mathcal{S}$, $i = 1, \ldots, m$ represented for example by the coordinates $\boldsymbol{s}_i$ of their centroids. We assume that each plot has the same number of observations $y(\boldsymbol{s}_{ij})$, $j = 1, \ldots, n$, located at the same positions relative to the centroid of the plot (see Fig. 1 for a graphical representation of such a structure; in

Supplement 6 an additional example is provided), which is, in fact, a prevalent structure in forest inventories. More precisely, let $\boldsymbol{s}_{ij}$ denote the location associated with observation $y(\boldsymbol{s}_{ij})$, then $\forall i, k \in 1, \ldots, m$ and $\forall j \in 1, \ldots, n$, the equality $\boldsymbol{s}_{ij} - \boldsymbol{s}_i = \boldsymbol{s}_{kj} - \boldsymbol{s}_k$ holds. We refer to such designs as fixed plot designs with different scales.

2.2. A Spatial Regression Model for Fixed Plot Designs

To incorporate spatial variation in a regression model for fixed plot designs, we consider the model equation

$$y(\boldsymbol{s}_{ij}) = \boldsymbol{x}(\boldsymbol{s}_{ij})'\boldsymbol{\beta} + \gamma^b(\boldsymbol{s}_i) + \gamma^w(\boldsymbol{s}_{ij} - \boldsymbol{s}_i) + \gamma^{int}(\boldsymbol{s}_i, \boldsymbol{s}_{ij} - \boldsymbol{s}_i) + \varepsilon_{ij} \tag{1}$$

where $y(\boldsymbol{s}_{ij})$ and $\boldsymbol{x}(\boldsymbol{s}_{ij})$ represent information on the response variable and the q-dimensional vector of covariates, respectively, $\boldsymbol{\beta}$ are corresponding regression coefficients, and ε_{ij} is an i.i.d. error term. The overall spatial variation is represented by the sum of three spatial effects

$$\gamma(\boldsymbol{s}_i, \boldsymbol{s}_{ij} - \boldsymbol{s}_i) = \gamma^b(\boldsymbol{s}_i) + \gamma^w(\boldsymbol{s}_{ij} - \boldsymbol{s}_i) + \gamma^{int}(\boldsymbol{s}_i, \boldsymbol{s}_{ij} - \boldsymbol{s}_i)$$

that corresponds to the spatial variation between plots on the large spatial scale ($\gamma^b(\boldsymbol{s}_i)$ being a function of the centroid locations alone), spatial variation within the plots ($\gamma^w(\boldsymbol{s}_{ij} - \boldsymbol{s}_i)$ being a function of the distance to the centroid alone), and their potential interaction ($\gamma^{int}(\boldsymbol{s}_i, \boldsymbol{s}_{ij} - \boldsymbol{s}_i)$ being a function of both sources of spatial information).

In this way, the overall spatial dependence implied by the composed spatial process $\gamma(\boldsymbol{s}_i, \boldsymbol{s}_{ij} - \boldsymbol{s}_i)$ can be much more complex than the spatial dependence of each of the individual components. The idea is to first account for fine-scale spatial structure within the plots via $\gamma^w(\boldsymbol{s}_{ij} - \boldsymbol{s}_i)$. Since this structure does not account for additional large-scale spatial correlation between plots, we superpose the spatial effect $\gamma^b(\boldsymbol{s}_i)$. The superposition of spatial effects allows us to explain both fine- and large-scale spatial dependence, without recurring to more complex and computationally intensive non-stationary spatial models (see, e.g., Lindgren et al. 2011; Nychka et al. 2015). Finally, any remaining interactions between and within plots are accounted for by an additional spatial process $\gamma^{int}(\boldsymbol{s}_i, \boldsymbol{s}_{ij} - \boldsymbol{s}_i)$. More details on the structure of each spatial effect are provided in Sect. 2.3.

2.3. Variance Partitioning Priors

Rather than assigning independent priors to the different quantities in model (1), we distribute the variance between spatial effects in a variance partitioning multi-level model (VPMM) specified as

$$y(\boldsymbol{s}_{ij}) = \boldsymbol{x}(\boldsymbol{s}_{ij})'\boldsymbol{\beta} + \tau \left(\sqrt{a_b}\gamma^b(\boldsymbol{s}_i) + \sqrt{a_w}\gamma^w(\boldsymbol{s}_{ij} - \boldsymbol{s}_i) + \sqrt{a_{int}}\gamma^{int}(\boldsymbol{s}_i, \boldsymbol{s}_{ij} - \boldsymbol{s}_i) + \sqrt{a_\varepsilon}\varepsilon_{ij}\right) \tag{2}$$

where $\tau > 0$ represents the overall variation, while the weights $0 \leq a_b, a_w, a_{int}, a_\varepsilon \leq 1$, subject to $a_b + a_w + a_{int} + a_\varepsilon = 1$ distribute this variation across the four sources of variability (see Fuglstad et al. 2020; Franco-Villoria et al. 2022, for similar variance partitioning specifications). One can think of the weight vector $\boldsymbol{a}$ as implying a joint prior for the nugget effect ε_{ij} and the three spatial effects. Using a joint prior here makes sense because (1) the main and interaction spatial effects in Eq. (2) are typically not independent and (2) for small spatial ranges between and within areas, some components of the main effect in Eq. (2) will approximately behave like the nugget. Moreover, from the stand-point of interpretability, interpretation of the relative contribution of each effect is facilitated and the resulting prior is more intuitive to elicit.

Assuming that data are organized according to the multi-level structure, Eq. (2) can be rewritten in matrix notation as

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \tau\left(\sqrt{a_b}\boldsymbol{Z}\boldsymbol{\gamma}^b + \sqrt{a_w}\boldsymbol{\gamma}^w + \sqrt{a_{int}}\boldsymbol{\gamma}^{int} + \sqrt{a_\varepsilon}\boldsymbol{\varepsilon}\right) \tag{3}$$

with the vector of observations $\boldsymbol{y}$, the design matrix $\boldsymbol{X}$, the block-diagonal matrix $\boldsymbol{Z} = \text{blockdiag}(\mathbf{1}_n, \ldots, \mathbf{1}_n)$, and the vector of residuals $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \boldsymbol{I})$ appropriately defined (e.g., $\boldsymbol{\varepsilon} = (\varepsilon_{11}, \ldots, \varepsilon_{1n}, \varepsilon_{21}, \ldots, \varepsilon_{mn})'$ and similar definitions for the other quantities).

For the different components in the VPMM, we now make more specific distributional assumptions where zero mean Gaussian random fields (GRFs) will be considered for all spatial effects, besides $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \boldsymbol{I})$. More concretely, the GRF $\boldsymbol{\gamma}^w = (\gamma^w(\boldsymbol{s}_{11} - \boldsymbol{s}_1), \ldots, \gamma^w(\boldsymbol{s}_{1n} - \boldsymbol{s}_1), \gamma^w(\boldsymbol{s}_{21} - \boldsymbol{s}_2), \ldots, \gamma^w(\boldsymbol{s}_{mn} - \boldsymbol{s}_m))'$ describing the spatial variation within each plot is a priori assumed to not be correlated between areas such that

$$\boldsymbol{\gamma}^w \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}_m \otimes \boldsymbol{R}^w) \tag{4}$$

where $\boldsymbol{R}^w$ is the correlation matrix of size $n \times n$ based on the positive-definite exponential covariance function $\text{Cor}(\boldsymbol{s}_{ij} - \boldsymbol{s}_i, \boldsymbol{s}_{il} - \boldsymbol{s}_i) = \exp\left(-\kappa_w\|\boldsymbol{s}_{ij} - \boldsymbol{s}_{il}\|\right)$, $i = 1, \ldots, m$ and $j, l = 1 \ldots, n$, where κ_w is related to the spatial range ρ_w of the GRF within the plot (see Chapter 2 in Gelfand et al. 2010). The spatial range is defined as the minimum distance at which the spatial correlation between locations is smaller than or equal to 0.05. Note that in the evaluation of the correlation function, the location of the plot centroid cancels out such that only relative distances within a plot play a role.

The GRF $\boldsymbol{\gamma}^b = (\gamma^b(\boldsymbol{s}_1), \ldots, \gamma^b(\boldsymbol{s}_m))'$ acts as a random intercept for area $\mathcal{S}_i$ with

$$\boldsymbol{\gamma}^b \sim \mathcal{N}(\mathbf{0}, \boldsymbol{R}^b), \tag{5}$$

where $\boldsymbol{R}^b$ is a correlation matrix of size $m \times m$ based on the positive-definite exponential covariance function $\text{Cor}(\boldsymbol{s}_i, \boldsymbol{s}_k) = \exp\left(-\kappa_b||\boldsymbol{s}_i - \boldsymbol{s}_k||\right)$, $i, k = 1 \ldots, m$, where κ_b is related to the spatial range ρ_b of the GRF between plots.

Lastly, the interaction term $\boldsymbol{\gamma}^{int} = (\gamma^{int}(\boldsymbol{s}_1, \boldsymbol{s}_{11} - \boldsymbol{s}_1), \ldots, \gamma^{int}(\boldsymbol{s}_m, \boldsymbol{s}_{mn} - \boldsymbol{s}_m))'$ is such that

$$\boldsymbol{\gamma}^{int} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{R}^b \otimes \boldsymbol{R}^w). \tag{6}$$

The covariance $\boldsymbol{R}^b \otimes \boldsymbol{R}^w$ is positive definite since it results from the Kronecker product of two positive-definite matrices (see Theorem 9 in Horn and Johnson 2012). The Kronecker product represents the interaction between the two spatial scales, as it assumes that the spatial dependence structure within each plot depends on the spatial dependence pattern between all plots. More concretely, it accounts for additional correlation among observations from different plots but close to each other relative to the plots' origin. Such interactions make sense in designs in which the environmental conditions (e.g., soil type) change identically in space within each plot or when an external factor, like wind from one direction or fences, affect all plots in the same manner. In the application, we consider plots located in line transects along an altitude gradient, such that the same locations in different plots have similar inclination and exposition (Bässler et al. 2010).

In the following, we denote $\gamma^b(\boldsymbol{s}_i) + \gamma^w(\boldsymbol{s}_{ij} - \boldsymbol{s}_i)$ the spatial main effects and $\gamma^{int}(\boldsymbol{s}_i, \boldsymbol{s}_{ij} - \boldsymbol{s}_i)$ the spatial interaction effect. For $a_{int} = 0$, the VPMM model implies the correlation structure

$$\operatorname{Cor}(y_{ij}, y_{kl}) = \begin{cases} a_b \boldsymbol{R}^b[i,k] & i \neq k, \\ a_b \boldsymbol{R}^b[i,k] + a_w \boldsymbol{R}^w[j,l] & i = k, j \neq l, \\ a_b \boldsymbol{R}^b[i,k] + a_w \boldsymbol{R}^w[j,l] + a_\varepsilon & i = k, j = l. \end{cases} \tag{7}$$

Thus, for observations in the same plot, we always have within-correlation, but if the plots are different this correlation is zero. If both the plots and locations within an area are different, we still have between-plot correlation.

The spatial interaction effect implies the pointwise correlation structure $\operatorname{Cor}(\gamma^{int}(\boldsymbol{s}_i, \boldsymbol{s}_{ij} - \boldsymbol{s}_i), \gamma^{int}(\boldsymbol{s}_k, \boldsymbol{s}_{kl} - \boldsymbol{s}_k)) = \boldsymbol{R}^b[i,k]\ \boldsymbol{R}^w[j,l]$. Consequently, for $a_{int} \neq 0$, we add $a_{int} \boldsymbol{R}^b[i,k]\ \boldsymbol{R}^w[j,l]$ to every case in Eq. (7)

2.4. Relation to Other Designs

In space-time contexts, one can follow a similar method to the one above. For example, in the case of one spatial resolution and one time resolution, one can adapt Eq. (2) to

$$y(\boldsymbol{s}_i, t_j) = \boldsymbol{x}(\boldsymbol{s}_i, t_j)'\boldsymbol{\beta} + \tau \left(\sqrt{a_s} \boldsymbol{\gamma}^s(\boldsymbol{s}_i) + \sqrt{a_t} \gamma^t(t_j) + \sqrt{a_{int}} \gamma^{int}(\boldsymbol{s}_i, t_j) + \sqrt{a_\varepsilon} \varepsilon_{ij} \right)$$

where $i = 1, \ldots, m$ indexes the plots, $j = 1, \ldots, n$ is the time index, and $(s_i, t_j) \in \mathbb{R}^2 \times \mathbb{R}$, $\forall i, j$. Moreover, in matrix notation (as introduced in the previous section) $\boldsymbol{\gamma}^s \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{R}^s \otimes \boldsymbol{I}_n)$, $\boldsymbol{\gamma}^t \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_m \otimes \boldsymbol{R}^t)$ and $\boldsymbol{\gamma}^{int} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{R}^s \otimes \boldsymbol{R}^t)$, where $\boldsymbol{R}^s$ is a spatial correlation matrix and $\boldsymbol{R}^t$ is a temporal correlation matrix. The novelty in a space-time context is that the computational trick that we introduce in Sect. 3.2 can also be used here to reduce the run-time complexity of factorizing the covariance function of the associated (partly marginalized) likelihood to $O(m^3 + n^3)$.

3. INFERENCE

3.1. Prior Structure

Consider the vector of all structural model parameters $\boldsymbol{\vartheta} = (\boldsymbol{\beta}, \tau^2, \boldsymbol{a}', \boldsymbol{\kappa}')'$, where $\boldsymbol{\kappa} = (\kappa_b, \kappa_w)'$. We use diffuse priors for $\boldsymbol{\beta}$, i.e., $\beta_0 \sim N(0, 100^2)$ and $\beta_v \sim N(0, 10^2)$ for $v = 1, \ldots, q$ with higher uncertainty attached to the intercept. For τ^2, we adopt a weakly informative inverse gamma distribution $IG(c = 0.001, d = 0.001)$ following the common practice of using $c = d$, with both values approaching zero, as a weakly informative choice for variance parameters (see Sect. 4.4 of Fahrmeir et al. (2013)). To sample within $\mathbb{R}$, we sample the logarithmic counterpart $\log(\tau^2)$ and change the density accordingly, following the change of variable theorem.

We assign a joint Dirichlet prior with parameters $\alpha_1, \ldots, \alpha_4 > 0$ to the weights $\boldsymbol{a}$. For notational simplicity, we replace here $(a_b, a_w, a_{int}, a_\varepsilon)$ by $(a_1, \ldots a_4)$ such that

$$p(\boldsymbol{a}) = \frac{1}{B(\alpha_1, \ldots, \alpha_4)} \prod_{p=1}^{4} a_p^{\alpha_p - 1}, \quad \boldsymbol{a} = (a_1, \ldots, a_4) \in \Delta^4$$

where $B(\cdot)$ is a multivariate beta function and Δ^4 is the 3-simplex. If any of the weights is 0 or 1, then the density is 0. We set $\alpha_1 = \alpha_2 = \alpha_3 = \alpha_4 = 1$ such that the prior is uniform and represents no preference for any of the random effects. Furthermore, as we do not sample $\boldsymbol{a}$ directly but sample on the equipotent $\mathbb{R}^3$, we need to perform a change of variable transformation. The transformation b_p and the so-called break proportions c_p can be defined element-wise as

$$b_p = \operatorname{logit}(c_p) + \log\left(\frac{1}{4-p}\right) \text{ where } c_p = \frac{a_p}{1 - \sum_{p'=1}^{p-1} a_{p'}}, \quad \text{for } p \geq 2,$$

where $\boldsymbol{b} = (b_2, b_3, b_4)' \in \mathbb{R}^3$, $\boldsymbol{c} = (c_2, c_3, c_4)' \in \mathbb{R}^3$ (see Stan Development Team 2022, Sect. 10.7).

For the parameters κ_b, κ_w, we sample the logarithmic counterpart $\theta_b = \log(\kappa_b)$ and $\theta_w = \log(\kappa_w)$. The densities are changed accordingly. In what follows, we describe the prior structure for κ_b, but the same logic applies to κ_w. We assume a normally distributed prior $\theta_b \sim N(\mu_{\kappa_b}, \sigma_\kappa^2)$. Then, given that for the exponential correlation function the spatial range satisfies $\rho_b \approx 3/\kappa_b$, from the properties of the log-normal distribution we obtain $\rho_b \sim \text{Log-normal}(\log(3) - \mu_{\kappa_b}, \sigma_{\kappa_b}^2)$. The p-quantiles of the log-normal distributions for the correlation range are

$$\rho_b(p) = 3\exp(-\mu_{\kappa_b} + \sigma_{\kappa_b}\Phi^{-1}(p)) \tag{8}$$

where $0 \geq p \geq 1$, and $\Phi(\cdot)$ is the cumulative distribution function for the standard normal distribution (see Ingebrigtsen et al. 2015, for a similar method). To choose priors, we specify two quantiles of the prior for ρ_b. In our case, we focus on the median and 0.95-quantile and then solve the corresponding two equations. We illustrate the prior's behavior in Fig. 2, which is based on the settings used in the simulation study and part of the real data application.

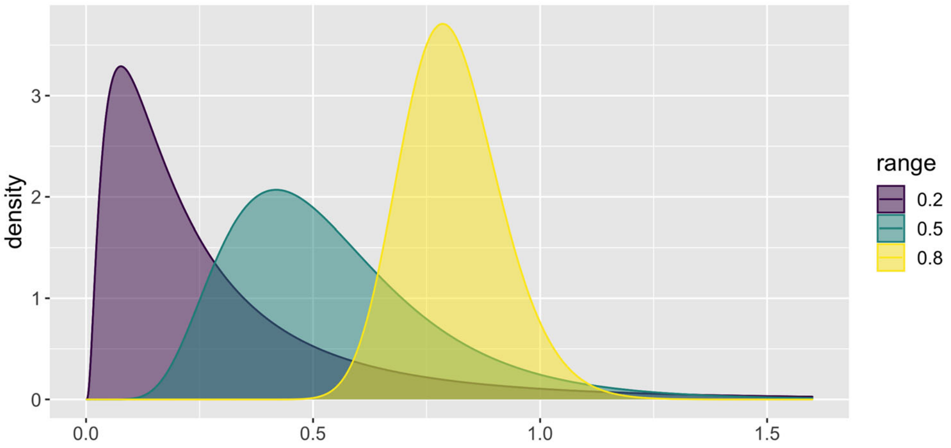

Figure 2. Density of the prior for spatial ranges. We consider $\rho_b(0.95) = 1$ and test different values for the range $\rho_b(0.5)$ as shown in the legend. It follows similarly for ρ_w.

In the figure, we consider $\rho_b(0.95) = 1$ and test different values for the range $\rho_b(0.5)$. It follows similarly for ρ_w.

3.2. EFFICIENT INFERENCE

This section introduces the technique of Stegle et al. (2011) in the context of our model in order to reduce computational complexity. Consider the marginalized likelihood following Eq. (3) where

$$\boldsymbol{y}|\boldsymbol{\beta}, \tau^2, \boldsymbol{a}, \boldsymbol{\kappa} \sim N\left(\boldsymbol{X}\boldsymbol{\beta}, \tau^2\left(a_\varepsilon \boldsymbol{I}_{mn} + a_b \boldsymbol{Z}\boldsymbol{R}^b\boldsymbol{Z}' + (a_w \boldsymbol{I}_m + a_{int}\boldsymbol{R}^b) \otimes \boldsymbol{R}^w\right)\right). \quad (9)$$

By integrating out the GRF in a spatial regression model, we typically achieve faster convergence in MCMC samplers (Finley et al. 2015). However, the cost of factorizing the covariance in Eq. (9) is cubic in mn.

By instead considering the likelihood with unmarginalized between main effect $\boldsymbol{\gamma}^b$, we can exploit the structure of $a_b \boldsymbol{I}_{mn} + (a_{int}\boldsymbol{I}_m + a_\varepsilon \boldsymbol{R}^b) \otimes \boldsymbol{R}^w$ to reduce computational complexity using a technique introduced in Stegle et al. (2011). With $\boldsymbol{\gamma}^b$ not marginalized, we obtain

$$\boldsymbol{y}|\boldsymbol{\beta}, \tau^2, \boldsymbol{a}, \boldsymbol{\kappa}, \boldsymbol{\gamma}^b \sim N\left(\boldsymbol{X}\boldsymbol{\beta} + \sqrt{\tau^2 a_b}\boldsymbol{Z}\boldsymbol{\gamma}^b, \tau^2\left(a_\varepsilon \boldsymbol{I}_{mn} + (a_w \boldsymbol{I}_m + a_{int}\boldsymbol{R}^b) \otimes \boldsymbol{R}^w\right)\right). \quad (10)$$

The evaluation of this multivariate normal distribution requires the calculation of the determinant and inverse of covariance which is a $mn \times mn$ matrix with costs $O(m^3n^3)$. These tasks can be accomplished more efficiently by further exploiting the properties of the Kronecker product.

Consider $\boldsymbol{Y} \in \mathbb{R}^{n \times m}$ with n rows and m columns. We define $\text{vec}(\boldsymbol{Y}) = \boldsymbol{y}$ to be the vector obtained by concatenating the columns of $\boldsymbol{Y}$. A Kronecker product plus a constant diagonal term can then be rewritten as

$$\begin{aligned}\tau^2\left(a_\varepsilon \boldsymbol{I}_{mn} + (a_w \boldsymbol{I}_m + a_{int}\boldsymbol{R}^b) \otimes \boldsymbol{R}^w\right) &= \tau^2 a_\varepsilon \boldsymbol{I}_{mn} + (\tau^2 a_w \boldsymbol{I}_m + \tau^2 a_{int}\boldsymbol{R}^b) \otimes \boldsymbol{R}^w \\ &= (\boldsymbol{U}_b \otimes \boldsymbol{U}_w)(\tau^2 a_\varepsilon \boldsymbol{I}_{mn} + \boldsymbol{S}_b \otimes \boldsymbol{S}_w)(\boldsymbol{U}'_b \otimes \boldsymbol{U}'_w)\end{aligned}$$

where $\boldsymbol{U}_b\boldsymbol{S}_b\boldsymbol{U}'_b$ is the eigenvalue decomposition (EVD) of $\tau^2 a_w \boldsymbol{I}_m + \tau^2 a_{int}\boldsymbol{R}^b$ and $\boldsymbol{U}_w\boldsymbol{S}_w\boldsymbol{U}'_w$ is the EVD of $\boldsymbol{R}^w$. By exploiting the identity $(\boldsymbol{U}_b \otimes \boldsymbol{U}_w)\text{vec}(\boldsymbol{Y}) = \text{vec}(\boldsymbol{U}'_w\boldsymbol{Y}\boldsymbol{U}_b)$, we can re-formulate the likelihood $\mathcal{L}$ in Eq. (10) such that

$$\begin{aligned}\mathcal{L} = -\frac{mn}{2}\log(2\pi) - \frac{1}{2}\log(|\tau^2 a_\varepsilon \boldsymbol{I}_{mn} + \boldsymbol{S}_b \otimes \boldsymbol{S}_w|) - \\ \frac{1}{2}\text{vec}(\boldsymbol{U}'_w\boldsymbol{Y}\boldsymbol{U}_b)'(\tau^2 a_\varepsilon \boldsymbol{I}_{mn} + \boldsymbol{S}_b \otimes \boldsymbol{S}_w)^{-1}\text{vec}(\boldsymbol{U}'_w\boldsymbol{Y}\boldsymbol{U}_b).\end{aligned}$$

This can now be interpreted as a multivariate normal distribution with diagonal covariance matrix $\tau^2 a_\varepsilon \boldsymbol{I}_{mn} + \boldsymbol{S}_b \otimes \boldsymbol{S}_w$ and rotated data $\text{vec}(\boldsymbol{U}'_w\boldsymbol{Y}\boldsymbol{U}_b)$ (Stegle et al. 2011).

The factorization of the diagonal covariance matrix implies a lower run-time complexity than that of the dense counterpart. Moreover, although we need to calculate two eigenvalue decompositions, in general, we can perform factorizations on the smaller matrices, reducing costs to $O(m^3)$ and $O(n^3)$, respectively. These two operations can additionally be parallelized. Ultimately, without parallelization, this reformulation has computational complexity of $O(n^3 + m^3)$, rather than $O(n^3m^3)$ in a global spatial model and in the scenarios we are interested in; i.e., scenarios with $n \geq 2$ and $m \geq 2$, $n^3 + m^3 < n^3 m^3$ are guaranteed.

3.3. Sampling

In the partially marginalized formulation of VPMM introduced in the previous section, we update $\boldsymbol{\gamma}^b$ and $\boldsymbol{\vartheta}$ with an alternating scheme:

Update of $\boldsymbol{\vartheta}$. For efficient sampling, we use proposals based on Hamiltonian dynamics with a subsequent Metropolis–Hastings correction known as Hamiltonian Monte Carlo (HMC, Neal 2011). In each case, the step size and the mass vector are learned during warm-up. We find that in some data settings the gradient of the unnormalized log-posterior with respect to $\log(\kappa_b)$ and $\log(\kappa_w)$ is numerically unstable and better results are obtained when removing those parameters from the HMC step and instead sample them with the Metropolis–Hastings algorithm using random-walk proposals. Similar to the HMC-based sampler, the step size of the random-walk proposals is tuned during warm-up.

Update of $\boldsymbol{\gamma}^b$. Here, we use Gibbs sampling and draw $\boldsymbol{\gamma}^b$ from the full conditional (see Supplement 1).

3.4. Software

The model is implemented in Python using the novel Liesel framework for Bayesian computation (Riebl et al. 2022). In particular, we use Goose, the MCMC library of Liesel. Goose provides a set of efficiently implemented and well-tested samplers capable of learning some tuning parameters, such as the step size, during warm-up. Different samplers can be associated with different parts of the parameter vector, allowing us to implement the sampling procedure described in Sect. 3.3 with minimal effort. Liesel facilitates using gradient-based samplers (e.g., HMC and NUTS) by taking advantage of automatic differentiation, which allows us to implement only the unnormalized log-posterior. However, using Liesel, we can—where necessary—integrate dedicated implementations incorporating the computational tricks discussed.

4. SPATIAL PREDICTIONS

4.1. Predictions at New Locations Within the Observed Plots

These types of predictions seem particularly valuable as foresters could thin out their data collection process within each plot or compensate any missing values within a plot. Consider observations y_{ij} available in each plot $i = 1, \dots, m$ at the same locations indexed with $j = 1, \dots, n$ and predictions at $t \in \mathbb{N}$ new locations in each plot indexed with $j = n+1, \dots, n+t$. For notational clarity, we write $y_{i,j}$ instead of $y(\boldsymbol{s}_{i,j})$ in the remaining part of this section. To predict a random $mt \times 1$ vector $\boldsymbol{y}_0 = (y_{1,n+1}, \dots, y_{1,n+t}, \dots, y_{m,1}, \dots, y_{m,n+t})'$ associated with a $mt \times p$ matrix of predictors, $\boldsymbol{X}_0$, we start with the joint distribution of $\widetilde{\boldsymbol{y}} = (y_{1,n+1}, \dots, y_{1,n+t}, y_{1,1} \dots, y_{1,n}, \dots, y_{m,n+1}, \dots, y_{m,1}, \dots, y_{m,n})'$. Moreover, we have $\boldsymbol{y}_1 = (y_{1,1}, y_{1,2}, \dots, y_{m,1}, \dots, y_{m,n})'$. The matrices $\widetilde{\boldsymbol{X}}$, $\widetilde{\boldsymbol{Z}}$, and $\widetilde{\boldsymbol{R}}^w$, shall denote the design matrix, projection matrix, and within-correlation matrix similar to $\boldsymbol{X}$, $\boldsymbol{Z}$, $\boldsymbol{R}^w$, but augmented such that they include the new values associated with $\boldsymbol{y}_0$. Now, the joint distribution of $\widetilde{\boldsymbol{y}}$ given the model parameters $\boldsymbol{\vartheta}$ and the between area effect $\boldsymbol{\gamma}^b$ is

$$\widetilde{\boldsymbol{y}}|\boldsymbol{\vartheta}, \boldsymbol{\gamma}^b \sim N\left(\widetilde{\boldsymbol{X}}\boldsymbol{\beta}+\sqrt{\tau^2 a_b}\widetilde{\boldsymbol{Z}}\boldsymbol{\gamma}^b, \tau^2\left((a_w \boldsymbol{I}_m+a_{int}\boldsymbol{R}^b) \otimes \widetilde{\boldsymbol{R}}^w+a_\varepsilon \boldsymbol{I}_{m(n+t)}\right)\right). \quad (11)$$

The $(n+t) \times (n+t)$ correlation matrix $\widetilde{\boldsymbol{R}}^w$ can be expressed as a block-matrix

$$\widetilde{\boldsymbol{R}}^w = \begin{bmatrix} \boldsymbol{R}^w_{00} & \boldsymbol{R}^w_{01} \\ \boldsymbol{R}^w_{10} & \boldsymbol{R}^w_{11} \end{bmatrix}$$

with the correlation matrices describing the within-correlation of the new observations and the old observations on the diagonal and the correlation matrix between those on the off-diagonal. The conditional distribution of the predictions is given by

$$\boldsymbol{y}_0|\boldsymbol{y}_1, \boldsymbol{\vartheta}, \boldsymbol{\gamma}^b \sim N\left(\boldsymbol{\mu}_0 + \boldsymbol{\Sigma}_{01}\boldsymbol{\Sigma}_{11}^{-1}(\boldsymbol{y}_1 - \boldsymbol{\mu}_1), \boldsymbol{\Sigma}_{00} - \boldsymbol{\Sigma}_{01}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{10}\right). \quad (12)$$

Here, $\boldsymbol{\mu}_0$ and $\boldsymbol{\mu}_1$ refer to the components of the mean vector in Eq. (11) suitable to express the mean of $\boldsymbol{y}_0$ and $\boldsymbol{y}_1$, respectively. Similar, the blocks $\boldsymbol{\Sigma}_{kl}$, $k, l = 0, 1$ arise from the covariance matrix in Eq. (11) referring to the conditional covariance of $\boldsymbol{y}_k$ and $\boldsymbol{y}_l$. Note, $\boldsymbol{\Sigma}_{11}$ is equal to the covariance matrix in Eq. (10) and can be efficiently inverted using Stegle's method (see Sect. 3). Thus, run-time complexity is $O(m^3 + n^3)$ rather than $O(m^3 n^3)$, with potential for parallelizing the eigendecompositions.

Bayesian prediction proceeds by sampling from the posterior predictive distribution $p(\boldsymbol{y}_0|\boldsymbol{y}) = \int p(\boldsymbol{y}_0|\boldsymbol{y}, \boldsymbol{\vartheta}, \boldsymbol{\gamma}^b)p(\boldsymbol{\vartheta}, \boldsymbol{\gamma}^b|\boldsymbol{y})\mathrm{d}\boldsymbol{\vartheta}\mathrm{d}\boldsymbol{\gamma}^b$. For each posterior sample of $(\boldsymbol{\vartheta}', (\boldsymbol{\gamma}^b)')'$, we draw $\boldsymbol{y}_0$ from the corresponding distribution (see Eq. (12)).

4.2. Predictions at New Plots

Predictions can also be constructed for new plots. Suppose we want to predict $t \in \mathbb{N}$ new plots. To predict a random $tn \times 1$ vector $\boldsymbol{y}_0 = (y_{m+1,1}, y_{m+1,n}, \dots, y_{m+t,1}, \dots y_{m+t,n})'$ associated with a $tn \times p$ matrix of predictors, $\boldsymbol{X}_0$, we start with the joint distribution of $\widetilde{\boldsymbol{y}} = (y_{m+1,1}, \dots, y_{m+1,n}, \dots, y_{m+t,n}, y_{1,n}, y_{1,2}, \dots, y_{m,n})'$, where $\boldsymbol{y}_1 = (y_{1,1}, \dots, y_{1,n}, \dots, y_{m,n})'$. The matrices $\widetilde{\boldsymbol{X}}, \widetilde{\boldsymbol{Z}}, \widetilde{\boldsymbol{R}}^b, \widetilde{\boldsymbol{\gamma}}^b$ shall denote the design matrix, projection matrix, and between correlation matrix similar to $\boldsymbol{X}$, $\boldsymbol{Z}$, $\boldsymbol{R}^b$, but augmented such they include the new values associated with $\boldsymbol{y}_0$. We can factorize

$$p(\boldsymbol{y}, \widetilde{\boldsymbol{\gamma}}^b, \boldsymbol{\vartheta}) = \underbrace{p(\boldsymbol{y}_0|\boldsymbol{y}_1, \widetilde{\boldsymbol{\gamma}}^b, \boldsymbol{\vartheta})p(\boldsymbol{\gamma}_0^b|\boldsymbol{\gamma}_1^b, \boldsymbol{\vartheta})}_{\text{predictive density}}\underbrace{p(\boldsymbol{y}_1|\boldsymbol{\gamma}_1^b, \boldsymbol{\vartheta})p(\boldsymbol{\gamma}_1^b, \boldsymbol{\vartheta})}_{\text{model density}}.$$

Assuming $t < m$, both predictive and model density can be evaluated in $O(m^3 + n^3)$ using Stegle's method. Both factors of the predictive distribution relate to a conditional normal distribution. The second term conditions on an m-dimensional vector, thus requiring the factorization of an $m \times m$ matrix, which can be done in $O(m^3)$. The first term is more involved. Consider, the joint distribution of $\widetilde{\boldsymbol{y}}$ given the model parameters $\boldsymbol{\vartheta}$ and $\widetilde{\boldsymbol{\gamma}^b}$

$$\widetilde{\boldsymbol{y}}|\boldsymbol{\vartheta}, \widetilde{\boldsymbol{\gamma}}^b \sim N\left(\widetilde{\boldsymbol{X}}\boldsymbol{\beta}+\sqrt{\tau^2 a_b}\widetilde{\boldsymbol{Z}}\widetilde{\boldsymbol{\gamma}}^b, \tau^2\left((a_w\boldsymbol{I}_{m+t}+a_{int}\widetilde{\boldsymbol{R}}^b) \otimes \boldsymbol{R}^w+a_\varepsilon\boldsymbol{I}_{(m+t)n)}\right)\right). \quad (13)$$

The $(m + t) \times (m + t)$ correlation matrix $\widetilde{\boldsymbol{R}}^b$ can be expressed as a block-matrix

$$\widetilde{\boldsymbol{R}}^b = \begin{bmatrix} \boldsymbol{R}_{00}^b & \boldsymbol{R}_{01}^b \\ \boldsymbol{R}_{10}^b & \boldsymbol{R}_{11}^b \end{bmatrix}$$

The conditional distribution of the predictions follows similar to the previous section, with the blocks $\boldsymbol{\Sigma}_{kl}$, $k, l = 0, 1$, forming the covariance matrix from Eq. (13). Once again, all terms involving the inversion of $\boldsymbol{\Sigma}_{11}$ can be efficiently computed using Stegle's method, thus reducing the computational complexity from $O(m^3n^3)$ to $O(m^3 + n^3)$.

In the absence of an interaction effect, the predictive distribution of $\boldsymbol{y}_0$ allows the insight that the expected value of $\boldsymbol{y}_0$ is constant within each plot. Moreover, the predictive distribution suggests a potential reduction in the uncertainty of the predicted values in the presence of an interaction effect. Therefore, an improvement in the predictions compared to a model focusing only on between-plot effects is expected if a considerable share of the total variance is attributed to the interaction.

5. SIMULATION

In this section, we present a simulation study that evaluates the performance of the VPMM in terms of the bias of all estimated parameters, as well as the corresponding number of MCMC effective samples (Geyer 2011). We assume that the Data Generation Model (DGM) and the Data Analysis Model (DAM) are identical and follow the VPMM. We evaluate the performance of the VPMM for: (1) increasing sample sizes m and n; (2) different true weight vectors $\boldsymbol{a}$; and (3) increasing variance τ^2. Objective (1) is to find thresholds for m and n at which the parameters of the model can be accurately estimated. Objective (2) aims to identify potential identification problems between the multiple spatial effects, or any tendency of the spatial effects to degenerate into i.i.d. processes, even if the priors on the range parameters avoid small values. Finally, in (3) we investigate how different variances affect the estimated parameters.

5.1. DGM

We expect the models to perform well for n relatively smaller than m, since the observations within-plot have m replicates. Given this, we consider $m \in \{30, 40\}$ and $n \in \{10, 25\}$. These are also close to the sample sizes in Sect. 6 (see also Supplement 6). We consider $\tau^2 \in \{1, 2\}$ and the partitionings of the variance $\boldsymbol{a} = (a_b, a_w, a_{int}, a_\varepsilon)'$ are such that $\boldsymbol{a} \in \{(0.35, 0.35, 0.2, 0.1)', (0.25, 0.55, 0.05, 0.15)', (0.70, 0.05, 0.05, 0.20)'\}$. The first vector of weights represents a well-behaved scenario that we expect should be easy to estimate for any reasonable sample size. The second vector of weights sets the interaction weight close to zero, a scenario that is realistic for data structures that do not lead to stronger correlation for the same locations in different plots. The scenario with $a_{int} = a_w \approx 0$ represents a standard model used in forest sciences for inventory data, where one simply has a random intercept for the plots, although this spatial effect is typically not spatially correlated. This scenario also aims at identifying any potential identification issues between spatial effects or tendency to degenerate, e.g., the within effect degenerates to white noise by having low values for the spatial range, instead of being assigned a weight of zero. Some parameters are kept fixed: $\kappa_b = 3/0.5$, $\kappa_w = 3/0.7$, $\beta_1 = 1$, and $\beta_2 = 0.5$. Moreover, $x(\boldsymbol{s}_{ij}) \sim N(0, 1)$. We consider 50 replicates.

5.2. DAM

The prior hierarchy follows Sect. 3.1. Since we resize every $\mathcal{S}$ and $\mathcal{S}_i$ such that $\mathcal{S} \subset [0, 1] \times [0, 1]$ and $\mathcal{S}_i \subset [0, 1] \times [0, 1]$ $\forall i$, we set $\rho_b(0.95)$ and $\rho_w(0.95)$ in Eq. (8) to the maximum diameter of the corresponding space; i.e., $\rho_b(0.95) = \rho_w(0.95) = 1$, and $\rho_b(0.5) = \rho_w(0.5) = 0.5$. We run two MCMC chains, each with 2000 MCMC samples and with a warm-up of 2000 samples. Convergence is confirmed by verifying that the R-hat (Gelman and Rubin 1992) is smaller than 1.1, as well as by checking the smallest effective sample size out of all the model's parameters, based on the median effective sample size for all MCMC samples (Geyer 2011; Gelman et al. 2013).

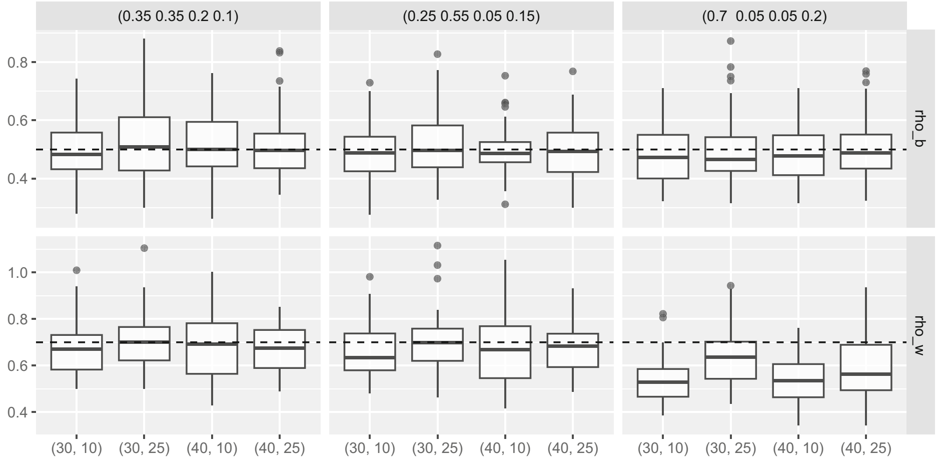

Figure 3. Boxplots of the estimated posterior mean of ρ_w and ρ_b calculated over 50 replicates for scenarios with $\tau^2 = 2$. On the x-axis, we show the different sample sizes (m, n). The columns show three different scenarios with different true weights $\boldsymbol{a} = (a_b, a_w, a_{int}, a_\varepsilon)'$ and the rows show the estimated values for each range ρ. The dashed lines show the true values.

5.3. Results

Results are summarized in Figs. 3 and 4. The main conclusions are the following:

Sample size and weights: Scenarios with $\boldsymbol{a} = (0.35, 0.35, 0.2, 0.1)'$ lead to unbiased estimates of all parameters for all sample sizes. The same is true for $\boldsymbol{a} = (0.25, 0.55, 0.05, 0.15)'$, except for $n = 10$ where there is a slight tendency for the within weight to be underestimated and the interaction weight to be overestimated. In the same direction, for $\boldsymbol{a} = (0.70, 0.05, 0.05, 0.20)'$ the within-plot range is underestimated for $n = 10$, although it remains far from zero. This underestimation ultimately leads to a slightly biased weight for the within and nugget weights, suggesting some tendency for the within effect to behave similarly to the nugget for situations in which it has a low weight and n is small. However, the priors used for the range prevent the degeneration of the within effect to white noise. Given n, both values of m behave similarly well, indicating that $m = 30$ is already large enough to recover all true model parameters. All in all, for $n = 10$, some parameters might be slightly biased for less well-behaved scenarios (some weights close to zero), but a sample size $n = 25$ is sufficient to recover unbiased estimates of all parameters.

Variance: The two values for variance τ^2 lead to nearly identical results for the distribution of the bias of all parameters, except for the dispersion of β_2, which is larger for larger τ^2. Thus, we restrain from presenting these results in the main text (see Supplement 5).

Convergence: The smallest median effective sample size is far above 100 for all scenarios. We follow the argumentation of (Gelman et al. 2013, p. 267), considering it enough for “reasonable posterior summaries” and, in particular, for posterior mean estimates. The R-hat value is also below 1.1 for all the results presented. Note that no thinning was used.

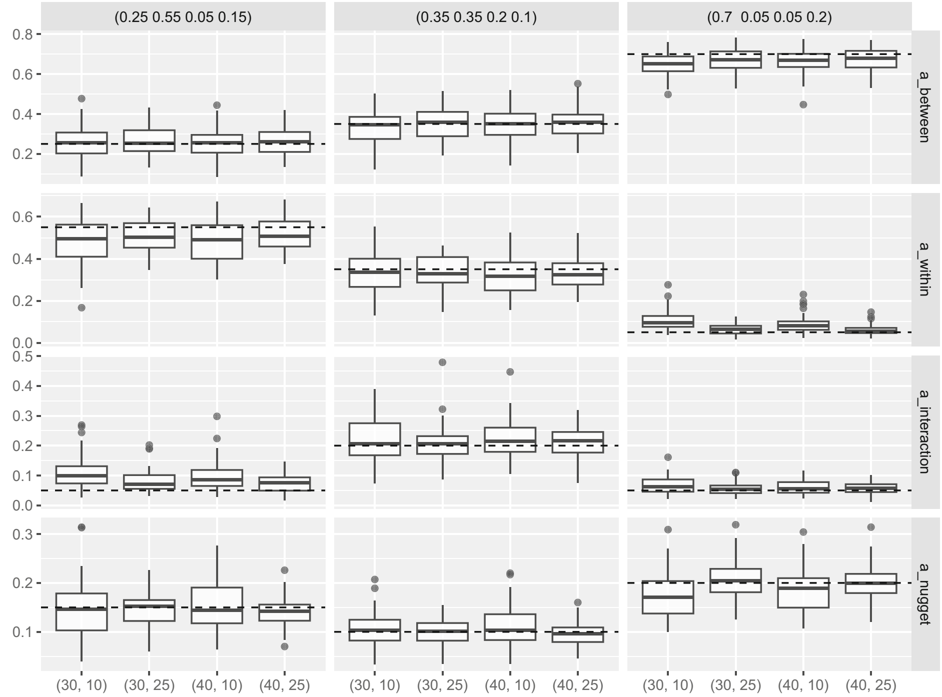

Figure 4. Boxplots of the estimated posterior mean of $\boldsymbol{a} = (a_b, a_w, a_{int}, a_\varepsilon)'$ calculated over 50 replicates for scenarios with $\tau^2 = 2$. On the x-axis, we show the different sample sizes (m, n). The columns show three different scenarios with different true weights $\boldsymbol{a}$ and the rows show the estimated values for each weight. The dashed lines show the true values.

6. APPLICATION

We consider a German forest inventory dataset from the BIOKLIM Project.[2] We model forest cover of *Vaccinium myrtillus*, also known as European blueberry. The data were collected in the Bavarian Forest National Park in $m = 30$ plots of $200m^2$. In each plot, there are $n = 8$ observations distributed on a circle and equally spaced. The structure of the data within and between plot is shown in Fig. 1.

The plots are distributed along four straight transects following the altitude gradient, such that the inclination should be roughly similar for the same location in different plots (as implied by the spatial effect $\gamma^{int}(\cdot)$ in Eq. (2)). The existence of distribution patterns along altitudinal gradients at large spatial scales remains disputed, partly because most models to date ignore potential spatial dependencies (Bässler et al. 2010). However, data collected along a transect with neighboring sampling points are likely to be spatially correlated. Thus, it makes sense to account for spatial dependence at both larger and smaller scales.

[2] See, e.g., https://keep.eu/project-ext/26176/. The dataset is available from the corresponding author on reasonable request.

6.1. General Setting

For the application, the VPMM follows the structure

$$f(y(\boldsymbol{s}_{ij}))=\beta_0+\beta_{elev}x_{elev}(\boldsymbol{s}_i)+\tau\left(\sqrt{a_b}\gamma^b(\boldsymbol{s}_i)+\sqrt{a_w}\boldsymbol{\gamma}^w(\boldsymbol{s}_{ij})+\sqrt{a_{int}}\gamma^{int}(\boldsymbol{s}_{ij})+\sqrt{a_\varepsilon}\varepsilon_{ij}\right) \tag{14}$$

where $y(\boldsymbol{s}_{ij})$ is the forest cover, which is subject to a transformation $f(\cdot)$. Particularly, $f(y(\boldsymbol{s}_{ij})) = (h \circ g)(y_{ij})$, such that $g(y(\boldsymbol{s}_{ij})) = \log(y(\boldsymbol{s}_{ij}) + 1)$ and $h(\cdot)$ additionally standardizes $g(y(\boldsymbol{s}_{ij}))$. We include standardized elevation (elev) as covariate in the model. Elevation is only available at the plot level and thus it is indexed by $\boldsymbol{s}_i$.

For the sake of comparison, we also run a non-spatial multi-level model which is commonly used for forest inventory data and specified as

$$f(y(\boldsymbol{s}_{ij})) = \beta_0 + \beta_{elev}x_{elev}(\boldsymbol{s}_i) + \tau_b b_i + \tau_\varepsilon \varepsilon_{ij} \tag{15}$$

where $f(\cdot)$ is the same transformation as in Eq. (14), $\boldsymbol{b} \sim N(0, \boldsymbol{I}_m)$, the i.i.d. errors follow $\varepsilon_{ij} \sim N(0, 1)$, and τ_b^2 and τ_ε^2 are variance parameters.

The prior hierarchy follows Sect. 3.1 and $\tau_b^2, \tau_\varepsilon^2 \sim IG(0.001, 0.001)$. We convert longitude and latitude to Universal Transverse Mercator (UTM) coordinates in kilometers and resize $\mathcal{S}$ and every $\mathcal{S}_i$ such that $\mathcal{S} \subset [0, 1] \times [0, 1]$ and $\mathcal{S}_i \subset [0, 1] \times [0, 1]\ \forall i \in \{1, \ldots, m\}$. Consequently, we consider $\rho_b(0.5) = \rho_w(0.5) = 0.5$, which represents a less informative prior while still avoiding values that go far beyond the edge length of the unit square (see Sect. 3.1). Moreover, $\rho_w(0.5) = 0.38$ and $\rho_w(0.95) = 0.72$ since dependence within the plot seems to take place mostly between the direct neighbors (see Fig. 5).

We run two MCMC chains, each with 5000 MCMC samples, including a warm-up of 2000 samples. Convergence is confirmed by verifying that the R-hat (Gelman and Rubin 1992) is smaller than 1.1 and by checking the smallest number of effective samples out of all parameters.

6.2. Evaluation Criteria

To assess the quality of the predictions for new locations and plots, we consider the mean squared error (MSE) in a leave-t-out cross-validation (CV) setting. Additionally, we also consider logarithmic score. Consider the case of new locations within a plot. The case of new plots follows similarly. To obtain the CV-MSE, the data are divided into training and test data by randomly selecting t from the n available within the plot locations for the test data. The remaining locations are used for training. This is repeated until there are fewer than t observations available that were not previously used for testing. The quality of the predictions is assessed using the posterior mean of the MSE with respect to the conditional mean (CV mean) and the posterior predictions (CV sample) (see Sect. 4). We choose $t = 1$ for within-plot predictions which implies roughly 12.5% of the data is used for testing. For predictions on new plots we use $t = 3$ which corresponds to 10% of the plots. Additionally, we consider cases of the VPMM with $a_{int} = 0$ and $a_{int} = a_w = 0$ with adjusted prior for $\boldsymbol{a}$.

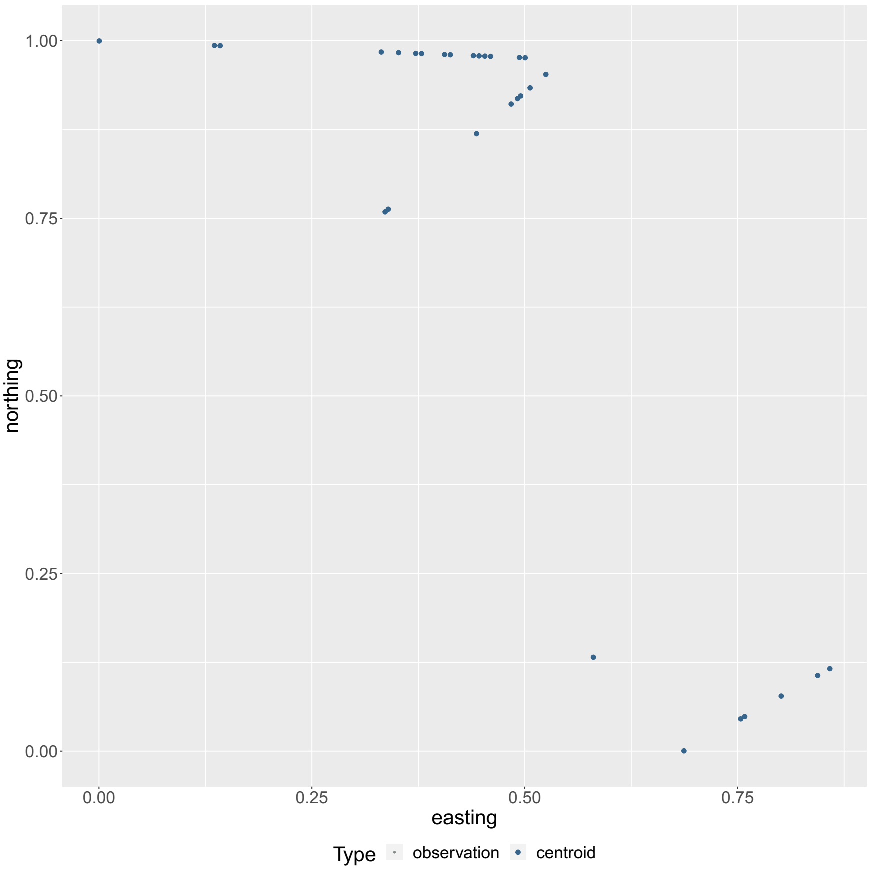

Figure 5. We rescaled the domain $\mathcal{S}$ to the unit-square for interpretability purposes.

The full-sample logarithmic score (log score) follows log $\left(\frac{1}{S}\sum_{s=1}^{S} p\left(\boldsymbol{y}|\boldsymbol{\vartheta}^{(s)}\right)\right)$, where $\int = 1, \ldots, S$ are MCMC samples, and $\boldsymbol{\vartheta}^{(s)}$ denotes the s-th MCMC sample of $\boldsymbol{\vartheta}$. Compared to Eq. (10) we also marginalize $\gamma^{b}(\cdot)$. The between effect is also marginalized in the model from Eq. (15). The full-sample log score omits the leave-one-out idea, as it has been shown that the full-sample option can have a better small-sample model discrimination ability than the cross-validated one (Krnjajić and Draper 2014), and it is computationally cheaper than doing CV.

6.3. Results

Recall that we resize $\mathcal{S}$ and every $\mathcal{S}_i$ $\forall i \in \{1, \ldots, m\}$ (see Sect. 6.1 and Fig. 5). The results are shown in Tables 1 and 2. In the VPMM, the posterior mean of τ^2 is 1.15. The results indicate that approximately 15% of the variance is attributed to the between effect, 35% to the within effect, 15% to interaction effect, and 34% to the nugget. The spatial range is 0.52 for the within effect. In Fig. 6, one can confirm that the spatial dependence within plot is mostly present for direct neighbors. Large-scale dependence is also present, as the

Table 1. Posterior mean estimates and equal-sided 90% credible interval for VPMM and non-spatial multi-level model

Parameters	VPMM	non-spatial
β_{elev}	−0.03[−0.23, 0.14]	0.09 [−0.05, 0.23]
a_b	0.15 [0.04, 0.31]	
a_w	0.35 [0.04, 0.70]	
a_{int}	0.15 [0.01, 0.34]	
a_ε	0.34 [0.04, 0. 69]	
τ^2	1.15 [0.94, 1.44]	
τ_b^2		0.09 [0.02, 0.24]
τ_ε^2		0.90 [0.78, 1.07]
ρ_b	0.30 [0.14, 0.58]	
ρ_w	0.52 [0.37, 0.69]	
Log score	−336.51	−337.79

The last row shows the log score

Table 2. Mean and sample-based CV criteria for the models, where $a_{int} = 0$ and $a_w = a_{int} = 0$ correspond to the VPMM with these weights set to zero

Parameters	VPMM	$a_{int} = 0$	$a_w = a_{int} = 0$	Non-spatial
CV-MSE mean (new locations)	1.04	1.01	1.02	1.04
CV-MSE sample (new locations)	1.80	1.76	2.23	1.93
CV-MSE mean (new plots)	1.02	1.02	1.01	1.07
CV-MSE sample (new plots)	1.93	2.02	2.00	2.09

New locations and new plots refer to predictions of the type presented in Sects. 4.1 and 4.2, respectively

model leads to a spatial range of 0.30 (approximately one-third of the edge length of the unit square) for the between effect, which covers most of each respective transect. Moreover, as expected, the interaction effect plays a relevant role. Since the plots are located along altitude gradients, the same locations on different plots are thought to have similar inclinations, thus inducing spatial correlation that can be explained by the space–space interaction.

Concerning the non-spatial multi-level model, the mean variance of the random intercept on the non-spatial model is 0.09 and thus rather small, given that the response is standardized. The remaining variance is attributed to the nugget. The credible interval (C.I.) of elevation includes zero in both models. Thus, when interpreting the results, the non-spatial multi-level model seems rather inappropriate for these data, since most of the behavior is explained by the nugget.

The evaluation criteria also point in the direction of a better performance of the VPMM. Indeed, the log score is higher for the VPMM and all CV-MSEs are lower (or equal at one instance) for the VPMM compared to the non-spatial model. In general, while the VPMM often does not outperform the three competitors in terms of the mean CV-MSE, significant differences are visible for the sample version. This might be due to the fact that the within and interaction effects are marginalized in our model, such that the sample version more

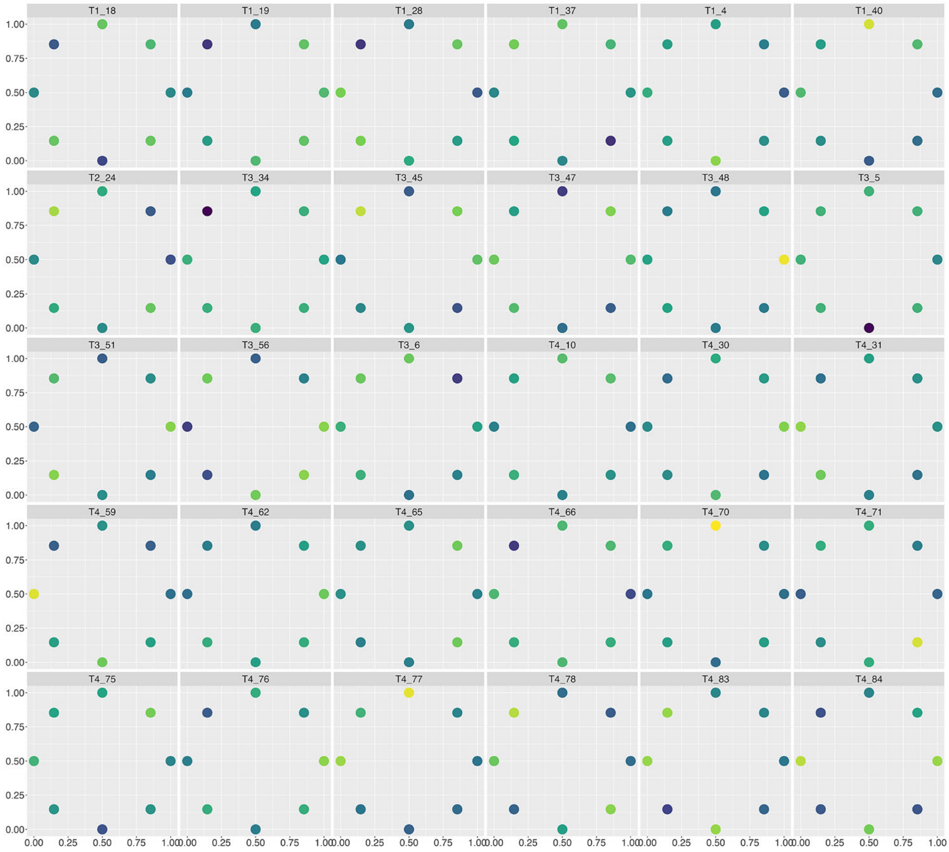

Figure 6. Data within each of the 30 plots in application. From each observation, we removed the mean of each corresponding plot .

clearly shows the differences in the two models. As speculated in Sect. 4.2, the interaction effect is particularly helpful in predictions for new plots.

7. DISCUSSION

In this paper, we develop a computationally feasible multi-level spatial model which accounts for dependence at multiple spatial scales—the VPMM. The model presented includes a data-driven approach to determine which (spatial) effects are relevant for a specific dataset. The results of the simulation study show that we can recover all true parameters of the VPMM, given a sufficiently large within-plot sample size (shown for $n \geq 25$). In the applications, we also demonstrate how the VPMM fulfills its purpose of improving interpretability of irregular spatial data, by providing separate range of parameters for different scales.

Future work should consider additional extensions to the VPMM. First and foremost, the current version of the model assumes the same set of locations within each plot. This assumption should be extended to flexibly deal with any sampling design in continuous space by, for example, using basis functions approaches within plot such as in Lindgren et al. (2011); Morris et al. (2019). Such an extension would also make predictions at new plots or different locations within each plot much more flexible. We suggest first steps in the Supplement 4.

Second, forest inventory data are often collected coarsely over time. Therefore, an extension of the VPMM toward space-time which further exploits the method in Stegle et al. (2011) could be investigated. A first tentative outline is presented in the Supplements 2 and 3. Indeed, in general, the technique used to reduce the computational complexity of the model by reformulating the normal likelihood could also be used in a space-time context.

Concerning the prior structure, it would make sense to extend the joint prior for the random effects to the fixed effects. There is, however, a need to rethink the concept of total variance since the amount of variance explained by fixed effects is determined by their coefficients, not their variances. It is worth noting that, although we present a forestry example, the resulting methods can be applied to potentially many areas of research where data of a similar structure are collected (e.g., agriculture).

ACKNOWLEDGEMENTS

The authors gratefully acknowledge the Deutsche Forschungsgemeinschaft for funding the project within the Research Training Group 2300 “Enrichment of European beech forests with conifers: impacts of functional traits on ecosystem functioning (EnriCo)”. Many thanks to Dr. Christian Ammer and Dr. Jörg Müller for proving and discussing the datasets considered. We also thank two anonymous reviewers for their valuable feedback.

Funding Open Access funding enabled and organized by Projekt DEAL.

Declarations

Conflict of interest No conflict of interests.

Code availability Example code is provided in https://github.com/isammarques/JABES-VPMM.

[Received February 2023. Revised April 2023. Accepted April 2023. Published Online May 2023.]

REFERENCES

Bässler C, Müller J, Dziock F (2010) Detection of climate-sensitive zones and identification of climate change indicators: a case study from the bavarian forest national park. Folia Geobot 45:163–182

Fahrmeir L, Kneib T, Lang S, Marx B (2013) Regression. Springer, Berlin

Finley AO, Banerjee S, Gelfand AE (2015) spBayes for large univariate and multivariate point-referenced spatio-temporal data models. J Stat Softw 63(13):1–28

Finley AO, Banerjee S, MacFarlane DW (2011) A hierarchical model for quantifying forest variables over large heterogeneous landscapes with uncertain forest areas. J Am Stat Assoc 106(493):31–48

Finley AO, Banerjee S, McRoberts RE (2009) Hierarchical spatial models for predicting tree species assemblages across large domains. Ann Appl Stat 3(3):1052

Franco-Villoria M, Ventrucci M, Rue H (2022) Variance partitioning in spatio-temporal disease mapping models. Stat Methods Med Res 31(8):1566–1578

Fuglstad G-A, Hem IG, Knight A, Rue H, Riebler A (2020) Intuitive joint priors for variance parameters. Bayesian Anal 15(4):1109–1137

Gelfand AE, Diggle P, Guttorp P, Fuentes M (2010) Handbook of spatial statistics. CRC Press

Gelman A, Rubin DB (1992) Inference from iterative simulation using multiple sequences. Stat Sci 1:457–472

Gelman A, Stern HS, Carlin JB, Dunson DB, Vehtari A, Rubin DB (2013) Bayesian data analysis. Chapman and Hall/CRC

Geyer CJ (2011) Introduction to Markov chain Monte Carlo. Handbook of Markov chain Monte Carlo. Chapman and Hall/CRC, New York, pp 3–48

Horn RA, Johnson CR (2012) Matrix analysis. Cambridge University Press, Cambridge

Ingebrigtsen R, Lindgren F, Steinsland I, Martino S (2015) Estimation of a non-stationary model for annual precipitation in southern Norway using replicates of the spatial field. Spatial Stat 14:338–364

Junttila V, Finley AO, Bradford JB, Kauranne T (2013) Strategies for minimizing sample size for use in airborne lidar-based forest inventory. For Ecol Manag 292:75–85

Katzfuss M (2017) A multi-resolution approximation for massive spatial datasets. J Am Stat Assoc 112(517):201–214

Knorr-Held L (2000) Bayesian modelling of inseparable space-time variation in disease risk. Stat Med 19(17–18):2555–2567

Krnjajić M, Draper D (2014) Bayesian model comparison: log scores and DIC. Stat Probab Lett 88:9–14

Lindgren F, Rue H, Lindström J (2011) An explicit link between Gaussian fields and Gaussian Markov random fields: the stochastic partial differential equation approach. J R Stat Soc: Ser B (Stat Methodol) 73(4):423–498

Morris SA, Reich BJ, Thibaud E (2019) Exploration and inference in spatial extremes using empirical basis functions. J Agric Biol Environ Stat 24:555–572

Neal RM (2011) MCMC using Hamiltonian dynamics. Handbook of Markov Chain Monte Carlo. Chapman and Hall/CRC, New York, pp 139–188

Nychka D, Bandyopadhyay S, Hammerling D, Lindgren F, Sain S (2015) A multiresolution Gaussian process model for the analysis of large spatial datasets. J Comput Gr Stat 24(2):579–599

Riebl, H., Wiemann, P. F., and Kneib, T. (2022). Liesel: a probabilistic programming framework for developing semi-parametric regression models and custom Bayesian inference algorithms. arXiv preprint arXiv:2209.10975

Stan Development Team (2022). Stan reference manual (version 2.3)

Stegle O, Lippert C, Mooij JM, Lawrence N, Borgwardt K (2011) Efficient inference in matrix-variate Gaussian models with iid observation noise. In: Shawe-Taylor J, Zemel R, Bartlett P, Pereira F, Weinberger K (eds) Advances in neural information processing systems 24 (NIPS 2011). Curran Associates Inc