Received: 20 August 2021 | Accepted: 10 September 2022

DOI: 10.1111/2041-210X.13998

Methods in Ecology and Evolution BRITISH ECOLOGICAL SOCIETY

RESEARCH ARTICLE

Bayesian models for spatially explicit interactions between neighbouring plants

Cristina Barber[1] | Andrii Zaiats[1] | Cara Applestein[1,2] | Lisa Rosenthal[3] | T. Trevor Caughlin[1]

[1]Biological Sciences, Boise State University, Boise, Idaho, USA

[2]Forest and Rangeland Ecosystem Science Center, U.S. Geological Survey, Boise, Idaho, USA

[3]Department of Plant Pathology, University of California, Davis, California, USA

Correspondence
Cristina Barber
Email: cristinabarberal@u.boisestate.edu

Funding information
National Science Foundation, Grant/Award Number: 1415297 and 2207158

Handling Editor: Robert Freckleton

Abstract

1. Interactions between neighbouring plants drive population and community dynamics in terrestrial ecosystems. Understanding these interactions is critical for both fundamental and applied ecology. Spatial approaches to model neighbour interactions are necessary, as interaction strength depends on the distance between neighbouring plants. Recent Bayesian advancements, including the Hamiltonian Monte Carlo algorithm, offer the flexibility and speed to fit models of spatially explicit neighbour interactions. We present a guide for parameterizing these models in the Stan programming language and demonstrate how Bayesian computation can assist ecological inference on plant–plant interactions.
2. Modelling plant neighbour interactions presents several challenges for ecological modelling. First, nonlinear models for distance decay can be prone to identifiability problems, resulting in lack of model convergence. Second, the pairwise data structure of plant–plant interaction matrices often leads to large matrices that demand high computational power. Third, hierarchical structure in plant–plant interaction data is ubiquitous, including repeated measurements within field plots, species and individuals. Hierarchical terms (e.g. 'random effects') can result in model convergence problems caused by correlations between coefficients. We explore modelling solutions for these challenges with examples representing spatial data on plant demographic rates: growth, survival and recruitment.
3. We show that ragged matrices reduce computational challenges inherent to pairwise matrices, resulting in higher efficiency across data types. We also demonstrate how metrics for model convergence, including divergent transitions and effective sample size, can help diagnose problems that result from complex nonlinear structures. Finally, we explore when to use different model structures for hierarchical terms, including centred and non-centred parameterizations. We provide reproducible examples written in Stan to enable ecologists to fit and troubleshoot a broad range of neighbourhood interaction models.

4. Spatially explicit models are increasingly central to many ecological questions. Our work illustrates how novel Bayesian tools can provide flexibility, speed and diagnostic capacity for fitting plant neighbour models to large, complex datasets. The methods we demonstrate are applicable to any dataset that includes a response variable and locations of observations, from forest inventory plots to remotely sensed imagery. Further developments in statistical models for neighbour interactions are likely to improve our understanding of plant population and community ecology across systems and scales.

KEYWORDS
big data, Hamiltonian Monte Carlo, hierarchical structures, neighbour interactions, optimization, pairwise matrix, plant–plant interactions, Stan

1 | INTRODUCTION

Interactions between neighbouring plants impact how plants grow, survive and reproduce. Although these interactions occur at the scale of individuals, their consequences shape population and community structure. Plants tend to do worse in single-species neighbourhoods than in many-species neighbourhoods (Feng et al., 2022; Sortibrán et al., 2014), an individual-level dynamic that helps explain how plant biodiversity is maintained across ecosystems, from montane deserts (Adler et al., 2010) to tropical rainforests (Comita & Stump, 2020). Plants can also facilitate the growth and survival of their neighbours, particularly in disturbed or stressful environments (Miriti et al., 2001). Managing plant neighbourhoods, from thinning dense stands of trees (Lechuga et al., 2017) to planting species that will facilitate their neighbours (Gómez-Aparicio, 2009), is a cornerstone of forestry, restoration and agriculture. The importance of neighbour interactions across basic and applied ecology underscores the need for statistical approaches that can quantify how plant neighbourhoods impact plant demography. Such analyses must account for space, as plants interact more with closer neighbours than with neighbours further away (Figure 1).

As an approximation to spatially explicit models, many studies have used plant density in a fixed radius (LaManna et al., 2017), which assumes plant neighbours at varying distances have equivalent interaction strength. Spatially explicit models enable a more realistic representation of individual plant relationships (Zambrano et al., 2020). However, fitting spatially explicit neighbourhood models requires accounting for the distance between all pairwise combinations of neighbours, which can be computationally expensive. A common simplification is to assume that the effects of distant neighbours are zero, creating an effective neighbourhood radius (Muller-Landau et al., 2004). Effective neighbourhood radii result in matrices with many zero elements, as most plant neighbours are distant enough to have negligible interactions. Matrices rich in zeros, known as sparse matrices, are found in a wide range of disciplines (Dokmanic et al., 2015). While there are existing methods to optimize computation on sparse matrices (Chalauri et al., 2018), these methods have not yet achieved wide use in ecology.

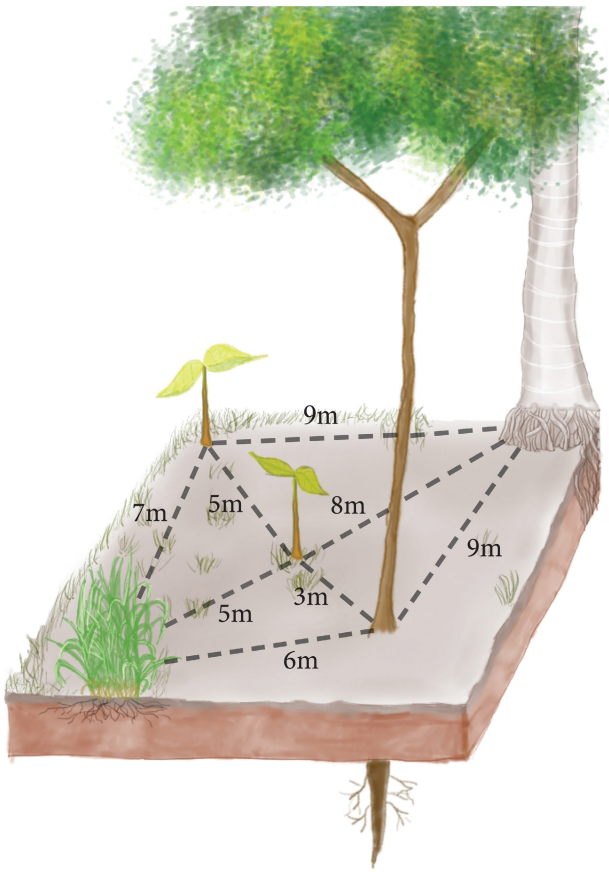

FIGURE 1 Spatial structure of a plant neighbourhood. The seedling in the centre of the plot experiences a range of neighbour interactions, depending on the neighbour's species identity, size and physical distance.

Another barrier to fitting spatially explicit neighbourhood models is that interaction strength is almost always nonlinear. There are a wide range of possible nonlinear response functions that can approximate spatial relationships between neighbouring individuals (Bolker, 2008). The downside of nonlinear models is that they are prone to identifiability problems, meaning it is difficult to define a single solution for the equation (Ogle, 2009). In

some cases, assuming linearity may be an acceptable solution to fit neighbourhood models. For example, the sum of basal area within circular plots is widely used as a neighbourhood competition index in linear models (e.g. Yang et al., 2022). Nevertheless, the capacity to fit a wide range of interaction functions, including nonlinear functions that cannot be easily transformed into linear functions, will improve links between theoretical models that rely on nonlinear functions (e.g. Bolker & Pacala, 1999) and enable better forecasting models for ecological processes (e.g. Clark et al., 1999). The challenges of fitting nonlinear models point to the need for flexible methods for model parameterization, including diagnostic metrics to assess model fit.

Hierarchical structures that violate assumptions of independence between observations can also complicate parameter estimation in statistical models. Hierarchical structures are ubiquitous in ecological data, including individuals representing different genotypes within species (Zaiats et al., 2020) or within sites (Caughlin et al., 2015; Schneider et al., 2006). Hierarchical models also present challenges for model estimation, for example, correlations between the variance and estimates of group-level ('random') effects are common. These correlations between group-level parameters mean that different parameter combinations have similar likelihood estimates, limiting the ability of the sampler to efficiently explore the probability surface. Solutions to this pathology have not yet been explored in the context of nonlinear models for neighbour interactions. Bayesian methods present a powerful tool for fitting spatially explicit plant interaction models with well-developed protocol for assessing divergences that may result from nonlinearity and hierarchical structure (Gelman et al., 2020). Nevertheless, guidance for fitting Bayesian models for large and sparse spatial datasets for neighbour interactions remains scarce.

In this paper, we provide a roadmap for how Bayesian methodology can expand opportunities to fit spatially explicit models for neighbour interactions. Our work builds off a recent advance in Bayesian inference, the Hamiltonian Monte Carlo (HMC) algorithm, which has improved sampling efficiency relative to older algorithms (Monnahan et al., 2017). The Stan software package provides an interface to HMC, including model assessment tools, with high value for fitting neighbour interaction models (Stan Development Team, 2019a). Using examples of plant demographic rates, we explore computationally efficient strategies for sparse matrices and alternative parameterizations that can help overcome computational time challenges when hierarchical structures are present in a statistical model. Our guide to fitting a range of spatially explicit neighbour interaction models will enable broader use of these powerful models in ecology.

2 | MATERIALS AND METHODS

We begin with the fundamental building block of a neighbourhood model, an interaction kernel. Following Canham and Uriarte (2006), we assume that the kernel alters the expected value (μ) of a demographic rate measured at a target plant p, with individual-specific covariates (e.g. crown area) described by the function g(p). For $j = 1, \ldots, n$ neighbouring plants (x_j), the function $f(x_j)$ describes the relationship between neighbours and the target plant.

$$\mu(p) = g(p) \sum_{j=1}^{n} f(x_j). \quad (1)$$

Equation (2) represents a simple example of an interaction kernel:

$$f(\mathbf{D}_{i,j}) = \sum_{k=1} \frac{1}{a_1 \mathbf{D}_{i,k}}, \quad (2)$$

where $\mathbf{D}_i$ is a pairwise matrix that contains the distance between plant i and the plants within its effective neighbourhood radius and a_1 is a parameter representing the strength of distance decay as the distance between neighbouring plants increases.

2.1 | Optimization of sparse matrices using ragged matrices in a neighbour interaction model

A common simplifying assumption for pairwise matrices (e.g. $\text{distance}_{i,j}$ in Equation (2)) is that neighbours beyond an effective neighbourhood radius do not interact. We set values beyond this radius to zero, thus transforming the pairwise matrix into a sparse matrix. In the example below, we explore how different sizes of the effective neighbourhood radius alter statistical results. Sparse matrices can be simplified further by representing them as ragged matrices (Chalauri et al., 2018). Ragged matrices allow different numbers of elements in each row, which reduces computer processing time but limits the use of linear algebra operations, such as matrix multiplication. By representing the position of non-zero elements in the original matrix with index vectors, which contain the row and column number of each element, the ragged matrix efficiently preserves information on matrix structure. In Stan, the built-in function 'segment()' creates a ragged matrix, representing elements in a vector and their position in the pairwise matrix using two index vectors (Figure 2; Stan Development team, 2022).

2.1.1 | Example 1: Plant growth

To demonstrate how ragged matrices can improve computation time for neighbour models, we simulated a spatially explicit dataset representing plant growth. We model plant growth as a function of intrinsic growth (i.e. growth in isolation) and neighbourhood characteristics (i.e. neighbour size and proximity; Equation (3)). To evaluate how choosing an effective neighbourhood radius could introduce bias in parameter estimation, we fit six models with effective neighbourhood radii of 5, 10, 15 and 20 m. The 'true' effective radius of this simulated data is 10 m. With real data, the decision for radius size should be based on biological knowledge, for example, root zone area (Zaiats et al., 2020).

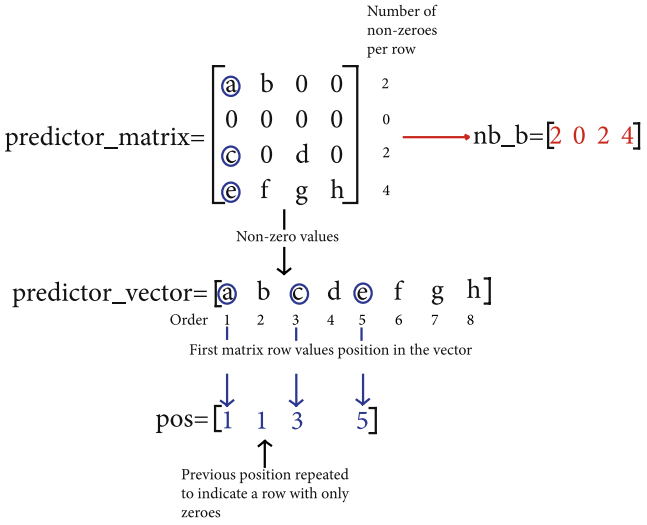

FIGURE 2 A demonstration of how the segment function creates a ragged matrix. The predictor_matrix above includes matrix elements labelled a:H, as well as zero elements. The nb_b vector contains the number of non-zero elements per row, the predictor_vector vector contains the non-zero elements and the pos_vector contains the position of non-zero elements.

Additional spatial information, such as crown allometry, can also enable the construction of biologically meaningful plant neighbourhoods without the need for an arbitrary decision on the effective neighbourhood radius (Zambrano et al., 2019). Alternately, the effective neighbourhood radius can be chosen by testing predictive performance of different sized radii (Zambrano et al., 2020). Beyond predictive performance of models with different effective neighbourhood radii, inference on effects of neighbourhood radii on ecological processes can be highly sensitive to the choice of neighbourhood radii (see Zambrano et al., 2020 for a functional traits example).

Equation (3) represents our generative model for neighbour-dependent growth:

$$\begin{aligned} \boldsymbol{S}_{t+1,i} &\sim \text{normal}\left(\mu_i, \sigma\right) \\ \mu_i &= \alpha + \beta \boldsymbol{S}_{t,i} + a_3 \sum_{k=1}^{n} \mathbf{SN}_{i,k}^{a_1} \bullet \frac{1}{\exp\left(\mathbf{D}_{i,k}^2 a_2\right)} \end{aligned} \quad . \tag{3}$$

In Equation 4, $\boldsymbol{S}_{t,i}$ and $\boldsymbol{S}_{t+1i}$ are the sizes of plant i at time $t = 0$ and 1, respectively, $\mathbf{D}_i$ are the distance between i and the plants within its effective neighbourhood radius, and $\mathbf{SN}_i$ represents the size at time $t = 0$ of the plants located within the effective neighbourhood radius of plant i. For demographic rates that involve two censuses, such as our growth example, we assume that the size of the neighbour at the first census determines the demographic response measured as the difference between the first and second censuses. α is the intercept, and β is the effect of $\boldsymbol{S}_{t_0,i}$. In the interaction kernel, parameter a_1 mediates the effect of neighbouring plant size, parameter a_2 determines the scale of distance decay, and a_3 represents the overall effect of the neighbourhood term.

We fit models in Stan using either a sparse or ragged matrix. These approaches share a large amount of code, with two main differences. For the sparse matrix code, the data block in Stan includes two sparse matrices containing the distances and sizes of the plants within the effective neighbourhood radius. A nested for-loop that iterates through every neighbouring individual then defines the interaction kernel (Stan Development Team, 2019a; for full code check the data availability statement): In contrast, the ragged matrix code includes the three vectors required to use the `segment()` function: neighbour size, distance and an index vector for non-zero entries of the matrix (Figure 2). By referencing only non-zero elements using index vectors, the ragged matrix approach reduces computationally expensive iterations of the nested for loop over zero-valued entries. In this example, we went from iterating through a matrix containing 250,000 elements to iterating through a ragged matrix containing 69,448 elements.

2.1.2 | Example 2: Plant recruitment

To further explore the application of the ragged matrix as an optimization strategy in neighbour interaction models, we parametrized a model using real data on seedling abundance of invasive strangler fig trees, *Ficus macrocarpa*, in Florida, USA. We analysed data from Caughlin et al. (2012), which includes the total number of strangler fig seedlings in 52 plots of 30 m at a single time point. Distances to adult fig trees within an effective neighbourhood radius of 300 m were recorded. We modelled seedling abundance for all plots i using a negative binomial distribution, with a mean (μ) and an over dispersion parameter (φ; Equation 4):

$$\text{abundance}_i \sim \text{negative binomial}\left(\mu_i, \varphi\right)$$

$$\mu_i = \begin{cases} a + b \sum_{k=1}^{n} \frac{1}{c + \mathbf{D}_{i,k}} \text{CP} & \text{if} \quad n > 0 \\ a & \text{if} \quad n = 0 \end{cases} , \tag{4}$$

where a is the global intercept and b describes the strength of the interactions kernel, which decays as a function of c and the distance from plot i to the adult fig trees within its effective neighbourhood radius for n total adult trees per plot. Similar to other strangler figs, *F. microcarpa* begins its life cycle by germinating in the canopy of a host tree. The number of potential host trees in the 30 m plots, CP, is multiplied by the kernel as an offset, assuming that more host trees create more opportunities for fig tree seedlings to recruit.

The original study exponentiated a, b and c to keep the parameters positive. To replicate the previous results, fit with maximum likelihood estimation in Caughlin et al. (2012), we ensured non-negative values for the mean of the negative binomial distribution by constraining parameters a, b and j to positive values. However, we note that the log-link is the canonical link-function for the negative binomial distribution and is a better choice for future studies (for full code check the data availability statement). In this example, there are seedling plots that do not have any adult strangler fig trees

nearby, resulting in zeroes in the n_nb vector (second row of predictor matrix in Figure 2). In some interaction kernels, these zeroes cause the denominator to become zero and hence undefined. A solution is to use an ifelse statement, so if nb_b is zero then μ is equal to the equation without the interaction kernel ($\mu_i = a$), and if nb_b is not zero then μ is as stated in Equation 4 (for full code, check the data availability statement). The consequence of the ifelse statement is that trees with no neighbours within the radius do not provide information on interactions with other trees but can still inform other parameters in the model.

2.2 | Centred and non-centred parametrization for random effects in neighbour interaction models

2.2.1 | Example 3: Seedling germination

To demonstrate how hierarchical models for plant neighbour interactions can be fit in a Bayesian context, we analysed a dataset on seedling germination that includes multiple individuals nested within field plots. These data represent the outcome of a seed addition experiment, an experimental design commonly used to study density dependence during early plant life stages (Clark et al., 2007). The objective of this study was to quantify how the density of seedling and adult tree neighbours impacted the probability of seed germination (Caughlin et al., 2015). In this study, the effective neighbourhood radius was set at 10 m. Germination success of all seeds in plot k was estimated by modelling the number of germination events, given the total number of added seeds (n), and probability of germination, (p). This study represents data at two time points, the initial number of seeds added and the proportion of those seeds that germinated after several months. We model the germination probability of seeds using the binomial distribution (Equation (5)):

$$\begin{aligned} \text{germination} &\sim \text{binomial}(\text{n}, \text{p}) \\ \text{logit}(p_i) &= \mu + b\,\text{Seedling}\,s_i + a\sum_{k=1}^{n} \boldsymbol{SN}_{i,k} \bullet \frac{1}{\boldsymbol{D}_{i,k}^{g}} + \omega_{k[i]} \\ \omega_k &\sim \text{normal}(0, \sigma) \\ a &\sim \text{normal}(0, 1) \\ \sigma &\sim \text{normal}(0, 1) \end{aligned}, \quad (5)$$

where the input to the binomial distribution includes the total number of seeds added to each plot (n) and the probability of successful germination events (p). μ is the global intercept, b is the effect of seedling density, Seedlings_i is the number of conspecific seedlings to represent the crowding effect, a is the total effect of neighbouring adult trees size and distance on recruitment, g is the distance decay of the effect of neighbour size and distance, and ω is the random effect of plot k, to account for non-independence between seeds in the same plot. **SN** is a matrix containing the size of the adult trees within the effective neighbourhood radius of germinating seed i, and **D** is the pairwise matrix containing the distance between the adults within the effective neighbourhood radius of germinating seed i.

Inclusion of group-level effects, such as the plot-level intercept ω in Equation 6, often leads to correlations between the variance (σ) and estimates of random effects (ω_k). These correlations can limit the ability of samplers to explore probability surfaces thoroughly, resulting in poor model convergence (Neal, 2011). One solution is to reparametrize the model, creating a linear model structure to decouple variance from random effect estimates (McElreath, 2020). This solution is often referred to as the 'non-centred parameterization', in contrast to the 'centred parameterization' in which the levels of the random effects have a common prior, in this case with mean 0 and standard deviation σ.

To create the non-centred parameterization, we re-write the random effects as a deterministic sum of the mean and scaled group variances, $\omega_{k[i]} = c + \sigma z_{k[i]}$ and sample z from a unit normal prior (McElreath, 2020). This new parameterization causes z to be orthogonal to the variance, reducing correlation between coefficients.

2.3 | Model performance across simulated datasets with varying sample size

As a final demonstration of the utility of our ragged matrix approach, we simulated a large dataset, including nonlinear interactions, a large number of neighbours and hierarchical effects. Simulated sample sizes are derived from one of the world's most extensive tree demographic datasets, the 50 ha plot from Barro Colorado Island (BCI; Davies et al., 2021). As the large, long-term forest dynamics plot design has become more common worldwide (Davies et al., 2021), the need for scalable methods for spatially explicit analysis has also grown. We evaluated the ragged matrix approach using a range of realistic sample sizes of individual trees, from 466 to 235,338 (the yearly mean number of live trees >1 cm Diameter at Breast Height in the BCI plot). Given this range of sample sizes, we simulated survival, growth and recruitment data as a function of tree neighbours, using the following interaction kernel:

$$a\sum_{k=1}^{n-1} \mathbf{SN}_{i,k} \bullet \frac{1}{e^{\frac{1}{\rho^2}\mathbf{D}_{i,k}^2}}. \quad (6)$$

In Equation (6) above, $\mathbf{D}_i$ is the distance between the target plant and the plants within its effective neighbourhood radius, and $\mathbf{SN}_i$ represents the size of the plants within plant i effective neighbourhood radius at the beginning of the census interval. The parameter ρ represents distance decay of neighbourhood interaction, and a represents the overall strength of neighbourhood effects on demography. We assumed a neighbourhood interaction radius of 50 m.

To simulate hierarchical structure in demographic rates, we subdivided our simulated 50 ha plot into ten 5-ha subplots and modelled subplot identity as a normally distributed group-level effect.

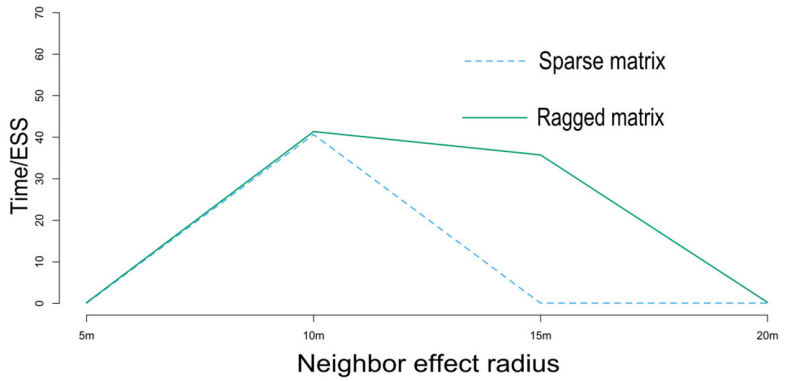

FIGURE 3 For the models fit using the simulated growth data, the ragged matrix is more efficient than the sparse matrix for models of all neighbour effects, especially for the 15 m. the dashed blue line shows the change in efficiency using the sparse matrix and the solid green line shows the change in efficiency using the ragged matrix. Greater values of ESS/time represent increased efficiency and lower values represent decreased efficiency.

An interpretation of this group-level effect is that trees within the same subplot tend to experience similar environmental conditions. However, our model code is easily adaptable to other group-level effects, such as species identity or health status.

To simulate tree growth and survival data, we generated data from a normal distribution with an identity link (growth) and a binomial distribution with a logit link (survival), incorporating tree size at the initial census, the subplot group-level effect, and the neighbourhood effect as additive terms within the link function. To simulate recruitment, we generated 100 m transects in the centre of each plot, divided into twenty 5 × 5 quadrats, resulting in 200 seedling monitoring plots. We simulated recruit counts using a negative binomial distribution with a log-link. In contrast to the growth and survival simulations, the number of plots remained constant throughout simulation runs however, the number of potential neighbouring trees ranged from 466 to 235,338.

For all three demographic rates, we initialized simulations by randomly locating neighbouring trees across the plot. After simulating data, we fit statistical models using either the ragged or sparse matrix approach. We then quantified how run time changed as the number of trees increased as well as the ability of statistical models to recapture the 'true' parameter values from the simulation (for full code, check the data availability statement).

2.4 | Models assessment

We estimated the model fitting efficiency by dividing the sum of the effective sample size (ESS) all the chains by the elapsed time to run 1000 iterations excluding the warmup time. ESS is an estimate of how much the autocorrelation within the chains increases uncertainty in estimates. Higher ESS indicates lower autocorrelation (Stan Development Team, 2019b). We also checked whether estimated parameter intervals recover the parameters used to generate the data. Lastly, we checked common diagnostic metrics to evaluate convergence, including $\hat{R}$, ESS, divergences and Bayesian fraction of missing information. We considered convergence when the $\hat{R}$ was lower than 1.01, all the chains mixed without any divergences and the ESS was over 10% (Gelman et al., 2020).

To compare the centred and non-centred parametrization, in addition to convergence metrics above, we graphically explored how well the model sampled the correlated area between the variance and group-level effects. We also assessed goodness-of-fit by calculating the mean absolute error (MAE) between the model predictions and observed data.

3 | RESULTS

3.1 | Comparison between sparse matrix and ragged matrix performance

3.1.1 | Example 1: Plant growth

For the models assessing simulated growth data, the ragged matrix was more efficient than the sparse for all effective neighbourhood radii. Ragged matrices enable faster exploration of models with different neighbour effect radii (Figure 3). When the effective neighbourhood radius was 10 m, both matrix types were able to recover the true parameters. For all other radii, the ragged matrix provided consistently tighter credibility intervals than the sparse matrix (Figure 4). There were no divergent transitions using either matrix approach for radii 10 m and 15 m. For the sparse matrix at 10 m radius and for the ragged matrix at 10 and 15 m radii, the $\hat{R}$ for all the parameters was lower than 1.01 and the ESS was over 10%, indicating convergence (Supplemental Tables S1–S3, and Supplemental Figure S1). Models fit with other radii showed divergence transitions, low ESS and high $\hat{R}$, indicating poor convergence for the models using both the sparse and the ragged matrices (Supplemental Tables S4–S8, and Supplemental Figure S2).

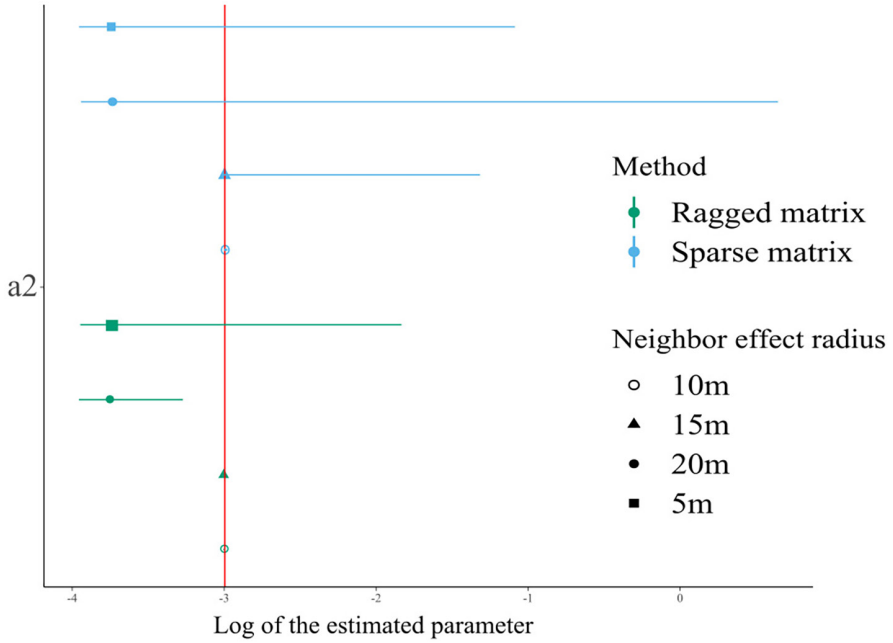

FIGURE 4 This figure shows parameter estimates for a2 (Equation (3)), which estimates distance decay in the interaction strength between growth and distance. The green shapes are the ragged matrix parameter estimates. The blue shapes are the sparse matrix parameter estimates. Each of the shapes corresponds to a different effective neighbourhood radius. The red line is the true parameter used in the simulation, and horizontal lines represent 95% credibility intervals (CI). Note that the CIs for the 10, and 15 m radii are not visible.

3.1.2 | Example 2: Seedling abundance

For the strangler fig case study, model efficiency was greater when using the ragged matrix. The model fit with the ragged matrix was 3.6 times more efficient than the model fit with the sparse matrix, with an efficiency of 3.32e07 ESS/Time(s) for the sparse matrix relative to an efficiency of 1.19e08 ESS/Time(s)for the ragged matrix. However, parameter estimates were the same for both model parameterizations (Supplemental Figure S3 and Figure 5), and produced estimates similar to the frequentist maximum likelihood estimation presented in the original analysis (Figure 5). Both Bayesian models converged well (Supplemental Tables S9 and S10).

3.2 | Comparison between centred and non-centred parametrization performance

3.2.1 | Example 3: Seed germination with plot-level random effects

For the germination case study, the centred parametrization had an efficiency of 1.32e07 ESS/time, while the non-centred parametrization sampled nearly 1,12 times more efficiently, with 1.48e07 ESS/time (Supplemental Figure S4). Both parametrizations converged well, with no divergences, ESS over 10%, and $\hat{R}$ lower than 1.01. This model had to run for 30,000 iterations to converge for both parametrizations, a number comparatively higher than the rest of the other models (Supplemental material, Appendices I–XII). The probability surface of the centred and non-centred parametrization shows that both parametrizations explored the probability surface and that there was no funnel shape (Supplemental material Figure S5).

The parameter estimates were similar for the centred and non-centred parametrizations (Supplemental Figure S6). For both parameterizations, the models slightly underestimated germination (Supplemental Figure S7). Overall error was comparable between the two parameterizations, with MAE = 1.196 (95% CI: 0.002–4.453) for the centred parametrization and MAE = 1.194 (95% CI: 0.000–4.376) for the non-centred parametrization.

3.3 | Model performance across simulated datasets with varying sample size

For all simulated datasets, the ragged matrix approach was more efficient than the sparse matrix approach. The difference in run time between the two approaches varied as sample size increased, from an initial difference of 0 ESS/Time(s), 9,22e5 ESS/Time(s)(survival), and 4.64e6 ESS/Time(s) (recruitment) at a sample size of 466 neighbouring trees to a maximum difference of 4.61e4 ESS/Time(s) for a sample size of 189,560 trees (recruitment). Despite these differences, parameter estimates were indistinguishable between the two model fitting approaches (Supplemental Figures S8, S10, and S12). Nevertheless, for sample sizes of >200,000 trees, computational demands rendered the sparse matrix approach infeasible for spatially explicit models with hierarchical structure, while the ragged matrix approach provides a scalable method even for large datasets (Figure 6 and Supplemental Figures S9 and S11).

4 | DISCUSSION

We have demonstrated how to leverage contemporary Bayesian methods to estimate spatially explicit plant neighbour interactions. The pairwise data structure of matrices representing neighbour interactions often leads to large datasets that present computational challenges. Our work shows that ragged matrices greatly increase

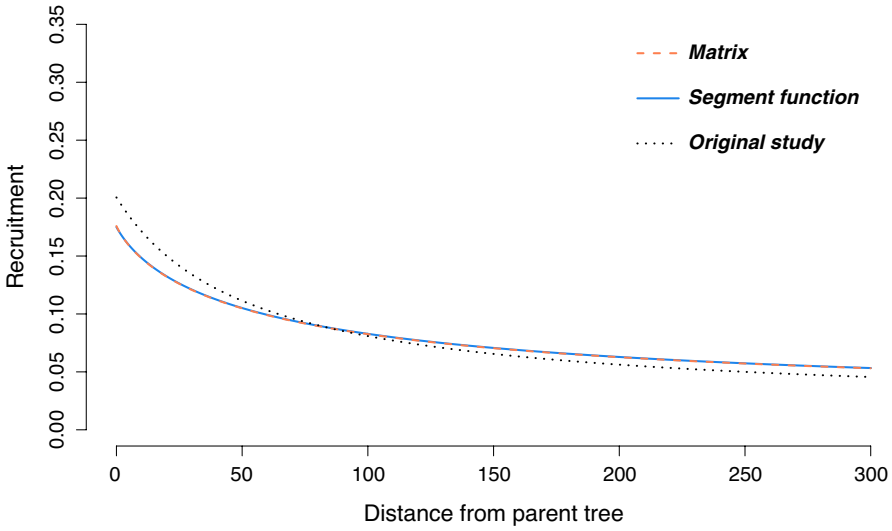

FIGURE 5 The ragged matrix and the sparse matrix approaches obtained similar estimates of the relationship between recruitment and the distance from a single parent tree. Curves show the relationship between recruitment and distance from parent tree parametrized using the sparse matrix, the segment function, and a frequentist maximum likelihood model. Shaded areas represent 95% credibility intervals (CI). The sparse matrix CI is the shaded orange and the ragged matrix is shaded blue.

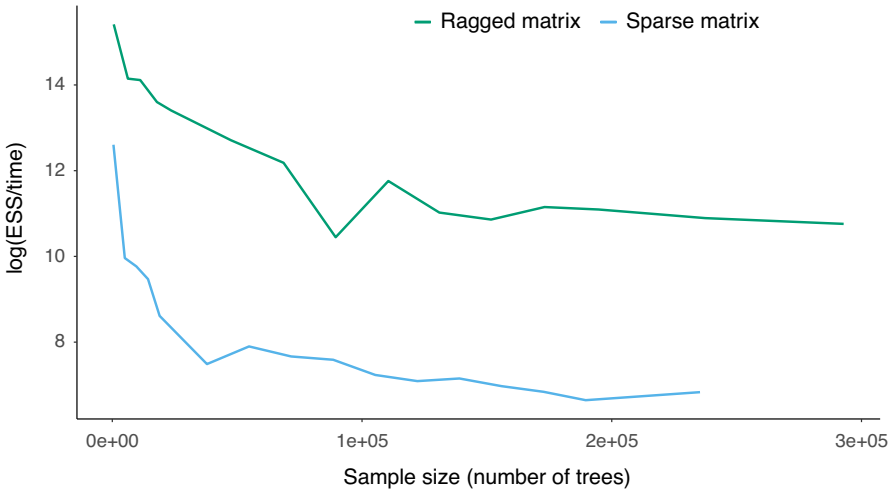

FIGURE 6 Effective sample size divided by time in the log scale of models with simulated data on tree recruitment across a range of sample sizes.

computational efficiency, relative to full, yet sparse pairwise matrices without changing the posterior estimates. We have also demonstrated how Bayesian models can include hierarchical structures in models of neighbour interactions, such as pseudoreplication between individuals of the same species or in the same plot (Schneider et al., 2006). Correlations between random effect parameters are inherent to many hierarchical models, and our work illustrates how HMC provides the means to efficiently parametrize complex statistical models, including diagnostic tools sensitive to detect sampling pathologies. As big data become more common in ecology, computational limits are expected to become an increasing bottleneck for analyses (Farley et al., 2018). We have demonstrated an algorithmic solution, ragged matrices, that increases the efficiency of spatially explicit analyses. Altogether, we expect that fitting neighbour models with contemporary Bayesian software packages, such as the Stan programming language, will open up new opportunities for ecological inference involving large, spatial datasets.

Spatially explicit neighbour matrices are frequently simplified using an effective neighbour radius that sets effects of neighbours beyond the radius to zero (Muller-Landau et al., 2004). This simplifying assumption creates a sparse matrix structure, with many zeros for non-interacting plant neighbours, that can be computationally inefficient. Models using ragged matrices were more computationally efficient relative to those using entire sparse matrices for a range of neighbourhood effect radii Built-in functions in the Stan programming language enable sparse representations of a matrix that improve storage efficiency but are limited in improving the sampling speed (Stan Development Team, 2019b). Our results show that ragged matrices can significantly improve computational speed in addition to storage requirements. Beyond plant neighbourhood analyses, ragged matrices present a solution for big data that can be generally applied to spatial ecological questions, ranging from landscape graph-theoretic connectivity (Urban & Keitt, 2001) to pairwise relatedness analysis between individuals (Hardy, 2003).

Parameter estimation depended on the size of effective neighbourhood radius for the ragged and the sparse matrices. A model fit to simulated data revealed that (1) the most accurate parameter estimates corresponded to the 'true' effective neighbourhood radius, (2) accuracy decreased minimally for slightly bigger radii than the 'true' radius and (3) accuracy decreased more for radii smaller than the

'true' radius. This result is similar to previous frequentist models, in which larger radii provided estimates with lower biases than smaller radii (Canham & Uriarte, 2006). However, we found that neighbour effect radii that were much larger than the true radius resulted in poor model convergence, which was straightforward to identify using Stan's built-in diagnostic metrics. An alternate approach could include estimating the neighbour effect radius as a parameter in the model. Such an approach would enable propagation of uncertainty from leaving some plants left out into model output (Uriarte et al., 2004).

4.1 | Hierarchical modelling

We assessed the centred and non-centred parametrization of random effects by checking model convergence and uncertainty (Gelman et al., 2020). Our results suggest that the advantages of one parametrization over another are highly case specific and depend on the properties of the dataset. Although the centred parametrization converged and the metrics did not show any sampling problem that indicated the correlation problems, we observed lower efficiency exploring the probability surface. In models that present stronger correlation problems in the hierarchical structures, we would expect less reliable parameter estimates and convergence problems. The diagnostic metrics provided by Stan can help to decide the appropriate parametrization, and we would recommend comparing ESS/time for both parametrizations to decide on the appropriate parametrization. The diagnostic metrics also allowed us to decide for how long to run the model to obtain reliable estimates. An interesting question for future research will be to explore how the choice of effective neighbourhood radius (e.g. Zambrano et al., 2020) potentially impacts the performance of different parametrizations for hierarchical models.

Further research across a range of data structure and study systems will be necessary to develop concrete recommendations for when the non-centred parameterization should be used (Gorinova et al., 2020). As the range of potential hierarchical data structure for neighbour interactions increases, including temporal (Valenta et al., 2020), spatial (Pu et al., 2020) and phylogenetic autocorrelation (Zaiats et al., 2020; Zambrano et al., 2017), developing efficient ways to fit these models should be a research priority. Automatic parametrization algorithms that build efficient sampling schemes from the data are a promising research avenue that could be used to parametrize neighbour interaction models (Gorinova et al., 2020).

4.2 | Research perspectives

An ever-growing body of literature seeks to understand population, community and ecosystem dynamics through individual-based models (Deangelis et al., 2020; Hardy, 2003; Romero-Mujalli et al., 2019; Seidl et al., 2012). Statistical models that incorporate spatial information are critical for developing individual-based models (Canham & Uriarte, 2006; Zhang & DeAngelis, 2020). Fortunately, the number of datasets that include data on plant locations is growing. Any dataset with location coordinates of plant individuals has potential to benefit from neighbourhood interaction models, and many are publicly available. As data sharing becomes the cultural norm, an increasing number of existing experimental and observational datasets could be used to fit neighbour interaction models (Soranno et al., 2015). Some examples include common garden experiments (Madsen et al., 2020; Zaiats et al., 2020) and forest inventories on permanent plots (Gillerot et al., 2021; Lieberman & Lieberman, 2007). Our case studies represent a limited time frame, with measurements at one (seedling abundance) or two time points (growth and seedling germination). Understanding the demographic impacts of plant–plant interactions and resultant consequences for population and community dynamics will require measurements over longer time periods (Butterfield et al., 2010; Caughlin et al., 2015; Miriti et al., 2001). As time-series data on plant–plant interactions continue to increase (Davies et al., 2021), we anticipate that sparse matrices will play an even more important role in computationally efficient analyses of these growing data.

The increasing volume of remote sensing data at the resolution of individual plant canopies also represents novel opportunities to fit neighbour interaction models. Individual plant canopies may be identified using remote sensing data from aerial lidar, unoccupied aerial systems, and high-resolution satellite imagery (Caughlin et al., 2016; Shen et al., 2020). High-resolution remotely sensed data offer opportunities to parameterize individual-based models for vegetation at unprecedented scales. However, we expect that increased uncertainty in identifying individual plants from air or space may require statistical models that can disentangle measurement from process error (Brack et al., 2018).

We have demonstrated how contemporary Bayesian algorithms, such as HMC sampling implemented in Stan, provide a flexible and efficient way to fit plant neighbourhood models. The flexibility of the Stan programming language provides new opportunities to apply Bayesian methods to large datasets, including optimization of sparse matrices. In addition, uncertainty and model assessment metrics provided in the Bayesian framework allow a more intuitive implementation of hierarchical structures (e.g. random effects; Monnahan et al., 2017, Ogle & Barber, 2020) in nonlinear models with non-normal error structures. We hope that these guidelines, together with new ongoing improvements in model parametrizations and the increasing availability of spatially explicit data, will help to advance the study of population, community and ecosystem dynamics.

AUTHOR CONTRIBUTIONS

None of the authors have any conflict of interest. Cara Applestein, Trevor Caughlin, Andrii Zaiats and Cristina Barber conceptualized the idea. Andrii Zaiats parametrized the growth simulations, Lisa Rosenthal parametrized the recruitment simulation, and Cristina Barber translated the code from Caughlin et al. (2012) and Caughlin et al. (2015) in Stan, parametrized the mortality simulation and performed the analysis. Trevor Caughlin and Cristina Barber led the writing of the manuscript. Cara Applestein, Andrii Zaiats, Lisa

Rosenthal reviewed and contributed to the writing of the manuscript. All authors contributed critically to the drafts and gave final approval for publication.

ACKNOWLEDGEMENTS

Support for was provided by the National Science Foundation under grant #1415297 in the SBE program and grant #2207158 in DEB. We thank Jonah Gabry for his helpful comments and the participants in the StanCon2020 for the engaging discussions.

CONFLICT OF INTEREST

None of the authors have any conflict of interest.

PEER REVIEW

The peer review history for this article is available at https://publons.com/publon/10.1111/2041-210X.13998.

DATA AVAILABILITY STATEMENT

The code to simulate the growth data can be found at https://mc-stan.org/users/documentation/case-studies/plantInteractions.html . The data for fig tree recruitment can be found in: Caughlin et al. (2012), https://doi.org/10.1890/11-1694.1. The data for the seedlings recruitment can be found in: Caughlin et al. (2015), https://doi.org/10.1098/rspb.2014.2095. The code can be found in: https://github.com/Cristinabarber/Spatial-plant-interactions/releases/tag/V1.0.0, https://doi.org/10.5281/zenodo.7093676.

ORCID

Cristina Barber https://orcid.org/0000-0002-3077-3130

T. Trevor Caughlin https://orcid.org/0000-0001-6752-2055

REFERENCES

Adler, P. B., Ellner, S. P., & Levine, J. M. (2010). Coexistence of perennial plants: An embarrassment of niches. *Ecology Letters*, *13*, 1019–1029.

Bolker, B. (2008). *Ecological models and data in R*. Princeton University Press.

Bolker, B. M., & Pacala, S. W. (1999). Spatial moment equations for plant competition: Understanding spatial strategies and the advantages of short dispersal. *The American Naturalist*, *153*, 575–602.

Brack, I. V., Kindel, A., & Oliveira, L. F. B. (2018). Detection errors in wildlife abundance estimates from unmanned aerial systems (UAS) surveys: Synthesis, solutions, and challenges. *Methods in Ecology and Evolution*, *9*, 1864–1873.

Butterfield, B. J., Betancourt, J. L., Turner, R. M., & Briggs, J. M. (2010). Facilitation drives 65 years of vegetation change in the Sonoran Desert. *Ecology*, *91*, 1132–1139.

Canham, C. D., & Uriarte, M. (2006). Analysis of neighborhood dynamics of Forest ecosystems using likelihood methods and modeling. *Ecological Applications*, *16*, 62–73.

Caughlin, T. T., Ferguson, J. M., Lichstein, J. W., Zuidema, P. A., Bunyavejchewin, S., & Levey, D. J. (2015). Loss of animal seed dispersal increases extinction risk in a tropical tree species due to pervasive negative density dependence across life stages. *Proceedings of the Royal Society B: Biological Sciences*, *282*, 20142095.

Caughlin, T. T., Graves, S. J., Asner, G. P., van Breugel, M., Hall, J. S., Martin, R. E., Ashton, M. S., & Bohlman, S. A. (2016). A hyperspectral image can predict tropical tree growth rates in single-species stands. *Ecological Applications*, *26*, 2367–2373.

Caughlin, T. T., Wheeler, J. H., Jankowski, J., & Lichstein, J. W. (2012). Urbanized landscapes favored by fig-eating birds increase invasive but not native juvenile strangler fig abundance. *Ecology*, *93*, 1571–1580.

Chalauri, G., Laluashvili, V., & Gelashvili, K. (2018). Jagged non-zero submatrix data structure. *Transactions of A. Razmadze Mathematical Institute*, *172*, 7–14.

Clark, C. J., Poulsen, J. R., Levey, D. J., & Osenberg, C. W. (2007). Are plant populations seed limited? A critique and meta-analysis of seed addition experiments. *The American Naturalist*, *170*, 128–142.

Clark, J. S., Silman, M., Kern, R., Macklin, E., & HilleRisLambers, J. (1999). Seed dispersal near and far: Patterns across temperate and tropical forests. *Ecology*, *80*, 1475–1494.

Comita, L. S., & Stump, S. M. (2020). Natural enemies and the maintenance of tropical tree diversity: Recent insights and implications for the future of biodiversity in a changing World1. *Annals of the Missouri Botanical Garden*, *105*, 377–392.

Davies, S. J., Abiem, I., Abu Salim, K., Aguilar, S., Allen, D., Alonso, A., Anderson-Teixeira, K., Andrade, A., Arellano, G., Ashton, P. S., Baker, P. J., Baker, M. E., Baltzer, J. L., Basset, Y., Bissiengou, P., Bohlman, S., Bourg, N. A., Brockelman, W. Y., Bunyavejchewin, S., … Zuleta, D. (2021). ForestGEO: Understanding forest diversity and dynamics through a global observatory network. *Biological Conservation*, *253*, 108907.

Deangelis, D., Gross, L., Wolff, W., Fleming, D., Nott, M., & Comiskey, E. (2020). *Individual-based models on the landscape: applications to the everglades* (pp. 199–211). CRC Press.

Dokmanic, I., Parhizkar, R., Ranieri, J., & Vetterli, M. (2015). Euclidean distance matrices: Essential theory, algorithms and applications. *IEEE Signal Processing Magazine*, *32*, 12–30.

Farley, S. S., Dawson, A., Goring, S. J., & Williams, J. W. (2018). Situating ecology as a big-data science: Current advances, challenges, and solutions. *Bioscience*, *68*, 563–576.

Feng, Y., Schmid, B., Loreau, M., Forrester, D. I., Fei, S., Zhu, J., Tang, Z., Zhu, J., Hong, P., Ji, C., Shi, Y., Su, H., Xiong, X., Xiao, J., Wang, S., & Fang, J. (2022). Multispecies forest plantations outyield monocultures across a broad range of conditions. *Science*, *376*, 865–868.

Gelman, A., Vehtari, A., Simpson, D., Margossian, C. C., Carpenter, B., Yao, Y., Kennedy, L., Gabry, J., Bürkner, P.-C., & Modrák, M. (2020). Bayesian Workflow. arXiv:2011.01808.

Gillerot, L., Forrester, D. I., Bottero, A., Rigling, A., & Lévesque, M. (2021). Tree Neighbourhood diversity has negligible effects on drought resilience of European beech, silver fir and Norway spruce. *Ecosystems*, *24*, 20–36.

Gómez-Aparicio, L. (2009). The role of plant interactions in the restoration of degraded ecosystems: A meta-analysis across life-forms and ecosystems. *Journal of Ecology*, *97*, 1202–1214.

Gorinova, M. I., Moore, D., & Hoffman, M. D. (2020). Automatic reparameterisation of probabilistic programs:10.

Hardy, O. J. (2003). Estimation of pairwise relatedness between individuals and characterization of isolation-by-distance processes using dominant genetic markers. *Molecular Ecology*, *12*, 1577–1588.

LaManna, J. A., Mangan, S. A., Alonso, A., Bourg, N. A., Brockelman, W. Y., Bunyavejchewin, S., Chang, L.-W., Chiang, J.-M., Chuyong, G. B., Clay, K., Condit, R., Cordell, S., Davies, S. J., Furniss, T. J., Giardina, C. P., Gunatilleke, I. A. U. N., Gunatilleke, C. V. S., He, F., Howe, R. W., … Myers, J. A. (2017). Plant diversity increases with the strength of negative density dependence at the global scale. *Science*, *356*, 1389–1392.

Lechuga, V., Carraro, V., Viñegla, B., Carreira, J. A., & Linares, J. C. (2017). Managing drought-sensitive forests under global change. Low competition enhances long-term growth and water uptake in Abies pinsapo. *Forest Ecology and Management*, *406*, 72–82.

Lieberman, M., & Lieberman, D. (2007). Nearest-neighbor tree species combinations in tropical forest: The role of chance, and some consequences of high diversity. *Oikos*, *116*, 377–386.

Madsen, C., Potvin, C., Hall, J., Sinacore, K., Turner, B. L., & Schnabel, F. (2020). Coarse root architecture: Neighbourhood and abiotic environmental effects on five tropical tree species growing in mixtures and monocultures. *Forest Ecology and Management*, *460*, 117851.

McElreath, R. (2020). *Statistical rethinking: A Bayesian course with examples in R and Stan*. CRC Press.

Miriti, M. N., Joseph Wright, S., & Howe, H. F. (2001). The effects of neighbors on the demography of a Dominant Desert shrub (ambrosia Dumosa). *Ecological Monographs*, *71*, 491–509.

Monnahan, C. C., Thorson, J. T., & Branch, T. A. (2017). Faster estimation of Bayesian models in ecology using Hamiltonian Monte Carlo. *Methods in Ecology and Evolution*, *8*, 339–348.

Muller-Landau, H. C., Dalling, J. W., Harms, K. E., Wright, S. J., Condit, R., Hubbell, S. P., & Foster, R. B. (2004). Seed dispersal and density-dependent seed and seedling survival in Trichilia tuberculata and Miconia argentea:42.

Neal, R. (2011). MCMC using Hamiltonian dynamics. In S. Brooks, A. Gelman, G. Jones, & X.-L. Meng (Eds.), *Handbook of Markov chain Monte Carlo*. Chapman and Hall/CRC.

Ogle, K. (2009). Hierarchical Bayesian statistics: Merging experimental and modeling approaches in ecology. *Ecological Applications*, *19*, 577–581.

Ogle, K., & Barber, J. J. (2020). Ensuring identifiability in hierarchical mixed effects Bayesian models. *Ecological Applications*, *30*, e02159.

Pu, X., Umaña, M. N., & Jin, G. (2020). Trait-mediated neighbor effects on plant survival depend on life stages and stage-specific traits in a temperate forest. *Forest Ecology and Management*, *472*, 118250.

Romero-Mujalli, D., Jeltsch, F., & Tiedemann, R. (2019). Individual-based modeling of eco-evolutionary dynamics: State of the art and future directions. *Regional Environmental Change*, *19*, 1–12.

Schneider, M. K., Law, R., & Illian, J. B. (2006). Quantification of Neighbourhood-dependent plant growth by Bayesian hierarchical modelling. *Journal of Ecology*, *94*, 310–321.

Seidl, R., Rammer, W., Scheller, R. M., & Spies, T. A. (2012). An individual-based process model to simulate landscape-scale forest ecosystem dynamics. *Ecological Modelling*, *231*, 87–100.

Shen, L., Yan, M., Wu, G., & Su, X. (2020). Individual tree location detection by high-resolution RGB satellite imagery in urban area. In *Proceedings of the 2020 International Conference on Computing, Networks and Internet of Things* (pp. 139–143). ACM.

Soranno, P. A., Bissell, E. G., Cheruvelil, K. S., Christel, S. T., Collins, S. M., Fergus, C. E., Filstrup, C. T., Lapierre, J.-F., Lottig, N. R., Oliver, S. K., Scott, C. E., Smith, N. J., Stopyak, S., Yuan, S., Bremigan, M. T., Downing, J. A., Gries, C., Henry, E. N., Skaff, N. K., … Webster, K. E. (2015). Building a multi-scaled geospatial temporal ecology database from disparate data sources: Fostering open science and data reuse. *GigaScience*, *4*, s13742-015.

Sortibrán, L., Verdú, M., & Valiente-Banuet, A. (2014). Nurses experience reciprocal fitness benefits from their distantly related facilitated plants. *Perspectives in Plant Ecology, Evolution and Systematics*, *16*, 228–235.

Stan Development Team. (2019a). Stan modeling language users guide and reference manual, 2.26.

Stan Development Team. (2019b). Stan reference manual. Version 2.23.

Stan Development team. (2022). 8.2 Ragged data structures. Page Stan user's guide 2.8.

Urban, D., & Keitt, T. (2001). Landscape connectivity: A graph-theoretic perspective. *Ecology*, *82*, 1205–1218.

Uriarte, M., Condit, R., Canham, C. D., & Hubbell, S. P. (2004). A spatially explicit model of sapling growth in a tropical forest: Does the identity of neighbours matter? *Journal of Ecology*, *92*, 348–360.

Valenta, M. D., Golluscio, R. A., Frey, A. L., Garibaldi, L. A., & Cipriotti, P. A. (2020). Short-term responses to sheep grazing in a Patagonian steppe. *The Rangeland Journal*, *42*, 1.

Yang, X., Angert, A. L., Zuidema, P. A., He, F., Huang, S., Li, S., Li, S.-L., Chardon, N. I., & Zhang, J. (2022). The role of demographic compensation in stabilising marginal tree populations in North America. *Ecology Letters*, *25*, 1679–1689.

Zaiats, A., Lazarus, B. E., Germino, M. J., Serpe, M. D., Richardson, B. A., Buerki, S., & Caughlin, T. T. (2020). Intraspecific variation in surface water uptake in a perennial desert shrub. *Functional Ecology*, *34*, 1170–1179.

Zambrano, J., Beckman, N. G., Marchand, P., Thompson, J., Uriarte, M., Zimmerman, J. K., Umaña, M. N., & Swenson, N. G. (2020). The scale dependency of trait-based tree neighborhood models. *Journal of Vegetation Science*, *31*, 581–593.

Zambrano, J., Fagan, W. F., Worthy, S. J., Thompson, J., Uriarte, M., Zimmerman, J. K., Umaña, M. N., & Swenson, N. G. (2019). Tree crown overlap improves predictions of the functional neighbourhood effects on tree survival and growth. *Journal of Ecology*, *107*, 887–900.

Zambrano, J., Iida, Y., Howe, R., Lin, L., Umana, M. N., Wolf, A., Worthy, S. J., & Swenson, N. G. (2017). Neighbourhood defence gene similarity effects on tree performance: A community transcriptomic approach. *Journal of Ecology*, *105*, 616–626.

Zhang, B., & DeAngelis, D. L. (2020). An overview of agent-based models in plant biology and ecology. *Annals of Botany*, *126*, 539–557.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Barber, C., Zaiats, A., Applestein, C., Rosenthal, L., & Caughlin, T. T. (2022). Bayesian models for spatially explicit interactions between neighbouring plants. *Methods in Ecology and Evolution*, *13*, 2788–2798. https://doi.org/10.1111/2041-210X.13998