SPATIAL ESTIMATION USING APPROXIMATE BAYESIAN COMPUTATION

ESTIMATING THE SPATIAL DYNAMICS OF PLANT RECRUITMENT USING APPROXIMATE BAYESIAN COMPUTATION

By JENNIFER FREEMAN, B.Sc.

A Thesis Submitted to the School of Graduate Studies in Partial Fulfillment of the Requirements for the Degree Master of Science

McMaster University MASTER OF SCIENCE (2023) Hamilton, Ontario (Computational Science and Engineering)

TITLE: Estimating the Spatial Dynamics of Plant Recruitment using Approximate Bayesian Computation

AUTHOR: Jennifer Freeman, B.Sc. (McMaster University)

SUPERVISOR: Professor B.M. Bolker

NUMBER OF PAGES: ix, 43

Lay Abstract

Ecosystems can be characterized by underlying processes that create observed patterns in space; by understanding these patterns we can learn about the ecological mechanisms at play. Plant *recruitment*, the process by which new individuals join a plant community, can be decomposed into two parts: (1) the dispersal of seeds across the landscape and (2) the environmental conditions that determine the success of those seeds becoming seedlings. We learn about the spatial dynamics behind recruitment by analyzing observed seed and seedling patterns with a computational statistics tool (Approximate Bayesian Computation, ABC). This tool can provide strong advantages over more classic statistical methods. Our method can serve as a general framework for understanding unobserved spatial processes that give rise to observed spatial patterns in any ecological system.

Abstract

We present a general statistical framework to infer processes (underlying ecological mechanisms) from patterns (observed arrangements in nature) in spatial ecology. We demonstrate our method by investigating the process of plant *recruitment*, how new individuals join a plant community, combining *seed dispersal* and environmental factors that determine the success of seeds germinating and surviving to juvenile maturity (*establishment*). Observed data includes seed and seedling counts at discrete spatial locations for the tree species slash pine (*Pinus elliottii*). The patterns in the data are described by their spatial correlation and we incorporate these correlations into historically used spatial models. We use a Bayesian simulation-based inference algorithm (Approximate Bayesian Computation, ABC) to estimate model parameters. Interest in ABC and Bayesian inference methods is growing in ecology, but they still remain behind classic approaches. Our results highlight techniques to validate the method to ensure accuracy and detect issues. Simulation tools are discussed to improve computational efficiency. We conclude with ABC parameter estimates that capture valuable spatial information ecologists can interpret. A small comparison study with classic likelihood-based parameter estimates is performed to illustrate the flexibility and informativeness of ABC. Our method is purposefully kept general to make it applicable to many spatial ecological problems.

Acknowledgements

First, I wish to express gratitude to my supervisor, Ben Bolker, for his constant support and encouragement throughout this journey. Thank you for introducing me to the interesting fields of spatial ecology and Bayesian statistics. Your passion for research has inspired me.

I would like to thank the members of my examining committee, Dr. Jonathan Dushoff and Dr. Sash Vaid, for their helpful feedback.

Thank you to all the members of the MacTheobio Lab for providing a welcoming space to ask questions, make mistakes, and grow as a researcher.

Finally, thank you to my family and friends for always lifting me up.

Contents

List of Figures

List of Tables

List of Abbreviations and Symbols

ABC	Approximate Bayesian Computation
CPU	Central Processing Unit
CSE	Computational Science and Engineering
CV	Coefficient of Variation
GLMM	Generalized Linear Mixed Model
GN	Generalized Normal (Distribution)
GP	Isotropic Stationary Gaussian Process
HDI	Highest Density Interval
HPC	High Performance Computing
IBM	Individual-based Model
MAD	Median Absolute Deviation
MC	Monte Carlo
MCMC	Markov Chain Monte Carlo
MLE	Maximum Likelihood Estimation
MVN	Multivariate Normal (Distribution)
N	Normal (Distribution)
PPD	Prior Predictive Check
RMSRE	Root Mean Square Relative Error
SBC	Simulation Based Calibration
SMC	Sequential Monte Carlo
U	Uniform (Distribution)

1 Introduction

Ecosystems consist of a complex web of interactions between living organisms (biotic processes), their environment (abiotic processes), and external influences (biotic/abiotic disturbances). Ecologists have simplified this story into two pieces; 1) patterns - the observable arrangements in nature and 2) processes - the underlying systems that give rise to patterns. Ecologists try to understand the link between the two.

The spatial or temporal scale, on which patterns are observed determines what we can learn about processes. For example, a terrestrial animal might forage locally and migrate seasonally over large ranges in response to environmental and biological cues. Processes often act over multiple scales. In predation, a song bird preys on an earthworm, yet these animals perceive and interact with their neighbourhood on vastly different scales. We cannot avoid thinking about scales in ecology, and no single scale provides the optimal vantage point.

Scales in ecological research, either imposed or chosen, are characterized by both the fine-scale resolution in measurements (the *grain*) and the large-scale range of the study (the *extent*). From macroecology and landscape ecology, to population and community ecology we move from global, to regional, to local extents. These subfields do not describe isolated systems but instead form a hierarchy of nested and overlapping scales in which local dynamics can effect global patterns. In study design and analysis we have to think carefully about the scales of interest, what choices we have, and how they influence what we can infer.

1.1 Problem

The general goal of this research is to quantify aspects of spatial ecological processes that are not directly observed and understand how they lead to the development of spatial patterns. We hope this general statistical framework will lend itself to many ecological studies investigating patterns from processes.

Exploring this relationship fully involves an inferential perspective. As scientists, making predictions is essential, especially in the ever-changing natural world. Sometimes this comes at the cost of making hard and fast rules about which variables matter and how they relate to one another. Here, we want to move away from these practices and emphasize learning how we can better improve our models and make more informed choices.

Our method combines existing models in spatial ecology with Bayesian inferential techniques. These techniques have been growing in popularity in this field, but much is still to be learned.

Under the umbrella of population and spatial ecology we apply our framework to a spatial data set to learn about the process of *plant recruitment*; how new plant individuals join an existing community. Seeds travel from their parent source to a final site through *seed dispersal* and seedlings emerge through germination and survival success to reach juvenile maturity (*establishment*). Factors such as distance from the parent, wind and dispersing animals contribute to dispersal. *Environmental filtering* effects such as soil quality, moisture, light, presence of pathogens, and herbivory in turn determine the establishment probability. Tracking all of these parts and more as they change through space and time would be impractical. This is why we think of ecological processes as being "hidden" or "unobserved"; however we can uncover elements and potentially learn about which variables to measure and on what scales by exploring the patterns they create.

We observe the spatial patterns of plant recruitment through a data set for the tree species *Pinus elliottii*(slash pine), a coniferous tree native to south eastern United States. The data set consists of seed and seedling counts recorded at discrete spatial locations. In the field, seeds were collected in seed traps and seedlings were counted in quadrats according to a pre-specified sampling design. Seeds in the same seed trap and seedlings in the same quadrat are mapped to the same spatial coordinate given in Eastings and Northings. The dimensions of traps and quadrats are unknown, defining the spatial grain as the smallest distance between traps and between quadrats; therefore, we cannot quantify minimal distances between individuals in the same trap or quadrat. The range of the experiment covers a 1500 km^2 region in Florida with data observed in 10 plots over this region. Plot sizes (in meters) specify our spatial extent because we ignore between plot effects, assuming that the scale between plots is too large to detect processes of interest.

1.2 Background

1.2.1 Spatial Ecology

In spatial ecology, we can characterize patterns by assessing the strength of relationship between observations. A defining feature of many spatial patterns is clustering or association where nearby measurements are more similar than farther measurements, called positive correlation in statistics. Negative correlation manifests in segregated patterns such as an evenly spaced grid of observations. Not surprisingly, the scale on which patterns are viewed reflect the strength and scale of relationships we measure.

A popular view of species organization is through *patch dynamics* and *metapopulation models*. Patches describe suitable habitat islands on the landscape in which species exist reflecting the assembly patterns we see in the world. Specifying the separate dynamics of within and between patch mechanisms explicitly defines the relevant scales on the order of patch size and overarching population extent (Levin, 1992). Instead of using spatial correlation to describe the gradient of environmental factors that make patches suitable for habitation, these scales fix the environment into the dichotomy of suitable and non-suitable regions. These models provide a simple mathematical tool and ecological intuition, but they can be too restrictive about habitat layout and the spatial variability of biotic processes (North and Ovaskainen, 2007).

Classical population models use differential equations describing the rates of demographic changes like births and deaths and the process mechanisms like dispersal and establishment, often with a temporal component. Historically, spatial contributions are ignored in these models because there is assumed to be sufficient population mixing to average out the effects on mean population density.

Stochastic simulation-based models can incorporate the best of both worlds in which populations can be simulated through continuous space and time governed by dynamics that can be spatially variable. Individual-based models (IBM), as a subcategory, track discrete individuals but they require large data sets and their complexity can make inference a challenge (Bolker, 2003; North and Ovaskainen, 2007).

Seabloom et al. (2005) use theoretical spatial correlation structures as a basis for understanding how seed dispersal, environmental heterogeneity, and competition individually and jointly create spatial patterns in plant communities at different scales. They use second-order moment models to improve upon the classical differential equation approach, to incorporate both mean population densities and the covariances between densities over space in addition to rates of reproduction, recruitment and mortality. They use simulations of individuals in a two-species community with a competition-dispersal tradeoff to construct hypotheses about how the spatial patterns of aggregation and segregation evolve over time. In field studies with grassland plants they confirm hypotheses and estimate scales of patterns through distance dependent correlation functions. At smaller distances these functions describe local species behaviour. As the function approaches zero with increasing distance we learn about the scales of pattern (patch size). Their results show that environmental heterogeneity acting alone can rapidly create aggregation patterns within species and that the scales of aggregation increased over time. All three processes have the combined effect of breaking up clusters into more segregated patterns. Large-scale disturbances, such as fire, can remove spatial dependencies.

In related theoretical work, North and Ovaskainen (2007), use an IBM with the same ecological processes as in Seabloom et al. (2005) to understand how environmental heterogeneity affects reproductive rates and establishment probability on a continuous time-space surface. The heterogeneity is classified by both patch size and “level of heterogeneity”, which determines the difference in habitat quality within and outside of patches (North and Ovaskainen, 2007). They use randomly spaced points on the surface to represent patch centers and a kernel function to describe the distribution of favourable environmental resources around those points. Over time, new patches form and others disappear. By varying these two environmental components, they learn that the optimal patch size leading to highest population density is intermediate. In small patches, more seeds disperse to poor habitat quality areas and patches are more likely to go extinct. In large patches, the population density begins to decline again as dispersers have difficulty reaching patches farther away. Population density increased with the magnitude of heterogeneity; that is, higher quality patches can theoretically increase population size despite the sacrifice of making poor quality regions worse. This was likely due to increased aggregation in patches and given the assumption of dispersal processes being independent of habitat type.

Many empirical studies encountered, Asefa et al. (2017); He and Biswas (2019); Raduła et al. (2020), made assumptions about probable abiotic and biotic variables contributing to patterns and processes and measured these on sequences of spatial scales. The ability to detect correlation in these variables and its strength can inform us about which scales are important in these systems. For example, Mod et al. (2020), use a sequence of nested sampling grains to record presence-absence plant species counts. Soil, moisture, light and temperature measurements were recorded on grains larger or equal to response grains. They use a joint species distribution model to calculate correlations between each species and the environment. Correlations of model residuals are assumed to be due to biotic interactions and a Bayesian inference method was used to parameterize the model. Traditionally, ecologists have assumed that environmental filtering acts first on larger scales, and biotic interactions perform further filtering as an individual relates to their neighbourhood on smaller scales. Mod et al. (2020) propose a further refinement to biotic interactions suggesting competition effects should decrease at larger scales and facilitation, the positive effects of co-existing species, should be independent of scale. Species were more strongly correlated with the environmental variables as the sampling grain increased. The biotic interaction results were less clear, but in general competitive effects seemed to decrease more than facilitative effects with increasing scale.

1.2.2 Bayesian Inference

There are two major fields of thought in statistics, Frequentist versus Bayesian. Frequentists describe data as a random sample from a population and wish to quantify unknown fixed population parameters. In the context of parameter estimation, this means we assume there is a true fixed value and its expectation is the average computed from an infinite sequence of random samples. Confidence intervals express our uncertainty that this true value is contained in an interval for a fraction of these theoretical experiment repetitions.

In contrast, Bayesians assume data are fixed observations and parameters are random variables represented by probability distributions. We describe our uncertainty about parameters before our experiment, and we use the information from our data to update our uncertainty (Crome et al., 1996). In this way we can make case-specific probability statements. *Credible intervals*, the Bayesian analog of confidence intervals, represent a direct measure about our uncertainty of a parameter by specifying an interval in the domain of a probability distribution.

Ellison (1996) argues that the repeatability of experiments in ecology is not a realistic assumption. Field experiments cannot be fully isolated from external influences and even two organisms of the same species are not identical (Ellison, 1996). Further, evolutionary changes mean that even if a true fixed parameter value exists it is likely to change over time (Ellison, 1996)

A recent study by Lines et al. (2020), uses Bayesian inference to investigate tree dynamics through forest models with count data. They are interested more specifically in forest recruitment at large scales, because early dynamics determine long term forest patterns. Due to the limited availability of early-life forest data over small scales, forest models often contain unrealistic simplifications of early dynamics. Instead of directly quantifying spatial patterns, they estimate demographic rates of recruitment, growth and mortality in juveniles. These parameter estimates are then used with large-scale adult forest data sets in forest models to assess their predictive capability.

1.3 Outline

We want to extend the work of Seabloom et al. (2005) and North and Ovaskainen (2007) by using spatial correlation to understand how underlying processes create patterns, with an emphasis on keeping methods general and not specific to a particular ecological community, species, or data set.

We use the terminology of patches to describe regions of high probability density for a population. This is a more realistic ecological view than the metapopulation dichotomy of an organism being either within a patch or outside of a patch. The term “patch size” will refer to the ecological scale of heterogeneity so that differences in patch sizes express that the strength of spatial correlation occurs over different distances on the landscape.

We use a similar view of environmental heterogeneity to North and Ovaskainen (2007), that is continuous on the landscape and characterized spatially by patch size and the smoothness of transition between high and low density regions. We view this surface as a continuous surface of establishment probabilities because favorable environmental areas will create higher chances of seeds establishing. We use the same continuous structure for the seeds, where the surface now describes probability densities of seeds on the landscape.

We are interested in differentiating between these two sources of spatial heterogeneity originating from the variability in biotic interactions (*endogenous*) and variability from the environment (*exogenous*) as in Bolker (2003).

Since our data is aggregated at the level of the grain we are not considering individual dynamics; because we observe only one time step between the two data sets, we ignore temporal scales. Possible predictor variable measurements were not provided with this data set, so we do not attempt to tie the spatial pattern to particular covariates.

We focus on parameter estimation via Bayesian inference as in Mod et al. (2020) and Lines et al. (2020). Instead of assuming a set of fixed scales as in Mod et al. (2020), we first describe our beliefs about plausible patch sizes and use our method to estimate the spatial scales of environmental filtering and seed densities.

We introduce the spatial model in Section 2, followed by the Bayesian inference tool (Section 3). We quantify the uncertainty about our parameters (Section 4) before performing model simulations (Section 5). Following model validation (Section 6) we give our inference results in Section 7. We conclude with a discussion (Section 8).

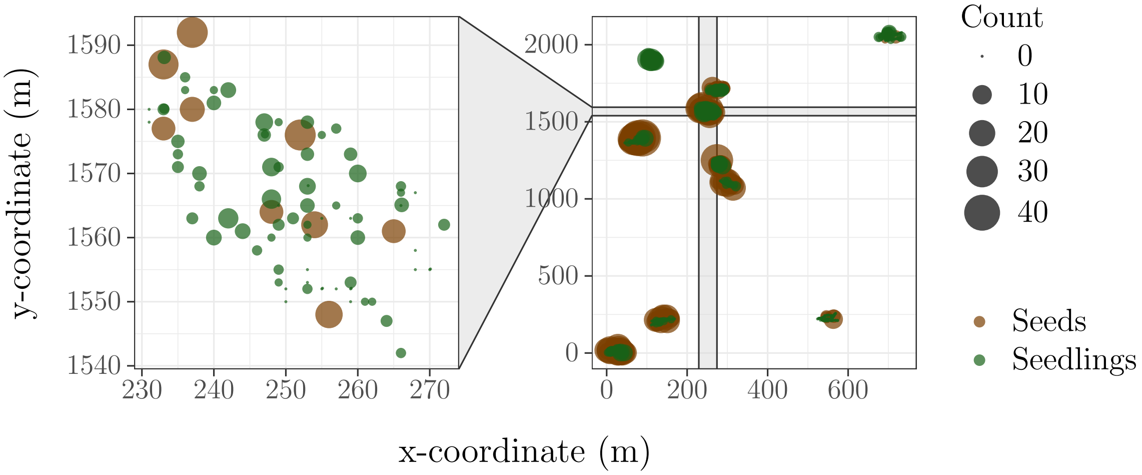

Figure 1: Slash pine data set. Right panel: The data set containing 107 seed counts collected in 9 plots and 749 seedling counts collected in 10 plots. Left panel: Detail showing a single plot.

2 Geostatistical Data and Models

Geostatistical data is described by a set of $i = 1, ..., n$ observed response values y_i at discrete spatial locations x_i. Typically x_i is a spatial coordinate pair like latitude and longitude, while y_i is a realization of Y_i, a random variable whose distribution depends on an underlying continuous spatial signal $S(x)$ over the region (Diggle and Ribeiro, 2007). We cannot directly observe $S(x)$ but we attempt to gain insight from it through the response $Y = (Y_1, ..., Y_n)$ (Diggle and Ribeiro, 2007). The systems we wish to model are typically noisy, so that the response is really a perturbed version of the true signal (Diggle and Ribeiro, 2007). This formulation would extend in the multivariate case to observing more than one response at each location.

The system we are describing contains two geostatistical data sets. The response variables are counts of seeds and seedlings observed at irregularly spaced locations. Eastings and Northings coordinates were transformed to units of meters relative to the southwestern corner of the study region, see Figure 1. The underlying spatial signals are seed dispersion and the environmental establishment process.

It is assumed the sampling design was chosen by ecological experts to be appropriate to estimate characteristics about dispersal and establishment. The irregularity of the design may in particular be better for parameter estimation than for spatial prediction (De Oliveira and Han, 2022). As usual, in geostatistics we assume that the design is statistically independent of the underlying spatial signals (Diggle and Ribeiro, 2007).

These continuous spatial signals when incorporated in a geostatistical model are usually represented as an *isotropic stationary Gaussian process* (GP) (Gelfand and Banerjee, 2017). A GP assumes that the signal $S(x)$ is multivariate Gaussian. Stationarity means $S(x)$ has a constant mean and variance independent of x, and the correlation between $S(x_j)$ and $S(x_i)$ only depends on a difference metric $u = x_j - x_i$ between locations x_j and x_i (Diggle and Ribeiro, 2007). Isotropic stationarity is achieved when the difference metric is symmetric (Diggle and Ribeiro, 2007). An obvious and popular choice for the difference metric is Euclidean distance, $u = ||x_j - x_i||$. This radial spatial symmetry simplifies our correlation structure to depend on only one value, the distance between spatial locations.

Interpreting GPs in the seed dispersal system, we imagine a continuous surface of seed densities, with constant expectation and variance of seed counts over the region. We know that seeds disperse close to their parent plant, naturally leading to positive association (Bolker and Pacala, 1999). Over the landscape this creates a heterogeneous pattern of seed counts. Negative spatial autocorrelation describes a self-avoiding pattern, which can occur when incorporating competing species, but is not expected to occur from seed dispersal alone (Bolker and Pacala, 1999).

For plant establishment, the GP is a surface representing the environmental filtering that determines whether a seedling successfully establishes. Although we do not use explicit patch dynamics, we can imagine this surface as a heterogenous pattern of suitable habitat patches to describe this filtering effect. Patch size describes regions of high suitability and gaps between patches describe low suitability areas (North and Ovaskainen, 2007). For example, slash pine seeds are particularly successful at germinating if the soil contains ample moisture (Burns, 1983). Soil moisture, like most environmental variables, typically exhibits positive spatial autocorrelation. If a seedling establishes at one location due to sufficient moisture, it is likely that the moisture conditions will be more similar nearby and less similar as we move away from this seedling, generating a heterogeneous pattern in established seedlings. The seedling pattern contains signatures from both endogenous and exogenous forms of heterogeneity.

Geostatistical models (Equation 1) are a form of *Generalized Linear Mixed Model* (GLMM) with a known link function $h(\cdot)$, spatial variables d_k and unknown regression parameters β_k (Diggle and Ribeiro, 2007).

$$h(E[Y_i|S(\cdot)]) = S(x_i) + \sum_{k=1}^{p} \beta_k d_k(x_i) \tag{1}$$

The spatial variables, or predictors, would be additional variables measured at sampling locations that are associated with the response. In this experiment we have no environmental covariates so we remove these terms. We could include additional fixed terms to describe spatial trends, such as a spatial varying mean as a linear or quadratic polynomial of the x and y coordinates.

It is convenient to split the variability in our models into two components 1) spatial variation contained in our GP $S(x)$ and 2) non-spatial residuals included as $\epsilon(x)$ in Equation 2 (Gelfand and Banerjee, 2017).

$$h(E[Y_i|S(\cdot)]) = S(x_i) + \epsilon(x) \tag{2}$$

The residuals $\epsilon(x)$, can also be interpreted as spatial variation on scales smaller than the smallest distance between observed locations in conjunction with measurement error (Diggle and Ribeiro, 2007). In the geostatistical framework, this type of variation is called the *nugget effect* with variance labelled as τ^2 (Diggle and Ribeiro, 2007). Given a random Normal deviate $Z_i \sim N(0,1)$ we can update our model to reflect this as in Equation 3.

$$h(E[Y_i|S(\cdot)]) = S(x_i) + \tau Z_i \tag{3}$$

The spatial variance component of a GP is described by a covariance function $\gamma(u)$ where u is distance and the unconditional variance at any location $\gamma(0) = \sigma^2$ is constant. Specific families of functions are typically used to describe spatial covariance in geostatistics because they possess the necessary property of positive definiteness (Diggle and Ribeiro, 2007).

$$\rho(u) = \frac{1}{2^{\kappa-1}\Gamma(\kappa)}(\frac{u}{\phi})^{\kappa} K_{\kappa}(\frac{u}{\phi}) \tag{4}$$

The *Matérn* family is commonly used in geostatistics. The Matérn covariance function, Equation 4, takes a shape parameter $\kappa > 0$ and scale parameter $\phi > 0$ (in units of distance u). ϕ determines the rate of correlation decay with distance, while κ determines the smoothness of the signal S (Diggle and Ribeiro, 2007). Specifically, $S(x)$ is $\lceil \kappa \rceil - 1$ mean-square differentiable and K is the modified Bessel function of the second kind (Diggle and Ribeiro, 2007).

A plot of the Matérn function is shown in Figure 2 for two different scale and shape parameters. The function continuously decreases with distance, which makes it an ideal choice to describe the decrease in positive autocorrelation in our models. The

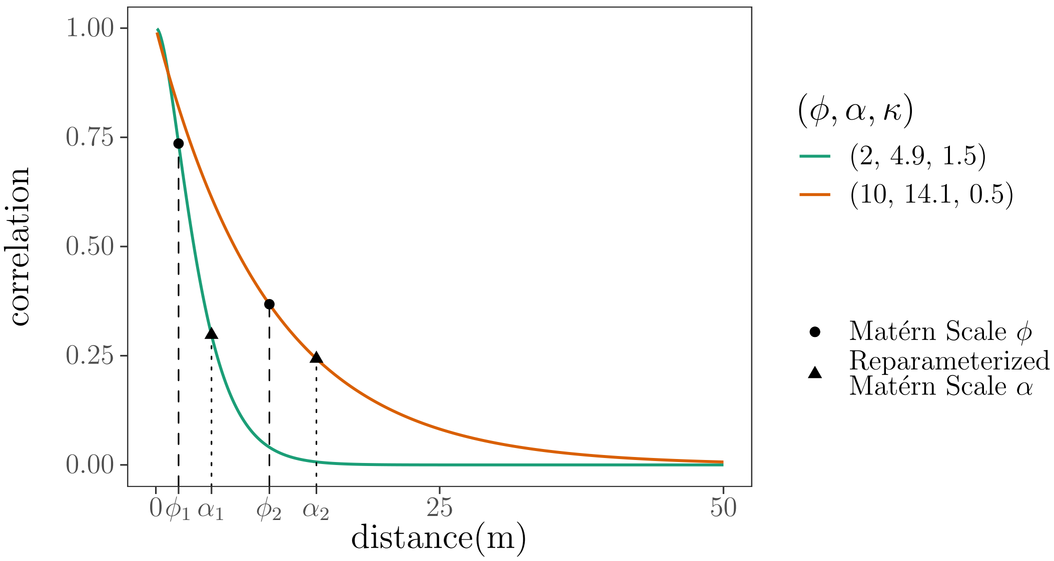

Figure 2: The Matérn correlation function for different scale and shape parameters. The function has an approximately constant correlation of 0.25 when using the reparameterized scale parameter α (triangle points). When the unaltered Matérn scale parameter ϕ is used, we get much larger differences in correlation among curves (circular points).

function with a larger scale parameter of 10m asymptotically approaches 0 more slowly than the function with a scale of 2m.

In the standard parameterization, the Matérn scale and shape parameters are non-orthogonal meaning that they are dependent (Diggle and Ribeiro, 2007). A fixed scale value will correspond to different rates of correlation decay depending on the specified shape value, see Figure 2. We want the ecological scale (patch size) to depend on an approximately constant value of correlation because otherwise we cannot compare the size of the patch with the level of spatial clustering we observe. We choose to reparameterize as in Diggle and Ribeiro (2007) so that the reparameterized scale $\alpha = 2\phi\sqrt{\kappa}$ can be interpreted as approximately independent of the shape. We can then interpret α as the approximate patch size in meters, capturing a spatial autocorrelation value of approximately 0.25 (Figure 2).

The changes in the shape parameter can be best visualized in the context of GPs where they govern the smoothness of transition from high density signal areas to low, as shown in Figure 3.

The *spline correlogram* is a modified version of the *spatial correlogram* which estimates the autocorrelation function in a spatial data set using binned distance groups

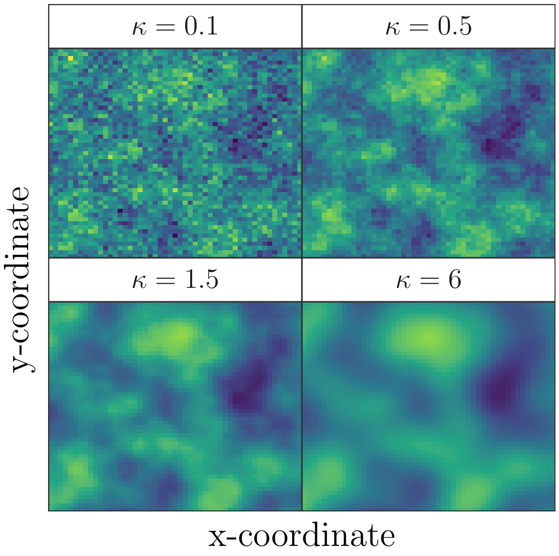

Figure 3: Example Gaussian processes for simulated response data over a regular grid with varying values of the shape parameter κ while all other parameters are held fixed ($\alpha = 1$, $\sigma^2 = 0.01$, $x = \{0, 0.1, 0.2, ..., 5\}$, $y = \{0, 0.1, 0.2, ..., 5\}$).

and without assuming any particular function characteristics (Bjørnstad and Falck, 2001). This flexibility of the correlogram allows it to be used across a wide range of spatial systems (Bjørnstad and Falck, 2001). The spline correlogram improves on the spatial correlogram by providing a smooth positive definite autocorrelation curve for which confidence bands can be estimated (Bjørnstad and Falck, 2001). The spline correlogram should closely resemble the true covariance function; if no autocorrelation is present the curve will be flat and approximately zero (Bjørnstad and Falck, 2001).

We use the spline correlogram as an initial non-parametric check of our observed data. Later we will use values of the spline correlogram from simulated data sets to summarize the scales of correlation in the data into a sequence of lower-dimensional statistics.

The *ncf* package in R includes a `spline.correlog` function which computes the univariate or multivariate spline correlogram (Bjørnstad, 2022). The spline correlograms for our observed data are shown in Figure 4; they agree with our model assumptions that correlation declines with distance.

The *correlation length* is a statistic derived from the spline correlogram that is defined as the first distance at which the spline correlogram is zero (Bjørnstad and Falck, 2001). Responses measured at distances larger than the correlation length are no more similar than by chance (Bjørnstad and Falck, 2001). Since the Matérn covariance

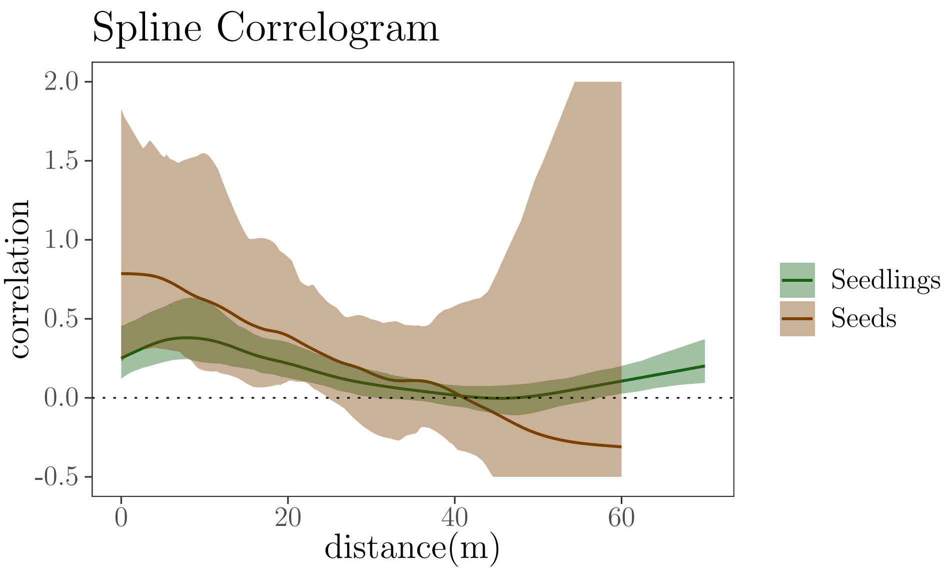

Figure 4: Spline correlogram of observed data with 8 degrees of freedom. The maximum within-plot distance was used as an upper bound. The default function estimates the correlation at 300 points. A 95% confidence envelope is shown by the shaded ribbon.

function never reaches zero but instead asymptotically approaches zero the correlation length can instead be defined as the distance at which the correlation reaches some small value (Bjørnstad and Falck, 2001). The correlation length for both data sets is approximately 40 m.

We combine the base geostatistical model in Equation 3 and the Matérn covariance function in Equation 4 to form the models for our two systems. Going forward we will use the notation that μ represents sample means, σ^2 represents signal variance, and τ^2 represents nugget variance.

2.1 Seed Dispersal

We observe (or simulate) a seed count of N_i at location x_i. The underlying signal is the continuous seed density surface $S_{seed}(x)$ represented by a GP $\sim$ MVN$(\mu_{seed}, \Sigma_{seed})$. μ_{seed} is the constant seed count mean over the region independent of location. Σ_{seed} is the covariance matrix with constant variance σ^2_{seed} along the diagonal and the (i-th,j-th) entry the Matérn correlation function multiplied by the variance $\rho_{i,j}(u)\sigma^2_{seed}$ with u, the Euclidean distance $u = \sqrt{(x_i - x_j)^2}$. To account for measurement error and non-spatial variation we add a nugget effect of τ^2_{seed}.

A natural choice for count data is the Poisson log-linear model where our seed counts N_i follow a Poisson distribution with location and scale parameter, λ_i, and a log link function. Our model for seed dispersion is described in Equation 5.

$$\begin{aligned} N_i &\sim \text{Poisson}(\lambda_i) \\ \log[\lambda_i] &= S_{seed}(x_i) + \tau_{seed} Z_{seed,i} \end{aligned} \tag{5}$$

In the absence of a nugget effect the Poisson counts will be highly variable when the GP mean is small and will more closely resemble the underlying surface when the GP mean is large (Diggle and Ribeiro, 2007). We can interpret the nugget effect as capturing this extra Poisson variation or *overdispersion*, as we expect noisy data.

2.2 Establishment Probability

We form a similar model for seedling establishment using a Binomial logistic-linear model. The number of trials is the seed count N_i from the seed dispersal model, from an observed or simulated data set. The probability of each trial is the mean establishment probability p_i at location x_i. The numbers of successful trials are the seedling counts M_i. Using the log odds (logit) link function is a natural choice because we want establishment to be on the probability scale; the inverse logit function transforms real values to $(0, 1)$. The underlying signal S_{est} is the establishment probability surface or the environmental filtering surface.

$$\begin{aligned} M_i &\sim \text{Binomial}(N_i, p_i) \\ \text{logit}[p_i] &= S_{est}(x_i) + \tau^2_{est} Z_{est,i} \end{aligned} \tag{6}$$

3 Approximate Bayesian Computation

For ecological models, usually generating data from the model for a given set of parameters is relatively easy but solving the inverse problem, identifying parameters that would generate data similar to the observed values, is hard (Lintusaari et al., 2016).

By regarding the parameters as random instead of fixed variables, we can use Bayesian inference. This method of inference includes using subject matter knowledge to specify our uncertainty about the parameters through priors (Ellison, 1996). We then update our prior probability conditional on a data set through a likelihood function (Lintusaari et al., 2016). With careful prior specification we achieve statistical regularization by reducing parameter estimates when the data set contains less information about the parameters because it is small or noisy (Lemoine, 2019).

Bayesian inference is derived from Bayes' Theorem (Equation 7). With parameters θ and data x we get the relationship that the posterior $p(\theta|x)$ is the product of the likelihood $p(x|\theta)$ and the prior $\pi(\theta)$ divided by the marginal likelihood $p(x) = \int p(x|\theta)\pi(\theta)d\theta$.

$$p(\theta|x) = \frac{p(x|\theta)\pi(\theta)}{p(x)} \tag{7}$$

Because the denominator in Bayes' Theorem is a constant, it can be convenient to think of the relationships in terms of a proportionality where the likelihood revises our prior information into posterior expectations (Ellison, 1996):

$$\text{posterior} \propto \text{likelihood} \times \text{prior}$$

Computing posterior probabilities, to estimate parameters then requires the evaluation of the likelihood function. In cases where there are hidden latent states (i.e. unobserved processes) the likelihood is often intractable and methods called "likelihood-free inference" such as Approximate Bayesian Computation (ABC) may be necessary (Beaumont, 2010).

ABC is a process of sampling from the posterior distribution by finding parameter combinations that lead to model-generated data similar to the observed values, thus arriving at parameter sets that are likely to describe our observed data (Lintusaari et al., 2016). ABC was first introduced in the population genetics literature by Tavaré et al. (1997) and Pritchard et al. (1999), and later named by Beaumont et al. (2002). ABC's simplicity makes it useful in a wide array of applications (Lintusaari et al., 2016).

The simplest ABC algorithm is *rejection ABC* (Algorithm 1). A set of summary statistics $S(\cdot)$ is chosen to be computed from a simulated data set, x_i, and observed data set, y. We choose a stopping tolerance criteria ϵ and distance metric $d(\cdot, \cdot)$ to arrive at a set of N approximate posterior samples. This approximation is due to the tolerance condition ϵ. An exact posterior sample would be achieved in the limit as $\epsilon \to 0$ (Robert, 2016).

Algorithm 1 Rejection ABC

```
$i = 1$
**while** $i \leq N$ **do**
    Sample parameter $\theta_i \sim \pi(\theta)$
    Using the data-generating model, simulate a dataset $x_i \sim p(x|\theta_i)$
    **if** $d(S(x_i), S(y)) > \epsilon$ **then**
        Reject $\theta_i$
    **else**
        Accept $\theta_i$
        $i = i + 1$
    **end if**
**end while**
```

Reducing data to a set of summary statistics so that the dimension of the summary statistics is much less than the dimension of the data naturally reduces the informative power of the data in the model, decreasing the accuracy of estimates. We can achieve identical summary statistics for non-identical but similar data sets (Lintusaari et al., 2016). Selecting a distance metric such as Euclidean distance and deciding on an appropriate tolerance and approximate posterior sample size are also limiting in the sense that they determine the computational efficiency of the algorithm (Lintusaari et al., 2016). There is a trade-off between these choices to balance accuracy versus computational burden.

Rejection ABC is basic to program and execute, and simulating a single data set is typically cheap enough that this step can be repeated many times (Sisson et al., 2018). However, we need a large number of simulations because most parameter sets will be rejected (Lintusaari et al., 2016). If we increase the number of summary statistics to include more information from our data we increase the dimension of the distance computation, accepting fewer parameter sets and needing to increase the number of simulations, the "curse of dimensionality" (Lintusaari et al., 2016).

More complicated ABC algorithms can help address some of these problems. The

Markov Chain Monte Carlo (MCMC) variant is derived from the Metropolis-Hastings MCMC algorithm and differs from simple rejection sampling because as the chain iterates, parameter values are drawn from noisy posteriors instead of the prior (Lintusaari et al., 2016). Sequential Monte Carlo (SMC) ABC uses successively smaller tolerances and iterative proposal distributions defined by weighted parameters that were accepted in the previous step (Lintusaari et al., 2016). Both MC methods speed up the computation by narrowing down the posterior parameter space, lowering the ABC rejection rate, and thereby requiring fewer numbers of simulations (Lintusaari et al., 2016).

A larger tolerance widens the approximate posterior because a greater variability of parameter sets are accepted, but the advantage is a lower computational cost (Lintusaari et al., 2016). Often the tolerance is not specified directly but is instead specified as a percentage of the simulations to be accepted (Lintusaari et al., 2016). This percentage will be referred to as the *acceptance rate* and is the default tolerance function argument in the *abc* R package used to perform ABC in this study (Csillery et al., 2012).

Choosing the right number of informative summary statistics can be challenging. Sisson et al. (2018) advocates for the use of *sufficient* statistics of low dimension. A statistic s is sufficient if the distribution "$\pi(y|s, \theta)$ is invariant to θ" (Sisson et al., 2018). In most models, Sisson et al. (2018) explains these ideal statistics are not available and a minimal set of insufficient statistics must be used instead. Lintusaari et al. (2016) and Sisson et al. (2018) describe a number of methods that can be used to help focus on a set of informative statistics.

When multiple summary statistics are used, they should be rescaled before the distance is computed so they have equal contribution in the rejection/acceptance step (Lintusaari et al., 2016). The default behaviour in the `abc` package normalizes the simulated summary statistics by dividing each statistic, s, by its *median absolute deviation* (MAD) given in Equation 8 after removing missing values and before evaluating distances. The observed summary statistics are similarly handled using the MAD of the simulated statistics.

$$MAD(s) = \text{median}|s - \text{median}(s)| \tag{8}$$

The MAD, like the standard deviation, is a measure of dispersion but is, more robust to the presence of extreme observations (Leys et al., 2013). Since our distributions of summary statistics before the ABC rejection/acceptance step are representative of a wide range of simulated data sets, they can be highly skewed. Although scaling by the MAD is a popular choice in ABC (Robert, 2016), in our experiments it did not provide enough scaling in the presence of outliers. Instead, all observed and simulated statistics were scaled by the standard deviation for each individual statistic.

3.1 Choice of Summary Statistics

A set of sixteen summary statistics were selected, eight for each system. To reflect within plot distances, 6 log-scaled breaks between 5 and 50m (for seeds) and between 5 and 60m (for seedlings) on the spline correlogram were chosen to capture the spatial components of the data. Values of the correlation for larger distances were discarded because the tail end of the spline correlogram is typically noisy. The correlation length was not chosen as a summary statistic due to the high number of simulated data sets in which the spline correlogram did not cross the x-axis. A short visualization of ten randomly simulated spline correlograms with setting varying degrees of freedom, led to a choice of 8 degrees of freedom in the simulations. The mean and standard deviation of simulated seed and seedling data sets were natural summary statistic choices to reflect the global characteristics.

4 Choice of Priors

4.1 Parameters

We have ten total parameters (Equations 5 and 6) for which to specify priors, five for each system. Each system is parameterized by a mean, i.e. mean seed count and mean establishment probability.

We chose to parameterize the variance parameters in terms of relative measures because these unitless measures are easier to conceptualize. The nugget proportion represents the proportion of total system variance attributed to the nugget effect (i.e. small scale variation). The total system coefficient of variation (CV) incorporates measurement error, small-scale and large-scale variation. We use the approximation from Lewontin (1966) to express the CV in terms of the variance of the log transformed data. For CV's less than 0.3, we assume $\mathrm{CV}^2 \approx \ln(\mathrm{Var}(x))$ (Lewontin, 1966). We can express this as,

$$\mathrm{CV} = \sqrt{\sigma_{log}^2 + \tau_{log}^2}$$

Since nugget proportion is a relative measure, we can again use the variance of the log transformed data to define

$$\tau_{prop} = \frac{\tau_{log}^2}{\sigma_{log}^2 + \tau_{log}^2}$$

We can then solve for the signal and nugget variance to be used in our geostatistical models.

$$\begin{aligned} \sigma_{log}^2 &= \mathrm{CV}^2(1 - \tau_{prop}) \\ \tau_{log}^2 &= \mathrm{CV}^2(\tau_{prop}) \end{aligned}$$

4.2 Prior Information

Prior distributions summarize subject matter knowledge about the unknown parameters of interest in our Bayesian model. We may know more or less about plausible parameter values, but we will always have some information about plausible values.

Priors can be categorized based on how much information they include. An *uninformative* or *vague* prior describes a wide distribution assuming large parameter variance. These distributions are often flat, assigning equal probabilities to all values in an unbounded or bounded domain (Sarma and Kay, 2020). A uniform prior is one distribution that might be used in this category amd communicates we know little to no information about the parameter except its overall feasible range. Conversely, an *informative* prior contains relevant information about the parameter (Sarma and Kay, 2020).

A third category of *weakly informative* priors, intentionally contain less information that we know in order to achieve a balance between how the data and prior jointly contribute to the posterior (Sarma and Kay, 2020).

Uninformative priors are used when the user wants the data to speak for itself (Lambert et al., 2005). Informative priors have the opposite effect; they can dominate the information in the model introducing too much subjectivity. Since the size of the data set communicates how much information the data contains about the unknown parameters, the sample size can also have a strong effect on the computed posterior (Lemoine, 2019). It is often recommended that weakly informative priors are used to achieve a balance in inferences obtained from all available data (Lemoine, 2019).

The process of *prior elicitation*, a challenging step in Bayesian statistics, formulates suitable prior distributions for parameters that reflect the range of possible values (Crome et al., 1996). A common method is to review priors from previous studies or to survey experts with diverse opinions.

Typically statements about proposed centrality and range limits for a parameter can be translated to prior distributions by specifying means and standard deviations for common probability distributions such as the Gaussian or Student's t (Sarma and Kay, 2020). For example, a Gaussian distribution contains 95% of the data within 2 standard deviations of the mean. We can use this knowledge to take proposed parameter limits $[l, u]$ and solve for the distributions standard deviation σ so that 95% of the data lie between the bounds $[l, u]$. Given $l = \mu - 2\sigma$ and $u = \mu + 2\sigma$, we get that $\sigma = (u - l)/4$.

It is important to elicit priors before looking at the data, because so called *data-dependent* priors inappropriately incorporate information from the data twice in our model (Berger, 2006). Forming a univariate prior is often easier to conceptualize than a multivariate (Sarma and Kay, 2020). The reparameterization of the Matérn scale allowed for the specification of two independent priors for the scale and shape parameters, instead of concerning ourselves with specifying a joint scale-shape prior.

Visualizing priors can be a useful tool in prior elicitation. If the parameters are difficult to understand, *prior predictive checks* (PPD) can help understand the implications of the chosen priors in the model (Gelman et al., 2020). After forming priors, PPD is performed by simulating data from the model using parameter draws from the priors and visualizing the simulated data in some capacity, possibly by summarizing the data in lower dimensions.

The priors chosen in the following sections were selected to be in the category of weakly informative and were loosely formed with a combination of information from literature and subject matter knowledge. We use standard transformations to ensure parameters are specified on the appropriate ranges and to increase interpretability. The inverse logit or log-odds scale ensures real values are mapped to [0,1], while the log scale ensures values are strictly positive and can be helpful when skewed distributions are required.

4.3 Seed Dispersal

Parameter (Unit) $\sim$ Distribution	Reasoning
Mean log Seed Count (log(count)) $\sim N\left(\log(10), \frac{\log(40)-\log(5)}{4}\right)$	We use a log-normal distribution to ensure mean seed counts are positive. We also expect higher probabilities for lower means of seed counts so a right-skewed prior is appropriate. A value of 10 seems to be a plausible high frequency seed count value and we choose 95% of the prior to range from 5 to 40.
Matérn Shape (unitless) $\sim U(0.5, 4)$	We use a uniform prior because it is difficult to understand plausible ecological values. Matérn shape values of 0.5, 1.5 and 2.5 are typically used in geostatistical models (Diggle and Ribeiro, 2007) so we want to incorporate this range of values. As the parameter tends to infinity the Matérn approaches the Gaussian which leads to an overly smooth GP as is seen in Figure 3 when $\kappa = 6$.

Parameter (Unit) $\sim$ Distribution	Reasoning
Matérn Reparameterized Scale (m) $\sim \exp(GN(\mu, \sigma))$ $\mu = \frac{\log(3)+\log(50)}{2}$ $\sigma^2 = \frac{\alpha\Gamma(3/\beta)}{\Gamma(1/\beta)} \approx 0.453$	A Generalized Normal (GN) was used so that the heaviness of the tails could be specified. A tail probability of 5% and centre probability of 55% was specified in the `get_gnorm` function (Bolker, 2022). These choices led to computed values of $\alpha \approx 1.39$ and $\beta \approx 4.76$ in the GN which creates lighter tails and more density in the centre. Based on the distribution of interpoint distances between seed trap locations (see Figure 5) we chose to set the 95% range of the data between 3m and 50m because there are not many distances outside of this range. This also agrees with slash pine seed disperal literature that indicates 90% of seeds fall within 45.7m of the parent tree (Burns, 1983). The mean and standard deviation are on the log-scale to form a right-skewed distribution because we expect seed dispersal to operate on smaller scales. We exponentiate to get the units back on the original scale in meters.
Nugget Proportion (unitless) $\sim N\left(\text{logit}(0.1), \frac{\text{logit}(0.4)-\text{logit}(0.01)}{4}\right)$	The proportion of variance described by small scale noise and measurement error should be small, less than 50%, because we are expecting to be able to detect large-scale spatial variation. We choose the mean nugget proportion to range from 1% to 40% with an average of 10%. We use the log-odds scale to restrict the nugget proportion to between 0 and 1.

Parameter (Unit) ∼ Distribution	Reasoning
Total CV (unitless) $\sim N\left(\log(0.25), \frac{\log(0.95)-\log(0.05)}{4}\right)$	Since it can be challenging to guess suitable CV values for seed dispersal, we performed a version of PPD by using prior draws from the seed count mean and CV distributions and each parameter pair was used to sample from the seed count distribution. For each sample size of 1000 from the seed count distribution, the 10% and 90% quantiles were computed. After 1000 prior draws, the density of these quantiles were visualized for different values of the CV location and scale hyperparameters. A Mean CV value of 0.25 with a range from 0.05 to 0.95 was chosen because quantile distributions captured reasonable seed count densities and ranges. This PPD check also confirmed that using a crude approximation by equating the CV to the variance of the log transformed data did not adversely affect the results. The PPD distributions can be seen in Figure 6.

Table 1: Seed Dispersal Priors

4.4 Establishment

Parameter (Unit) ∼ Distribution	Reasoning
Mean Establishment Probablity (unitless) $\sim N\left(\text{logit}(0.1), \frac{\text{logit}(0.5)-\text{logit}(0.01)}{4}\right)$	The probability of a seed establishing into a seedling should be less than 50% because trees disperse many more seeds than result in established plants. We specify 95% of the density to range from 1% to 50% on the probability scale with a mean of 10%. The log-odds scale is used to restrict the parameter to be in the probability range [0,1].
Matérn Shape (unitless) $\sim U(0.5, 4)$	We use the same prior as the seed dispersal system for the same reasons.

Parameter (Unit) $\sim$ Distribution	Reasoning
Matérn Reparameterized Scale (m) $\sim \exp(GN(\mu, \sigma))$ $\mu = \frac{\log(3)+\log(50)}{2}$ $\sigma^2 = \frac{\alpha\Gamma(3/\beta)}{\Gamma(1/\beta)} \approx 0.42$	The scale on which establishment operates is again restricted to our experimental design and a similar prior was formed. A GN was used so that the heaviness of the tails could be specified. A tail probability of 0.5% and centre probability of 55% was specified in the `get_gnorm` function. These choices led to computed values of $\alpha \approx 1.33$ and $\beta \approx 12.24$ in the GN which creates lighter tails and more density in the centre. As in Seed Matérn Scale we chose to set the 95% range of the data between 3m and 50m as in Figure 5. We exponentiate to get the units back on the original scale in meters.
Nugget Proportion (unitless) $\sim N\left(\text{logit}(0.1), \frac{\text{logit}(0.4)-\text{logit}(0.01)}{4}\right)$	We use the same prior from the seed dispersal system for the same reasons.
Total CV (unitless) $\sim N\left(\log(1), \frac{\log(3.5)-\log(0.5)}{4}\right)$	This prior is defined on the log-odds scale because we are parameterizing the standard deviation of establishment probability which is also defined on the log odds scale. A change of 1 on the log-odds scale is moderate, a change of 3.5 is large, and change of 0.5 is small. These values were selected to capture this knowledge and a PPD (as in the seed CV prior) was performed to verify that they produced reasonable distributions in Figure 6. The log transformation ensures this prior is always positive.

Table 2: Plant Establishment Priors

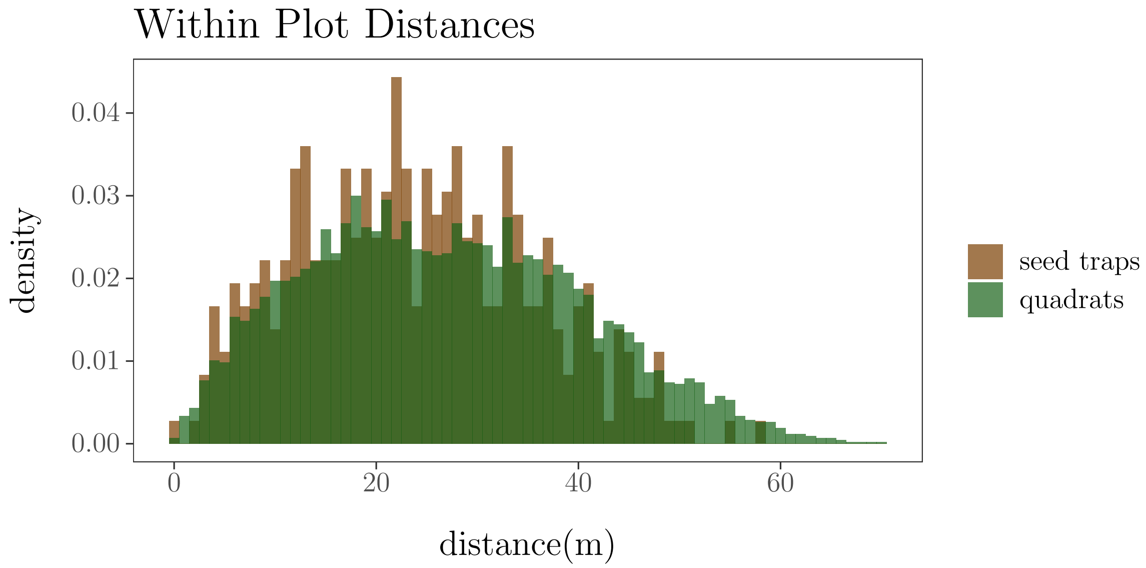

Figure 5: Distribution of distances among seedling quadrats and among seed traps within plots.

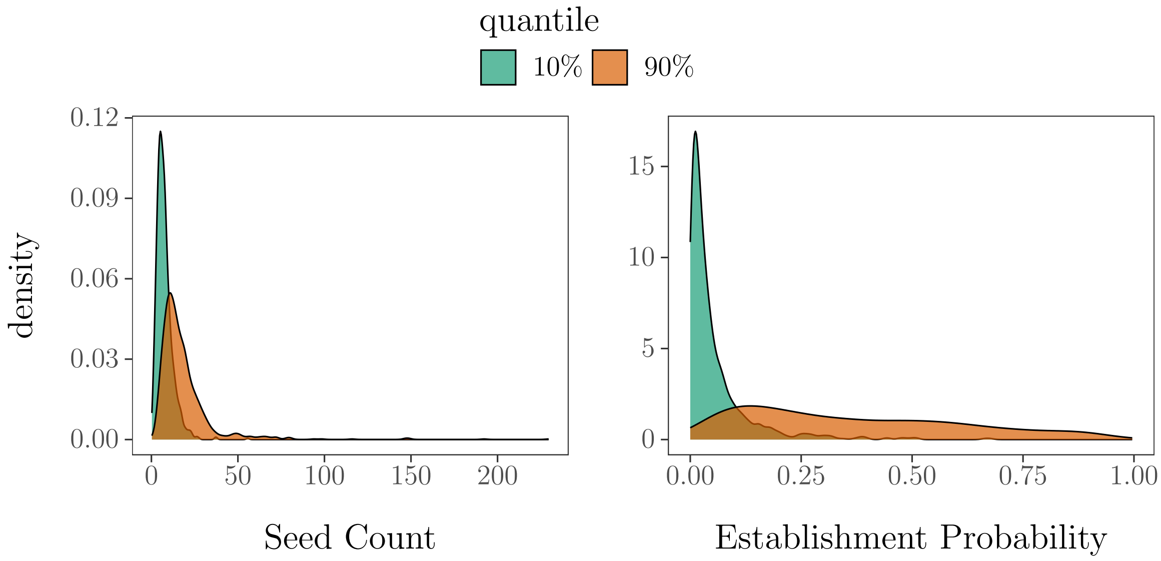

Figure 6: 1000 random prior draws from the mean and CV priors for each system. The 10% and 90% quantiles of the resulting seed count and establishment probability distribution were visualized to assess the appropriateness of the priors.

5 Simulations

The ABC algorithm can be implemented computationally in two steps. First a large database of prior draws and associated data sets is generated and saved. The ABC step is then performed by appropriate filtering of this database to arrive at a set of approximate posterior samples. Once the database is saved, the ABC step can be implemented many times over to validate the model. Since the prior draws are independent, the first step can be implemented using parallel computing. To save space, we only need to save the summary statistics associated instead of the data set itself.

We chose to simulate 1 million parameter and summary statistic pairs to have a large enough posterior sample size after a decision on the acceptance criteria was made (Section 6.1). The for-loop that generates this set of one million can be partioned into smaller chunks so the computational workload is distributed to individual CPU cores to compute, significantly reducing the total run time.

R has a number of packages that exploit parallelism. Experiments with parallezing the ABC for-loop were done using the R packages *doParallel* (Microsoft and Weston, 2022a) and *foreach* (Microsoft and Weston, 2022b). However, due to the communication overhead between parallel workers and master processes, it makes more sense to keep the code serial and parallelize via multiple serial runs.

The META-Farm package was chosen (Mashchenko, 2023) due to its ease of use and availability on all Digital Research Alliance of Canada high performance computing (HPC) research clusters. This package implements a form of computational task distribution called *job farming*. Here "job" refers to a computing task that can be allocated to a computing resource via a scheduling system called a *scheduler*. This package allows for *serial* job farming. In this way we execute a serial script and parallelize at the level of the scheduler instead of writing parallelized code.

A test run of 10,000 simulations was executed locally to estimate total wall clock time per simulation. Partitioning the set of 1 million iterations into 100 chunks of size 10,000 generated all results in a reasonable amount of time, on the order of several hours. Each chunk of the for-loop received a different pseudo random seed number to ensure no two chunks used the same sequence of prior draws. META-Farm handled how each chunk was distributed to individual computing resources on the Graham HPC cluster. This flexibility meant that chunks may be distributed to different cores to compute, or some chunks waited in queue until an available core was ready. Regardless, the serial farming led to significant speedup than generating all simulations serially. The META-Farm has additional features such as capturing job exit status and resubmitting failed jobs, which were useful when learning how to use the package. The post-script processing feature was used to aggregate all simulations from the 100 serial jobs.

6 Model Validation

6.1 Acceptance Rate Validation

Model validation can be performed by selecting a known parameter set θ^* from the joint prior, simulating data sets y^* from the the model using θ^* and assessing centrality and spread of the computed posterior values with θ^*. This method can also be applied to tune hyperparameters such as the ABC acceptance criteria (Lintusaari et al., 2016). Differences in average error between θ^* and the approximate posterior mean or mode determine how well centrality is captured (Lintusaari et al., 2016). Assessment of spread can be done by checking the coverage property (Lintusaari et al., 2016).

As mentioned in Section 5 our database of prior draws and summary statistic pairs can be used for model validation many times over as any one of these records can be used as our "observed" data set with known prior parameters before applying ABC to the remaining records.

Acceptance Rate	Number of Batches	Batch Size	Posterior Sample Size
0.01	1	100000	1000
0.01	5	20000	200
0.01	10	10000	100
0.01	50	2000	20
0.01	100	1000	10
0.05	5	20000	1000
0.05	25	4000	200
0.05	50	2000	100
0.05	250	400	20
0.05	500	200	10
0.1	10	10000	1000
0.1	50	2000	200
0.1	100	1000	100
0.1	500	200	20
0.1	1000	100	10

Table 3: Experimental design for tuning ABC acceptance rate.

The experimental design chosen to tune the acceptance rate and validate the model consisted of 10 randomly chosen parameter sets from a database of 1 million parameter and corresponding summary statistic pairs. The 1 million records were partitioned into 10 sets of 100,000 to match the number of observed parameter sets. Each set of 100,000 was further partitioned into batches of varying size with three acceptance rates being considered (1%, 5% and 10%) so that the size of the computed posterior was the same across acceptance rates. Table 3 gives the summary of the experimental design computed for each parameter set.

The Root Mean Square Relative Error $\left(\text{RMSRE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\frac{\theta_i-\theta^*}{\theta^*}\right)^2}\right)$ was used to validate the differences between the computed posterior samples and the observed parameter for each batch in the experimental design. The RMSRE was then averaged over each of the 10 sets and averaged for each of the three acceptance criterias (Figure 7).

RMSRE can be interpreted as approximately proportional change. Over the 10 parameters, 6 of them generated the lowest RMSRE at the 1% acceptance rate. The RMSRE was large (> 1) for half of the parameters; however, in these cases the 1% acceptance rate led to the lowest RMSRE for all parameters except establishment mean. The reasoning was sufficient to choose a 1% acceptance rate for ABC going forward.

Large RMSRE values could be due to the fact that each prior draw in the data base was used to generate one simulated data set from which summary statistics were computed. Some simulated data sets are less informative than others. Another approach that could prove to work better would be to generate multiple data sets for each parameter draw and average the summary statistics.

Coverage was computed using the same database of 1 million simulations and using 500 random draws as observed data. The 90% HDI credible interval was computed for each set of computed posterior samples. The percentage of time each true parameter was in the credible interval for each of the 500 draws is the computed percent coverage (Figure 8).

All but two parameters had a coverage in the 90% binomial confidence interval. Both establishment mean and seed CV had a higher percent coverage, indicating we are underestimating the coverage in these cases. This could be due to the fact that the random draws of observed parameters were slightly more concentrated in the higher density regions of these priors. Given Binomial($n = 10, p = 0.9$), we can compute the probability that eight or less parameters would lie in this confidence interval (P[$X \leq 8$] ≈ 0.26). This is a reasonably large probability, suggesting we could also have achieved this coverage result due to chance.

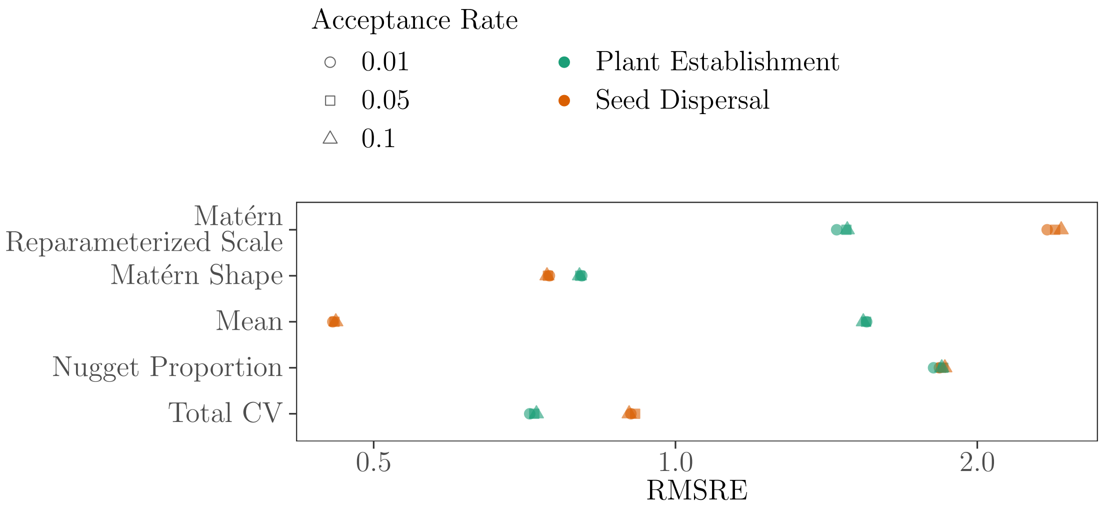

Figure 7: ABC acceptance rate validation results using RMSRE. The 1% acceptance rate (circular points) provided the lowest RMSRE for the majority of parameters, but within parameters RMSRE was relatively constant.

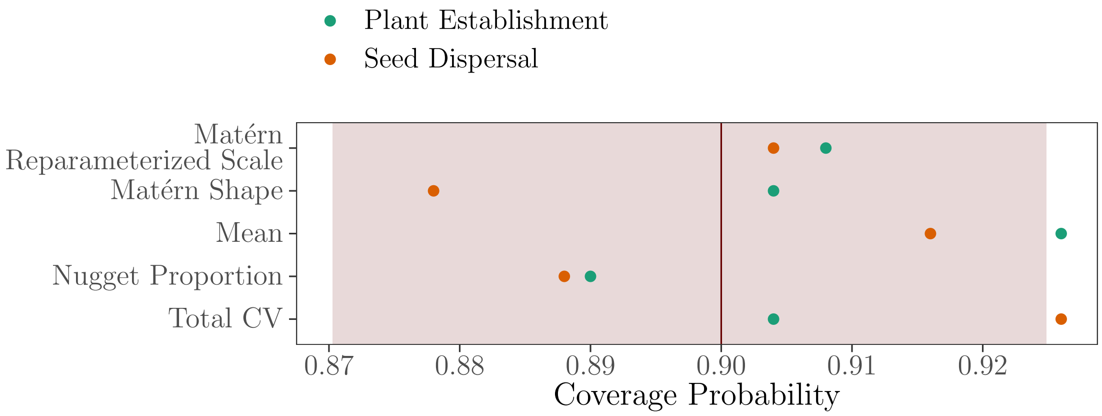

Figure 8: Coverage probability using 500 observed parameter and data sets with a computed posterior sample size of 100,000. The shaded region is the 90% Binomial confidence interval.

6.2 Simulation Based Calibration (SBC)

The problem with validation procedures that rely on defining a true parameter θ^* and simulating data from the model using θ^* to generate pseudo-observed data as done previously is that using one or multiple true parameters will not speak to how the model performs in general. The model may perform better for some subsets of the parameter space (Talts et al., 2020). The programmed algorithm may also contain bugs. To be confident that our algorithm is giving us a valid result we want to validate over the entire parameter space. In the Bayesian context the space of meaningful parameter values should be fully described by the prior.

Simulation Based Calibration (SBC) is a validation method that can be applied to any Bayesian computational model that computes approximate or true posterior samples (Talts et al., 2020). The only assumption of SBC is that we have a data generating model (Talts et al., 2020). SBC cannot detect if the model is appropriate for the system being described; this is up to the modeller (Talts et al., 2020).

Given a prior parameter draw $\tilde{\theta}$ and data generated from this draw and the model $\tilde{y}$ we define the *data averaged posterior*, $\pi(\theta|\tilde{y})\pi(\tilde{y}|\tilde{\theta})\pi(\tilde{\theta})$ (Talts et al., 2020). When integrated over all possible prior parameter draws and data sets, we get back the prior; $\pi(\theta) = \int \mathrm{d}\tilde{y}\, \mathrm{d}\tilde{\theta}\pi(\theta|\tilde{y})\pi(\tilde{y}|\tilde{\theta})\pi(\tilde{\theta})$. This self-consistency condition means we can use simulated data from our model to ensure it is sufficiently calibrated (Talts et al., 2020). Differences between the data-averaged posterior and the prior indicate inference problems (Talts et al., 2020).

Theorem 1 *Let $\tilde{\theta} \sim \pi(\theta)$, $\tilde{y} \sim \pi(y|\tilde{\theta})$, and $\theta_1, ..., \theta_L$ sampled independently from $\pi(\theta|\tilde{y})$ for any joint distribution $\pi(y, \theta)$. The rank statistic of any one-dimensional random variable over θ is uniformly distributed over the integers* $[0, L]$

Algorithm 2 SBC (Talts et al., 2020)

```
for n in N do
    Draw $\theta^* \sim \pi(\theta)$
    Simulate a data set $y^* \sim \pi(y|\theta^*)$
    for  i in L do
        Draw $\theta_i \sim \pi(\theta|y^*)$
    end for
    Compute the rank statistic $r_n(\{\theta_1, \ldots, \theta_L\}, \theta^*)$
end for
Plot a histogram of all rank statistics $r$ and assess for uniformity
```

The SBC procedure, based on Theorem 1, is defined in Algorithm 2. For multi-dimensional parameter sets θ^* the rank statistics are computed for each single parameter, so that we have one histogram for each individual parameter. Using b equal sized histogram bins, there is a $\frac{1}{b}$ probability of being in each bin. Given N rank statistics we can use the Binomial distribution, Binomial$(N, \frac{1}{b})$ to form confidence intervals to help assess uniformity. One common choice is to use $b = L + 1$ bins.

The drawback of SBC is the computational resources required to compute N rank statistics (Talts et al., 2020). Since all the steps are independent this algorithm can be implemented in parallel, reducing the computational time (Talts et al., 2020).

Non-uniform patterns in the histogram can help diagnose bias or mis-calibrated posteriors (Talts et al., 2020). When the histogram is weighted to extreme large and small values, this can indicate correlation among posterior samples (Talts et al., 2020). We do not expect to observe this pattern for this data because rejection ABC, unlike MCMC methods, uses independent samples from the prior. $\cup$-shaped and $\cap$-shaped histograms indicate underdispersion and overdispersion between the prior and data-averaged posterior (Talts et al., 2020). This means on average the computed posterior will be narrower or wider than the true posterior (Talts et al., 2020). Histograms that are skewed to one side indicate bias in the opposite direction between the computed and true posterior (Talts et al., 2020).

The initial SBC run showed $\cap$-shaped histograms for some parameters indicating over-dispersion. This led us to revisit how scaling was performed on the summary statistics in ABC. After a different scaling was done, by dividing by the standard deviation, SBC was re-run and the results are show in Figure 9.

If the model passes the SBC procedure, with a 99% Binomial confidence interval we expect that less than 1 bin on average will stray from this interval. The results dramatically improved from the previous run, although we still observe deviations from uniformity for most parameters. In particular the scale parameters and seed CV show some $\cap$-shape and possible skew. We decided to accept this level of overdispersion and bias in the model going forward.

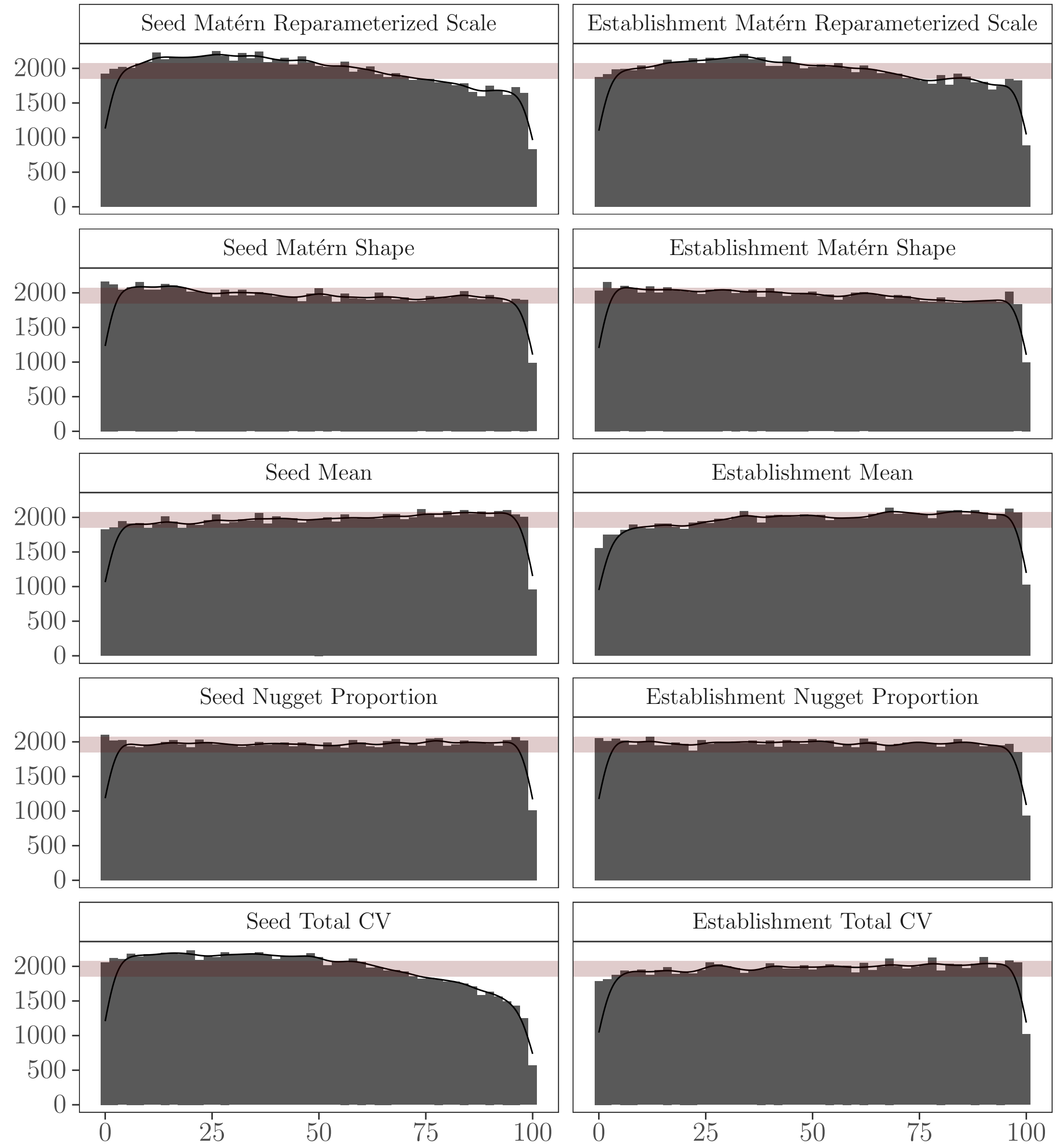

Figure 9: SBC was performed for $N = 100000$, $L = 100$, $b = 51$ and a 99% Binomial confidence interval is shown by the horizontal band. Establishment scale, seed scale and seed CV show deviations from uniformity with characteristic ∩-shapes indicating overdispersion and right skew indicating bias.

7 Parameter Estimates

7.1 ABC

The prior and posterior parameter distributions from the observed data are shown in Figure 10 with summary information given in Table 4. In all parameters, except possibly seed nugget proportion, the data is informing our estimates as we see obvious differences between priors and posteriors.

Parameter	Mean	Median	95% Credible Interval
Seed Matérn			
Reparameterized Scale	15.6	12.9	(2.45, 36.9)
Seed Matérn Shape	2.02	1.90	(0.50, 3.78)
Mean Seed Count	15.5	13.7	(3.52, 31.3)
Seed Nugget Proportion	0.13	0.09	(0.003, 0.36)
Seed CV	0.27	0.25	(0.06, 0.538)
Establishment Matérn			
Reparameterized Scale	18.1	15.6	(3.49, 38.1)
Establishment Matérn Shape	1.51	1.20	(0.50, 3.48)
Mean Establishment Probability	0.18	0.14	(0.01, 0.46)
Establishment Nugget Proportion	0.10	0.08	(0.003, 0.27)
Establishment CV	1.44	1.32	(0.42, 2.75)

Table 4: ABC posterior statistics. These should not be viewed as point estimates but as posterior summary information from Figure 10.

7.2 Frequentist Method

To cross check our Bayesian estimates we estimate the seed dispersal parameters using the classic frequentist approach. We expect to be able to estimate seed parameters because we can view our seed count data set as a random sample from a population. In contrast, we do not have a data set of establishment probabilities because observed seed and seedling data were not collected from the same locations. In ABC, we simulate seed counts at seedling locations so we can infer establishment probabilities.

To model seed dispersal we closely follow our geostatistical model by fitting a Poisson GLMM on the seed count data with a log-link function. We incorporate the unobserved seed dispersal signal $S(x)$ as a zero-mean multivariate Gaussian with Matérn

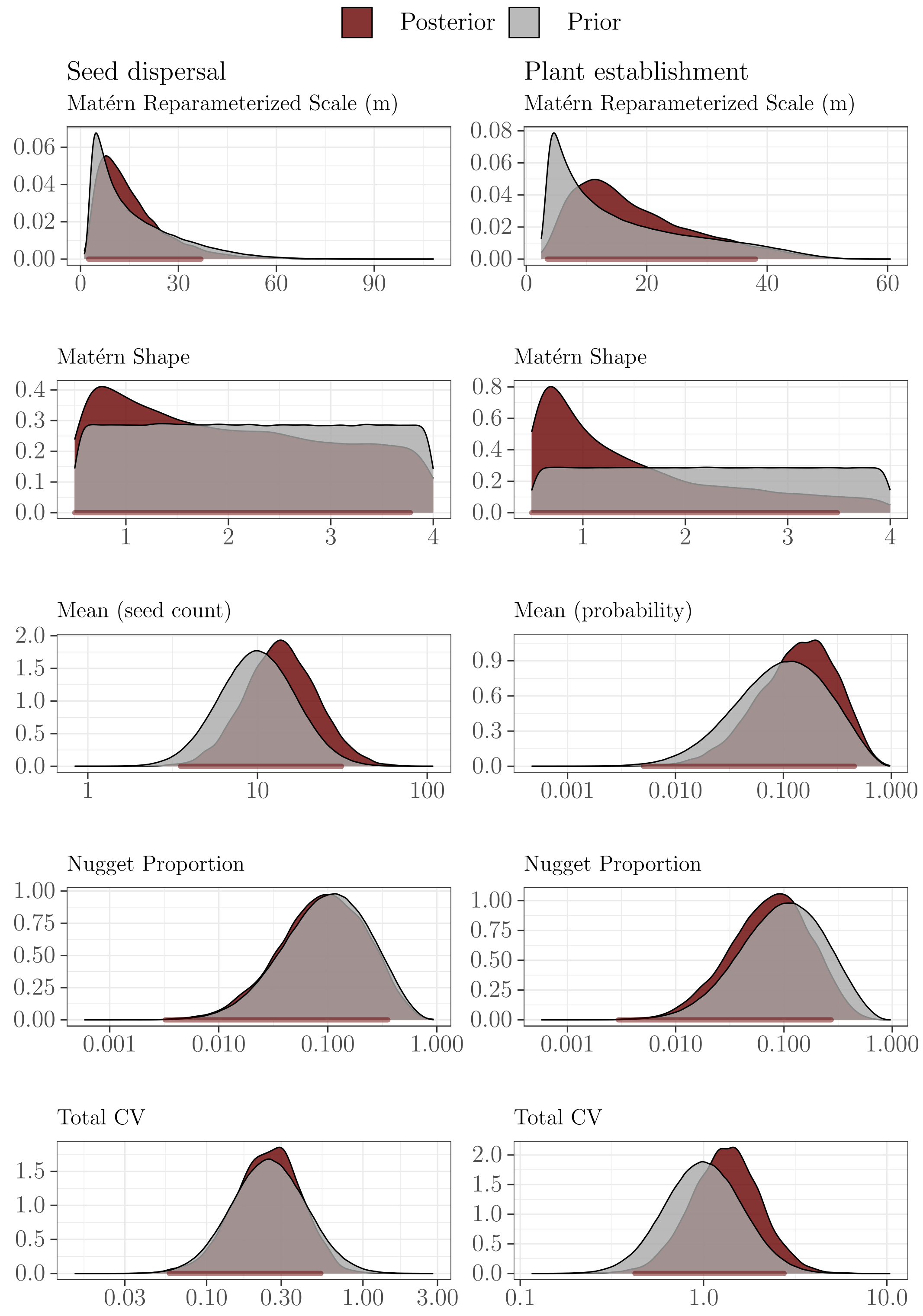

Figure 10: Priors and posteriors using observed data with 95% posterior credible intervals shown by the horizontal bar (prior sample size = 1,000,000, posterior sample size = 100,000).

covariance structure as a random effects. We create a seed plot grouping variable to ensure we only consider within plot covariance, as done previously. A nugget effect becomes a random effect by treating each individual observation as its own group. The intercept, mean seed count, is the only fixed term and the random effects are independent of this.

Using Maximum Likelihood Estimation (MLE) provided by the *glmmTMB* R package (Brooks et al., 2017), we arrive at the parameter estimates and 90% confidence intervals given in Table 5. We again use the approximation to interpret the standard deviation for the log transformed data as the CV on the original scale.

Parameter	Estimate	95% Confidence Interval
Matérn Scale	356.42	(9.31, 13641.05)
Matérn Shape	0.22	(0.08, 0.55)
Mean Seed Count	7.79	(4.42, 13.74)
Nugget Variance	3.64e-4	NA
Seed CV	0.99	(0.67, 1.45)

Table 5: MLE seed dispersal parameter estimates

Further exploration using the *profile log-likelihood* was done to investigate the Matérn Scale estimate in Table 5. The estimate is much larger than the maximum within seed plot distance of approximately 60m and the confidence interval is very wide. The profile log-likelihood allows us to fix a value for the Matérn Scale and optimize the likelihoood over all other parameters (Kreutz et al., 2013). By considering a range of scale parameters, we can reduce the likelihood to a one dimensional vector to investigate the identifiability of this parameter (Kreutz et al., 2013).

The profile likelihood, computed using the *TMB* package in R, is shown in Figure 11 (Kristensen et al., 2016). The limited range of negative log-likelihood values suggest that although there exists a global minimum for the scale parameter at 356.42 m, the likelihood is likely relatively flat in this region possibly making optimization challenging.

The nugget variance is very small and had a very large standard error suggesting we are learning little about this parameter. Fitting the profile likelihood for this parameter in Figure 12 we see it is flat for nearly all values suggesting parameter identifiability issues.

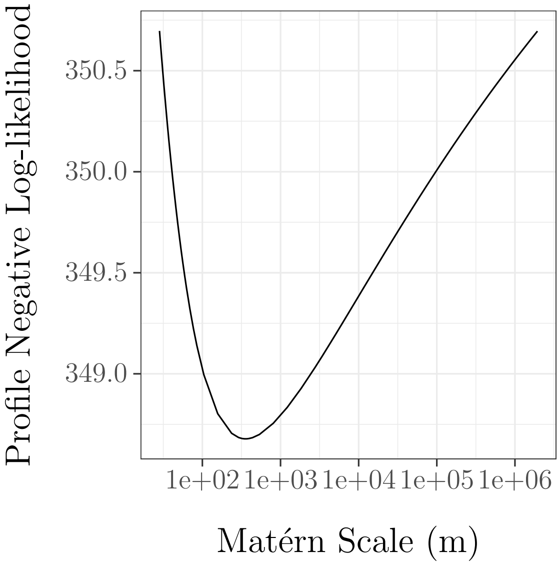

Figure 11: Profile likelihood for a range of Matérn Scale parameters.

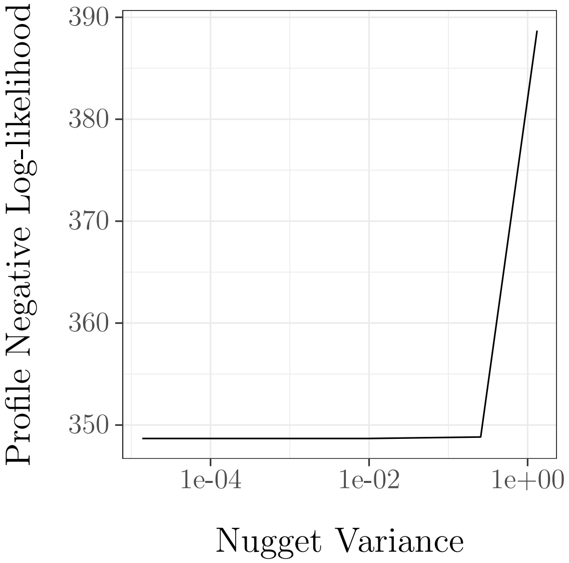

Figure 12: Profile likelihood for a range of nugget variances.

8 Discussion

The majority of the confidence intervals (Table 5) are narrower than the credible intervals (Table 4), but this is not suprising because Bayesian uncertainty is propagated from all known sources (priors and data) (Doll and Feiner, 2022). ABC can also create wider interval estimates because we lose information by summarizing the data (Lines et al., 2020). The obvious advantage of the ABC method is we learn more about our parameters through probability distributions instead of point estimates and when frequentist methods can not be used (establishment parameters).

As with the Matérn scale parameter, identifying an optimal set of parameters can end up in unrealistic regions in the parameter space. In contrast, the incorporation of priors in any Bayesian inference method prevents unrealistic parameter estimates, even if the data is not informative enough. The ABC Reparameterized Matérn scale estimates show we are learning about the size of seed density patches, with most ranging from 2.5 to 37 m. Environmental filtering effects are detected on slightly larger scales, from 3.5 to 38 m.

Matérn shape parameters are often arbitrarily fixed in advance because it is assumed the data provides little information about them (De Oliveira and Han, 2022). There is also evidence that jointly estimating all Matérn parameters (shape, scale and signal variance) using MLE can have identifiability issues "leading to ridges or plateaus in the log-likelihood surface" (Diggle and Ribeiro, 2007). Diggle and Ribeiro (2007) suggests the profile likelihood can be used to decide on a shape value from a small candidate list. From the Bayesian perspective, we note it is difficult to hypothesize ecologically about the smoothness of these processes when forming priors. Despite these challenges, in both estimation procedures we learn about the shape parameter with ABC being more informative. De Oliveira and Han (2022) shows the amount of information we learn about the shape parameter increases when we have a less regular sampling design (our case) and when strong correlation is present (larger Matérn scales). The credible intervals incorporate much of the prior range but the drastic differences in distribution shape between priors and posteriors suggest our underlying surfaces are more rough than smooth. This is an important result as we hope to convince readers that this inferential method can lead to models that make more realistic predictions.

The estimates for the mean seed count are modelled on the log seed count scale and we exponentiate the individual frequentist estimate and the entire Bayesian posterior distribution to view on the scale of ecological interest. The mean point estimate (Table 5) is really the geometric mean of seed counts which is smaller than the arimetic mean (Table 4). This highlights another advantage to Bayesian posteriors because we can transform entire samples instead of worrying if our point estimate is invariant to transformations.

We learn about mean establishment probability with a 95% credible interval range from 1 to 46%.

We had challenges estimating the proportion of variance attributed to the nugget effect using MLE and additional investigation is required here. The ABC results suggest the seed data has limited information about the seed dispersal nugget effect. We do learn about the nugget in the establishment system perhaps because we have more data (seed and seedling counts) informing our model as opposed to the seed dispersal system, and this noise becomes more detectable.

The MLE seed CV estimate of approximately 1 is a plausible ecological value for spatial variability in seed counts given an estimate of approximately 8 for the mean of seed counts. The ABC seed CV estimate incorporates all levels of variation including the nugget effect so we cannot make a direct comparison to the MLE estimate. We do however see an increase in relative variability from the seed dispersal system to the establishment system.

The results show, even with a relatively small data set, ABC can provide valuable ecologically meaningful parameter estimates. We advocate for its use because it is relatively straightforward and comes with the informative power of Bayesian inference. ABC can also come to the rescue when likelihood-based methods fail or can not be applied. Our method highlights the flexibility of geostatistical models and ABC to investigate spatial ecological problems.

9 Public Access

All data and code for this research has been made publicly available at doi:10.5281/zenodo.8002390.

Bibliography

Asefa, M., Cao, M., Zhang, G., Ci, X., Li, J., and Yang, J. (2017). Environmental filtering structures tree functional traits combination and lineages across space in tropical tree assemblages. *Scientific Reports*, 7(1):132.

Beaumont, M. A. (2010). Approximate Bayesian Computation in Evolution and Ecology. *Annual Review of Ecology, Evolution, and Systematics*, 41(1):379–406.

Beaumont, M. A., Zhang, W., and Balding, D. J. (2002). Approximate Bayesian Computation in Population Genetics. *Genetics*, 162(4):2025–2035.

Berger, J. (2006). The case for objective Bayesian analysis. *Bayesian Analysis*, 1(3).

Bjørnstad, O. N. (2022). *ncf: Spatial Covariance Functions.* R package version 1.3-2.

Bjørnstad, O. N. and Falck, W. (2001). Nonparametric spatial covariance functions: estimation and testing. *Environmental and Ecological Statistics*, 8:53–70.

Bolker, B. M. (2003). Combining endogenous and exogenous spatial variability in analytical population models. *Theoretical Population Biology*, 64(3):255–270.

Bolker, B. M. (2022). Power exponential prior: R. Available at `https://github.com/bbolker/bbmisc/blob/master/Rmisc/powexp_prior.R`.

Bolker, B. M. and Pacala, S. W. (1999). Spatial Moment Equations for Plant Competition: Understanding Spatial Strategies and the Advantages of Short Dispersal. *The American Naturalist*, 153(6):575–602.

Brooks, M. E., Kristensen, K., van Benthem, K. J., Magnusson, A., Berg, C. W., Nielsen, A., Skaug, H. J., Maechler, M., and Bolker, B. M. (2017). glmmTMB balances speed and flexibility among packages for zero-inflated generalized linear mixed modeling. *The R Journal*, 9(2):378–400.

Burns, R. M. (1983). Silvicultural Systems for the Major Forest Types of the United States. *U.S. Department of Agriculture Forest Service Agriculture Handbook No. 445.*

Crome, F., Thomas, M., and Moore, L. (1996). A novel Bayesian approach to assessing impacts of rain forest logging. *Ecological Applications*, 6(4):1104–1123.

Csillery, K., Francois, O., and Blum, M. G. B. (2012). abc: An R package for Approximate Bayesian Computation (ABC). *Methods in Ecology and Evolution.*

De Oliveira, V. and Han, Z. (2022). On Information About Covariance Parameters in Gaussian Matérn Random Fields. *Journal of Agricultural, Biological and Environmental Statistics*, 27(4):690–712.

Diggle, P. and Ribeiro, P. J. (2007). *Model-based geostatistics*. Springer series in statistics. Springer, New York, NY. OCLC: ocm71284654.

Doll, J. and Feiner, Z. (2022). Think like a Bayesian and avoid pitfalls from our frequentist past. *Ideas in Ecology and Evolution*, 15.

Ellison, A. M. (1996). An introduction to Bayesian inference for ecological research and environmental decision-making. *Ecological applications*, 6(4):1036–1046.

Gelfand, A. E. and Banerjee, S. (2017). Bayesian Modeling and Analysis of Geostatistical Data. *Annual Review of Statistics and Its Application*, 4(1):245–266.

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

He, D. and Biswas, S. R. (2019). Negative relationship between interspecies spatial association and trait dissimilarity. *Oikos*, 128(5):659–667.

Kreutz, C., Raue, A., Kaschek, D., and Timmer, J. (2013). Profile likelihood in systems biology. *FEBS Journal*, 280(11):2564–2571.

Kristensen, K., Nielsen, A., Berg, C. W., Skaug, H., and Bell, B. M. (2016). TMB: Automatic differentiation and Laplace approximation. *Journal of Statistical Software*, 70(5):1–21.

Lambert, P. C., Sutton, A. J., Burton, P. R., Abrams, K. R., and Jones, D. R. (2005). How vague is vague? A simulation study of the impact of the use of vague prior distributions in MCMC using WinBUGS. *Statistics in Medicine*, 24(15):2401–2428.

Lemoine, N. P. (2019). Moving beyond noninformative priors: why and how to choose weakly informative priors in Bayesian analyses. *Oikos*, 128(7):912–928.

Levin, S. A. (1992). The Problem of Pattern and Scale in Ecology: The Robert H. MacArthur Award Lecture. *Ecology*, 73(6):1943–1967.

Lewontin, R. C. (1966). On the measurement of relative variability. *Systematic Zoology*, 15(2):141–142.

Leys, C., Ley, C., Klein, O., Bernard, P., and Licata, L. (2013). Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median. *Journal of Experimental Social Psychology*, 49(4):764–766.

Lines, E., Zavala, M., Ruiz-Benito, P., and Coomes, D. (2020). Capturing juvenile tree dynamics from count data using Approximate Bayesian Computation. *Ecography*, 43(3):406–418.

Lintusaari, J., Gutmann, M. U., Dutta, R., Kaski, S., and Corander, J. (2016). Fundamentals and Recent Developments in Approximate Bayesian Computation. *Systematic Biology*, page syw077.

Mashchenko, S. (2023). Meta-farm. Available at `https://docs.alliancecan.ca/wiki/META-Farm`.

Microsoft and Weston, S. (2022a). *doParallel: Foreach Parallel Adaptor for the 'parallel' Package*. R package version 1.0.17.

Microsoft and Weston, S. (2022b). *foreach: Provides Foreach Looping Construct*. R package version 1.5.2.

Mod, H. K., Chevalier, M., Luoto, M., and Guisan, A. (2020). Scale dependence of ecological assembly rules: Insights from empirical datasets and joint species distribution modelling. *Journal of Ecology*, 108(5):1967–1977.

North, A. and Ovaskainen, O. (2007). Interactions between dispersal, competition, and landscape heterogeneity. *Oikos*, 116(7):1106–1119.

Pritchard, J. K., Seielstad, M. T., Perez-Lezaun, A., and Feldman, M. W. (1999). Population growth of human Y chromosomes: a study of Y chromosome microsatellites. *Molecular Biology and Evolution*, 16(12):1791–1798.

Raduła, M. W., Szymura, T. H., Szymura, M., Swacha, G., and Kacki, Z. (2020). Effect of environmental gradients, habitat continuity and spatial structure on vascular plant species richness in semi-natural grasslands. *Agriculture, Ecosystems & Environment*, 300:106974.

Robert, C. P. (2016). Approximate Bayesian Computation: A Survey on Recent Results. In Cools, R. and Nuyens, D., editors, *Monte Carlo and Quasi-Monte Carlo Methods*, volume 163, pages 185–205. Springer International Publishing, Cham. Series Title: Springer Proceedings in Mathematics & Statistics.

Sarma, A. and Kay, M. (2020). Prior Setting in Practice: Strategies and Rationales Used in Choosing Prior Distributions for Bayesian Analysis. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–12, Honolulu HI USA. ACM.

Seabloom, E. W., Bjørnstad, O. N., Bolker, B. M., and Reichman, O. J. (2005). Spatial Signature of Environmental Heterogeneity, Dispersal, and Competition in Successional Grasslands. *Ecological Monographs*, 75(2):199–214.

Sisson, S. A., Fan, Y., and Beaumont, M. (2018). *Handbook of Approximate Bayesian Computation*. CRC Press.

Talts, S., Betancourt, M., Simpson, D., Vehtari, A., and Gelman, A. (2020). Validating Bayesian Inference Algorithms with Simulation-Based Calibration. arXiv:1804.06788 [stat].

Tavaré, S., Balding, D. J., Griffiths, R. C., and Donnelly, P. (1997). Inferring Coalescence Times From DNA Sequence Data. *Genetics*, 145(2):505–518.