A Scalable Algorithm for Dispersing Population *

Sathish Govindarajan † Mike Dietze ‡ Pankaj K. Agarwal § James S. Clark ¶

Abstract

Models of forest ecosystems are needed to understand how climate and land-use change can impact biodiversity. In this paper we describe an ecological dispersal model developed for the specific case of predicting seed dispersal by trees on a landscape for use in a forest simulation model. We present efficient approximation algorithms for computing seed dispersal. These algorithms allow us to simulate large landscapes for long periods of time. We also present experimental results that (i) quantify the inherent uncertainty in the dispersal model and (ii) describe the variation of the approximation error as a function of the approximation parameters. Based on these experiments, we provide guidelines for choosing the right approximation parameters, for a given model simulation.

*Work has been supported by NSF under grants CCR-00-86013 EIA-98-70724, EIA-99-72879, EIA-01-31905, and CCR-02-04118 and by a grant from the U.S.-Israeli Binational Science Foundation.

†Department of Computer Science, Box 90129, Duke University, Durham, NC 27708, USA. E-mail: gsat@cs.duke.edu

‡Department of Biology, Duke University, Durham, NC 27708, USA. E-mail: mcd7@duke.edu

§Department of Computer Science, Box 90129, Duke University, Durham, NC 27708, USA. E-mail: pankaj@cs.duke.edu

¶Department of Biology, Duke University, Durham, NC 27708, USA. E-mail: jimclark@duke.edu

1 Introduction

Motivation. The movement of organisms, or dispersal, determines biodiversity and how species will respond to climate change and habitat loss. Predicting dispersal can be computationally challenging, because it involves interactions between individual organisms and factors that control change in a population's range. The effect of dispersal on species interactions can involve calculations at fine spatial scales. For trees, most seeds fall close to the parent [12, 4] and thus tends to promote spatial aggregation. Migration in response to habitat loss and climate change may depend on rare long-distance dispersal events (kilometers) [5]. A prediction grid, or map, of dispersing organisms involves a number of calculations that is quadratic in map area or population size. This quadratic relationship results from calculating the contribution of dispersing propagules from each reproductive individual to each location on the landscape.

Not only landscape size, but also the complexity of the process itself challenges dispersal prediction. Our example comes from forest trees, but the challenges apply to any population with individual movement. Complexity may depend on densities of different species, the heterogeneous arrangement of adults, the variability among trees in terms of seed production, and the stochastic nature of the dispersal process itself. Spatial and temporal variability in seed dispersal can have a number of ecological consequences. For example, variability in seed production determines how populations respond to environmental fluctuations [3]. Some species may exploit transient climatic conditions conducive to establishment or rare dispersal to favorable habitats. Through periodic variation in dispersal, a phenomenon known as masting, trees can temporarily satiate their predators, resulting in greater mean survival than would occur from constant seed production. These effects that result from different sources of stochasticity suggest that dispersal prediction must accommodate a range of processes.

To be useful to scientists and managers, forest models must be sufficiently detailed to capture processes that affect the establishment of trees and yet sufficiently broad to admit landscape and atmospheric processes. This is particularly true for dispersal, where fine scale temporal and spatial variability is important to understanding key ecological processes like the maintenance of biodiversity and migration. However, the important effects of dispersal unfold over large spatial scales and over a long time.

In this paper we present a dispersal model developed for the specific case of predicting seed dispersal by trees. Our dispersal model is part of a individual-based, spatially explicit forest model [9]. However, the principles developed here are general to many dispersal problems faced by ecologists.

Related work Within the context of forest models, concern about dispersal ability and seed availability is a relatively recent phenomenon. Early forest models assumed that there was a constant rain of seeds from all species, regardless of whether those species were actually present in a stand or the surrounding area . Species do not go extinct in such models, because seedlings appear even if there are no adults to produce them. Later models included seed production by adults, but ignored spatial arrangement or dispersal capacity. Clearly, neither assumption is appropriate for understanding spatial and temporal heterogeneity in dispersal, the impact of long-distance dispersal, or the ability of forests to migrate in response to climate change or anthropogenic disturbance. The SORTIE forest model [11] was the first forest simulator to include dispersal explicitly. The algorithm used in the SORTIE model involves drawing a dispersal distance for each propagule and from a dispersal kernel. If there are n individuals each producing an average of p propagules, this requires $O(np)$ calculations. If the number of propagules is small then this approach can be much faster than the alternative pair-wise dispersal calculation between n individuals to all locations on a landscape of area A, i.e., $O(nA)$ calculations. In the SORTIE model p is small, because seeds are ignored, and "dispersal" involves a small number of established saplings. Whereas seed production rates of 10^6 seeds per year is not uncommon for many tree species [6], only a few may survive to the sapling stage. In order to explicitly model interactions that occur between seed and sapling stage, we needed to model dispersal of seed itself. Our algorithms are developed to address the more challenging goal of dispersal, where the number of propagules can be large. It applies not only to forests, but also to ecological 'invasions', where an introduced species might spread rapidly as a result of high seed production and dispersal capacity.

Our approach. We develop efficient approximation algorithms based on a full parameterization of uncertainty and variability in seed production and dispersal. These algorithms allow approximation errors that speed computation. The degree of acceptable error is gauged through explicit comparison with different sources of stochasticity parameterized from data. Detailed description of statistical computation is presented in [6, 7].

We further exploit spatial coherence to design efficient algorithms for dispersal calculations. We use a hierarchical data structure to represent the forest at various spatial scales. Using the multi-resolution nature of the quad tree [8, 13], we make spatial approximations, depending on the required accuracy. To compute dispersal, we use the Monopole approximation [1] to aggregate seed dispersal from distant trees. This yields an efficiency-accuracy tradeoff scheme to compute dispersal.

Our results. For reasonable error, our algorithm achieves a speedup of an order of magnitude. We have performed a series of experiments that quantify the stochasticity in the dispersal process. We have also performed a series of experiments that evaluate the variation of approximation error with the approximation parameter. Based on these experiments, we provide guidelines that help the user to choose the appropriate approximation parameter for a given forest simulation.

2 Overview of Our Model

We first give a brief overview of our forest model. A detailed description of this model can be found in [9]. The forest consists of a landscape $\mathbb{L}$ and a population of trees; the latter is modeled as a family of densities and a family of individuals. The landscape remains fixed over time but the population evolves with time.

Landscape. Our model considers the landscape $\mathbb{L}$ of the forest as a planar region whose boundary is a closed polygonal curve. The area of the landscape varies from a few hectares to few hundreds

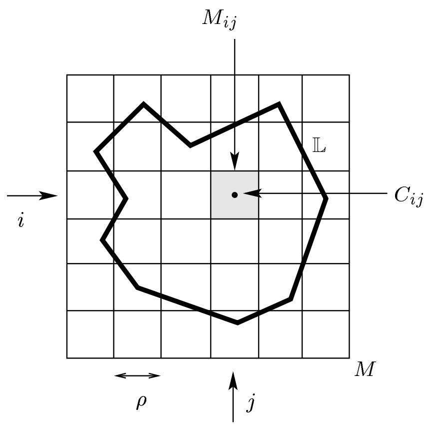


Figure 1: Landscape $\mathbb{L}$ and the underlying mesh M.

of hectares. We discretize $\mathbb{L}$ by enclosing it with a square and overlaying a uniform grid (mesh) M. Each grid cell M_{ij} of M is a square with side length ρ; we refer to ρ as the *resolution* of M. We use C_{ij} to denote the center of M_{ij}, and we associate an elevation (height) $z_{ij} \in \mathbb{R}$ with C_{ij}. By interpolating the heights at other points $\mathbb{L}$, we can view $\mathbb{L}$ as a terrain. We could also associate various geological and urban features such as rivers, lakes, roads, etc. with $\mathbb{L}$. Figure 1 shows an example of a landscape alongwith the underlying mesh.

Population. Our model is hybrid in the sense that we use both densities and individuals to model the population of the forest. The early stages of trees are modeled as densities, and after some growth, they are modeled as individuals with unique physical attributes. More precisely, we classify the population into five stages: *seed, yearling, seedling, sapling,* and *adult* (Figure 2). We further refine the stage seed into *seed rain* and *seed bank* — the former representing the seeds that are dispersed by trees and the latter representing the ones that are on the ground. The seeds that have germinated are called yearlings. We model seed rain, seed bank and yearling as densities, as they do not have any geometric attributes and all of them with the same species are identical. We assume that the density is uniform within each grid cell.

We model the next three stages — seedling, sapling, adult — as families of individuals. Each individual X has a physical location $\ell(X) \in \mathbb{R}^2$ and various physical attributes. Currently, we model each individual as a cylindrical trunk and a cylindrical crown sitting on top of the trunk; see Figure 3. Let $D_t(X), H_t(X)$ denote the diameter and height of the trunk at time t. The diameter and the height of the crown depend on $D_t(X)$, and the relationship can be found in [7]. An individual X is a seedling if $H_t(X) \leq \tau_D$, a sapling if $\tau_D \leq H_t(X) \leq \tau_P$, and an adult if $H_t(X) \geq \tau_P$, where τ_D and τ_P are threshold parameters.

Dynamics. The dynamics of our forest model consists of three parts — establishment of individuals, growth, and mortality. Individuals are established by dispersal of seeds. The adult trees

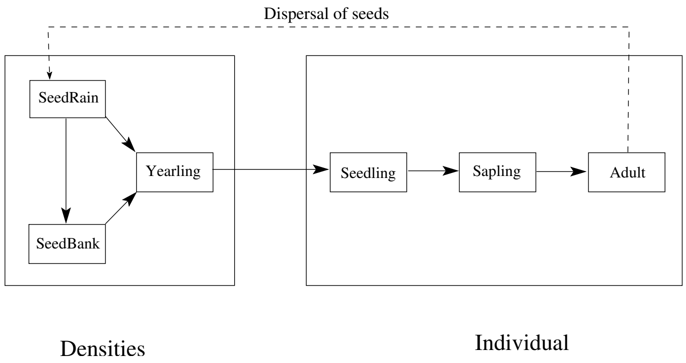


Figure 2: Evolution of densities of stages seed and yearling and growth of an individual from a seedling to an adult.

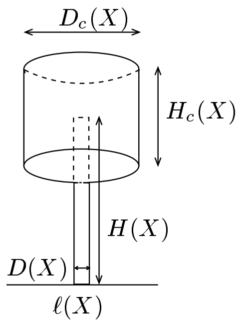


Figure 3: Geometric model of an individual.

produce seeds depending on $D_t(X)$ (also known as *diameter at breast height*) and these seeds are dispersed based on a dispersal kernel. The dispersal kernel accounts for both short and long distance dispersal. Growth of each of the stages is calculated based on resource availability and local density. Individuals are promoted from one stage to next based on the growth thresholds. An individual dies at the current time based on its *mortality probability*. The mortality probability is calculated based on the individual's growth suppression and natural disturbances.

Resources. The forest contains several resources like light, moisture, nitrogen, etc which are vital for the growth of the individuals in the forest. We model each resource as a separate submodel. Light is considered as one of the main resources in our model. We develop a sophisticated light model based on Casetti [2] to calculate the availability of understory light at each gridcell. Since the light model is computationally intensive, we use graphics hardware to accelerate the light computation.

3 Dispersal Model

The dispersal model determines the number of seeds that disperse into each grid cell of M. This quantity depends on:

- the number of seeds produced by each individual, denoted as *fecundity*.
- spatial distribution of seeds, which is defined by the *dispersal kernel*.

Fecundity and dispersal kernel are parameterized from study areas in the Duke Forest and Southern Appalachians. For each species, all parameters are simultaneously estimated using a hierarchical Bayesian framework [7].

Fecundity. The reproductive output of an individual is nonzero only if it is a female and reproductively mature. We choose a functional form of the fecundity composed of factors that depend on the species to which that individual belongs, the size of the individual, and a factor that captures the temporal variability. More precisely, $\beta_t(X)$, the fecundity of individual X at time t, has the following form:

$$\beta_t(X) = \chi(X) \cdot \Delta_t(X) \cdot 10^{a_0 + b(X) + \epsilon_t(X)} \cdot (D_t(X))^{a_1}, \tag{1}$$

where a_0 and a_1 are species-specific scaling parameters and $D_t(X)$ is the diameter of the trunk of individual X at time t. The functions $\chi(X)$ and $\Delta_t(X)$ are indicator functions, indicating the gender and reproductive maturity of X, respectively.

$$\chi(X) = \begin{cases} 1 & \text{if individual } X \text{ is female,} \\ 0 & \text{if individual } X \text{ is male.} \end{cases}$$

and

$$\Delta_t(X) = \begin{cases} 1 & \text{if } D_t(X) > \gamma(X), \\ 0 & \text{if } D_t(X) \leq \gamma(X). \end{cases}$$

$$\gamma(X) \sim \mathrm{Gamma}(m_0, m_1),$$

Here $\mathrm{Gamma}(m_0, m_1)$ is the Gamma distribution with species-specific maturity parameters m_0, m_1. $b(X)$ is an individual scaling parameter defined as:

$$b(X) \sim \mathrm{Normal}(0, \tau^2),$$

where $\mathrm{Normal}(0, \tau^2)$ is the Normal distribution with species-specific parameter τ. Finally, $\epsilon_t(X)$ is a temporally autocorrelated Gaussian stochastic process, defined as:

$$\epsilon_t(X) \sim \mathrm{Normal}(\nu \cdot \epsilon_{t-1}(X), \sigma^2),$$

where ν and σ are species-specific parameters.

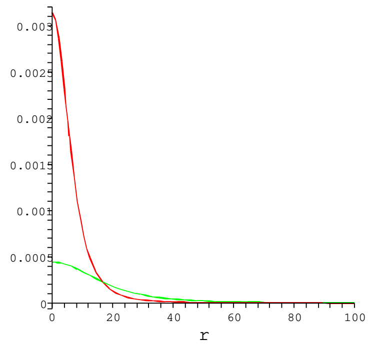


Figure 4: Dispersal kernel for parameters of species *Acer rubrum* p=1, u=101.3; *Liriodendron tulipifera* p=1, u=719.8. Parameter values are from Clark et al [7].

Dispersal Kernel. The dispersal kernel describes the spatial distribution of the scattering of seeds in the vicinity of the parent plant, as a function of distance r. We use a bivariate Student's t-distribution for the dispersal kernel, which has the following form:

$$f(r;u,p) = \frac{p}{\pi u \left[1+\frac{r^2}{u}\right]^{p+1}},$$

where p and u are species specific parameters. Figure 4 shows the graph of the dispersal kernel for the parameters of two species: *Acer rubrum* and *Liriodendron tulipifera* (Clark et al. [7]).

The actual number of seeds dispersed into the grid cell M_{ij}, denoted as s_{ijt} is drawn from a Poisson distribution

$$s_{ijt} \sim \text{Poisson}(q_t(C_{ij}) \cdot \rho^2),$$

where ρ is the side length of grid cell M_{ij} and $q_t(y)$ is the expected seed density in location y at time t,

$$q_t(y) = \sum_X \beta_t(X) \cdot f(\|y - \ell(X)\|; u, p), \tag{2}$$

where the sum is taken over all the individuals in the forest.

The functional form of the dispersal kernel was chosen from a number of potential functions using formal statistical techniques for model comparison (Clark *e*t al [4]). Dispersal parameters for seven species were derived empirically using field data from Duke Forest stand, which is located in the Blackwood Division of the Duke Forest in Chapel Hill, NC. Over an area covering 4 hectare every individual over 2m tall was identified to species, mapped, and its diameter was measured at 1.45m high, a common metric in forestry and ecology referred to as Diameter Breast Height (DBH). In total there were 52 species observed in this stand, but in this paper we will focus on seven species:

Species	σ^2	ν	τ^2	u	a_0	a_1	m_0	m_1
ACru	1.18	0.036	0.024	62	2.77	0.406	1.15	0.05
CAca	1.36	-0.025	0.020	50.5	1.53	0.923	16.11	2.12
CEca	0.37	0.272	0.040	163.9	1.62	0.757	2.84	0.7
FRam	1.70	0.105	0.015	34.7	2.28	0.425	3.10	0.14
LIst	0.57	-0.790	0.001	518	2.31	0.527	5.30	0.15
Litu	0.55	0.371	0.005	719.8	3.37	0.577	3.09	0.12
PIta	0.72	-0.349	0.145	1706.1	2.06	0.700	1.65	0.03

Figure 5: Dispersal parameters for seven species estimated from field experiments (Clark et al [7]).

*A*cer rubrum (ACru), *Carpinus caroliniana* (CAca), *Cercis canadensis* (CEca), *Fraxinus americana* (FRam), *Liquidambar styraciflua* (LIst), *Liriodendron tulipifera* (LItu), *Pinus taeda* (PIta). Figure 5 lists the parameter values that were estimated for each of these species.

4 Computing Dispersal

We developed a forest simulator based on the model described in [9] and summarized in Section 2, which takes an initial configuration of the forest and landscape as input and simulates dynamics of the forest at annual time steps. Figure 6 shows the flowchart of operations.

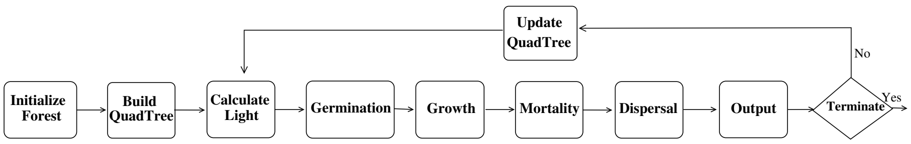
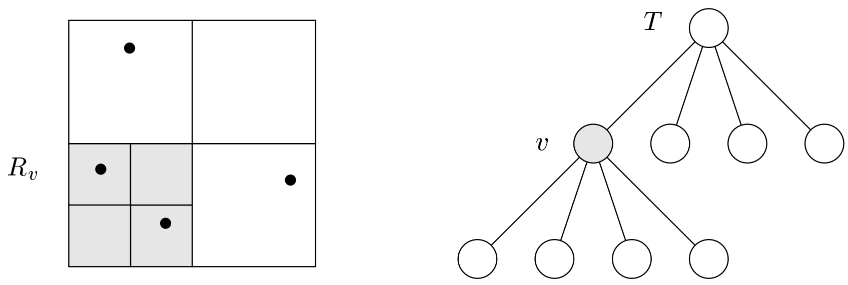


Figure 6: Flow chart of the sequence of operations performed by the simulator.

Since dispersal and light calculations are computationally intensive, and ecological experiments need to be performed on large landscapes (at least 1 sq. km.) and for long durations (up to several thousand years), performing exact calculations to simulate dynamics would take months (e.g. a time step on 512×512 landscape took 5 hours). We therefore expedite the simulation by performing calculations approximately — the approximation error can be controlled by the user (the approximation error is fixed such that it is within the inherent stochasticity of the model). We maintain a hierarchical (multi-resolution) representation of the forest using a quad-tree data structure and calculate dispersal approximating at spatial resolutions depending on the required accuracy.

We first describe the quad-tree data structure, and then the approximation algorithm to calculate dispersal. In describing the data structure and algorithm, we assume, for sake of simplicity, that all individuals in the forest belong to the same species. They can be easily extended for the case of multiple species. Finally, we present experimental results that show the performance of our algorithm.



(i) (ii)

Figure 7: (i) A recursive subdivision of the landscape. (ii) The quad-tree representing this subdivision.

Quad-tree data structure. For simplicity, we assume that $\mathbb{L}$, the landscape, is enclosed into a square of side-length 2^l and discretized into a $2^l \times 2^l$ mesh M, where $l \geq 0$ is an integer; we assume that $\rho = 1$. Let A denote the area of $\mathbb{L}$, which is also the number of grid cells of M. We denote the set of individuals (saplings, adults) in the forest and their locations as follows:

$$I = \{X \mid X \text{is an individual}\},$$

$$P = \{\ell(X) \mid X \text{is an individual}\}.$$

A quad-tree T on $\mathbb{L}$ is a 4-way tree that represents a hierarchical subdivision of $\mathbb{L}$. Each node v of T is associated with a square $R_v \in \mathbb{L}$, a subset $P_v = R_v \cap P$ of points and a set $I_v = \{X \mid X$ is an individual and $\ell(X) \in R_v\}$ of individuals. For the root u of T, $R_u = \mathbb{L}$, $P_u = P$ and $I_u = I$. If R_v is a grid cell of M or $|P_v| = 1$, v is a leaf. Otherwise, we partition R_v into four congruent squares by bisecting its two sides, and assigning the four squares to the four children of v. See Figure 7. If the depth of a node v is d, then the side-length of R_v is 2^{l-d}. The maximum depth of T is $\log_4 A$.[1]

In each node v, we store $|P_v|$, the total number of individuals of each species and $b_t(v) = \sum_{X \in I_v} \beta_t(X)$, the total fecundity of all the individuals of a specific species contained in R_v. We use this information to develop an approximation scheme to compute dispersal. At each leaf v of T, we store the sets I_v and P_v in a list.

Approximation scheme. The approximation scheme computes the expected number of seeds $q_t(M_{ij})$, from all the individuals of the forest, that fall into each grid cell M_{ij} at time t using 2. The exact computation takes $O(nA)$ time since we iterate over all grid cells and individuals, which is too expensive even for a moderately sized forest. For example, it takes about 90 minutes to compute dispersal exactly on a 1024×1024 sized forest with 500,000 individuals.

[1] In general, a node v of a quad-tree is a leaf only if $|P_v| = 1$, but in our application it suffices to stop the subdivision as soon as we reach a grid cell. This ensures that the depth of T is $\log_4 A$.

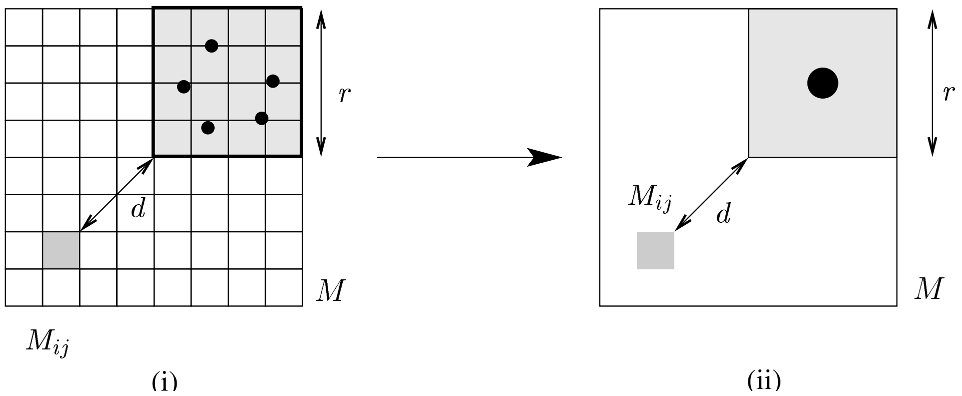


Figure 8: Monopole approximation for dispersal. (i) A 8×8 mesh M, grid cell M_{ij} (shaded) and a 4×4 region containing 5 individuals (shown as small circles). (ii) The mesh after monopole approximation. The "super individual" (shown as a large circle) is located at the center of the 4×4 region.

We rely on an approximation scheme to expedite the computation at a slight loss in accuracy. We first describe the intuition and then give a formal description. From figure 8, it is clear that if an individual is far away from a grid cell M_{ij}, then the expected number of seeds falling into M_{ij} is almost the same if we vary the location of X a little. We therefore cluster the grid cells that are far away from M_{ij} and move all individuals in a single cluster to a canonical location. We regard all these individuals as a single "super individual" X_v, whose fecundity $\beta_t(X_v) = b_t(v)$ (sum of the fecundity of all the individuals). The quad tree provides a natural way of computing this clustering. We refer to this approximation as *monopole approximation*.

We now describe the algorithm formally. For a node v of T, let C_v (resp. r) denote the center (resp. side-length) of R_v and let d be the shortest distance from grid cell M_{ij} to the boundary of square R_v. See figure 8. We set a threshold parameter μ called *monopole coefficient*. If $\frac{r}{d} \leq \mu$ (grid cell M_{ij} is far away from R_v as compared to side length of R_v), we perform the monopole approximation, i.e. replace all the individuals I_v with the "super individual" X_v, located at C_v. The seeds falling into M_{ij} due to individuals I_v is approximated by the seeds falling into M_{ij} due to X_v.

The approximation algorithm is a recursive procedure, starting at the root of the quad tree. We perform the following at each node v of T: At each node v of T, we check if the monopole approximation can be performed at v. If so, we approximate the expected number of seeds falling into M_{ij} due to individuals I_v by calculating the expected number of seeds falling into M_{ij} due to the "super individual" X_v. If the monopole approximation cannot be performed at v and v is a leaf, we calculate the expected number of seeds falling into M_{ij} due to individuals in I_v by summing the contribution from each individual in I_v. Else, we recurse on the children w of v. The algorithm is described more formally in Algorithm 1.

We now analyze the running time of our algorithm. The main idea behind the analysis is the

Algorithm 1 `Monopole Approximation`(v, M_{ij})

$r =$ side-length of square R_v
$d =$ distance between R_v and M_{ij}
if ($\frac{r}{d} \leq \mu$) **then**
  return $\beta_t(X_v) \cdot f(\|C_{ij} - C_v\|; u, p)$
else
  if v is a leaf **then**
    return $\sum_{X \in I_v} \beta_t(X) \cdot f(\|C_{ij} - \ell(X)\|; u, p)$
  else
    for each child w of v **do**
      return `Monopole Approximation`(w, M_{ij})

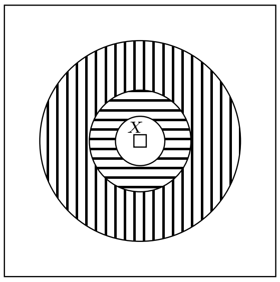


Figure 9: Concentric circles C_0, C_1, C_2 around gridcell M_{ij}.

claim that the monopole algorithm performs monopole approximation at constant number of quad tree nodes at each level. We illustrate this using Figure 10, which shows a 16×16 mesh and grid cell M_{ij}. For sake of simplicity, we assume that $\mu = 1$. It is easy to verify that for that 27 1×1 shaded squares and 27 2×2 checked squares and 7 4×4 bricked squares satisfy the monopole condition. From this illustration, we can intuitively see that the algorithm performs monopole approximation in at most 27 quad tree nodes v of depth i (R_v correspond to a $2^i \times 2^i$ square). Since the depth of the quad tree is $\log_4 A$, the number of nodes of the quad tree at which monopole approximation is performed is $O(\log_4 A)$.

We now present a more formal analysis of the algorithm. Let C be a circle of radius $1/\mu$, centered at the grid-cell M_{ij} and let m_{ij} is the number of individuals that lie inside circle C. The dispersal due to all plants in C is calculated by calculating the contribution from each individual . This is because $d < 1/\mu$ and thus $r/d > \mu$. The complexity of the dispersal calculation is $O(m_{ij})$.

Let $C_i, 1 \leq i \leq \log A$ be an annular ring, centered at the center of X and having inner radius $2^i/\mu$ and outer radius $2^{i+1}/\mu$. Figure 9 shows concentric rings $C = C_0, C_1, C_2$ The monopole condition is satisfied for all quad-tree regions R_v with side length 2^i, that are contained in C_i. Using a packing argument, the number of such regions R_v is at most $3/\mu^2$. In the illustration shown in Figure 10, the constant is 27.

For any gridcell M_{ij}, the forest with A gridcells is covered by atmost $\log A$ annular rings C_i.

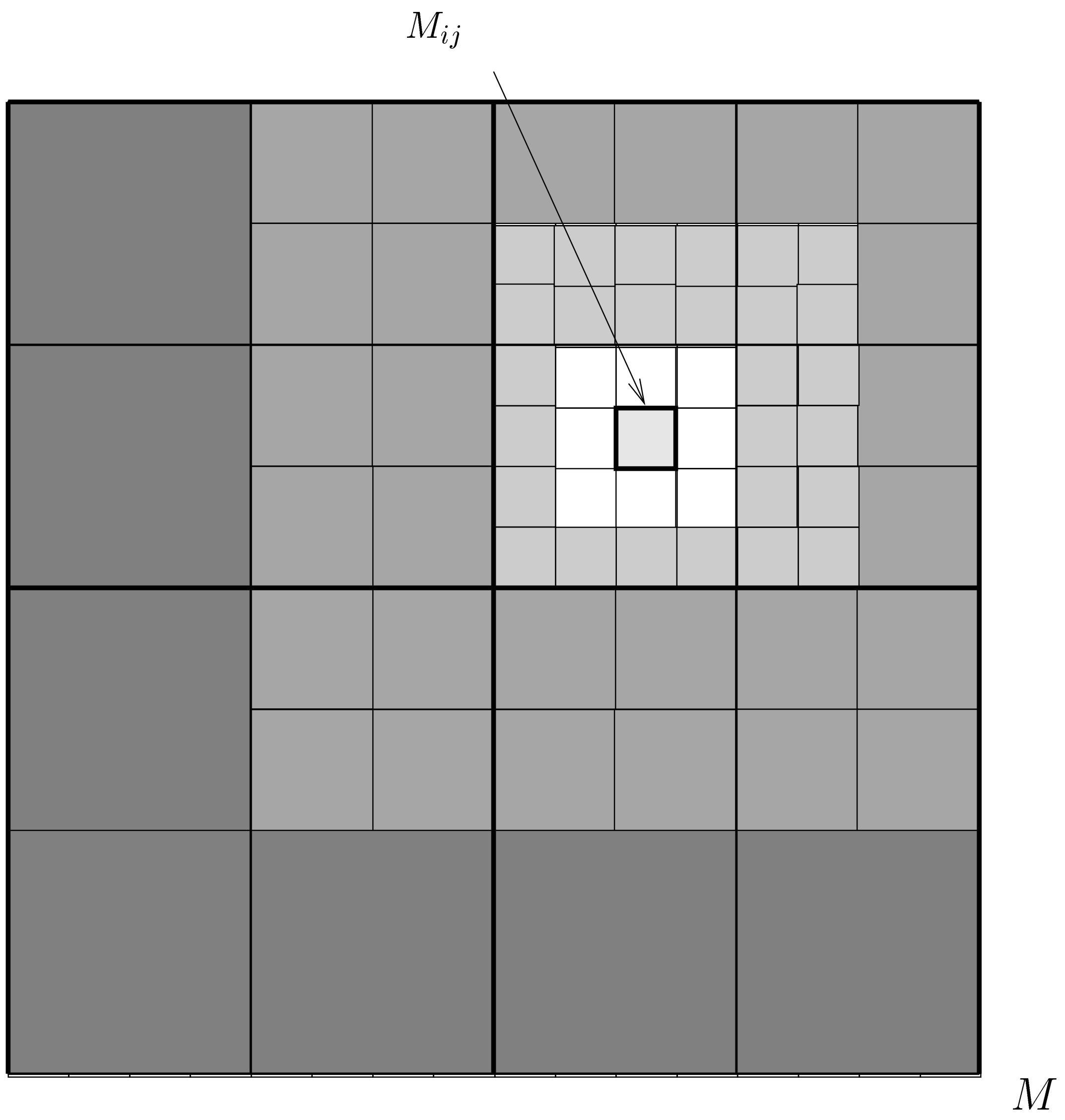


Figure 10: Figure shows a 16×16 grid. The monopole coefficient $\mu = 1$. For grid cell M_{ij}, the monopole approximation is performed at the 27 1×1 light shaded squares, 27 2×2 medium shaded squares and 7 4×4 dark shaded squares.

The number of monopole approximations performed in any such C_i is $3/\mu^2$. Thus the total complexity of the Monopole procedure is $O(\log A/\mu^2 + m_{ij})$.

We perform the Monopole procedure at each gridcell M_{ij}. The time to compute dispersal for the entire forest is thus given by $O(A \log A + \sum_{i,j} m_{ij})$. Note that $\sum_{i,j} m_{ij}$ denotes the total number of dispersal calculations performed exactly (monopole condition was not satisfied). For each individual X in the forest located in grid cell M_{ij}, dispersal due to X is calculated exactly for all grid cells that are at distance at most $1/\mu$. Thus each individual contributes to exact dispersal calculations in at most $1/\mu^2$ grid cells. The total number of individual dispersal calculations is thus at most n/μ^2. The total time to calculate dispersal is thus $O((A \log A + n)/\mu^2)$.

Experimental results. We have performed a set of experiments to evaluate the computational performance of the approximation algorithm as a function of the area of the forest and the monopole coefficient.

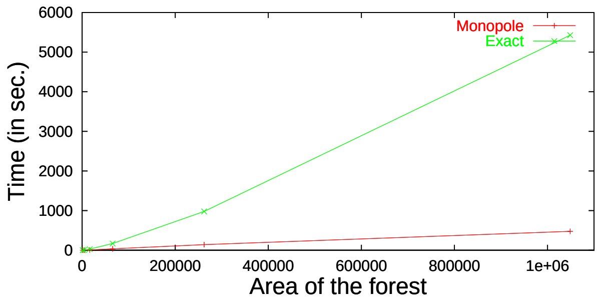


Figure 11: Running time of the exact algorithm and the approximation algorithm with monopole coefficient 0.1 for varying forest area sizes.

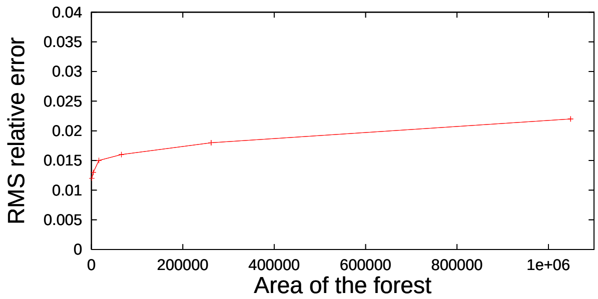


Figure 12: Relative error of dispersal algorithm with monopole 0.1 for varying forest area sizes.

We performed experiments on a forest initialized with the output of a 100 year simulation involving a single species. In our experiment, we varied the side length of the forest from 32 meters to 1024 meters. Figure 11 compares the running time of the exact algorithm with the monopole algorithm (for monopole coeffient 0.1). For a 1024×1024 m^2 forest, the monopole acheievs speedup of two orders of magnitude. Figure 12 plots the RMS error of seeds dispersed for monopole coeffient 0.1. Note that the error in seeds dispersed is less than 2% for the landscapes simulated.

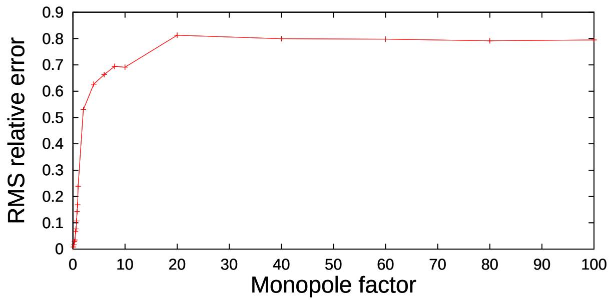


Figure 13: Relative error of dispersal algorithm for varying monopole coeffient. The forest is initialized with trees of species Acer rubrum from Duke forest site.

Finally, we performed an experiment to evaluate the effect of monopole coeffient (μ) on the error incurred. We initialized the forest with trees of species Acer rubrum from Duke forest site. Figure 13 plots the RMS error of seeds dispersed for monopole coeffient from 0 to 100. The RMS error increases rapidly for $\mu \leq 1$ and does not change for $\mu \geq 20$. The shape of the dispersal kernel for *Acer rubrum* (see figure 4) explains this behavior. Since the dispersal kernel is almost a constant after a 20 meter radius, any amount of monopole approximation involving distances above 20 meters does not affect the RMS error.

5 Inherent Variability

In this section, we quantify the inherent variability (stochasticity) of the dispersal model using statistical methods. We perform $N = 1000$ iterations of dispersal computation on the Duke Forest stand mentioned in Sec. 3. The Duke forest stand was approximately centered in a 512×512 landscape at $1m \times 1m$ resolution. We set the monopole coefficient to 0.125 so that the error in computation is negligible.

For each iteration, the simulator outputs the spatial map of seed rain, number of seeds dispersed, in the forest. Let s_{ijt} denote the seed rain at grid cell M_{ij} in iteration t, $1 \leq i, j \leq 512$, $1 \leq t \leq N$. From the N replicate dispersal maps, we calculate the mean seed rain, $\hat{s}_{ij} = \frac{1}{N}\sum_{t=1}^{N} s_{ijt}$, at grid cell M_{ij} for each species. We also calculate the coeffient of variation of seed rain CV_{ij}, using the following formula:

$$CV_{ij} = \left(\frac{1}{N} \sum_{t=1}^{N} \left(\frac{s_{ijt}}{\hat{s}_{ij}} - 1 \right)^2 \right)^{1/2}.$$

Figure 14, which depicts the spatial distribution of mean seed rain $\hat{s}_{ij}$, illustrates the spatial variability in seed rain and it's relation to parent trees, which are depicted as circles with diameter equal to their mean canopy diameter (Figure 14). Figure 15, which depicts CV, illustrates the spatial pattern of the temporal variability in seed rain

Two of the seven species clearly appear to be dispersal limited (CAca, CEca) while the other five species appear to be reasonably abundant at most, though not necessarily all, places on the

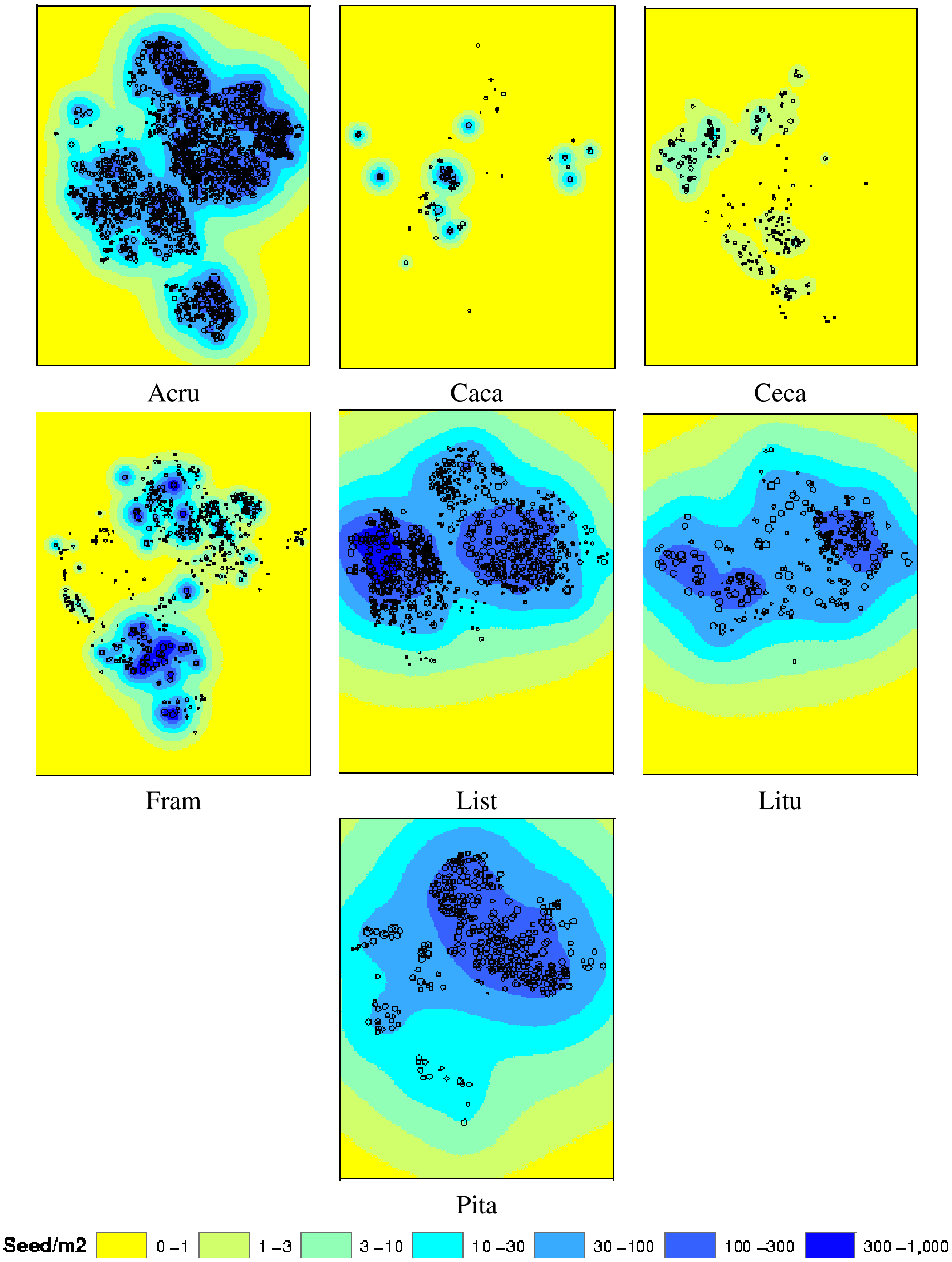


Figure 14: Spatial map of the mean number of seeds dispersed per sq. m for seven different species.

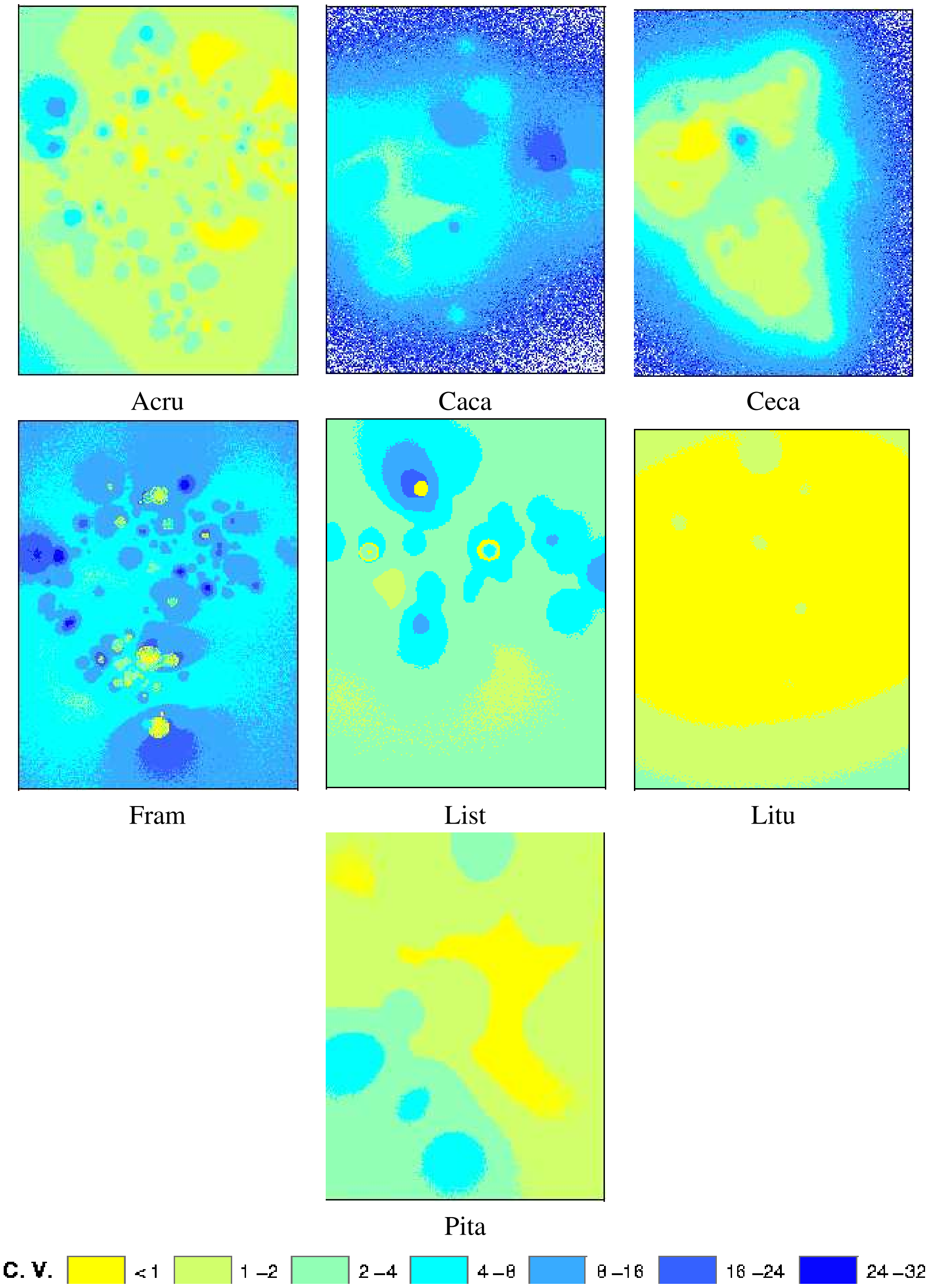


Figure 15: Spatial map of the coefficient of variation in the number of seeds dispersed for seven different species

Spp	$\hat{s}$	CV	RelAbund
ACru	17.64	4.56	0.23
CAca	0.45	65.12	0.01
CEca	0.21	6.22	0.03
FRam	7.22	24.05	0.06
LIst	21.18	10.51	0.11
Litu	13.55	1.45	0.03
PIta	19.81	2.73	0.04

Figure 16: Summary statistics for dispersal maps by species. $\hat{s}$ is the mean of $\hat{s}_{ij}$. CV is the RMS value of CV_{ij}. RelAbund is the relative abundance of each species on the landscape based on stem density. RelAbund does not sum to one because there are 45 other species in the forest which are not presented here.

landscape. The seed limited species all appear to be relatively less abundant (column RelAbund in Table 16) and have short dispersal distances (column u in Table 5), but they also span a spectrum of temporal variability due largely to their different levels of variability in fecundity. Of the remaining species with relatively high mean seed rain, they also span a range of variability from LItu, which is widely dispersed, highly fecund, and has trees that are well distributed across the landscape, to FRam, which has high levels of variability in fecundity and a short dispersal distance, leading to high spatial and temporal variability in seed rain. In between are species like ACru and PIta, which have similar mean seed rain and CV, but reach this in very different ways. PIta has a very long dispersal distance but is highly aggregated to one part of the landscape, leading to very smooth dispersal contours. ACru on the other hand has a much smaller dispersal distance, but is very abundant and distributed across the landscape, which leads to a much more complicated pattern of both mean dispersal and coefficient of variation map (Figures 14, 15).

For each species, the RMS value of CV_{ij}, denoted by CV, captures the inherent variability of the dispersal process for that species. The following formula expresses CV in terms of seed rain s_{ijt} and mean seed rain $\hat{s}_{ij}$:

$$CV = \left(\frac{1}{N \cdot A} \sum_{i,j,t} \left(\frac{s_{ijt}}{\hat{s}_{ij}} - 1 \right)^2 \right)^{1/2}$$

Also, the mean value of $\hat{s}_{ij}$, denoted by $\hat{s}$, represents the mean numbers of seeds dispersed by each species. Table 16 shows the values of $\hat{s}$ and CV for each of the seven species.

Four of the seven species (ACru, List, LItu, PIta), which had large dispersal distances (see Figure 14), have a large value of $\hat{s}$, as expected. Similarly, species FRam, CAca and LIst that show high and spatially distributed CV_{ij} values from Figure 15, have a large value of CV. The high value of CV indicates a high spatial variability in the dispersal process for this species.

6 Determining the appropriate monopole coefficient

In this section, we present experimental results to show how we should choose the value of monopole coefficient, the main user-controlled parameter that provides a tradeoff between the accuracy and efficiency of the algorithm. As in Section 5, all experiments are performed on a 512×512 landscape and use species parameters given in Table 5 of Sec 3.

The first experiment describes the variation in the running time of the approximation algorithm with respect to the monopole coefficient. The experiment was performed on seven species used in the previous section. We initialized the forest with data from the Duke forest stem map. We varied the monopole coefficient from 0.1 to 1.0. Figure 17 shows the running time of the approximation algorithm, as a function of monopole coefficient. As anticipated, the running time decreases sharply with the increase in the monopole coefficient. We can thus choose the monopole coefficient based on how much time we can spend on each step of the simulation.

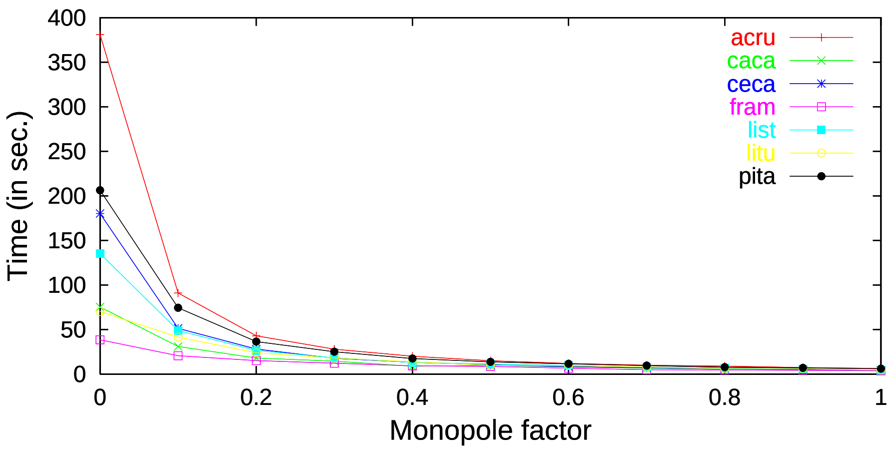


Figure 17: Running time of the approximation algorithm for seven different species as the monopole coefficient is varied from from 0 to 1

Next, we study how the relative error of seeds dispersed for different species, varies with the monopole coefficient. Let s_{ij} denote the actual seed rain that falls in grid cell M_{ij}, computed by the exact algorithm and $\tilde{s}_{ij}$ denote the seed rain in M_{ij} computed by the approximation algorithm. We define the relative error ε_{ij} to be

$$\varepsilon_{ij} = \left| 1 - \frac{\tilde{s}_{ij}}{s_{ij}} \right| .$$

Of course, ε_{ij} depends on the value of μ, the monopole coefficient. Figure 18 shows the spatial distribution of ε_{ij} with $\mu = 0.5$ for all the seven species. The circular arc patterns in the figure (for all species) is an artifact of our algorithm — the portion of the forest in which we perform the monopole approximation at the same level of the quad-tree is bounded by a circle. Figure 19 shows the RMS value of ε_{ij}, defined as:

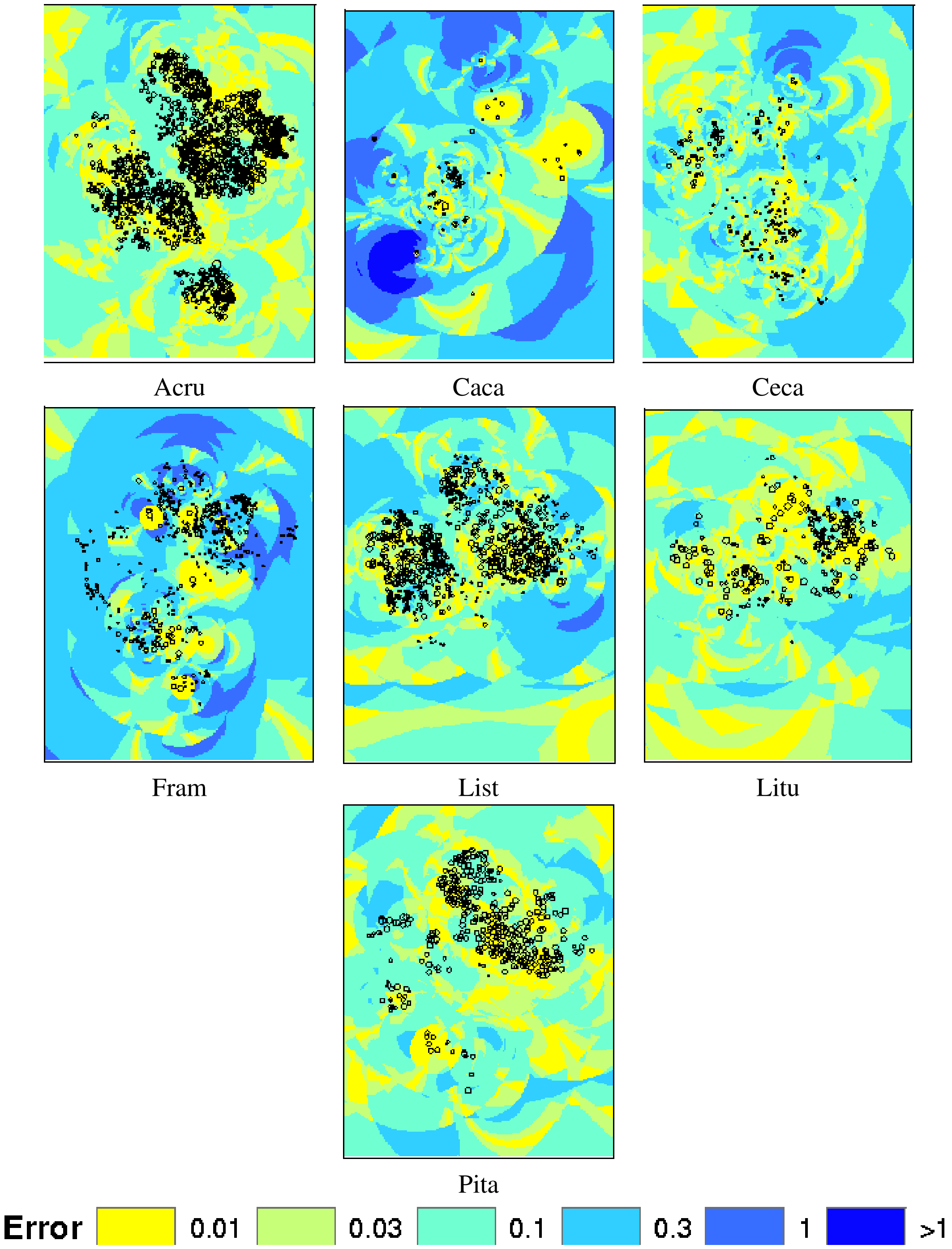


Figure 18: Spatial distribution of relative error for seven species. Monopole coefficient is set to 0.5.

$$E = \left(\frac{1}{A} \sum_{i,j} \left(1 - \frac{\tilde{s}_{ij}}{s_{ij}} \right)^2 \right)^{1/2},$$

for all the seven species. As anticipated, the RMS relative error increases with increase in μ. We can thus control the relative error by choosing an appropriate μ.

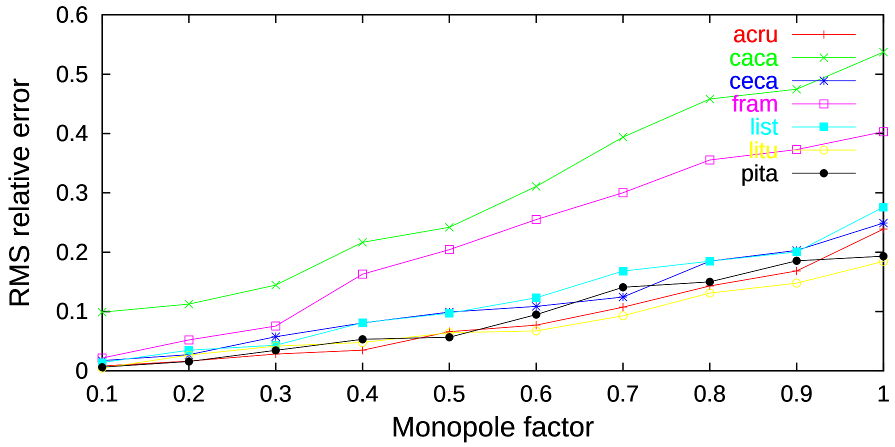


Figure 19: RMS relative error of the approximation algorithm for seven different species as the monopole coefficient is varied from from 0 to 1

Figure 20 indicates that the value of E for a fixed value of μ depends on the population density. We therefore studied how E depends on μ and the number of individuals. Figure 20 shows the variation in E for *Acer rubrum*; the monopole coefficient was varied from 0.1 to 1.0 and the number of individuals from 100 to 1000. It is not so surprising that, for a fixed value of μ, E decreases with the population density. Indeed for a given grid cell M_{ij}, the distance of the closest individual to the grid cell decreases as the number of individuals increase. Since a large fraction of seeds to M_{ij} come from nearby individuals and the monopole approximation error is less when the distance to the individual is small, the RMS relative error decreases. This experiment suggests that we would choose different values of μ for different species, depending on their density, in order to obtain the same relative error.

The previous experiment raised the question of whether the relative error depends only on the number of individuals or also on how uniformly these individuals are distributed. We therefore performed an experiment in which we studied how the relative error and the running time varied as the percentage of forest area that contained individuals was varied from 5% to 100%. We fixed μ to 0.1.

For a given occupancy, we obtain the initial tree distribution as follows: randomly select $16X16$ regions of appropriate number in the forest and fill each $16X16$ region with constant (in this experiment 20 individuals) number of individuals placed randomly inside the region. Figure 21 plots the RMS relative error as the occupancy is varied. As anticipated, the RMS relative error decreases

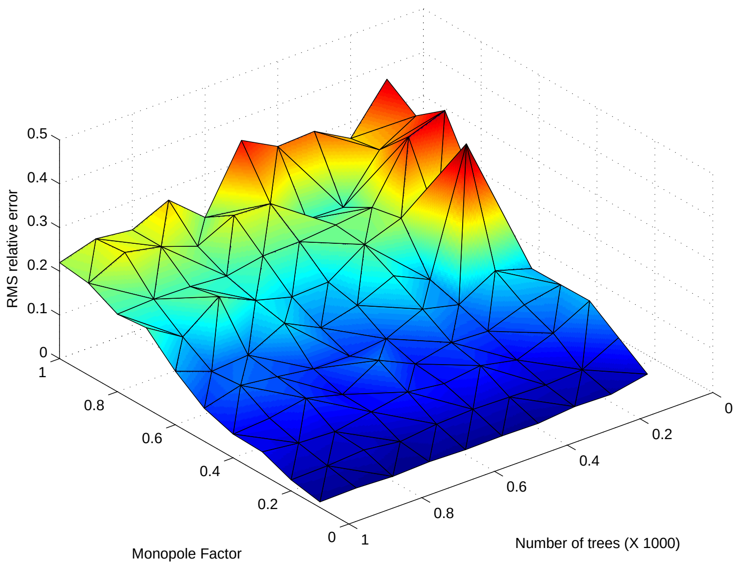


Figure 20: RMS relative error variation plotted as a surface. Monopole coefficient is varied from from 0 to 1 and number of individuals is varied from 100 to 1000.

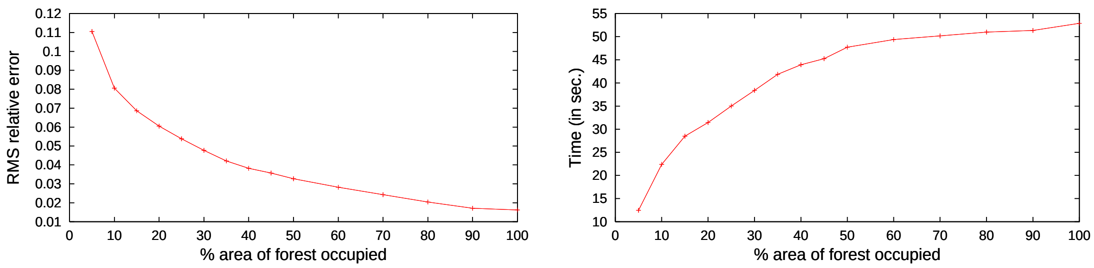


Figure 21: Variation of RMS relative error and running time for different occupancy values. The left figure plots the RMS relative error and the right figure plots the running time.

and running time increases with increase in occupancy. Two factors contribute to this behavior: (i) the number of individuals in the forest increases as occupancy increases; (ii) the distribution of individuals becomes more uniform as occupancy increases.

These experiments suggest how we can choose the value of μ depending on whether we want to control the running time or the relative error, which depends on the number of individuals and fragmentation. In scenarios where we want to simulate large landscapes and/or for long periods of time, it is critical to control the running time. In contrast, if we want to perform a highly sensitive ecological simulation, controlling the relative error is important.

In order to quantify the tradeoff between the running time and the relative error, we introduce a *tradeoff index* $\Phi_i(\mu)$, defined as follows:

$$\Phi_i(\mu) = e^{c \cdot E_i(\mu)} \cdot \frac{T_i(\mu)}{T_i(0)}, \tag{3}$$

where $E_i(\mu)$ is the RMS relative error and $T_i(\mu)$ is the running time of dispersal algorithm for monopole factor μ and c is a constant.

We calculate the cumulative index by summing the index of all the seven species. Figure 22 shows the variation of the cumulative index with monopole coefficient μ. It also shows the tradeoff index for species *Acer rubrum* and *Carpinus caroliniana*.

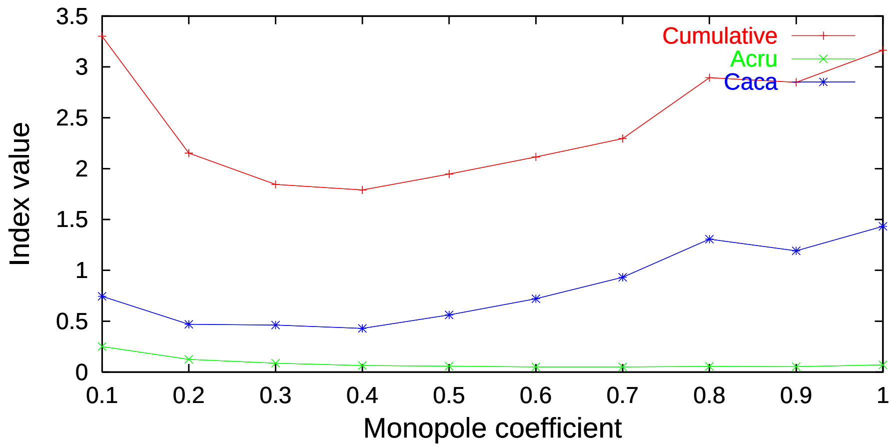


Figure 22: Plot of Φ for various monopole coefficients. Figure shows the cumulative index and the index for species *Acer rubrum* and *Carpinus caroliniana*.

7 Conclusion

We have developed efficient approximation algorithms to compute dispersal. For reasonable error, our approximation algorithm achieves a speedup of an order of magnitude. Based on our experiments that (i) quantify the inherent variability of the dispersal model and (ii) study the variation of

RMS relative error with monopole coefficient, we provide guidelines that would help the user to choose the right monopole coefficient for a given simulation.

Work presented in this paper is part of an ongoing inter-disciplinary project to study forest ecosystems using simulation. Some of the interesting algorithmic issues and future direction include:

- Our current dispersal algorithm has complexity O($n \log n$). We plan to adapt the linear time *Multipole algorithm* of Greengard [10] to calculate dispersal.
- Our dispersal algorithm utilizes spatial coherence to obtain good approximation. We plan to develop algorithms that utilize temporal coherence as well.
- Our algorithms are efficient for internal memory. When simulating very large scale forests, increased I/O operations will lead to poor performance. We plan to extend our data structures and algorithms to minimize data transfer between main memory and disk.

A Appendix: Notation Index

Landscape

Landscape	$\mathbb{L}$
Mesh bounding the landscape	M
Grid cells	M_{ij}
Center of grid cell M_{ij}	C_{ij}
Side-length of each gridcell	ρ

Spatial Attributes of Individual X

Position	$\ell(X)$
Trunk diameter	$D(X)$
Trunk height	$H(X)$
Crown height	$H_c(X)$
Crown diameter	$D_c(X)$

Dispersal Model

Expected number of seeds dispersed in gridcell M_{ij} at time t	$q_t(C_{ij})$
Number of seeds dispersed in gridcell M_{ij} at time t	s_{ijt}
Indicator variable for sex of individual at time t X	$\chi_t(X)$
Indicator variable for reproductivity of individual X	$\Delta(X)$
Dispersal Kernel - Student t-distribution with parameters p, u	$f(r; u, p)$
Fecundity of individual X at time t	$\beta_t(X)$

References

[1] J. Barnes and P. Hut, A hierarchical $O(n \log n)$ force-calculation algorithm, *Nature*, 324 (1986), 446–449.

[2] A. Cescatti, Modelling the radiative transfer in discontinuous canopies of asymetric crowns. i. model structure and algorithms, *Ecol. Modeling*, 101 (1997), 263–274.

[3] P. Chesson, General theory of competative coexistance in spatially-varying environments, *Theoretical Population Biology*, 58 (2000), 211–237.

[4] J. Clark, M. Silman, R. Kern, E. Maclin, and J. HilleRisLambers, Seed dispersal near and far: patterns across temparate and tropical forests, *Ecology*, 80 (1999), 1475–1494.

[5] J. S. Clark, B. Beckage, J. HilleRisLambers, I. Ibanez, S. LaDeau, J. MacLachlan, J. Mohan, and M. Rocca, Dispersal and plant migration, in: *Encyclopedia of Global Environmental Change*, Vol. 3.

[6] J. S. Clark, M. Dietze, I. Ibanez, and J. Mohan, Coexistence: how to identify trophic tradeoffs, *Ecology*, 84 (2003), 17–31.

[7] J. S. Clark, S. LaDeau, and I. Ibanez, Fecundity of trees and the colonization-competition hypothesis, in press (2004).

[8] R. A. Finkel and J. L. Bentley, Quad trees: a data structure for retrieval on composite keys, *Acta Inform.*, 4 (1974), 1–9.

[9] S. Govindarajan, M. Dietze, P. K. Agarwal, and J. Clark, A scalable simulator for forest dynamics, *Proc. 20th ACM Symposium on Computational Geometry*, 2004, p. Accepted.

[10] L. Greengard, *The rapid evaluation of potential fields in particle systems*, MIT Press, Cambridge, 1988.

[11] S. W. Pacala, C. D. Canham, and J. A. S. Jr., Forest models defined by field measurements: 1 the design of a northeastern forest simulator, *Can. Journal For Res.*, 23 (1993), 1980–1989.

[12] E. Ribbens, J. A. SIlander, and S. W. Pacala, Seedling recruitment in forests: calibrating models to predict patterns of tree seedling dispersal, *Ecology*, 75 (1994), 1794–1806.

[13] H. Samet, *Applications of Spatial Data Structures: Computer Graphics, Ima ge Processing, and GIS*, Addison-Wesley, Reading, MA, 1990.