Available online at www.sciencedirect.com
SCIENCE DIRECT®
Ecological Modelling 190 (2006) 205–222
ECOLOGICAL MODELLING
www.elsevier.com/locate/ecolmodel

Sequential sampling designs for catching the tail of dispersal kernels

A. Pielaat*, M.A. Lewis, S. Lele, T. de-Camino-Beck

University of Alberta, Department of Mathematical and Statistical Sciences, 632 Central Academic Building, Edmonton, Alta., Canada T6G 2G1

Received 17 September 2003; received in revised form 13 January 2005; accepted 28 February 2005
Available online 17 June 2005

Abstract

Methods to design a sampling strategy should depend on the research question involved when conducting the experiment. The objective of this study is to design a seed trap configuration surrounding a parent plant when the long distance component of the seed dispersal kernel is of interest. In particular, as a population's invasion speed depends mainly on the tail of the dispersal kernel, the sampling design in this study is based on calculating this quantity. The optimality criterion is to minimize the mean squared error (MSE) of the estimated invasion speed (using a limited number of traps) with respect to the "true" calculated invasion speed. Detailed procedures are given on how to calculate an invasion speed, both in a 1D and a 2D setting, with examples on how to implement the method to get a local optimal sampling strategy using *Calluna vulgaris* as a test system. Results show a trade-off between nearby sampling (many seeds, no long-distance dispersal measured) and distant sampling (few seeds, but long-distance dispersal measured).

Keywords: Sequential sampling; Traveling wave; Heather plants; Field data; Histogram estimator

1. Introduction

Dispersal is an important strategy for species survival (Murray, 1986). Establishment of species in their new environment affects ecosystem's dynamics by its influence on, e.g. biodiversity (Malanson, 1996) and competition (Jesson et al., 2000; Matsinos and Troumbis, 2002). Vegetation, in particular, can spread rapidly when huge numbers of seeds are being dispersed over long distances from the parent plant. The spread rate can be used as a measure of invasiveness and can be calculated using information on the tail of the dispersal kernel (Kot et al., 1996; Lewis et al., 2005).

Although work has been done to assess short distance dispersal (Jongejans and Telenius, 2001), experimental studies on measuring the tail of distribution kernels are rare. Still, see Greene and Johnson (1995, 1996), for long-distance wind dispersal research of tree seeds. See also Paradis et al. (2002), Nurminiemi et al.

* Corresponding author. Tel.: +31 30 2743711; fax +31 30 2744434.
E-mail address: annemarie.pielaat@rivm.nl (A. Pielaat).

0304-3800/$ – see front matter ©2005 Elsevier B.V. All rights reserved.
doi:10.1016/j.ecolmodel.2005.02.023

(1998) and Tackenberg (2003) for some methods to analyse dispersal data.

A dispersal kernel $k(x)$ is a probability density function associated with moving distance x in a single dispersal event. Knowledge of the dispersal pattern of seeds from a parent plant and growth and survival of these individuals at new sites gives insight in the spread rate of vegetation at population level. In this paper, we will focus on the sampling design for estimating vegetation spread into a new environment based on seed counts in traps surrounding a single point source. It will be shown that measurement of the "tails" of the dispersal kernels in the field plays a dominant role in the estimation of the population spread, but that optimal sampling effort as a function of distance from the source plant involves a trade-off between nearby sampling (many seeds, no long-distance dispersal measured) and distant sampling (few seeds, but long-distance dispersal measured).

Accurate field measurements are a prerequisite to understand the mechanisms behind long distance seed dispersal. A good sampling design is important to achieve this goal. An optimal spatial sampling design for studies on pollen dispersal was given by Assunção and Jacobi (1996). Their interest was in the shape of the dispersal curve. The kernel was estimated by a histogram, based on observed counts of individuals in the field. The optimal sampling design, in that case, minimized the error in estimating the shape of the kernel. This is equivalent to minimizing the area of the difference between the continuous dispersal curve and the histogram estimator. As a consequence, their design resulted in concentrating samples near the parent plant where most of the seeds fall. However, their algorithm is not applicable when the invasion speed of organisms is of interest, and information on long-distance dispersal events, as described by the tails of the dispersal kernels, is crucial. A different approach which emphasizes the measurement of the long distance component of the spatial spread in the field, is presented in this paper. Our sampling design is based on getting the best estimate for the invasion speed of species when only limited information on the seed dispersal kernel is available from seed trap data. The design will consist of placing seed traps at several distances from a parent plant such that the mean squared error (MSE) of the estimated wave speed with respect to the true wave speed is minimized. Such global optimal design is difficult computationally, hence we use a sequential design where seed traps are added one at a time. Although sequential sampling does not result in a global optimal design, it does ensure that each additional seed trap is placed at the locally optimal location of the remaining open sites, and thus approximates a global optimum.

The first step in obtaining a good sampling design is knowledge about the true dispersal kernel. As this is unavailable, a useful approach requires an initial guess on the dispersal kernel obtained from field data on, for example, a related species. The sampling design is then based on this initial estimate for the true kernel with the idea of subsequently getting step by step improvement towards an optimal design from repeated field experiments.

By way of example, our analysis will be based on work by Bullock and Clarke (2000), who measured dispersal for the heather plant *Calluna vulgaris* in the field. Heather plants have very light seeds which are dispersed by wind over long distances. As the bushes produce many seeds, this is a good plant to study seed dispersal. With this species, an accurate long distance dispersal pattern can be measured in the field. In this paper, we will first calculate the spread rate of *C. vulgaris* from the preliminary studies of Bullock and Clarke (2000) on the approximate shapes of dispersal kernels. This so called "true" spread rate will then be used to determine a sequential seed trap configuration when only a limited number of seed traps is available and gaining insight in the tail of the dispersal kernel is our goal.

First, we show how to calculate an invasion speed with an example on using seed dispersal data to calculate the speed. Then, the sampling design problem will be defined with a step-by-step approach on how to implement the method of sequential sampling based on the preliminary calculated invasion speed. Subsequently, a detailed explanation of the procedure to actually get the sampling design with detailed steps using seed dispersal as an example will be given followed by results for the specific *C. vulgaris* test system under consideration. Our goal is to provide a "user guide" for future applications of this method. A better insight in the long distance component of the dispersal pattern in the field is the basis for the development of mechanistic models for ecological processes. Hemerik et al. (2004), for example, stress the need for knowledge on the dispersal behaviour to predict the expansion velocity of the western corn rootworm into Europe. An improved

Table 1
Variables used in the procedures to obtain an optimal sampling design in this study

Variable	Explanation
$c(s)$	Theoretically derived true wave speed from the dispersal kernel $k(x)$
$c_h(s)$	Assumed "true" wave speed calculated from the histogram dispersal kernel, $k_h(x)$
$\overline{c}_h(s)$	Expected "true" wave speed calculated using multiple histogram dispersal kernels
$\hat{c}_h(s)$	Estimated average wave speed from limited seed trap data
$k(x)$	True dispersal kernel
$k_h(x)$	True dispersal kernel in the form of a histogram
$\hat{k}_h(x)$	Estimated dispersal kernel from limited field data in the form of a histogram
$M(s)$	Moment generating function (MGF) for the dispersal kernel, $k(x)$
$M_h(s)$	Moment generating function (MGF) for the histogram dispersal kernel, $k_h(x)$
R_0	Basic reproductive number
S	Total number of seeds over all sampling distances
D_n	The nth sampling distance with respect to the source
L_j	The furthest sampling distances j with respect to the source in a field with length L

sampling design would also help in the development of models in the field of risk assessment (Lewis et al., 1996; Reshetin and Regens, 2003). That is, a good insight in spatial spread associated with the dispersion of pathogenic and/or genetically modified microbes is of major importance to build models in cases where sampling needs to be limited from a hazard perspective. The general discussion will give some considerations on how to apply this method to other systems.

To help reading through the procedures in this paper a summary of the most frequently used variables is given in Table 1.

2. The theory of calculating invasion speeds

2.1. Spread in one spatial dimension

A full derivation of the theory of one-dimensional invasion speeds can be found in Kot et al. (1996). To be able to follow the arguments in this paper, however, only a general insight in the theory is required.

Consider a population that is in the initial stage of the spread into a habitat as part of the design of a field experiment. As the plant is being introduced in low densities into a habitat in which it has not previously been grown, the assumption of density-independent population dynamics is justified. We can, therefore, assume a population whose density (N) at location (x) changes in time (t) through reproduction and dispersal following:

$$N_{t+1}(x) = \int_{-\infty}^{\infty} k(x-y)\, R_0\, N_t(y)\, \mathrm{d}y, \tag{1}$$

where $R_0 > 1$ represents the basic reproductive number (the number of offspring produced by one parent organism which survive at least until the next reproduction time (Heesterbeek, 2002)). The function $k(x-y)$ is the dispersal kernel describing the relative frequencies of distances traveled (from y to x) as individuals disperse within one time step. This model assumes separate growth and dispersal events and discrete non-overlapping generations, although it also can be used to approximate growth and dispersal when generations overlap. Repeated application of Eq. (1) describes the expected spatial distribution of individuals as time progresses from one time step to the next. The spatial spread of $N_t(x)$ describes spread of the invading population. We consider an invading population whose leading edge is described by an exponentially decreasing function

$$N_t(x) = b\, \mathrm{e}^{-sx}, \tag{2}$$

where b represents the population density at point $x = 0$ at some given time t.

If organisms move forward with a constant distance c at each point in space (x), then

$$\begin{aligned} &N_{t+1}(x) = N_t(x-c) \quad \text{and so} \\ &N_{t+1}(x) = b\, \mathrm{e}^{-s(x-c)}. \end{aligned} \tag{3}$$

Fig. 1 demonstrates visually the population dynamics formulated by Eqs. (2) and (3). Substituting Eqs. (2) and (3) in Eq. (1) gives

$$b\, \mathrm{e}^{-s(x-c)} = \int_{-\infty}^{\infty} k(x-y)\, R_0\, b\, \mathrm{e}^{-sy}\, \mathrm{d}y. \tag{4}$$

Changing variables to $u = x - y$ and x yields

$$\begin{aligned} \mathrm{e}^{-sx}\, \mathrm{e}^{sc} &= R_0 \textstyle\int_{\infty}^{-\infty} k(u)\, \mathrm{e}^{-s(x-u)}\, (-\mathrm{d}u) \\ &= R_0 \textstyle\int_{-\infty}^{\infty} k(u)\, \mathrm{e}^{-sx}\, \mathrm{e}^{su}\, \mathrm{d}u \end{aligned} \tag{5}$$

and hence

$$\mathrm{e}^{sc} = R_0 M(s). \tag{6}$$

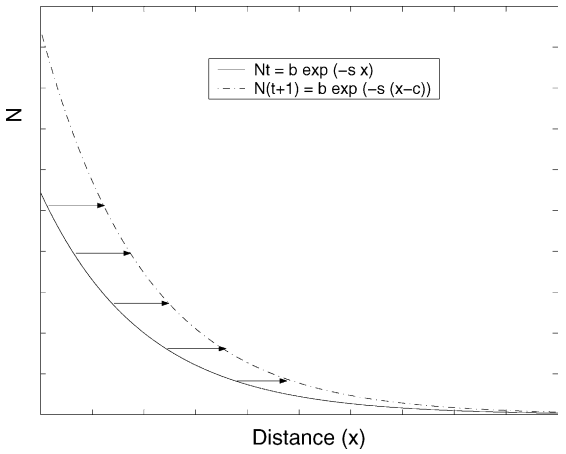

Fig. 1. Assume the numbers (N) in a population to decrease with exponentially bounded tails as the distance (x) from the source increases. And, each time step the population progresses with a constant speed (c) (Eqs. (2) and (3)).

Here,

$$M(s) = \int_{-\infty}^{\infty} k(u)\, \mathrm{e}^{su}\, \mathrm{d}u \tag{7}$$

is the moment generating function (MGF) for the dispersal kernel $k(u)$. This defines a dispersion relation between the speed c and the wave steepness s as

$$c(s) = \frac{1}{s} \ln(R_0 M(s)). \tag{8}$$

Eq. (8) indicates a relation between the shape of the wave and the invasion speed that involves the MGF of the dispersal kernel. That is, an invasion speed (c) can be calculated for every slope (s) of the wave (Eq. (8)). However, typically, the initial distribution of the invading plant will not decline exponentially, as described by Eq. (2). Rather, it will be confined to some finite region. Weinberger (1982) proved rigorously that, for such initial distributions, the asymptotic spread rate of the population is given by the minimum value of $c(s)$:

$$c = \min_{s>0} \left\{ \frac{1}{s} \ln(R_0 M(s)) \right\}. \tag{9}$$

This argument was explained heuristically by Kot et al. (1996).

Note that the moment generating function (Eq. (7)) gives exponentially increasing weight to the tails of the dispersal kernel k. Hence, the moment generating function is defined only for exponentially bounded functions. Given exponentially bounded tails, the "fatter" the tails, the larger the moment generating function for any given s and the larger the value of c in (9) and thus the faster the invasion process. When the tails of the dispersal kernel are not exponentially bounded, accelerating invasions, with infinite asymptotic speeds, result (Kot et al., 1996).

2.2. Spread in two spatial dimensions

Typical spread for a plant species will occur in two spatial dimensions, rather than the one spatial dimension assumed above. Analysis for the case of 2D spread is given in detail in Lewis et al. (2005). Here, it is assumed that dispersal need not be symmetric in all directions, but that it is translationally invariant (i.e. does not vary from one location to the next). The kernel $k(\mathbf{x})$ describes the probability density associated with dispersing x_1 units east and x_2 units north. It is then possible to calculate the spread rate for a "planar" wave front. The "planar" wave front refers to a well established invasion process, which has progressed to the point where one can approximately divide invaded and uninvaded locations with a straight line in x_1, x_2, space which is perpendicular to a unit vector describing movement in direction $\mathbf{w}$. The formula for the asymptotic rate of spread (9) remains unchanged, but the formula for the moment generating function is now

$$M(s) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} k(\mathbf{u}) \mathrm{e}^{(s\,\mathbf{u}\cdot\mathbf{w})}\, \mathrm{d}\mathbf{u}. \tag{10}$$

This is equivalent to calculating the moment generating function of a one dimensional kernel, which is the marginal distribution of an initial dispersal kernel in 2D given by $k(\mathbf{x})$. To see this observe

$$\begin{aligned} M(s) &= \textstyle\int_{-\infty}^{\infty} \left[\int_{-\infty}^{\infty} k(\mathbf{u})\, \mathrm{d}\eta \right] \mathrm{e}^{s\xi}\, \mathrm{d}\xi. \\ &= \textstyle\int_{-\infty}^{\infty} k_M(\xi)\, \mathrm{e}^{s\xi}\, \mathrm{d}\xi, \end{aligned} \tag{11}$$

where $\xi = \mathbf{u}\cdot\mathbf{w}$, $\eta = \mathbf{u}\cdot\mathbf{w}^{\perp}$ and $k_M(\xi) = \int_{-\infty}^{\infty} k(\mathbf{u})\, \mathrm{d}\eta$. Thus, integration of $k(\mathbf{x})$ in the direction perpendicular to the population spread $\mathbf{w}$ results in the 1D marginal distribution $k_M(\xi)$, and this kernel can then be used to calculate the MGF in Eq. (7).

3. Calculating invasion speeds in practice

3.1. Invasion speed from seed trap data

When the dispersal kernel of seeds is unknown, but their densities are being sampled by seed traps at various distances from a source plant, one natural distribution to employ is the histogram

$$k_h(x) = \begin{cases} f_i, & \text{if } \xi_{i-1} \le x \le \xi_i \quad \text{for } -L \le |x| \le L \\ 0, & \text{otherwise,} \end{cases} \tag{12}$$

where $k_h(x)$ is the histogram for the distribution of seeds from a parent plant at location $x = 0$ in a field of length $2L$, and f_i is the relative frequency of seeds on the interval i with end points ξ_{i-1}, ξ_i. Here, the first and last histogram points are $\xi_{L_0} = -L$ and $\xi_{L_j} = L$. We assume that $k_h(x)$ is the "true" histogram describing dispersal of a population of seeds. When we estimate $k_h(x)$ from limited field data we write $\hat{k}_h(x)$.

In order to calculate an invasion speed using the histogram (Eq. (12)) for the dispersal kernel, its moment generating function has to be derived first, that is

$$M_h(s) = \frac{1}{s} \sum_{i=1}^{L_j} f_i(\exp(s\xi_i) - \exp(s\xi_{i-1})) \tag{13}$$

This moment generating function is used in (9) to yield a speed of

$$c_h = \min_{s>0} \left\{ \frac{1}{s} \ln(R_0 M_h(s)) \right\}. \tag{14}$$

Here, the index h indicates that the dispersal kernel is taken to be a histogram. Of course, the histogram (Eq. (12)) does not exactly describe the true distribution of seeds in a field, and the histogram assumption can introduce some small bias in the calculation of the wave speed due to the arbitrary chosen number of "bins" (Clark et al., 2001). However, simulations show that, given at least 20 "bins", such bias is extremely small relative to errors arising from uncertainty associated with having very few observations in the tail of the dispersal kernel. Thus, this bias is of little practical significance when designing field studies. For the remainder of this paper, this source of bias is ignored.

When field data consists of seed counts on a lattice, a 2D wave speed can be obtained. The 2D planar wave front can be calculated from field data in a similar way as described above for the 1D case. That is, calculate the marginal distribution of the initial 2D dispersal kernel to get the MGF (see Section 2.2, Eq. (11)) and use its result in Eq. (14) to calculate the invasion speed.

3.2. Monte-Carlo methods

Suppose we have an estimate for the dispersal kernel $k(x)$. How do we calculate the corresponding invasion speed c? Analytical or numerical integration of the kernel is one choice to calculate the MGF (Eq. (7)) needed to calculate c. However, in higher dimensions Monte-Carlo simulation is easier to apply. For example, any arbitrary probability density function $k(x)$ can be approximated by forming a histogram derived of many independent and identically distributed (i.i.d.) random samples from k.

When $k(\mathbf{x})$ describes dispersal in two spatial dimensions, its marginal distribution in direction $\mathbf{w}$ can be approximated by a histogram $k_h(z)$, where z is the signed distance in direction $\mathbf{w}$ by the following procedure. To do this, first generate many i.i.d. random samples from k. Then, for each random sample, calculate $\xi = \mathbf{x} \cdot \mathbf{w}$. Lastly, generate a histogram of the ξ values and use Eqs. (13) and (14) to give the spread rate c_h. This method is used in Section 6.2 where the expected spread rate $\bar{c}_h$ is a Monte-Carlo estimate of c. While Monte-Carlo methods give inexact solutions they are simple to use and accurate enough for the question at hand.

4. "Optimal" sampling design for seed dispersal

As the following procedures apply to an optimal configuration of seed traps in the field, assumptions about the physical field characteristics have to be made explicit first. As a first approach, we assume the field is a flat terrain without any major vegetation growth and no predominant wind direction. The experimental set-up consists of a source of one or more parent plants bunched together as a single point source surrounded by seed traps in one or more wind directions (e.g. north–east, south–east, south–west and north–west). Collecting data from the field results in number of seeds per seed trap (i.e. per surface area) at various distances from the source.

4.1. Optimality criterion

4.1.1. An estimate for the invasion speed from limited field information

The first step into calculating an invasion speed from field data is to have a general insight in the reproduction and dispersal pattern of the test species. The next step is then to link these population characteristics to an invasion speed. Following the assumptions in Section 2, the calculation actually only needs two inputs (see Eqs. (12)–(14)), that is,

1. R_0 : basic reproductive number
2. $k_h(x)$: histogram of dispersal distances

This means that the extent of species invasiveness depends on the number of seeds that will germinate the next growing season at sites they were dispersed to from a parent plant.

If we knew the true spatial distribution of seeds from a parent plant, then we could immediately calculate the spread rate of the population using Eq. (9), and no field sampling would be necessary.

Here, we consider the case where we do not know the precise form of the dispersal kernel covering the long distance component. However, we assume some preliminary sampling has been done which provides us with preliminary estimates for the dispersal kernel spread rate c_h (14). The error in these estimates depends on the level of preliminary sampling. However, for reasons explained in Section 3.1, this source of bias is ignored for our test system and so we proceed as if c_h were the "true" wave speed. In the example given in Section 5, the quantity $\bar{c}_h$ represents the average of wave speeds c_h generated by Monte-Carlo methods. Here, each wave speed calculation for c_h used Eq. (14) where M_h was based on the histogram generated by the Monte-Carlo Simulation (Section 3.2).

We now ask how to distribute seed traps in subsequent sampling efforts so that we can use this new data to most closely estimate the "true" expected spread rate. We assume that this subsequent sampling gives an estimate for the expected spread rate, i.e. $\hat{c}_h$. Closeness to the "true" expected spread rate is achieved through minimizing the mean squared error of $\hat{c}_h$. In other words, our goal is to minimize the variance plus bias squared

$$\mathrm{MSE} = E(\hat{c}_h - c)^2 = \mathrm{Var}(\hat{c}_h) + (E(\hat{c}_h) - c)^2. \quad (15)$$

Of course, lack of information on the true spread rate c hampers our ability to calculate the bias. However, the expected spread rate $\bar{c}_h$ (Section 3.2) allows us to minimize an approximation to the MSE

$$\mathrm{MSE} = \mathrm{Var}(\hat{c}_h) + (E(\hat{c}_h) - \bar{c}_h)^2. \quad (16)$$

4.1.2. Practical considerations with an application to seed dispersal

The ideal seed trap configuration would consist of placing a maximum number of traps that fit in each transect surrounding the source plant (i.e. divide the length of the transect by the diameter of a seed trap to get this maximum number). However, filling all the transects with the maximum number of seed traps would, from a practical point of view, be impossible in most cases. From a trapping efficiency viewpoint, seed traps were chosen to be 10 cm in diameter and four transects were used each with a length of 100 m. This means a maximum of 1000 seed traps would fit in each direction, whereas a manageable number appeared to be at most 300 per transect. Therefore, the optimality criterion is to define the locations of at most 300 traps in the field in such a way that the mean squared error (Eq. (16)) of the estimated average invasion speed with a reduced number of seed traps, $\hat{c}_h$, with respect to the "true" calculated average invasion speed using the maximum number of seed traps, $\bar{c}_h$, is minimized. Results will show how much more information is gained, i.e. how much the MSE decreased, with each additional placed seed trap.

5. The procedure: dispersal in 1D

Eq. (14) shows that, in order to calculate a "true" wave speed, an initial guess on the distribution of seeds over a whole field transect has to be *made* (Eq. (12)). However, the number of seeds found in a particular seed trap will differ for separate dispersal events (i.e. from year to year). So, the first step into designing an "optimal" seed trap configuration is the generation of various data sets representing the *initially assumed* dispersal pattern of *C. vulgaris* over a transect using the histogram of Eq. (12). With those spatial distributions a "true" expected wave speed ($\bar{c}_h$) can be calculated (Fig. 2a). Then the "optimal" design procedure (Fig. 2b) includes the following steps: (i) start with an

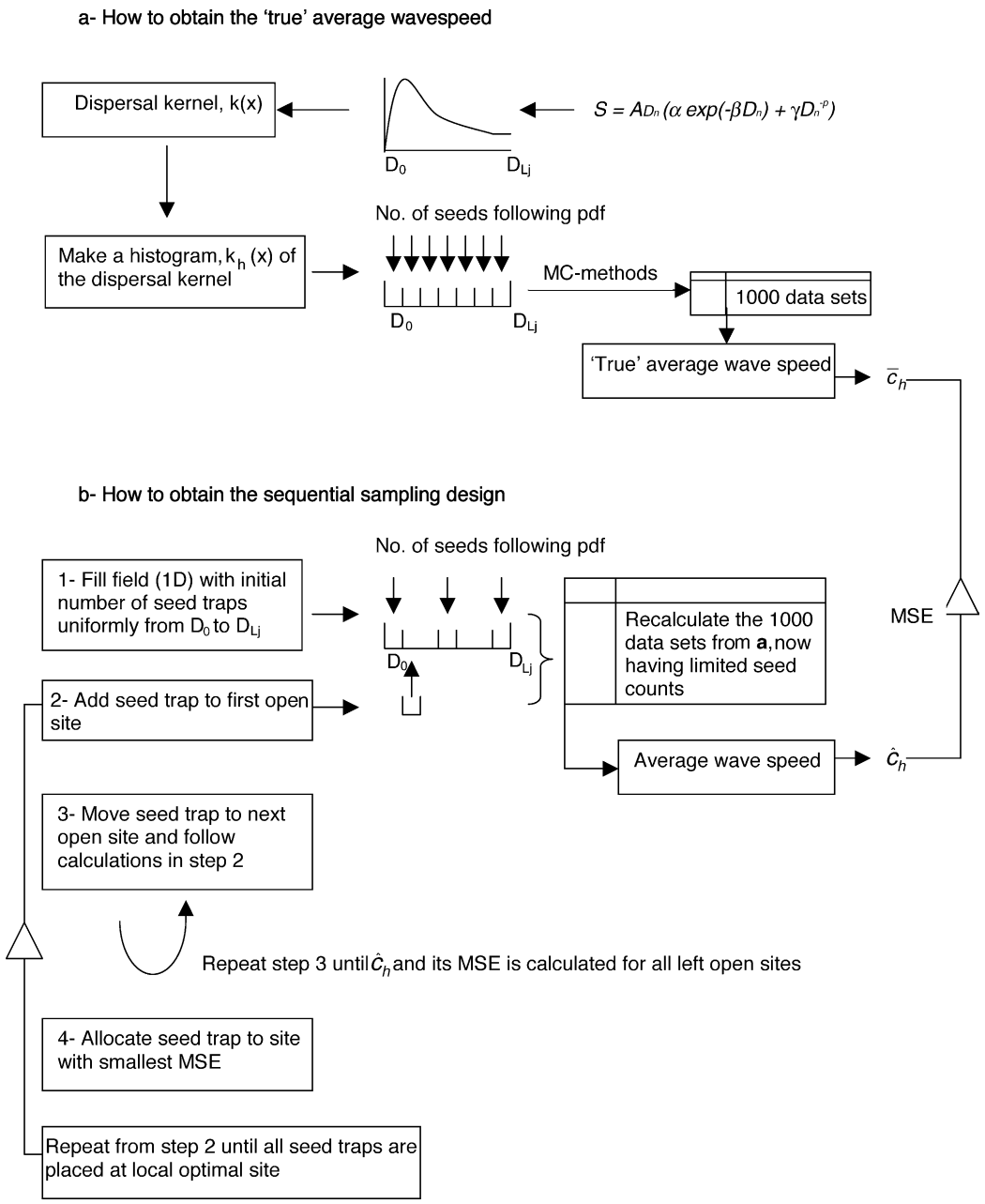

Fig. 2. Schematic representation of calculating a "true" average spread rate, $\bar{c}_h$ (a) and the sequential sampling design (b). (MC-methods refers to the Monte-Carlo simulations as described in Section 3.2.)

initial limited number of seed traps over the sampling domain. (ii) Find the location of the next trap in such a way that the mean squared error of the now estimated average wave speed (i.e. $\hat{c}_h$) is minimized with respect to the calculated "true" expected wave speed ($\bar{c}_h$). (iii) The sample location resulting in the smallest MSE is the location of the next seed trap in the field. (iv) This process should then be repeated until the maximum

number of samples that can be taken from a practical point of view are assigned to a location. A summary of the procedure is given in Fig. 2b and will be explained in more detail in the next two sections.

5.1. Calculating the "true" spread rate, c_h, for C. vulgaris

Recently, Bullock and Clarke (2000) sampled dispersal of *C. vulgaris* seeds over distances up to 80 m using field traps. They summarized their data with an empirically derived probability density function of the form

$$k(x) = \begin{cases} T_A\,(\alpha\,\mathrm{e}^{-\beta x+z} + \gamma\,(x+z)^{-\rho}), & z \le x \le 100\,\mathrm{m} \\ 0, & \text{otherwise,} \end{cases} \quad (17)$$

where T_A represents the total sample area at distance x, and α, β and γ are positive parameters. The variable z represents the edge of the point source and dispersal of seeds beyond 100 m was not measured. For the purpose of this discussion we assume no seeds disperse further than 100 m (see discussion later in this section). The power function $x^{-\rho}$ decreases much more slowly than $\mathrm{e}^{-\beta x}$ for large x. Thus, the term $x^{-\rho}$ accounts for the long-distance component of the dispersal pattern. Although this model lacks a mechanistic basis, terms scaling like $\mathrm{e}^{-\beta x}$ occur in simple mechanistic dispersal models (Neubert et al., 1995) and the term $x^{-\rho}$ also appears when the spatial spread of seeds is derived from the physics on wind flow patterns (Okubo and levin, 1989).

Of course, when some initial parametric form of the dispersal kernel is known, an invasion speed can be estimated using Eq. (9) instead of using the histogram estimator as presented in Section 3.1. However, often *limited* field data is all we have. Therefore, rather than constructing a histogram of their raw dispersal data (which is not published in their paper), we mimic the process of field sampling by generating 10,876 i.i.d. dispersal distances using Eq. (17), use this to form a histogram (Fig. 4), calculate a histogram spread rate c_h and repeat this many times to calculate a mean spread rate which is taken to be the "true" spread rate $\bar{c}_h$.

To calculate the "true" invasion speed for *C. vulgaris* both R_0 and the distribution of seeds in the field (Eq. (14)) have to be known. From a Lefkovitch matrix model R_0 was found to be approximately 2. One thousand data sets were used to generate histograms representing seed dispersal, following the empirical equation (Eq. (17)) (Bullock and Clarke, 2000). This equation can be extended to

$$S = \begin{cases} \sum_{n=1}^{L_j} S_{D_n} = \sum_{n=1}^{L_j} A_{D_n}(\alpha \exp(-\beta(D_n+z)) + \gamma(D_n+z)^{-\rho}), & \text{for } z \le D_n \le 100\,\mathrm{m} \\ 0, & \text{otherwise.} \end{cases} \quad (18)$$

The expression $(\alpha \exp(-\beta D_n) + \gamma D_n^{-\rho})$ in Eq. (18) gives the number of seeds found at distance D_n (n = 1, 2, ..., total number of sampling distances (L_j)) from the edge of the source (z). The parameters α, β, γ and ρ ($\rho > 1$) describe the shape of the seed distribution. in order to get long tails in the dispersal curve. Multiplying this expression with the sample area at distance D_n, A_{D_n}, results in the total number of seeds trapped at some distance from the source, S_{D_n}. Summing the number of seeds found at all distances, n = 1, 2, ..., L_j, results in the total number of seeds over all sampling distances, S. Values for the separate terms in this study were found from literature and from this particular field set-up (Table 2).

Seed traps were 10 cm in diameter in this study and only one trap was used per distance, so A_{D_n} is constant (i.e. $7.854 \times 10^{-3}\,\mathrm{m}^2$) at each distance. The sampling domain (from $-L$ to L) stretched over an area of 40,000 m^2 which made the furthest sampling distance from the edge of the source in the center of the field to be 100 m. This means a maximum of 1000 seed traps fitted in each direction, i.e. $n = 1, \ldots, 1000$. As

Table 2
Parameter values and variables used for Eq. (18) in this study (based on Bullock and Clarke (2000))

Parameters and variables	Value
S	10876
A_{D_n}	$7.854 \times 10^{-3}\,\mathrm{m}^2$
L_j	1000
α	72×10^6
β	8.42
γ	5098
ρ	1.46

no predominant wind direction is assumed, the average values over Tables 1–3 from Bullock and Clarke (2000) were used for the parameters α, β, γ and ρ as an initial approximation. Finally, the total number, S, was calculated to be 10,876 seeds being caught in the seed traps. This calculation was based on having the first trap being placed at the edge of the bush, i.e. having $D_n = D_n + 0.5$ m, since the bush was 1 m in diameter.

As heather bushes may produce different seed dispersal patterns in separate experiments, 1000 separate data sets were generated from Eq. (18) using Monte-Carlo methods and stored as histograms following Eq. (12) to simultaneously being used for optimization purposes. That is, each data set was simulated using a random generator together with the inverse CDF method, applied to the cumulative distribution function of Eq. (18), to assign a deposition site n to each released seed. This means 10,876 numbers between 0 and 1 were generated representing the seeds that were blown into the field from the heather bush. Then the seed trap distance D_n each seed would fly in was calculated following the CDF extracted from the PDF (i.e. Eq. (18)). As the sampling domain is restricted to some distance from the source, one should be aware that some seeds will always fly out of the domain. Simulations showed that, in this case, about 10 seeds would travel further than 100 m, i.e. 0.1% of the total number of "released" seeds.

The next step towards an "optimal" sampling design was to first calculate the expected "true" speed ($\bar{c}_h$, Eq. (14)) of the traveling wave evolving from the application of Eq. (18) for the histogram of seed dispersal. That is, a speed when all possible locations, $D_1, \ldots, D_{L_j}$, of the domain in a 1D setting are filled with seed traps continuously.

Fig. 2a gives an overview of how to obtain the "true" expected wave speed. A detailed description of the procedure is given in Appendix A.

The data sets generated using the CDF method (Appendix A and Fig. 2a) on the continuous function (Eq. (18)) were saved in the form of a histogram needed to describe the dispersal kernel (Eq. (12)). In this study, the transect was divided in $i = 100$ intervals each having a length of 1 m and so $n = 10$ seed traps were incorporated in an interval from ξ_{i-1} to ξ_i. Generated data sets with a number of seeds at each distance D_n (Eq. (18)) were distributed over the intervals in order to get the histogram for the dispersal kernel. The expected invasion speed, $\bar{c}_h$, for *C. vulgaris* from the 1000 generated histograms was calculated to be $\bar{c}_h = 358\,\text{cm year}^{-1}$.

5.2. The sequential sampling design

As a starting point towards the sequential sampling procedure, a limited number of seed traps ($t = 100$) were evenly spaced over the sampling region (i.e. in a line from the source plant up to the end of the domain with length L). Then, seed traps were added sequentially. The location of each new seed trap was chosen so as to minimize the MSE as given in Eq. (16). This meant that for every following seed trap, an entire distribution of average $\hat{c}_h$ values was calculated corresponding to every possible location of the new seed trap. In other words, for every possible new location, joint with the already occupied sites, an average $\hat{c}_h$ was calculated using the histogram formulas (Eqs. (12)–(14)) applied to the 1000 generated data sets (see Section 5.1). The difference with calculating a $\bar{c}_h$ is that now the histograms used to calculate $\hat{c}_h$ consisted of limited information, i.e. having seed counts only at sites where seed traps are available.

A detailed step-wise procedure for obtaining a sequential sampling design is given in Appendix B and is visualized in (Fig. 2b).

As wind is assumed to be uniform in all directions, the resulting location of the seed traps (following the procedure in Appendix B) can be put in any transect opposite to the parent plant.

6. The procedure: dispersal in 2D

The procedures presented up until now are based on 1D analysis, whereas, of course, seeds are being spread in 2D. This section describes two possible methods on how to proceed when an optimal design is to be obtained taking a 2D spread into account. The first method will show how the sampling algorithm can be changed relatively easily to allow for aggregation of seed traps, meaning to allow for more than one seed trap at the same distance from the source. More than one seed trap can be placed at the same distance from the source by filling adjacent sites along the circumference for a certain radius. In addition to this, a more elaborate change will show how to make a sequential sampling

design when the marginal distribution is derived from the dispersal kernel (Eq. (18)).

6.1. Aggregation of seed traps

Bullock and Clarke (2000) used a sampling design where the number of seed traps increased with distance from the source. The algorithm presented in this study so far, however, only allows for at most one seed trap at each distance from the source. Therefore, their sampling design could never be verified using an optimal design based on invasion speeds. To test whether the design would change towards a more aggregated form of pot placement similar to the design by Bullock and Clarke (2000), the procedure was adjusted to allow for more than one seed trap at any distance D_n. More specific, it means that instead of decreasing the number of open sites with increasing number of placed pots, the number of available sites stays constant (i.e. the maximum available) from the start until the end of the procedure. The algorithm for sequential sampling as presented in Section 5.2 will only change with respect to the number of places tested for each new seed trap. Whereas seed traps were only placed at open sites in the 1D setting, now all sites will be tested for each additional seed trap. This results, in this case, in $(T - t)$ is 200×1000 sites to be tested to find the sampling design instead of testing $900 \times 899 \times 898 \times \cdots \times 700$ sites in the 1D case.

6.2. Applying the marginal distribution

Not only the distance a seed is spread, but also its direction becomes important when an optimal sampling design is based on invasion speeds in 2D. That is, when a seed is spread parallel to the wave front this seed does not add anything to the invasion speed no matter how far it is being spread (Fig. 3). Subsequently, when the angle of spread with respect to the wave front increases up to 90° downwind, its contribution to the wave speed increases to a maximum. Therefore, the wave speed should be a weighted function of the angle under which seeds are being deposited from the source.

One approach to account for a weighted wave speed is calculating the marginal distribution of seed deposition in the direction of the wave front. For example, have seeds being spread in 2D (Fig. 3). And, assume the wave front is represented by a vector $\mathbf{w} = (1, 0)$. Then, integrating over y will result in a distribution kernel, which is now a kernel in 1D again with seeds being redistributed in the direction of the wave front (Fig. 3). This distribution kernel will differ qualitatively from the original model used by Bullock and Clarke (2000) (Eq. (18)). That is, an increased proportion of the original number of released seeds will be deposited at D_0 (because of spread parallel to the wave front) and, therefore, not contribute to the wave speed at all. In addition, the effect of including an angle in the deposition will increase towards the end of the domain. The ultimate distance being spread from the source (in 1D) will decrease relatively more for seeds that were initially deposited far from the source than for those that were deposited close by (in 2D) due to the application of the marginal distribution.

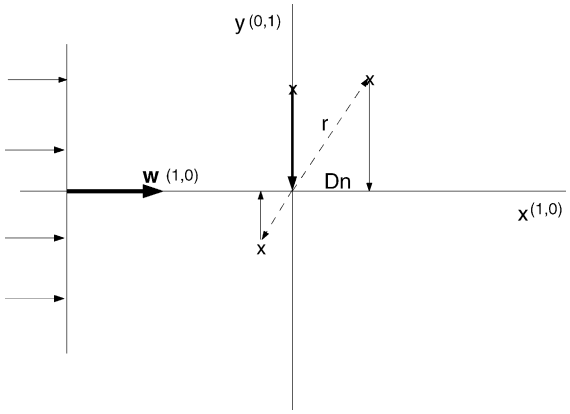

Fig. 3. Schematic representation of calculating a marginal distribution ($D_n = r\ \cos(\theta)$) of seeds originally being spread in 2D.

This theory can be applied to seed dispersal by calculating a marginal distribution in the x direction to find the dispersal of heather seeds from a point source. First step is to transform the 1D model (Eq. (18)) into a 2D model, i.e. write D_n in terms of a (x, y) coordinate. So,

$$S = \sum_{r=1}^{L_j} S_r = \sum_{r=1}^{L_j} A_r(\alpha\ \exp(-\beta r) + \gamma r^{-\rho}) \quad \text{for } r \leq 100, \tag{19}$$

where $r = \sqrt{x^2 + y^2}$. So, r is the distance from the source where seeds are deposited (Fig. 3). However, now having an actual position in a 2D plane assigned to it. With $S = 10{,}876$ seeds being released from the source and knowing the distance D_n each seed travels (see Section 5.1), an x, y coordinate can be assigned to each seed. Seeds will be randomly distributed in the field if no predominant wind direction is assumed. That is, draw S random numbers between 0 and 2π representing the angles (θ) assigned to each distance D_n in order to get accompanying values for r. Then $x = r\cos(\theta)$ and $y = r\sin(\theta)$. And so $x = D_n = r\cos(\theta)$ will give the marginal distribution of seeds in 1D when the wave front is parallel to the y direction.

The calculations then follow the procedure as explained in the previous sections and lead to an expected wave speed, $\overline{c}_h$ of 152 cm year^{-1}.

7. Results

7.1. Dispersal in 1D

Fig. 4 shows one of the histograms, $k_h(x)$, for seed dispersal used to obtain an optimal sampling design in 1D. The histogram is generated from the dispersal kernel fitted by Bullock and Clarke (2000) for field data of *C. vulgaris* (Eq. (18) and Table 2). Note that most of the 10,876 released seeds were deposited near the source and that the first two bars (corresponding to the first two meters) in Fig. 4 actually, on average, account for 10,812 of the dispersed seeds.

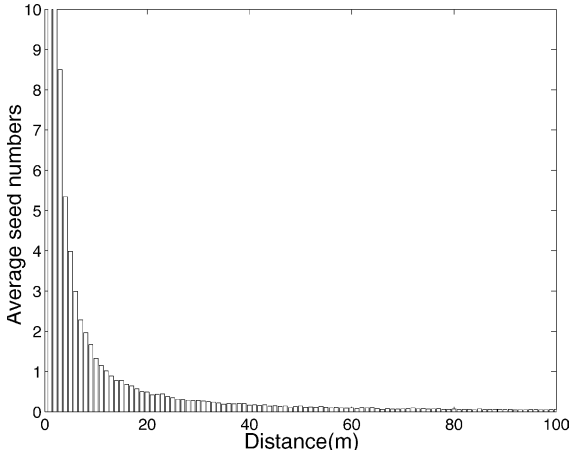

Fig. 4. One of the 1000 seed dispersal kernels, $k_h(x)$ (Eq. (12)) for the original field data of Bullock and Clarke (2000) using Eq. (18) and parameter values from Table 2. Note that the scale of the y-axis has been set to an upper limit of 10 in order to visualize seed numbers at all distances. The first two bars (2 m) actually account for over 95% of the dispersed seeds.

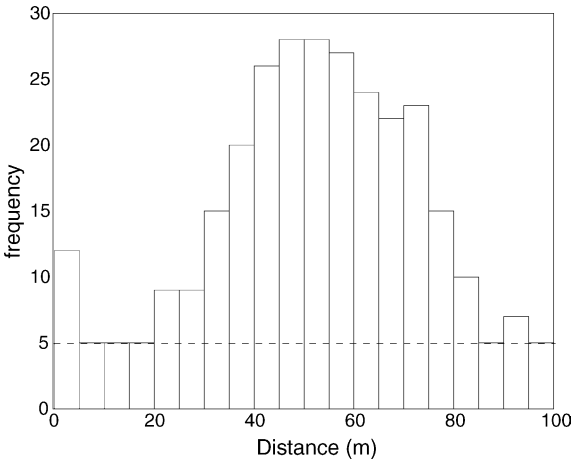

Fig. 5. Optimal seed trap location using histograms $k_h(x)$ (Eq. (12)) representing the original field data of Bullock and Clarke (2000) and $R_0 = 2$. The dashed line indicated the frequency of the initial 100 equally spaced pots.

The optimal location of seed traps when *C. vulgaris* seeds are dispersed from a point source is shown in Fig. 5. That is, the optimal placement of an extra 200 pots on top of an initial uniform distribution of 100 pots (indicated by the dashed line in Fig. 5) over a 1D domain of 100 m. The optimal design is based on implementing the dispersal pattern as shown in Fig. 4 and an R_0 of 2 in Eq. (14) followed by a minimization of the MSE with respect to the "true" wave speed in this setting.

The seed trap configuration in Fig. 5 can be explained both from a mathematical and biological point of view. As the tail of the distribution kernel gives the most information on invasiveness of species a majority of the seed traps should be put there. From a biological viewpoint it is known that the probability of finding a seed is very small at the very edge of the domain. Therefore, it is better to sample a little closer to the source so both qualitative and quantitative information is guaranteed.

Figs. 6 and 7 show the change in the components of the MSE (Eq. (16)) when the field is being filled with

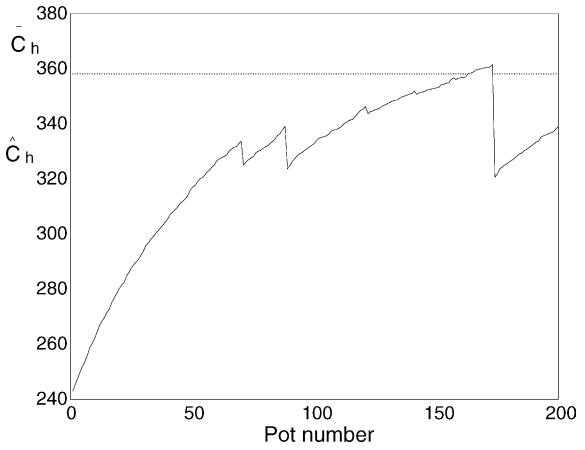

Fig. 6. Estimated average wave speed $\hat{c}_h$ (cm year^{-1}) approaching the "true" average wave speed $\bar{c}_h$ (358 cm year^{-1}) with increasing number of placed seed traps when the original field data of Bullock and Clarke (2000) and $R_0 = 2$ is used.

seed traps giving rise to the optimal seed trap configuration of Fig. 5. The figures show that there is a trade-off between having a small bias of the estimator $\hat{c}_h$ and decreasing the variance of $\hat{c}_h$ when it comes to choosing the best site for the next pot. In other words, a trade-off between precision and accuracy of the estimator. The relatively big increase in the bias of $\hat{c}_h$ when placing pot number 71, 89 and 174 (Fig. 6) is apparently overcompensated with a decrease in the Var($\hat{c}_h$) in order for the MSE to still decrease, Fig. 7. The relatively large decrease in the variance of $\hat{c}_h$ for these pots corresponds to placing a seed trap close to the source (i.e. at 60, 50 and 30 cm, respectively, not shown). Therefore, the estimated wave speed $\hat{c}_h$ dropped with respect to the "true" wave speed $\bar{c}_h$, but apparently gave rise to a significant decrease in Var($\hat{c}_h$) in order for the MSE to decrease.

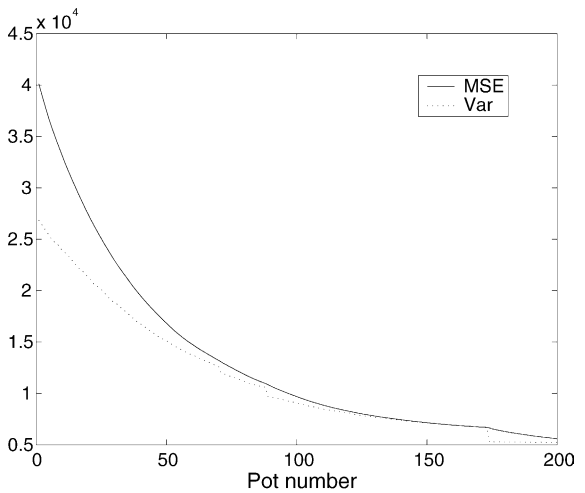

Fig. 7. Decrease in the mean squared error (cm^2) of $\hat{c}_h$ with respect to $\bar{c}_h$ and change in the variance of $\hat{c}_h$ with increasing number of placed seed traps when the original field data of Bullock and Clarke (2000) and $R_0 = 2$ is used.

7.2. Dispersal in 2D

Allowing for aggregation of seed traps, i.e. the possibility of placing more than one seed trap at the same distance from the source plant, did not change the optimal sampling design shown in Fig. 5. This indicates that spreading seed traps over more distances is preferred over aggregation when 300 seed traps are to be placed in a field where a maximum of 1000 would fit in 1D.

Calculating the marginal distribution from an initial seed dispersal in 2D when a random wind frequency distribution was applied resulted in the histogram representing the dispersal pattern in 1D as shown in Fig. 8. Note the change in scale on the y-axis for this figure compared to Fig. 4, indicating a seed

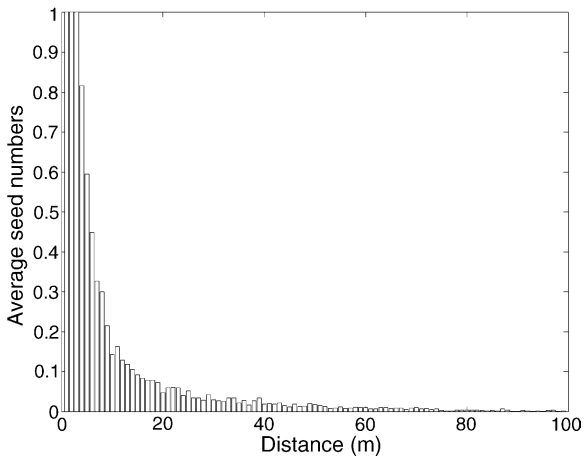

Fig. 8. One of the 1000 seed dispersal kernels, $k_h(x)$ (Eq. (12)) for the field data of Bullock and Clarke (2000) applying the marginal distribution to Eq. (19) with a random wind frequency distribution. Note that the scale of the y-axis has been set to an upper limit of 1 in order to visualize seed numbers at all distances. The first three bars (3 m) actually account for over 95% of the dispersed seeds.

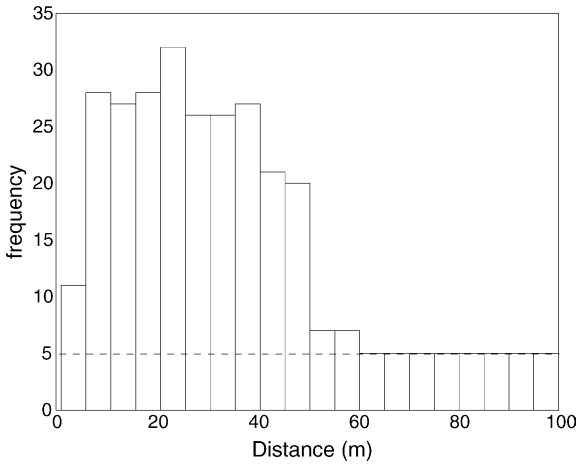

Fig. 9. Optimal seed trap location using histograms $k_h(x)$ (Eq. (12)) when the marginal distribution to Eq. (19) with a random wind frequency distribution is applied and $R_0 = 2$. The dashed line indicates the frequency of the initial 100 equally spaced pots.

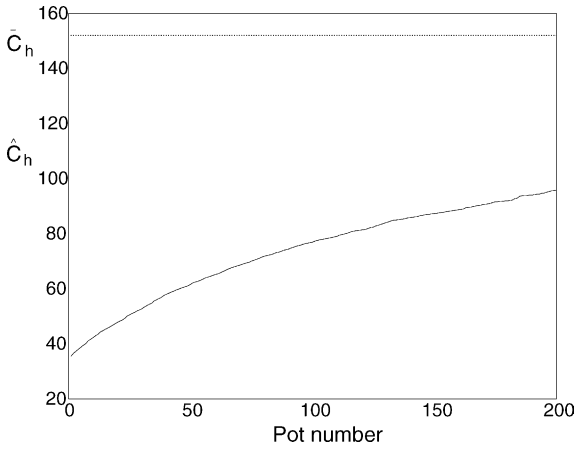

Fig. 10. Estimated average wave speed $\hat{c}_h$ (cm year^{-1}) approaching the "true" average wave speed $\bar{c}_h$ (152 cm year^{-1}) with increasing number of placed seed traps when the marginal distribution to Eq. (19) with a random wind frequency distribution and $R_0 = 2$ is used.

dispersal pattern which is concentrated even closer to the source. This resulted in a left shift in the optimal sampling design (Fig. 9). As very few seeds are being dispersed at larger distances, a sample effort near the source is preferred over seed traps at relatively long distances. Catching the tail of the seed dispersal kernel means putting seed traps relatively close to the source where the probability of finding a seed is actually significant.

Figs. 10 and 11 show the statistics resulting in the pot configuration presented by Fig. 9. In general, the algorithm chose to put subsequent seed traps further away from the source (not shown) causing the bias of the estimator $\hat{c}_h$ to decrease relatively more (Fig. 10) than its variance to increase in order to still have a decrease in the MSE (Fig. 11).

The application of the 1D model (Eq. (17)) compared to using a 2D model (Eq. (19)) shows clear differences in the preliminary calculated "true" invasion speed c_h. That is, the 1D model results in a "true" invasion speed of 358 cm year^{-1} (Fig. 6), a value which is more than twice as high than the true invasion speed calculated from the 2D model, which is 152 cm year^{-1} (Fig. 10).

The general pattern of seed trap configuration did not change in any of the simulations when different values of R_0 in the range from 1 to 10 were tested.

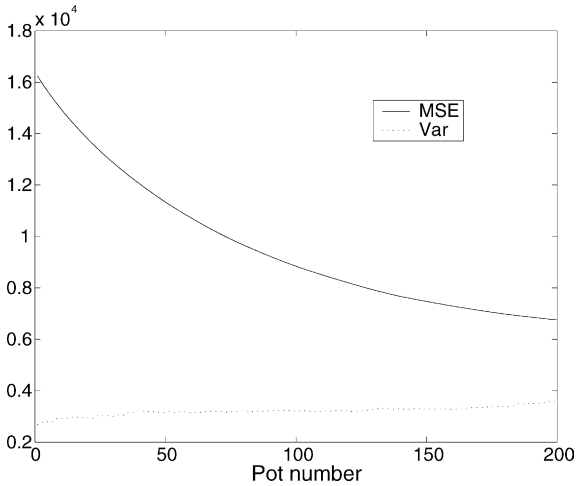

Fig. 11. Decrease in the mean squared error (cm^2) of $\hat{c}_h$ with respect to $\bar{c}_h$ and change in the variance of $\hat{c}_h$ with increasing number of placed seed traps when the marginal distribution to Eq. (19) with a random wind frequency distribution and $R_0 = 2$ is used.

8. Discussion

Long-distance seed dispersal is a relatively rare event during a plant's stage of seed spread. As most seeds will be deposited close to the parent plant, the probability of finding a seed in field experiments decreases with distance from the source due to spatial

aspects. Still, it is those long-distance dispersal events which facilitates plants to invade into new habitats and, therefore, enhance species survival. As a consequence, a statistically well justified sampling design is of major importance when gaining information on these highly unpredictable tails of a dispersal pattern from field experiments is the study objective.

The calculation of a species' invasion speed emphasizes what happens in the tail of a dispersal kernel. A sampling design based on this population characteristic is therefore an obvious approach. This calculation has the additional advantage that it only requires general insights in the growth and redistribution dynamics of the population (expressed in R_0 and a dispersal kernel $k(x)$ for the spatial spread, respectively). After the actual field experiment the sampling design can be adjusted according to improved insights into the population dynamics. Moreover, the simulations for this study revealed that the sampling design is generally insensitive to an R_0 in the range of 1–10.

Typical parametric dispersal kernels include the Gaussian, and Laplace (back-to-back exponential) kernels. Use of the histogram to calculate wave speeds gives an added advantage over typical parametric descriptions of $k(x)$ when the precise shape of the dispersal kernel is not known a priori. Although this method will not work if the data collector had no idea whatsoever of the distribution of dispersal distances, yet the histogram formulation allows for a high level of flexibility in the shape of the kernel. Indeed, the histogram can be considered to be a parametric formulation with a high number of L_j parameters, one corresponding to each "bin height" f_i . This high number of parameters allows one to accurately depict the shape of the tail of the dispersal kernel. Kot et al. (1996) show that simple parametric (one or two parameter) kernels used to calculate spread rates can introduce significant bias in the wave speed estimates. In the example shown in Kot et al. (1996), speeds varied by an order of magnitude when different parametric forms were fitted to classic insect dispersal data from Dobzhansky and Wright (1943). Related issues, including a non-parametric estimator are discussed in Clark et al. (2001).

In this paper we apply Monte-Carlo methods to create multiple seed dispersal histograms. For the Monte-Carlo simulations we used the dispersal kernel fitted to field data by Bullock and Clarke (2000) (Eq. (18)). However, when only field data are available Eq. (12) can be applied directly to create a histogram. Then, the method of bootstrapping can be used to generate multiple data sets to be used for further calculations.

The procedure for sequential sampling as described in this paper reveals why this method is actually suboptimal. To find the optimal sampling design in this setting the procedure should calculate the MSE of $\hat{c}_h$ with respect to $\overline{c}_h$ for all possible site combinations in which t seed traps could be placed in the field. That is, comparing the MSE of $\hat{c}_h$ with respect to $\overline{c}_h$ for all $\binom{t}{L_j}$ possible seed trap location combinations. As this is a computationally impossible task a suboptimal algorithm of sequential sampling is used. In this case, the MSE of $\hat{c}_h$ with respect to $\overline{c}_h$ for all $(L_j - (t+1))$ left locations is calculated each time a new seed trap is placed in the field.

The MSE of the estimator $\hat{c}_h$ is used as the actual statistic for the optimal pot configuration. This statistic has two components; the variance of the estimator ($\mathrm{Var}(\hat{c}_h)$) and its bias ($E(\hat{c}_h) - \overline{c}_h$). This means that with the attempt of approaching the "true" expected wave speed $\overline{c}_h$ when searching for the optimal position of the next pot in the field, also minimizing the error of $E(\hat{c}_h)$ with respect to the X $\hat{c}_h$ values calculated from the X generated dispersal histograms plays an important role in the optimal design. For this reason $\hat{c}_h$ does not approach $\overline{c}_h$ continuously in Fig. 6, but drops when the algorithm chooses to put a pot close to the source in order to obtain a relatively larger decrease in the variance (Fig. 7). This means that although emphasis is put on calculating a "true" wave speed, $\overline{c}_h$, from limited field data, still the algorithm makes sure that the variability in the actually estimated wave speed, $\hat{c}_h$, with this limited information is minimized. From a practical viewpoint this demonstrates that, while the "tails" of the dispersal kernel can play a dominant role in the estimation of population spread, in actual fact optimal sampling effort, given a fixed number of seed traps involves a trade-off between nearby sampling (many seeds, no long-distance dispersers) and distant sampling (few seeds, long-distance dispersers). This trade-off occurs because minimizing the mean squared error (bias squared plus variance) of the estimator requires a balance between accuracy (which comes from accurately measuring the tails) and the precision (which comes from having a large enough sample size).

Furthermore, the trade-off between precision and accuracy meant that the sequential sampling design assigns no seed traps (beyond the initial baseline level) to the furthest dispersal distances (see Figs. 5 and 9). In other words, given the finite number of seed traps, sampling of the tails of the dispersal kernels beyond 100 m is insignificant.

Although using invasion theory to obtain an optimal sampling design has a solid basis from a statistical viewpoint, experimental drawbacks will always exist. That is, gaining insight into the tail of a distribution kernel requires sampling at long distances from the source. However, as the sampling domain is always restricted in a practical sense, it is impossible to get all the information in the tail. As a consequence, the estimated wave speed, $\hat{c}_h$ is compared to a "true" wave speed, $\overline{c}_h$ on a restricted domain. The sampling design will, therefore, always only be an approximation for the real optimal seed trap configuration on an "infinite" domain. Of course, it depends on the shape of the initially assumed distribution kernel how far the domain should be extended and, with that, the introduced sampling error will depend on the population's dispersal pattern. In this case, using Eq. (18) as the initial dispersal kernel, only 10 out of about 10,000 released seeds traveled out of the domain ($L = 100$ m), i.e. about 0.1% calculated from simulations.

In addition, Figs. 6 and 10 show the huge difference in the calculated "true" wave speed $\overline{c}_h$ when dispersal in 2D is considered compared to dispersal in 1D. The population invades less than half as fast in the 1D versus the 2D case. Also, Fig. 10 shows a relatively huge gap between $\hat{c}_h$ and $\overline{c}_h$ compared to Fig. 6. Apparently there is not enough seed traps used to be able to capture the little information in the tails needed to reach an invasion speed of 152 cm year^{-1} (Fig. 8). As more seeds are found at long distances from the source in a 1D setting (Fig. 4) less seed traps have to be put at long distances to still be able to capture the tail of the dispersal kernel and so the gap between $\hat{c}_h$ and $\overline{c}_h$ is less (Fig. 6). This statistical error (as arising from a finite sample size) will decrease when the sample size N increases. In that case, the "variance" component of the mean squared error will approach zero.

Based on this study, an optimal sampling design to catch the tail of dispersal kernels involves sampling towards the end of the domain. However, sampling effort should depend on the amount of information gained and, therefore, on the thickness of the tail towards the end of the sampling domain. Therefore, a slight spread of samples about the main sample effort to catch the tails is needed to still gain enough information for further analysis.

Acknowledgments

M.A. Lewis gratefully acknowledges funding from NSERC (CRO and Discovery grants), the NSF (DEB 02-13698) and a Canada Research Chair. S. Lele acknowledges funding from NSERC. We are grateful to James Bullock for helpful discussion.

Appendix A. Calculation of the "true" expected wave speed

1. Make a histogram, $k_h(x)$ (Eq. (12)), of the preliminary known dispersal kernel $k(x)$ using Monte-Carlo methods, i.e.
   (a) Derive the cumulative distribution function (CDF) for the kernel $k(x)$. This results in a function which outcome $F(x)$ lies between 0 and 1 for every value of x.
   (b) Define S, the total number of seeds that will be released from the plant source in the field.
   (c) Use a random generator to generate S numbers between 0 and 1.
   (d) Each generated number represents an outcome $F(x)$ of the CDF.
   (e) Assign a distance x to the S generated values of $F(x)$ using the inverse CDF.
   (f) Save this data set containing S distances associated with each dispersed seed. This data set will be used again during the actual sequential design procedure.
   (g) Define the length ξ_{i-1} to ξ_i of each of the i intervals (bins) for the histogram. ($\xi_i - \xi_{i-1} = \frac{L}{i}$, where i is at least 20).
   (h) Create an array MaxBins of size i.
   (i) Form the histogram $k_h(x)$ (Eq. (12)) by this array MaxBins:
      i. Assign the appropriate number of seeds that fall in each element of MaxBins from the data set created in step 1e. For example, the number of seeds that will be assigned to the first

element is the total number of x values from the data set that lie between distance ξ_0 and ξ_1 (the length of the first bin).

2. Generate more (say, X) histograms $k_h(x)$ following steps 1c to 1(i)i X times.
3. For each generated histogram, calculate its moment generating function $M_h(s)$ following Eq. (13) using the X produced arrays MaxBins.
4. For each generated histogram, calculate an invasion speed c_h following Eq. (14) and save their values.
5. Calculate and average invasion speed $\overline{c_h}$ using the X invasions speeds calculated in step 4 ($\overline{c_h} = \frac{\sum_{j=1}^{X} c_{h_j}}{X}$) and save its value.

Appendix B. Procedure to obtain a sequential sampling design

1. Define the maximum number of seed traps (T) to be put in the field for the experiment.
2. Define the diameter ($\emptyset$) of the seed traps to be used.
3. Recall the bin size ($\xi_i - \xi_{i-1}$) used in Section 5.1, step 1g to create a histogram $k_h(x)$.
4. Calculate the maximum number of seed traps (MaxTraps) that would fit in the field with length L.
5. Create an array Max of length MaxTraps. This array will be used to create a histogram $\hat{k}_h(x)$.
6. Place an initial number of seed traps (t out of T) in the array Max at evenly spaced sites (i.e. at every $(\frac{\text{MaxTraps}}{t})$th element of Max starting at the first element). This will correspond to having t seed traps in the field having distances $x_1 = 0, x_2, \ldots, x_t$.
7. Assign a number of seeds to each of the t placed seed traps in Max following:
   (a) Recall the data set as saved in Section 5.1, step 1f.
   (b) Define the number of seeds that fall in each of the t placed seed traps. That is, count the number of x values from the data set that fall between $x_1 + \emptyset, x_2 + \emptyset, \ldots, x_t + \emptyset$ and save their values in the corresponding element of the array Max.
8. Make a histogram $\hat{k}_h$ in MaxBins from the limited (t) seed trap data in Max:
   (a) Define which of the elements of Max (that may or may not contain a seed trap and thus seed counts) go in which of the i bins defined in Section 5.1, step 1g. So,
      i. $e = \frac{T}{i}$ elements of Max per bin.
      ii. The sum of the first e elements go in the first element of MaxBins, the sum of the second e elements of Max go in the second element of MaxBins, and so on until the sum of the last e elements of Max, which go in the last element of MaxBins. In other words,
      iii. For $k = 1$ do $\sum_{j=1}^{e} \text{Max}_j = \text{MaxBins}_k$. For $k = 2$ to i do $\sum_{j=e+1}^{ke} \text{Max}_j = \text{MaxBins}_k$.
   (b) This results in a histogram $\hat{k}_h(x)$ (Eq. (12)) presented by MaxBins with seed counts based on the t seed traps.
9. Repeat steps 7b to 8b for all the data sets produced in Section 5.1, step 1f, which will give X histogram estimators $\hat{k}_h(x)$.

These X histograms, $\hat{k}_h(x)$, form the basis for the following procedure; the sequential sampling design. The following algorithm is a guide to selecting the optimal site in the field for the left ($T - t$) seed traps.

1. Load the first of the X histograms $\hat{k}_h(x)$ in the form of Max (in which each element contains the actual seed counts per seed trap) and load the associated data set from Section 5.1, step 1f.
2. Start with the next, $t = t + 1$, seed trap and put it at the first element of the array Max having no seed counts (open site). This will be the second element, as the first element has already been filled with seeds when the first t traps were assigned. The distance from the source in the field associated with this open site is the one next to $x_1 + \emptyset$ (the actual location in the field of the first element of the array Max). In other words, the site in the field of this new seed trap is $x_1 + \emptyset + \emptyset = x_1 + 2 \cdot \emptyset$.
3. Define the number of seeds that fall at this open site by counting the number of x values from the data set that fall between $x_1 + \emptyset$ and $x_1 + 2 \cdot \emptyset$ and save this additional value in the array Max.
4. Recalculate the histogram $\hat{k}_h(x)$, in the form of MaxBins now having one extra seed trap count, which is at the element in the array Max associated with seed trap $t = t + 1$. Use step 8(a)iii from the previous procedure to do this.

5. This will give a recalculated histogram $\hat{k}_h(x)$ (Eq. (12) in MaxBins with seed counts based on the $(t+1)$ seed traps.
6. Repeat steps 2–5 to recalculate the left $(X-1)$ histograms $\hat{k}_h(x)$. Note that $t = t+1$ in step 2 still refers to the first extra seed trap in addition to the t initially placed seed traps (just repeating the same procedure X times).
7. Calculate an average invasion speed for the X recalculated histograms based on a limited number of seed traps (i.e., $t+1$). Use Eqs. (13) and (14) for each histogram. The average invasion speed is $\hat{c}_h = \frac{\sum_{j=1}^{X} \hat{c}_{h_j}}{X}$.
8. Create a matrix SpeedSiteMSE and store the value of $\hat{c}_h$ in the first column. Store the location of the $(t+1)$th seed trap in the second column of this matrix (this is the element number in Max (same for all X arrays Max) where the $(t+1)$th seed trap was placed). The third column is reserved for the associated MSE which is about to be calculated.
9. Calculate the MSE following equation Eq. (16) using the $\hat{c}_h$ from step 7 and the $\overline{c}_h$ as calculated in Section 5.1 step 5 and save its value in the third column of SpeedSiteMSE.

In order to find the actual optimal site for this $(t+1)$th seed trap do:

1. Reposition the same seed trap to the next of the MaxTraps $-(t+1)$ left open "test" elements (in each of the X arrays Max associated with the X histograms $\hat{k}_h(x)$). By "repositioning" we mean: set the element of Max where this $(t+1)$th seed trap is placed at the moment, back to zero.
2. Start with step 3 of the previous procedure. However, note to enter the correct number of x values in each of the X Max arrays at the new position. So,
   (a) Load the appropriate data set from Section 5.1, step 1f with each of the X Max arrays.
   (b) Define between which distances (x_{begin} and x_{end}) from the source the repositioned seed trap would end up in the field.
   (c) Count the number of x values from the appropriate data set that lie between the end points of the new position x_{begin} and x_{end}.
   (d) Enter these counts in the appropriate Max array in the element associated with the new position of the seed trap.
3. Follow the previous procedure from step 4 to 9. Note that now the matrix SpeedSiteMSE does not need to be created again, just add the outcomes as a new entry.
4. Repeat steps 1–3 until every of the MaxTraps $-(t+1)$ left open sites in the field has been tested and thus associated with an average $\hat{c}_h$ value and a MSE saved in SpeedSiteMSE.
5. Finally the best site in the field for this $(t+1)$th seed trap has been located, which is the one having the smallest MSE in the matrix SpeedSiteMSE.
6. Put the seed trap which resulted in the smallest MSE (found in step 5) in Max. This means, get the seed counts associated with the location of this seed trap from the data set (Section 5.1, step 1f) and put them at the appropriate location in Max. Do this for all the X arrays Max.

Now $(t-1)$ seed traps still need to be placed at their optimal site in the field. The question is to find the optimal site for the next $t = t+1$ seed trap and repeat the procedure until $t =$ MaxTraps. In order to find the optimal site for each additional seed trap, repeat the last two procedures. Note that now $t = t+1$ in "Start with the next, $t = t+1$, seed trap …" of step 2 actually refers to adding an extra seed trap.

References

Assunção, R., Jacobi, C.M., 1996. Optimal sampling design for studies of gene flow from a point source using marker genes or marked individuals. Evolution 50, 918–923.

Bullock, J.M., Clarke, R.T., 2000. Long distance seed dispersal by wind: measuring and modelling the tail of the curve. Oecologia 124, 506–521.

Clark, J.S., Horvath, L., Lewis, M.A., 2001. On the estimation of spread rate for a biological population. Stat. Probability Lett. 51, 225–234.

Dobzhansky, T., Wright, S., 1943. Genetics of natural populations, X. Dispersion rates in *Drosophila pseudoobscura*. Genetics 28, 304–340.

Greene, D.F., Johnson, E.A., 1995. Long-distance wind dispersal of tree seeds. Can. J. Bot. 73, 1036–1045.

Greene, D.F., Johnson, E.A., 1996. Wind dispersal of seeds from a forest into a clearing. Ecology 77, 595–609.

Heesterbeek, J.A.P., 2002. A brief history of R-0 and a recipe for its calculation. Acta Biotheor. 50, 189–204.

Hemerik, L., Busstra, C., Mols, P., 2004. Predicting the temperature-dependent natural population expansion of the western corn rootworm, Diabrotica virgifera. Entomologia Experimentalis et Applicata 111, 59–69.

Jesson, L., Kelly, D., Sparrow, A., 2000. The importance of dispersal, disturbance, and competition for exotic plant invasions in Arthur's Pass National Park, New Zealand. N. Z. J. Bot. 38, 451–468.

Jongejans, E., Telenius, A., 2001. Field experiments on seed dispersal by wind in ten umbelliferous species (*Apiaceae*). Plant Ecol. 152, 67–78.

Kot, M., Lewis, M.A., van den Driessche, P., 1996. Dispersal data and the spread of invading organisms. Ecology 77, 2027–2042.

Lewis, M.A., Neubert, M.G., Caswell, H., Clark, J., Shea, K., 2005. A guide to calculating discrete-time invasion rates from data. In: Cadotte, M.W., McMahon, S.M., Fukami, T. (Eds.), Conceptual Ecology and Invasions Biology: Reciprocal Approaches to Nature.

Lewis, M.A., Schmitz, G., Kareiva, P., Trevors, J.T., 1996. Models to examine containment and spread of genetically engineered microbes. Mol. Ecol. 5, 165–175.

Malanson, G.P., 1996. Effects of dispersal and mortality on diversity in a forest stand model. Ecol. Model. 87, 103–110.

Matsinos, Y.G., Troumbis, A.Y., 2002. Modeling competition, dispersal and effects of disturbance in the dynamics of a grassland community using a cellular automaton model. Ecol. Model. 149, 71–83.

Murray, D.R., 1986. Seed Dispersal. Academic Press, Sydney, NSW.

Neubert, M.G., Kot, M., Lewis, M.A., 1995. Dispersal and pattern formation in a discrete-time predator–prey model. Theor. Popul. Biol. 48, 7–43.

Nurminiemi, M., Tufto, J., Nilsson, N.O., Rognli, O.A., 1998. Spatial models of pollen dispersal in the forage grass meadow fescue. Evol. Ecol. 12, 487–502.

Okubo, A., levin, S.A., 1989. A theoretical framework for data analysis of wind dispersal of seeds and pollen. Ecology 70, 329–338.

Paradis, E., Baillie, S.R., Sutherland, W.J., 2002. Modeling large-scale dispersal distances. Ecol. Model. 151, 279–292.

Reshetin, V.P., Regens, J.L., 2003. Simulation Modeling of Anthrax spore dispersion in a bioterrorism incident. Risk Anal. 23, 1135–1145.

Tackenberg, O., 2003. Modeling long-distance dispersal of plant diaspores by wind. Ecol. Monogr. 73, 173–189.

Weinberger, H.F., 1982. Long-time behavior of a class of biological models. SIAM J. Math. Anal. 13, 353–396.