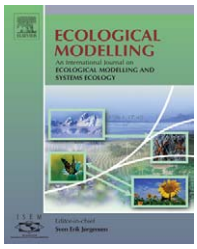
available at www.sciencedirect.com

ScienceDirect

journal homepage: www.elsevier.com/locate/ecolmodel



Managing sparse data in biological invasions: a simulation study

Brian Leung, David G. Delaney*

Department of Biology & School of Environment, McGill University, Montreal, Que. H3A 1B1, Canada

ARTICLE INFO

Article history:
Received 7 November 2005
Received in revised form
14 April 2006
Accepted 24 April 2006
Published on line 12 June 2006

Keywords:
Risk assessment
Probability
Dispersal
Survival analysis
Monte Carlo
Uncertainty

ABSTRACT

Despite recent progress, estimating species spread and the risk of different sites becoming invaded remains a major challenge in invasion biology. One of the most common problems is sparse data: we will have rarely sampled all areas for invaders and consequently do not know where all the invaded sites are. This is problematic because the entire system of sites determines invasion probability for any single site—unsampled sites can act as both sources and sinks of propagules. Thus, the difficulty is to predict invasions across the entire system with only sparse data. In this manuscript, we develop an approach to make use of partial data to forecast new invasions, and compare it with default ways of handling missing information in biological invasions. We demonstrate that it is possible to estimate spread with only a fraction of sites sampled. We find that reliability depends on the number of invaded sites sampled and that there is a tendency to underestimate the effect of propagule pressure when Allee effects are present.

1. Introduction

There is increasing concern that the extent and rate of dispersal, mediated through human activity, has increased to such a degree that global changes are occurring in which natural and human systems are unable to adapt (Lodge and Shrader-Frechette, 2003). Although most species will not establish and will not cause problems if they do (Williamson, 1996), others – termed invasive species – have the potential to cause immense environmental (e.g., loss of global biodiversity) and economic damages (Pimentel et al., 1999; Sala et al., 2000).

Due to the clear importance of invasive species, there is a need to predict where invaders are most likely to establish themselves. In order to establish, propagules need to be transported to new areas (termed propagule pressure) and they need to survive, reproduce, and form a self-sustaining population. The probability of forming a self-sustaining population is determined by population dynamical processes, such as Allee effects and stochasticity, which are important at small population sizes (i.e., the propagules reaching the new area). Unfortunately, population dynamics are typically studied long after establishment, when population sizes are large and when processes such as Allee effects are no longer important. Further, detailed information on vital rates are difficult to obtain for the early stages of invasions, given that there are not many individuals and they would be difficult to sample. Finally, given that we are typically interested in large spatial scales across many areas, estimating vital rates for the entire region of interest, based on one or two areas, requires caution.

Due to these difficulties, researchers have been developing predictive models to identify areas at risk, without explicit

* *Corresponding author*. Tel.: +1 514 398 6460; fax: +1 514 398 5069.
E-mail addresses: brian.leung2@mcgill.ca (B. Leung), david.delaney@elf.mcgill.ca (D.G. Delaney).
0304-3800/$ – see front matter © 2006 Elsevier B.V. All rights reserved.
doi:10.1016/j.ecolmodel.2006.04.015

incorporation of population dynamics (Lewis and Pacala, 2000; Drake and Bossenbroek, 2004; Leung et al., 2004; Muirhead and MacIsaac, 2005). For instance, instead of direct estimation of vital rates, researchers have related propagule pressure to successful and failed invasions (Kolar and Lodge, 2001), determined the probability of establishment for a given propagule pressure, and forecast establishment in new areas (Leung et al., 2004). Propagule pressure itself is difficult to measure directly. Thus, given that human activity is the major vector of non-indigenous species, researchers have modelled human movement patterns as a surrogate of propagule pressure, using so-called "gravity" models (Schneider et al., 1998; MacIsaac et al., 2004; Muirhead and MacIsaac, 2005; Leung et al., 2006). As the name suggests, gravity is used as an analogy for human traffic to a given area. Traffic is moderated by the mass or attractiveness of a body (e.g., a lake with fishing will be more attractive than a lake without fishing), the distance between bodies–people prefer to travel shorter distances than longer ones, and the other bodies that might "pull" traffic away (e.g., other lakes that people might visit). Gravity models have been used successfully in Geography for the past couple of decades to predict human behaviour (Thomas and Hugget, 1980). The integration of gravity models with models to estimate the probability of establishment has shown promise for well studied systems such as zebra mussels in Michigan (Leung et al., 2004).

Unfortunately, most species are not as well studied as zebra mussels, and typically, we know the invasion status only for a subsample of sites. For instance, there are over 250,000 lakes in Ontario, Canada; in a given watershed, there may be several thousand lakes. Generally, we cannot measure all sites and we therefore cannot know the entire pattern of invasion. However, we can feasibly sample several hundred sites. For instance, as part of a larger project in Canada (the Canadian Aquatic Invasive Species Network), there are plans to sample 500 lakes to determine the invasion status for *Bythotrephes longimanus*. However, it is not known whether this sampling effort will be useful, and to what extent we will be able to draw inferences regarding spread across an entire watershed or the entire province of Ontario. Intuitively, one might think that 500 data points would provide abundant statistical power. However, the problem goes beyond standard sampling theory because the entire system is dynamic and interconnected. Unsampled locations can act as both sources of propagules and alternative destinations for human vectors, modifying the propagule pressure to uninvaded sites. Put another way, it is the entire system of sites that determines the probability of invasions to any single site. Thus, techniques are needed to make predictions when only a fraction of sites have been sampled and invasion status of many sites is unknown.

In this manuscript, we extend techniques developed by Leung et al. (2004). Leung et al. (2004) estimated the probability of establishment of zebra mussels using relatively complete knowledge of invasion status. We now consider analysis given unsampled sites. The typical way to treat unsampled sites is to ignore them. This can be done either by treating sampled sites as the entire system (Schneider et al., 1998; MacIsaac et al., 2004; Muirhead and MacIsaac, 2005) or treating unknown locations as uninvaded (Anderson and Martinez-Meyer, 2004; Drake and Bossenbroek, 2004). The consequences of such simplifications remain unknown. Given the importance of forecasting invasions and the potential consequences of propagule pressure to unsampled sites, it is timely to examine the ability of these approaches to describe and forecast invasions.

We have several objectives: first, to develop a method to minimize the impact of missing data on the estimation of spread across a system; second, to examine the consequences of missing data on "default" methods and compare them to the method developed here. Default methods include treating sampled sites as the entire system, treating unsampled sites as uninvaded, and using only Monte Carlo simulations. If differences are not large, more complex, computer intensive methods may not be needed; third, to examine the factors affecting the accuracy of our approach to characterize the system (sensitivity analysis). For instance, we test whether it is the fraction or the number of sites sampled that is important; fourth, to suggest future directions to further improve our ability to forecast invasions.

2. Materials and methods

2.1. *Background model*

We used the modelling approaches developed in Leung et al. (2004) as our starting point, so that we could focus on the important extensions to the theory rather than recreating existing work. The primary extension was to minimize the impact of missing data (unsampled sites) on the estimation of spread across a system.

To begin, we first used gravity models as a mechanistic basis to generate vector movements. Gravity models are well known in Geography (Thomas and Hugget, 1980) and may be more applicable for discrete spatial units (e.g., lakes) with human vectors (e.g., boaters), compared to other methods of modeling spread such as reaction-diffusion and integro-difference models. Instead of modeling continuous landscapes and spread, gravity models simulate an interconnected matrix of sites and discrete spread and capture the following ideas: vectors are more likely to visit closer sites than ones farther away and also prefer some sites over others, due to characteristics such as lake size or fishing opportunities. The visitation rate to any single site is positively related to the vector populations around that site, but is moderated by other possible destinations that a vector might visit instead. Thus, the simulated system used in this paper mimicked systems such as inland lakes and boater vectors. Leung et al. (2004, 2006) provided detailed equations for gravity models and demonstrated the utility of this approach for estimating recreational boater movement patterns to predict biological invasions.

Second, we considered the probability of establishment of propagules that arrived to a new site. If propagules each have an independent chance of establishment, the total probability of establishment (E) is the complement of all propagules

failing to establish:

$$E(Q_{j,t}) = 1 - (1-p)^{Q_{j,t}} \quad (1)$$

where p is the probability of a single propagule establishing and Q is the number of propagules arriving at site (j) at time (t). Q is determined by the underlying gravity model. Eq. (1) is the same as the standard asymptotic curve

$$E(Q_{j,t}) = 1 - e^{-(\alpha Q_{j,t})} \quad (2)$$

where α is a shape coefficient and is equal to $-\ln(1-p)$. Additionally, processes such as the Allee effect might be present, and propagules might interact with one another such that there is disproportionately low probability of establishment at low population sizes (Dennis, 2002). This corresponds to a "lag phase" at low population sizes, and can be described through an additional shape parameter (c), using the Weibull function (Dennis, 2002)

$$E(Q_{j,t}) = 1 - e^{-(\alpha Q_{j,t})^c} \quad (3)$$

These equations were used successfully by Leung et al. (2004) to describe and predict invasions by zebra mussels. Through the use of theoretic simulations, we set parameters values for α and c to generate patterns of invasions (i.e., these were the "true" underlying parameter values). We then determined how well we could recapture these underlying values given different treatments of unsampled sites.

Third, where applicable, we used logic behind survival analysis and maximum likelihood techniques to recapture the underlying parameter values. Survival analysis allows the use of the entire data set available—the presence and absence of invaders and the timing of invasion. Maximum likelihood allows us to determine the parameter values that maximized the probability of generating the observed pattern of invaded and uninvaded lakes.

Specifically, as in Leung et al. (2004), we considered H_j, the probability given by the model of an empirical observation for site j. We assumed that there was a probability of invasion during each year dependent upon propagule pressure Q. For sites that were invaded, H_j was the joint probability of becoming invaded at time t but remaining uninvaded up until t, given the model.

$$H_j = E(Q_{j,t}) \prod_{i=1}^{t-1} [1 - E(Q_{j,i})] \quad (4)$$

where E was defined in Eq. (3). Thus, in this manuscript, we determined the probability of invasion of each site j given model values of $\hat{\alpha}$ and $\hat{c}$, which were estimates of the true parameter values used to generate the pattern of invasion. For locations that did not become invaded for the time frame (T), H_j was the joint probability of remaining uninvaded until T, given a model.

$$H_j = \prod_{i=1}^{T} [1 - E(Q_{j,i})] \quad (5)$$

The "best" values of $\hat{\alpha}$ and $\hat{c}$ were those that minimize the negative log-likelihood values (L) (i.e., the maximum likelihood values),

$$\min(L) = -\sum_{j=1}^{S} \ln(H_j) \quad (6)$$

where $\ln(H_j)$ was summed across all sites (S).

2.2. *Extensions: treatment of unknown sites and analysis*

We considered four treatments for sites with unknown invasion status.

2.2.1. *Treat sampled sites as whole system*

We treated the sampled sites as the whole system. Here, analysis was limited only to the subset of sites with known invasion status. Henceforth, this treatment will be denoted as WHOLE. This treatment has occurred repeatedly in the literature (e.g., Schneider et al., 1998; MacIsaac et al., 2004; Muirhead and MacIsaac, 2005). We applied the approach described above, estimating parameter values $\hat{\alpha}$ and $\hat{c}$ using only sites with measured known invasion status.

2.2.2. *Treat unsampled sites as uninvaded*

We treated unsampled sites as uninvaded. Henceforth, this treatment will be denoted as UNINVADED. This treatment has occurred in studies forecasting invasions over large geographical areas (Anderson and Martinez-Meyer, 2004; Drake and Bossenbroek, 2004). We applied the approach described above, estimating parameter values $\hat{\alpha}$ and $\hat{c}$ using all sites, but treating unsampled sites as uninvaded.

2.2.3. *Monte Carlo simulations*

As an intuitive approach, we simply simulated invasions from the initial invasion event to time T, for different values of $\hat{\alpha}$ and $\hat{c}$, using gravity models and Eqs. (1)–(3). This resulted in a series of invasions for each site j, which were compared to the "empirical" observations. This approach will be denoted SIM. Similar Monte Carlo approaches have appeared in the literature (e.g., Bossenbroek et al., 2001).

The simplest metric of fit was to compare the number of known invaded sites, between simulations and observed ("observed" was the invasions generated using the "true" values of α and c).

$$\min(L) = \frac{1}{G} \sum_{n=1}^{G} \text{abs} \left(\sum_{j=1}^{S} O_j - \sum_{j=1}^{S} I_j \right) \quad (7)$$

O_j was the actual invasion status of site j, and I_j was the invasion status predicted from simulations using $\hat{\alpha}$ and $\hat{c}$. O_j and I_j were equal to one if site j was invaded and zero if it was uninvaded. S was the number of sampled sites, G was the number of simulations, and L was the metric of fit. In each case, we chose the parameter values ($\hat{\alpha}$ and $\hat{c}$) that minimized the value of L, maximizing the correspondence between model predictions and observed inva-

sions. We averaged across G simulations. We also examined two other metrics—comparing the invasions status for each individual site $\left(\min(L) = (1/G)\sum_{s=1}^{G}\sum_{j=1}^{S}\text{abs}(O_j - I_j)\right)$ and comparing the year of invasion (Y_j) for each site j $\left(\min(L) = (1/G)\sum_{s=1}^{G}\sum_{j=1}^{S}\text{abs}(Y_j - Y_j)\right)$. All three metrics resulted in similar results so we only presented the first and simplest metric (Eq. (7)).

2.2.4. *Monte Carlo, maximum likelihood, survival analysis mix (MCMSAM)*

Arguably, we can improve upon the three default treatments of missing data discussed above. To do so, we integrated Monte Carlo simulations with survival analysis and maximum likelihood (MCMSAM). Monte Carlo approaches might improve upon the WHOLE and UNINVADED approaches by simulating invasions for unsampled sites rather than assuming that those sites either do not exist or are uninvaded. However, the use of Monte Carlo approaches in isolation (SIM approach) does not incorporate the available data to the maximum extent possible. The new invasions in a given time interval (t) will depend upon the invaded sites in $t-1$, which can act as sources of propagules. As such, we should use the real invasion status for all sampled sites, reserving the Monte Carlo simulations only for the unknown sites, and then determine the probability of observing the invasion pattern in the next time interval, for a given model. We modified the equations for survival analysis to incorporate Monte Carlo simulations with the entire available time series of invasions for sampled sites. We used maximum likelihood to find the best fitting parameters.

A formal recipe is as follows:

For each test parameter set ($\hat{\alpha}$ and $\hat{c}$):

(1) For time interval t, use Eqs. (1)–(3) to determine the probability of invasion ($E(Q_{j,t})$) at time t. Compare each known site that is uninvaded at time $t-1$, with observed invasion status in the current time interval ($O_{j,t}$). Thus, if a site is invaded ($O_{j,t} = 1$), we use the probability of invasion estimated by the model ($E(Q_{j,t})$); if the site is uninvaded ($O_{j,t} = 0$), we use the probability of being uninvaded ($1 - E(Q_{j,t})$). The probability (R_t) of the pattern of new invasions during time interval t given the model is the joint probability:

$$R_t = \prod_{j=1}^{N} \begin{matrix} E(Q_{j,t}) \\ 1 - E(Q_{j,t}) \end{matrix} \quad \text{if} \quad \begin{matrix} O_{j,t} = 1 \\ O_{j,t} = 0 \end{matrix} \tag{8}$$

where N is the number of sampled sites that are uninvaded at time t. We evaluate only uninvaded sites because previously invaded sites would have already been taken into account in previous time intervals. We evaluate invasion success or failure in the current time interval only, because we are interested in probabilities of invasions at time t given the system at $t-1$.

(2) Use $E(Q_{j,t})$ (Eq. (3)) and test parameters $\hat{\alpha}$ and $\hat{c}$ to simulate invasion of all sites in the system.

(3) After R_t has been calculated, force all known sites to the observed invasion status ($O_{j,t}$) at time t. For each of N uninvaded sites at time t,

$$I_{j,t} = O_{j,t} \tag{9}$$

where $I_{j,t}$ is the invasion status for sampled site j at time t used in the simulations.

(4) Iterate through steps 1–3 for all time intervals for which data exists ($t = 1$ to T).

(5) Calculate the log-likelihood value for simulation g:

$$U_g = \sum_{t=1}^{T} \ln(R_t) \tag{10}$$

In the special case of full knowledge, where invasion history is known for all sites, steps 1–5 results in identical calculations as Eqs. (4) and (5).

(6) Repeat steps 1–5, and take the average of all G simulations. Find the values of $\hat{\alpha}$ and $\hat{c}$ that minimizes the negative average log-likelihood.

$$\min(L) = -\frac{1}{G}\sum_{g=1}^{G} U_g \tag{11}$$

Below, we determined the validity of these four approaches.

2.3. *Simulations and tests*

2.3.1. *Bias and variability*

We examined the ability to recapture the underlying parameter values α and c, examining both bias and variability. We measured the deviation (δ) between the predictions based on methods described above and the real parameter values, using the formula:

$$\delta = \ln\left(\frac{\hat{p}}{p}\right) \tag{12}$$

where $\hat{p}$ was the predicted parameter value ($\hat{\alpha}$ or $\hat{c}$) estimated using any of the four treatments discussed above, and p was the true underlying parameter value (α or c, respectively). We used the log ratio so that deviations were proportional to each true parameter value and were comparable across parameter values. For instance, if $\alpha = 1$, $\hat{\alpha} = 0.5$ underestimated the true value α by a factor of two, and $\hat{\alpha} = 2$ overestimated α by a factor of two. Biologically consistent values of $\hat{\alpha}$ and $\hat{c}$ theoretically range between zero and infinity (increasing propagule pressure increases the probability of establishment or has no effect; probability of establishment ranges between zero and one).

We conducted W simulations, repeating the procedures for each approach described above. Bias (B) (i.e., the tendency to underestimate or overestimate results) was measured as the mean value of δ across all simulations W.

$$B = \frac{1}{W}\sum_{w=1}^{W} \delta_w \tag{13}$$

Variability (V) was measured using the coefficient of variation of $\hat{p}$ across all simulations W, and was a metric of the

reliability for any single run. We used the coefficient of variation to control for differences in the magnitude of $\hat{p}$.

$$V = \frac{\sigma_{\hat{p}}}{(1/W)\sum_{w=1}^{W}\hat{p}} \quad (14)$$

We compared bias and variability for each of the four treatments described above. We present values for true parameters $\alpha = 1$ and $c = 1$. We tested α and c values ranging from 0.3 to 2.3 but obtained similar findings, so do not present them (but see sensitivity analysis). We modelled a system $S = 1500$ sites with $K = 10$ vector populations over $T = 3$ time intervals, and re-simulated the system $W = 50$ times. For each of the simulations, we redistributed sites and populations randomly across the landscape using the gravity model. For each simulation, we examined seven different levels of knowledge or proportion of sites sampled ($\kappa = 1, 0.5, 0.25, 0.125, 0.062, 0.031, 0.015$). For each of W simulations of the system, and each of the seven levels of knowledge, for SIM and MCMSAM, we used $G = 10$ simulations to determine the optimizations in Eqs. (7) and (11).

2.3.2. *Forecasting*

We examined the ability to forecast new invasions for each approach described above. We also examined the ability to predict invasions in sampled sites, to allow comparison of all four methods. Predictions of invasions of unsampled sites were only possible using SIM or MCMSAM. We conducted $W = 100$ simulations. For each simulation, we simulated $S = 1500$ sites and $K = 10$ populations over $T = 3$ time intervals to fit the model parameters (see Eqs. (1)–(11)), and then generated the probability of invasion in the next time interval $t = 4$. We used $\alpha = 0.3$, $c = 1$, and $d = 2$, and $\kappa = 0.1$ (i.e., 10% of the sites had been sampled). These values were chosen such that there was a good mix of invaded and uninvaded sites, and we could best distinguish between predictive abilities. As our metric of predictiveness, we compared the actual invasions in year $T = 4$ to the probabilities predicted from each treatment, using Eq. (8). The expected probabilities for each treatment were based on the best fit $\hat{\alpha}$ or $\hat{c}$ for each approach and the known set of invaded and uninvaded sites in year $T = 3$. For SIM and MCMSAM that were based on simulations, we used the best fit $\hat{\alpha}$ or $\hat{c}$, simulated invasion histories 100 times using these values, and took the average probability of invasion to determine $E(Q_{j,t})$. We only considered sites (N) that were uninvaded at time $t - 1$, as we were only interested in forecasting ability.

2.4. *Sensitivity analysis*

We conducted a sensitivity analysis to determine which factors might affect the behaviour of our model. First, to separate out the effect of percent knowledge versus sample size, we examined different system sizes ($S = 500, 1000, 2000, 4000$) at four levels of knowledge ($\kappa = 1, 0.5, 0.25, 0.125$), using a baseline

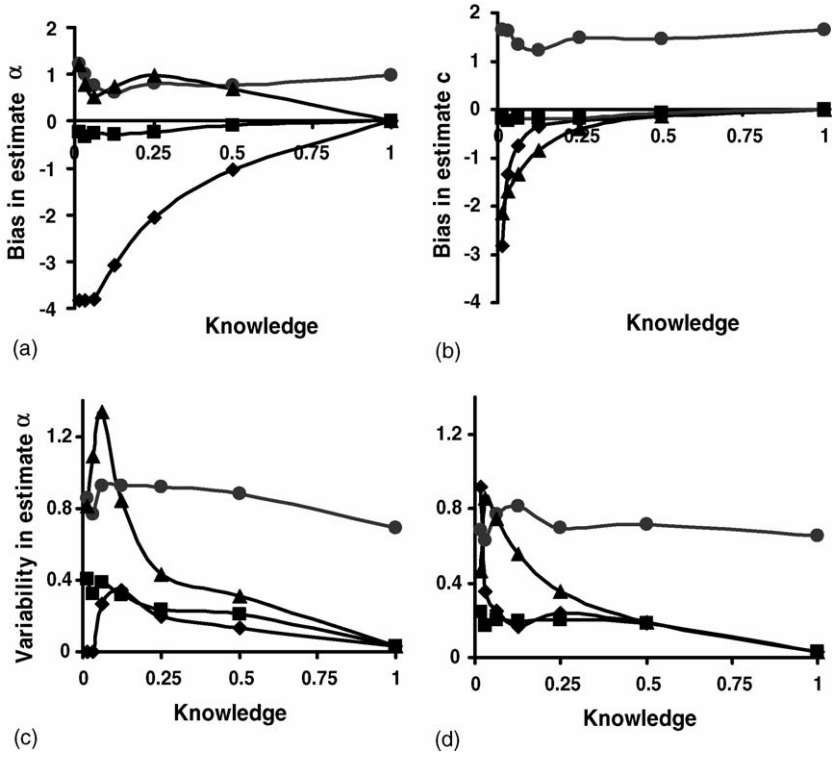


Fig. 1 – Bias and variability for approaches 1–4: WHOLE (diamond), UNINVADED (triangle), SIM (circle), MCMSAM (square). Degree of bias (Eqs. (12) and (13)) in estimation of parameter α (panel a) and c (panel b), and degree of variation (Eq. (14)) for the estimate of α (panel c) and c (panel d) are shown. Values closer to zero indicate less bias. Negative values indicate underestimation, and positive values indicate overestimation. Variation is estimated using the coefficient of variation. Smaller values indicate less variation. Approach MCMSAM generally has the lowest bias. Knowledge is the proportion of lakes sampled.

set of simulations ($\alpha = 1$, $c = 1$, $K = 10$, $G = 10$, $W = 30$). For example, we examined the question: does measurement of 25% of 1000 sites give us the same, less, or more accuracy compared to 25% of 2000 sites? Although the proportion of sites sampled is the same, the absolute sample size is twice as much with 2000 sites. We also examined the sensitivity to "true" values of α and c. We present results for α and c values, ranging between 0.3 and 2.3 (in steps of 0.5), at two levels of knowledge ($\kappa = 1$, 0.25).

3. Results

3.1. Bias, variability, and forecasting

We found that approaches WHOLE, UNINVADED, and SIM resulted in substantial bias compared to MCMSAM. SIM and WHOLE tended to overestimate parameter α, whereas UNINVADED tended to underestimate it (Fig. 1a). For comparison, using back-transformation (exp(B), see Eqs. (12) and (13)), bias for MCMSAM ranged from 0.71 to 1.01 (unity is unbiased), whereas bias for approach UNINVADED ranged from 1.01 to 3.35 (i.e., three times larger than the actual parameter value). SIM and WHOLE were even more biased in estimates of α.

Estimates of parameter c were less biased than those of α for approaches WHOLE, UNINVADED, and MCMSAM for knowledge levels down to 12.5% (Fig. 1b). Here, SIM overestimated c, whereas WHOLE, UNINVADED and MCMSAM tended to underestimate c. Again, however, MCMSAM was generally less biased than the other approaches. Using back-transformation, across the entire range examined, bias for MCMSAM ranged from 0.80 to 1.01 whereas bias for UNINVADED ranged from 0.12 to 1.01. As above, SIM and WHOLE were even more biased. Thus, MCMSAM provided relatively unbiased estimates of parameter values, whereas the other approaches did not. With full knowledge, WHOLE, UNINVADED, and MCMSAM were mathematically identical, and had little bias.

In the case where bias is large (e.g., for SIM, WHOLE, and UNINVADED), variability is less important as the approach is already flawed. Nevertheless, we presented changes in variability with knowledge for all approaches (Fig. 1c and d). With full knowledge, variability was low and increased as knowledge decreases. SIM generally had higher variability. WHOLE, UNINVADED and MCMSAM had similar variability, except at low levels of knowledge (% sites sampled). The coefficient of variation of MCMSAM was as high as 40 and 25% of the estimated value for α and c, respectively, at 1.5% knowledge.

The better fit of MCMSAM was reflected in the ability to forecast new invasions (Fig. 2). MCMSAM assigned higher probabilities of invasion to sites that actually became invaded and lower probabilities to sites which did not, compared to the other approaches, as reflected by better likelihood values. In comparison, the lowest likelihood value for MCMSAM was four to five times better than the lowest value for SIM or WHOLE and 30 times better than the lowest value for approach UNINVADED. On average, MCMSAM ($P = -65$) had likelihood values 20% better than WHOLE ($P = -80$), 50% better than SIM ($P = -99$), and two orders of magnitude better than UNINVADED ($P = -984$) (P-values closer to zero indicate better predictiveness).

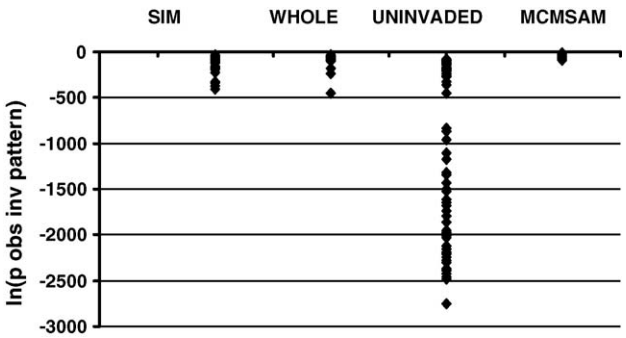


Fig. 2 – Scatter plot of ability to probabilistically forecast invasions in the future. Each point represents the joint probability (ln-transformed) of observing the pattern of invaded and uninvaded sites in the next time interval. Values closer to zero indicate a closer match to the pattern of invaded and uninvaded sites (i.e., higher probabilities assigned to sites which actually became invaded and lower probabilities assigned to sites which did not). Approach MCMSAM generally had the best forecasting ability.

Given these results, MCMSAM was clearly superior to the other simpler treatments. Comparisons across a range of other parameter values yielded similar conclusions (data not shown). Thus, the sensitivity analysis was presented only for MCMSAM.

3.2. Sensitivity analysis

We first examined the effect of system size (S) at each level of knowledge, to test the hypothesis that it is the number of sites sampled rather than the level of knowledge that affects reliability. Contrary to expectations, bias and variability increased in $\hat{\alpha}$ as system size increased (and concurrently number of sites sampled increased) (Fig. 3a and b). Nevertheless, in comparison with magnitudes of bias in the other approaches (Fig. 1), the degree of bias observed was small. For c, the degree of bias was small, and there was no obvious relation between bias or variability and system size (Fig. 3c and d). Because these results were based on simulation, and bias was an average deviation from the true parameter values, we repeated the simulations in order to see if the patterns were due to variability. We found similar relations that did not change our conclusions.

We conducted additional analyses to identify the potential mechanism for this unexpected result and found that the number of sampled invaded sites was most important (Fig. 4), rather than system size or total sample size (including both invaded and uninvaded sites). The higher bias for larger systems occurred because there were fewer invasions, due to initially more uninfested sources, a lower probability of visiting an infested source, a lower propagule supply to uninvaded sites, and a subsequent lower number of invasions (at least for the number of time intervals we examined). Thus, we next examined whether there was evidence that system size ($S = 1000$, 2000, 4000) was important after controlling for number of sampled invaded sites (Z) ($Z = 50$, 100, 200, 400, generated by beginning the analysis as soon as the appropriate number

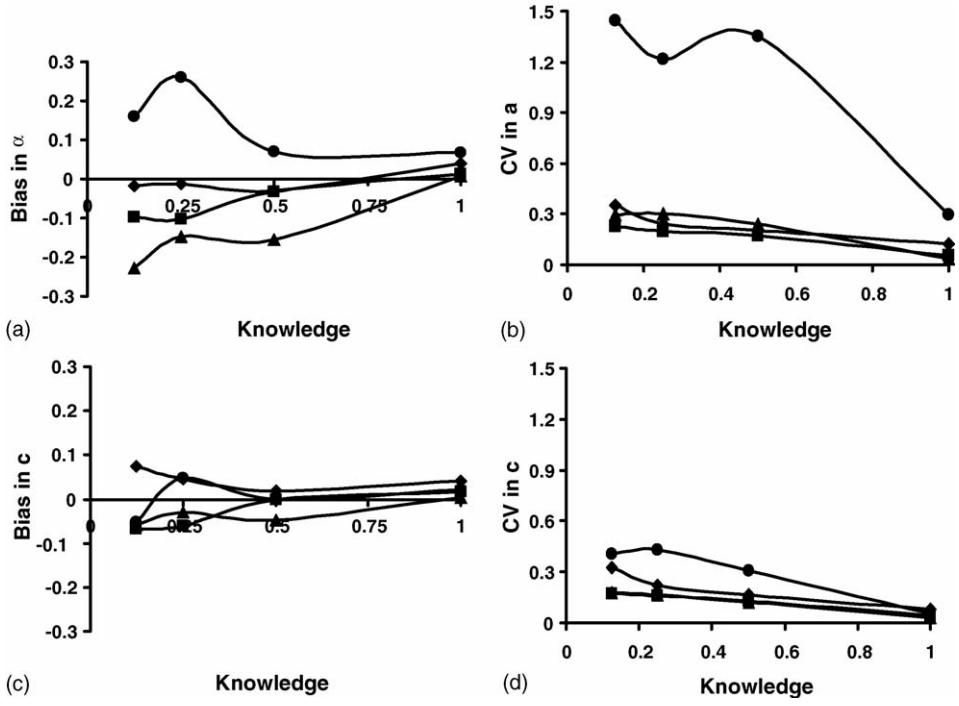


Fig. 3 – The effect of system size (S) on bias and variability across different levels of knowledge ($\kappa = 1, 0.5, 0.25, 0.125$), for α (a and b) and for c (c and d). Bias values closer to zero indicate no bias (calculated using Eqs. (12) and (13)). Variability was estimated using coefficient of variations (Eq. (14)). Values closer to zero indicate less variation. Sample sizes of 500 are denoted by diamonds, of 1000 by squares, of 2000 by triangles, of 4000 by circles. Knowledge is the proportion of lakes sampled.

of known sites became invaded). We used the same baseline set of simulations, but set knowledge levels to $\kappa = 0.5$ (Fig. 5a and b). We also examined the effect of knowledge level ($\kappa = 1$, 0.5, 0.25, 0.125) by setting system size $S = 4000$ (Fig. 5c and d). We found that reliability was not related to the system size (S) (Fig. 5a), but was related to knowledge level (κ) after controlling for number of known infested sites. Thus, we conducted the remainder of the analysis controlling for known infested sites ($Z = 250$), and determined the consequences of different α and c.

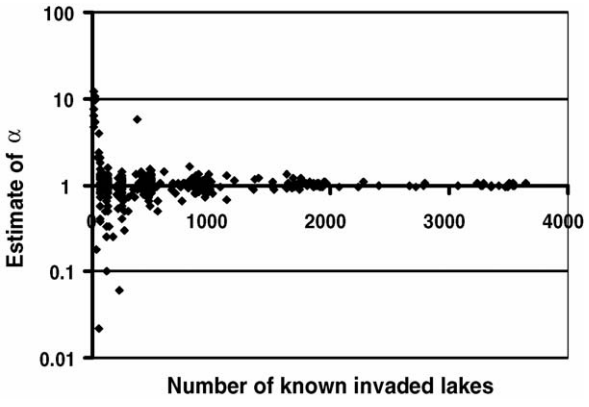


Fig. 4 – Scatter plot of estimated parameter value of $\alpha(\hat{\alpha})$ vs. number of known invaded sites, shown on a log scale. $\hat{\alpha} = 1$ indicates no bias (the true underlying parameter value was $\alpha = 1$).

The effect of real parameter values α and c on bias and variability was dependent upon knowledge level (Fig. 6). With complete knowledge, low values of c ($c = 0.3$) resulted in the highest bias and highest variability (Fig. 6a–d). $\hat{\alpha}$ was considerably more biased and variable than $\hat{c}$ at $c = 0.3$ (compare Fig. 6a and b with Fig. 6c and d). Using back-transformation, for our worst set of parameter values, $\hat{\alpha}$ underestimated α by 80%, whereas $\hat{c}$ underestimated c by only 30%. Variability of $\hat{\alpha}$ was about four times greater (max CV for $\hat{\alpha} = .57$) than $\hat{c}$ (max CV for $\hat{\alpha} = .16$), when $c = 0.3$. For other values of c, the degree of variability was similar for both $\hat{\alpha}$ and $\hat{c}$. For values $c > 0.3$ examined, the model performed very well, with full knowledge. In contrast, with more limited knowledge ($\kappa = 0.25$), biases (underestimates) could also occur at high values of c (i.e., when Allee effects were stronger, $c = 2.3$) (Fig. 6e–h). As above, $\hat{\alpha}$ had higher biases and greater variability than $\hat{c}$. As one might expect, variability was much greater than for full knowledge $\kappa = 1$.

These results were based on stochastic simulations where bias was estimated by taking an average across simulations. Thus, to examine whether the magnitude (and direction) of bias observed was actually due to the high variability, we re-ran simulations with $c = 0.3$ for full knowledge ($\kappa = 1$) and $c = 2.3$ for partial knowledge ($\kappa = 0.25$). For full knowledge, we obtained similar results. Thus, the estimated bias was not due to high variability with $\kappa = 1$, $c = 0.3$. However, for $\kappa = 0.25$, $c = 2.3$, the estimated bias was much more variable and typically larger than for other values of c. It usually underestimated the true values. Thus, we limited the interpretation of

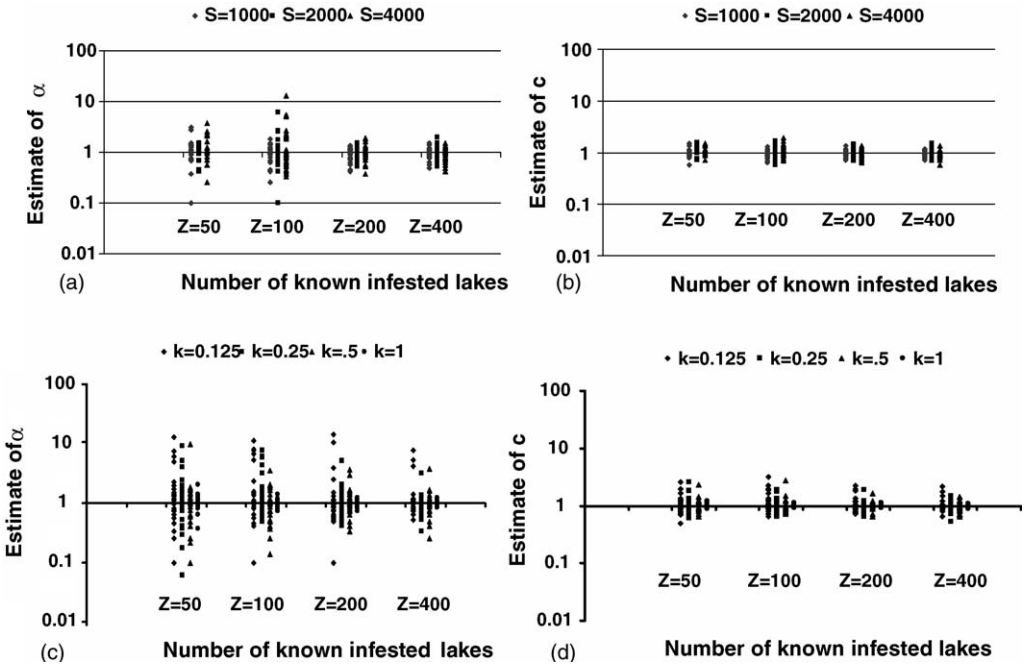


Fig. 5 – Scatter plot showing the effects of system size (*S*) and knowledge (κ) on estimated parameter value at different numbers of known invaded sites (*Z* = 1000, 2000, 4000). System size did not have a pronounced effect after controlling for the number of known invaded sites for either the estimate of α (a) or c (b). Knowledge (c and d) had a pronounced effect, with estimates closer to the true values with increasing knowledge. Knowledge is the proportion of lakes sampled. Estimates of $\alpha(\hat{\alpha})$ and $c(\hat{c})$ of 1 indicates no bias (the true underlying parameter values were $\alpha = 1$, $c = 1$).

general patterns for low κ and high c to less reliability (i.e., higher variability and/or bias).

4. Discussion

Predicting the risk of invasion is difficult because typically not all sites have been sampled. This can be especially problematic for biological invasions for the following reasons: invasions of new sites are dependent upon propagule pressure, which in turn is dependent upon the number of previously invaded sites. Unfortunately, we usually do not know all the sites that are invaded, and therefore we do not know propagule pressure, and we cannot easily calculate the probability of invasion to a new site. Yet, the importance of this lack of knowledge has not been formally examined. Before more complex approaches are used, we should make sure there are tangible benefits compared to simpler approaches. The three treatments of unknown invasion status that we chose as comparison points for our more complex approach (MCMSAM) were appropriate because they are simple and have been used commonly in the literature. Generally, these simpler treatments yielded substantially poorer results (discussed below).

Some techniques used to estimate invasion progress have often been parameterized by treating unsampled sites as uninvaded (UNINVADED) (Anderson and Martinez-Meyer, 2004; Drake and Bossenbroek, 2004). In studies estimating invasible environments, this assumption may work (Anderson et al., 2003; Anderson and Martinez-Meyer, 2004), given the absence of interactions between sites. However, for questions involving invasion progress, where unsampled sites can contribute propagules and influence new invasions, this assumption is inaccurate, and should be avoided. Our study suggests that the ability to capture underlying parameters is seriously reduced and the ability to forecast invasions is poor when equating unsampled with uninvaded sites.

The approach of treating the sampled sites as the entire system (WHOLE) (i.e., considering only sampled sites) also contained biases (Schneider et al., 1998; MacIsaac et al., 2004; Muirhead and MacIsaac, 2005). The ability to recapture underlying parameter values was compromised and forecasting new invasions in sampled sites was less accurate than MCMSAM. However, the major limitation was that extrapolation using estimated parameters to other unsampled lakes was not possible. For instance, Canadian researchers are planning a large scale sampling project for the invader, *Bythotrephes*, consisting of 500 lakes in Ontario (N Yan, personal communication, 2005). Yet, there are over 250,000 lakes in Ontario, and 2000 lakes in the 2EB watershed where *Bythotrephes* is most prevalent and where sampling will be conducted. Applying the WHOLE approach to the 500 lakes project, could give insight into new invasions within those 500 lakes, but would not permit extrapolation to the rest of the lakes, of which we are also interested. In contrast, the MCMSAM approach would provide better forecasts of invasions to those same 500 sampled lakes, as well as projections to other unsampled lakes.

The approach of simply simulating invasion progress numerous times (SIM) (Bossenbroek et al., 2001; Seymour et al., 2005) also yielded considerably poorer results than MCMSAM. Invasions are stochastic and previous invasions affect

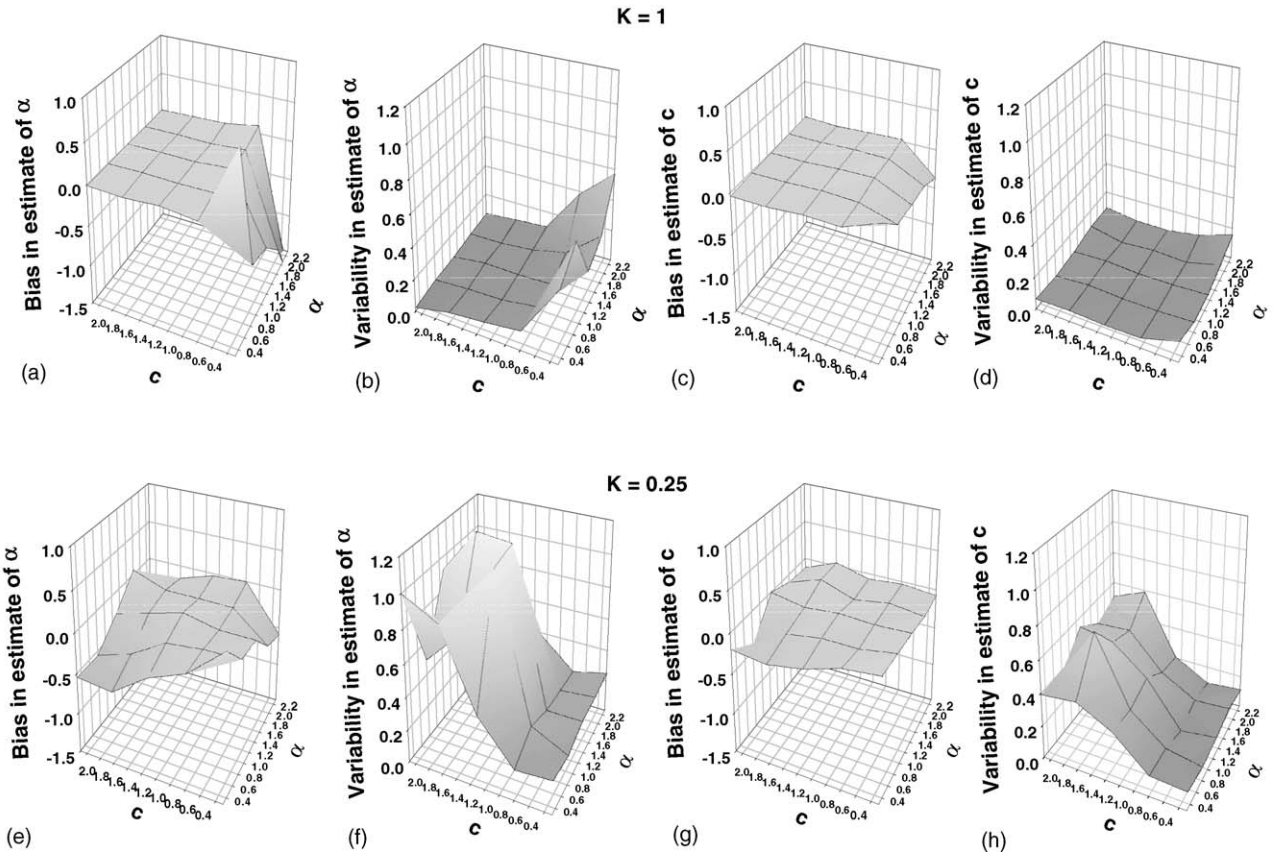


Fig. 6 – Sensitivity of bias and variability to values of estimate of α and c, shown at two levels of knowledge: $\kappa = 1$ (panels a–d) and $\kappa = 0.25$ (panels e–h). Knowledge is the proportion of lakes sampled. Bias was calculated using Eqs. (12) and (13). Variability, as estimated by the coefficient of variation, was calculated using Eq. (14).

future invasions. In contrast to SIM, MCMSAM used the actual invasion status across time for sampled lakes to estimate the probability of new invasions.

In comparison with the simpler approaches discussed above, MCMSAM yielded significant improvements. Thus, MCMSAM provided the best way to forecast invasions given limited data, even if the accuracy of estimates was not uniform across conditions. From a management point of view, this is highly relevant. For instance, the results suggest that we may be able to characterize the invasion dynamics in the 2EB watershed (2000 lakes) by sampling 500 lakes. Further, it suggests that it may be possible to forecast invasions in unsampled lakes as well as sampled ones. However, the results suggest that we cannot extrapolate to the entire lake system in Ontario (250,000 lakes), even if environmental conditions were homogeneous. This limits the geographical range in which sampling should be conducted.

The results of this study provide guidance for factors to consider. For example, the results suggested that it was the number of invaded sites sampled that was important, rather than the system size or total sample size (uninvaded and invaded) *per se*. The ramifications were: first, these results indicated that the problem of unsampled sites was not simply a sampling theory issue, given that sample size was not directly related model accuracy. Instead, other factors related to the invasion dynamics were more important to capture (i.e., it was the invasion pattern across the entire system that affected the probability of invasion to any single site). Second, some systems are very large. Nevertheless, it may be feasible to estimate spread across this entire system since system size is not itself a limiting factor. Second, due to sensitivity to the number of known infested sites, invasion progress may need to be advanced before we can reliably estimate establishment parameters. Research that increases reliability for low numbers of infested sites would be advantageous.

Parameter c, which describes the Allee effect, influenced the ability to capture the underlying parameters. Bias occurred when c was low ($c = 0.3$). However, the importance of this bias is questionable, as c values below unity do not have an obvious biological interpretation. Specifically, when $c = 1$, propagules have an independent chance of establishing a new population. When $c > 1$, the shape of the curve describes the Allee effects. Bias also occurred when Allee effects were present (high c) and knowledge was limited—there was a tendency to underestimate the effect of propagule pressure. However, the estimate of c was considerably more robust. Therefore, because we have confidence in c, and bias in α occurred only at high c, we could be reasonably confident in the model results when c is low.

4.1. *Future directions*

The approach described in this manuscript provides the next step in estimating establishment probability. Further research should be conducted in a number of areas: increasing our forecasting ability; determining conditions that may affect the reliability of forecasting approaches, in addition to those identified in this manuscript (i.e., knowledge level, number of invaded sampled sites, and Allee effects); and applying these approaches to the design of monitoring programs. This manuscript addressed the important issue of missing data. However, there are other issues that characterize real data sets that also need work, such as issues of detection (i.e., does absence of detection really indicate absence of invasion), and inconsistent sampling effort (i.e., different sites might be sampled a different number of times and in different years).

Bayesian approaches that incorporate the uncertainty distribution might improve forecasts. In some conditions, the likelihood profile may be dispersed and many parameter values may result in similar likelihoods. In such a case, maximum likelihood techniques may be more strongly affected by stochasticity, whereas Bayesian techniques would explicitly incorporate the vagueness of our information. Bayesian approaches could be implemented using Markov Chain Monte Carlo (MCMC) (Gilks et al., 1996). To apply MCMC, one would simulate the invasion process for each site across all time intervals for each test value of $\hat{\alpha}$ and $\hat{c}$. Unfortunately, given that thousands of simulations are required to estimate a single posterior distribution, and thousands of sets of simulations are required for sensitivity analysis (e.g., number of invaded lakes, true values of α and c, etc.), validating and determining the importance of Bayesian techniques would be a monumental computational task. Nevertheless, the results of this study will allow us to focus our analyses to areas where we think Bayesian techniques will provide an advantage (i.e., when maximum likelihood approaches are biased or too variable). In such cases, the added complexity of Bayesian techniques may be warranted.

In this study, we highlighted knowledge level, number of invaded sampled sites, and Allee effects as important considerations. Other factors may also influence forecasting reliability. For example, environmental heterogeneity (each site has a different α or c value) and spatial heterogeneity (the spatial distribution of sites might be clumped) likely occurs and is worth exploring. In this manuscript, we only included heterogeneity related to propagule pressure in terms of site attractiveness and distances to source populations.

Finally, this work should be integrated with empirical work to define sampling procedures. For instance, should one conduct a randomized stratified sampling protocol (Rand, 2000; Mac Nally and Horrocks, 2002)? Should one sample along the propagule pressure gradient? What are the ramifications of each approach? Likely, the invasion pattern will not be randomly distributed and different sampling strategies may provide greater power. Interactions between modellers and empiricists will keep the models relevant to real world problems, provide direction to empirical studies, and provide potential empirical tests with which to validate the expectations of the models.

Acknowledgements

This research has been supported by a grant from the NOAA Sea Grant College Program and from NSERC. We thank Catherine Pirkle and two anonymous reviewers for their help editing and improving this manuscript.

REFERENCES

Anderson, R.P., Lew, D., Peterson, A.T., 2003. Evaluating predictive models of species' distributions: criteria for selecting optimal models. Ecol. Model. 162, 211–232.

Anderson, R.P., Martinez-Meyer, E., 2004. Modeling species' geographic distributions for preliminary conservation assessments: an implementation with the spiny pocket mice (Heteromys) of Ecuador. Biol. Conserv. 116, 167–179.

Bossenbroek, J.M., Kraft, C.E., Nekola, J.C., 2001. Prediction of long-distance dispersal using gravity models: zebra mussel invasion of inland lakes. Ecol. Appl. 11, 1778–1788.

Dennis, B., 2002. Allee effects in stochastic populations. Oikos 96, 389–401.

Drake, J.M., Bossenbroek, J.M., 2004. The potential distribution of zebra mussels in the United States. Bioscience 54, 931–941.

Gilks, W.R., Richardson, S., Spiegelhalter, D.J., 1996. Markov Chain Monte Carlo in Practice. Chapman & Hall/CRC, New York.

Kolar, C.S., Lodge, D.M., 2001. Predicting invaders—response. Trends Ecol. Evol. 16, 546.

Leung, B., Bossenbroek, J.M., Lodge, D.M., 2006. Boats, pathways, and biological invasions: estimating dispersal potential with gravity models. Biol. Invasions 8, 241–254.

Leung, B., Drake, J.M., Lodge, D.M., 2004. Predicting invasions: propagule pressure and the gravity of allee effects. Ecology 85, 1651–1660.

Lewis, M.A., Pacala, S., 2000. Modeling and analysis of stochastic invasion processes. J. Math. Biol. 41, 387–429.

Lodge, D.M., Shrader-Frechette, K., 2003. Nonindigenous species: ecological explanation, environmental ethics, and public policy. Conserv. Biol. 17, 31–37.

Mac Nally, R., Horrocks, G., 2002. Proportionate spatial sampling and equal-time sampling of mobile animals: a dilemma for inferring areal dependence. Aust. Ecol. 27, 405–415.

MacIsaac, H.J., Borbely, J.V.M., Muirhead, J.R., Graniero, P.A., 2004. Backcasting and forecasting biological invasions of inland lakes. Ecol. Appl. 14, 773–783.

Muirhead, J.R., MacIsaac, H.J., 2005. Development of inland lakes as hubs in an invasion network. J. Appl. Ecol. 42, 80–90.

Pimentel, D., Lach, L., Zuniga, R., Morrison, D., 1999. Environmental and economic costs of nonindigenous species in the United States. Bioscience 50, 53–65.

Rand, T.A., 2000. Seed dispersal, habitat suitability and the distribution of halophytes across a salt marsh tidal gradient. J. Ecol. 88, 608–621.

Sala, O.E., Chapin, F.S., Armesto, J.J., Berlow, E., Bloomfield, J., Dirzo, R., Huber-Sanwald, E., Huenneke, L.F., Jackson, R.B., Kinzig, A., Leemans, R., Lodge, D.M., Mooney, H.A., Oesterheld, M., Poff, N.L., Sykes, M.T., Walker, B.H., Walker, M., Wall, D.H., 2000. Biodiversity—global biodiversity scenarios for the year 2100. Science 287, 1770–1774.

Schneider, D.W., Ellis, C.D., Cummings, K.S., 1998. A transportation model assessment of the risk to native mussel communities from zebra mussel spread. Conserv. Biol. 12, 788–800.

Seymour, A., Varnham, K., Roy, S., Harris, S., Bhageerutty, L., Church, S., Harris, A., Jennings, N.V., Jones, C., Khadun, A., 2005. Mechanisms underlying the failure of an attempt to eradicate the invasive Asian musk shrew Suncus murinus from an island nature reserve. Biol. Conserv. 125, 23–35.

Thomas, R.W., Hugget, R.J., 1980. Modeling in Geography. Barnes and Noble books, Totowa, NJ.

Williamson, M., 1996. Biological Invasions. Chapman & Hall/CRC, London.