Ecology, 97(1), 2016, pp. 194–204

Dynamic occupancy models for explicit colonization processes

Kristin M. Broms,[1,8] Mevin B. Hooten,[1,2,3] Devin S. Johnson,[4] Res Altwegg,[5,6] and Loveday L. Conquest[7]

[1]*Department of Fish, Wildlife, and Conservation Biology, Colorado State University, Fort Collins, Colorado 80523, USA*
[2]*U.S. Geological Survey, Colorado Cooperative Fish and Wildlife Unit, Fort Collins, Colorado 80523, USA*
[3]*Department of Statistics, Colorado State University, Fort Collins, Colorado 80523, USA*
[4]*National Marine Mammal Laboratory, Alaska Fisheries Science Center, NOAA, 7600 Sand Point Way NE, Seattle, Washington 98115-6349, USA*
[5]*Statistics in Ecology, Environment and Conservation, Department of Statistical Sciences, University of Cape Town, Rondebosch 7701, Cape Town, South Africa*
[6]*African Climate and Development Initiative, University of Cape Town, Rondebosch 7701, South Africa*
[7]*School of Aquatic and Fishery Sciences, University of Washington, Box 355020, Seattle, Washington 98161-2182, USA*

Abstract. The dynamic, multi-season occupancy model framework has become a popular tool for modeling open populations with occupancies that change over time through local colonizations and extinctions. However, few versions of the model relate these probabilities to the occupancies of neighboring sites or patches. We present a modeling framework that incorporates this information and is capable of describing a wide variety of spatiotemporal colonization and extinction processes. A key feature of the model is that it is based on a simple set of small-scale rules describing how the process evolves. The result is a dynamic process that can account for complicated large-scale features. In our model, a site is more likely to be colonized if more of its neighbors were previously occupied and if it provides more appealing environmental characteristics than its neighboring sites. Additionally, a site without occupied neighbors may also become colonized through the inclusion of a long-distance dispersal process. Although similar model specifications have been developed for epidemiological applications, ours formally accounts for detectability using the well-known occupancy modeling framework. After demonstrating the viability and potential of this new form of dynamic occupancy model in a simulation study, we use it to obtain inference for the ongoing Common Myna (*Acridotheres tristis*) invasion in South Africa. Our results suggest that the Common Myna continues to enlarge its distribution and its spread via short distance movement, rather than long-distance dispersal. Overall, this new modeling framework provides a powerful tool for managers examining the drivers of colonization including short- vs. long-distance dispersal, habitat quality, and distance from source populations.

Key words: Acridotheres tristis*; citizen science; colonization; Common Myna; dynamic occupancy model; extinction; invasive species; multi-season model; Southern African Bird Atlas Project; spatiotemporal processes; species distribution maps.*

Introduction

Invasive species are a problem worldwide: damaging crops, contributing to the loss of biodiversity, and causing disturbances. They are generally seen as the second biggest threat to biodiversity, after habitat destruction (Wilcove et al. 1998, Pejchar and Mooney 2009), and the economic costs to control them are great. For example, the Working for Water program in South Africa was recently given a 3-yr budget of R7.8 billion (~USD660 million) to control invasive plants near Cape Town (van Wilgen et al. 2012). Increased knowledge about the causes of invasive species' spread could reduce the damage incurred by giving managers an understanding of what is driving their expansion.

More broadly, ecologists have sought to understand colonization and extinction patterns for decades. The dynamic occupancy model (MacKenzie et al. 2003), alternatively called the multi-season occupancy model, has become a widely used model to learn about the colonization and extinction processes. Occupancy models rely on a hierarchical framework (either implicit or explicit) to account for species that may be present at a site but go undetected. The multi-season version of the model specifies occupancy probabilities as functions of colonization and extinction probabilities and the occupancy status of a site from the previous time step. However, the original version of the multi-season model does not contain an explicit spatial component nor a spatiotemporal interaction. Depending on the species and the dynamic process, it may be more

Manuscript received 10 March 2015; revised 15 June 2015; accepted 10 July 2015. Corresponding Editor: E. G. Cooch.
[8]E-mail: kristin.broms@rams.colostate.edu

appropriate to acknowledge within the model that the density of occupied sites and the distance between occupied sites will play a role in the colonization of unoccupied sites and the persistence of those already colonized.

Explicit spatiotemporal relationships have been recognized and incorporated in several recent variations of the multi-season occupancy model (Bled et al. 2011, 2013, Yackulic et al. 2012, Eaton et al. 2014, Sutherland et al. 2014). The Yackulic et al. (2012) and Eaton et al. (2014) models are most similar to the multi-season model first introduced by MacKenzie et al. (2003), but they add an autocovariate term to the colonization and extinction probability functions. The models incorporate and estimate coefficients for the autocovariates as if they were fixed effects, and are fit using PRESENCE (MacKenzie et al. 2003). The autocovariate is a weighted average of the occupancy probabilities of a site's neighbors from the previous time step. Eaton et al. (2014) expand on the work of Yackulic et al. (2012) by including a provision that the autocovariate is additionally weighted by the proportion of habitat available. For their study, they modeled the occurrence of an endangered marsh rabbit in the Lower Keys, Florida, USA, where the available habitat is limited and highly fragmented due to development and bodies of water. Therefore, habitat weighting was necessary for their species. Bled et al. (2011, 2013) incorporated a similar autocovariate, but used a Bayesian framework. Bled et al. (2011) weighted the neighbor occupancy status based on sines and cosines to account for directional spread, and also extended the MacKenzie et al. (2003) dynamic occupancy model to include separate colonization and recolonization parameters. Bled et al. (2013) extended the model by having two nested time frames during which there were separate colonization and extinction processes.

Sutherland et al. (2014) took a different approach to incorporating spatial information into the dynamic process by relating the colonization and extinction functions to age class abundance data and metapopulation theory. This approach leads to their colonization function having a different form than the MacKenzie et al. (2003) multi-season model. The Sutherland et al. (2014) model assumes that the species of interest has a fragmented metapopulation structure and therefore it relies on the availability of abundance data and variation in the distance between patches.

We present a model similar to those just described in that colonization probabilities are a function of the number of neighboring sites that are currently occupied, but we develop the colonization probabilities from diffusion and advection processes. This induces an explicitly mechanistic colonization process that is similar to that of Sutherland et al. (2014), but stems from different theory and utilizes different types of data. We adapt a continuous diffusion model to detection/non-detection data collected on discrete spatial and temporal units as in Hooten and Wikle (2010), but additionally account for imperfect detection. We explicitly account for two types of colonization: neighborhood colonization and long-distance dispersal.

We introduce the model, validate it through a simulation study, and then apply it to study the Common Myna (*Acridotheres tristis*, hereafter "myna") invasion using data from the second Southern African Bird Atlas Project (SABAP 2), which is a large database of bird detections/non-detections in southern Africa from 2007 to the present. The myna, a starling native to Asia, is one of the world's worst invasive species (Lowe et al. 2000). It was introduced to Durban, a city in the Southeast corner of South Africa, in 1902 (Peacock et al. 2007), stabilized in that region, and then underwent periods of rapid expansion. The myna is now widespread in the eastern half of South Africa. The myna's distribution has been noted anecdotally, but the drivers of its expansion have not been studied empirically or statistically. Our model is the first statistical model to determine what may be driving the myna's expansion and what its rates of colonization are. The myna may compete with native species; thus our inference has important conservation implications.

Methods

Models

We borrow the notation of other Bayesian occupancy models (e.g., Royle and Kéry 2007, Royle and Dorazio 2008), and let $y_{i,j,t}$ represent the detection of the species of interest on survey $j \in \{1, \ldots, \mathrm{J}_{i,t}\}$ of site $i \in \{1, \ldots, m\}$ during time period $t \in \{1, \ldots, T\}$, and $z_{i,t}$ be the true occurrence of the species at site i during time period t. $J_{i,t}$ is the number of surveys of site i during time period t; this number may vary among sites and time periods. M is the number of sites for which we will draw inference; only a subsample m of them need to be surveyed to gain inference for the entire region. We use T to denote the number of time periods. Each time period is assumed to be a closed season during which occupancies do not change. The probability of the species occurring at site i during time t is $P(z_{i,t} = 1) = \psi_{i,t}$, which may be a function of site-level covariates such as elevation. If the species does occur at site i, then $z_{i,t} = 1$ and the detection probability is $P(y_{i,j,t} = 1 | z_{i,t} = 1) = p_{i,j,t}$. The detection probability may be a function of site-level covariates and survey-level covariates, such as time of day of the survey. Assuming a logit link relationship between the detection probabilities and the covariates, the occupancy model is

$$y_{i,j,t} \sim \text{Bernoulli}(z_{i,t} p_{i,j,t}) \tag{1}$$

$$\text{logit}(p_{i,j,t}) = \mathbf{x}'_{i,j,t} \boldsymbol{\beta}_p \tag{2}$$

$$z_{i,t} \sim \text{Bernoulli}(\psi_{i,t}) \quad (3)$$

where $\mathbf{x}_{i,j,t}$ is the set of covariates that affect detection for survey j of site i at time t.

The function associated with the occupancy probabilities, $\psi_{i,t}$, varies depending on whether $t = 1$ (an initial condition) or if it is a subsequent time period. For the first time period, $t = 1$, the occupancy probabilities are modeled as in a spatially explicit single-season occupancy model such that

$$\text{logit}(\psi_{i,1}) = \mathbf{x}'_{\psi,i,1}\boldsymbol{\beta}_{\psi} + \eta_i. \quad (4)$$

In this case, $\mathbf{x}_{\psi,i,1}$ is a separate set of covariates from Eq. 2, although a site-specific covariate may be a member of both sets. Eq. 4 includes a spatial random effect (η_i) to account for residual spatial patterns that are not captured by the covariates. In analyzing the myna data, we included a restricted spatial regression (RSR) random effect (Hughes and Haran 2013, Johnson et al. 2013), which is similar to an intrinsic conditional autoregressive (ICAR) variable and is discussed further in the *Data* section.

We let occupancy probabilities in subsequent time periods depend on the occurrence patterns from the previous time period (Hooten and Wikle 2010). If site i was previously occupied, then the probability of it remaining occupied at time t is $\varphi_{i,t}$, the persistence probability. Often occupancy dynamics are written in terms of local extinctions; the persistence probability, alternatively called the survival probability, is the complement of the local extinction probability and is more commonly used in the Bayesian literature (MacKenzie et al. 2003, Kéry et al. 2013). If site i was not occupied in the previous time period and neither were any of its neighbors, then the probability of it becoming occupied at time t is $\gamma_{i,t}$, the long-distance dispersal probability. If site i was not occupied in the previous time period but at least one of its neighbors was, then the probability of it becoming occupied is $\bar{d}_{i,t}$, the neighborhood colonization probability. Thus, the occupancy probabilities are formulated as a mixture:

$$\psi_{i,t+1} = z_{i,t}\phi_{i,t} + \left(1 - z_{i,t}\right) I_{\mathcal{N}_{i,t}} \bar{d}_{i,t} + \left(1 - z_{i,t}\right)\left(1 - I_{\mathcal{N}_{i,t}}\right)\gamma_{i,t} \quad (5)$$

where $I_{\mathcal{N}_{i,t}}$ is an indicator variable that equals 1 if site i has at least one neighbor that was occupied in year t, and equals 0 otherwise (Hooten and Wikle 2010).

The persistence probability, $\varphi_{i,t}$, and the long-distance dispersal, $\gamma_{i,t}$, may be modeled as functions of a time- or space-varying covariate, for example:

$$\text{logit}(\gamma_{i,t}) = \beta_{\gamma,0} + \beta_{\gamma,1} x_{\gamma,i,t} \quad (6)$$

$$\text{logit}(\phi_{i,t}) = \beta_{\phi,0} + \beta_{\phi,1} x_{\phi,i,t}. \quad (7)$$

Additionally, $\varphi_{i,t}$ can be a function of the density of nearby occupied sites at the previous time step, similar to the neighborhood colonizations that we will describe. In studying the myna invasion, we assume that $\varphi_{i,t}$ and $\gamma_{i,t}$ are constant across sites and time periods (i.e., $\varphi_{i,t} = \varphi$ and $\gamma_{i,t} = \gamma$), but we allow the neighborhood colonization probability, $\bar{d}_{i,t}$, to vary among sites and time periods as a function of the number of neighbors that were occupied in the previous time step. The choice of neighborhood structure is project specific. For example, it may include all sites within a specified distance, or it may be all sites that share a border. In our application, the sites have a gridded design and we use the queen's definition of neighborhood (Fig. 1). In this design, most sites have eight neighbors. If a site exists on the edge of the area of inference, it will have fewer neighbors; although the model could be expanded to allow for neighborhood colonization from outside the study area. The probability of site i being colonized by its neighbors is a function of the neighbors previously occupied, such that

$$\bar{d}_{i,t} = 1 - \exp\left(\mathbf{z}'_{\mathcal{N}_{i,t}} \log\left(\mathbf{1} - \mathbf{d}_i\right)\right). \quad (8)$$

Neighborhood colonization (Eq. 8) is a function of the vectors $\mathbf{z}_{\mathcal{N}_{i,t}}$ and $\mathbf{d}_i$, both with length equal to the number of neighbors of site i (derived in Appendix S1). An element k of $\mathbf{z}_{\mathcal{N}_{i,t}}$ equals 0 if neighbor k was not occupied and equals 1 if the neighbor was occupied at time t. Each element k of the colonization vector, $\mathbf{d}_i$, represents the probability of site i being colonized by neighbor k. These probabilities are constant across time but may or may not be constant across space. They may be modeled in one of two ways: homogeneous or gradient-based colonization.

In the homogeneous neighborhood colonization model, we assume that local colonization patterns do not vary across the landscape. This dispersal pattern might indicate that the invasion generally happens along a latitudinal

	1	2	3	
	4	Site i	5	
	6	7	8	

FIG. 1. The $k = 8$ neighbors of a site for data with a gridded design, assuming a first-order, queen's definition of neighborhood, in which most sites have eight neighbors.

or longitudinal trend. The $\mathbf{d}_i$ vector is the same for each site i, so $\mathbf{d}_i \equiv \mathbf{d}$ for all i, but each element k of the vector may be a different probability. For example (see Fig. 1), if $\mathbf{d} = (0, 0, 0, 0.05, 0.05, 0.2, 0.2, 0.2)'$, it would indicate that the species is colonizing its northern neighbors with greater probability.

In the gradient-based neighborhood colonization model, the colonizations are functions of a covariate and whether site i has better or worse habitat than its neighbors:

$$\text{logit}(\mathbf{d_i}) = \beta_{d,0} + \beta_{d,1}\mathbf{x}_{\mathbf{d},i}. \quad (9)$$

The $\mathbf{d}_i$ vectors are different for each site because site i will have different habitat in relationship to its neighbors. The $\mathbf{x}_{\mathbf{d},i}$ contain the gradient of the habitat variable between site i and its k neighbors; each element, $x_{\mathbf{d},i,k}$, of the vector is equal to site i's kth neighboring covariate value minus site i's covariate value, divided by the distance between sites to account for the possibility of varying distances between neighbors:

$$x_{\mathbf{d},i,k} = \frac{x_k - x_i}{\text{dist}(i,k)}. \quad (10)$$

We use the differences in habitat as opposed to the habitat values themselves because the differences explicitly account for diffusive flow; it is a discrete approximation of the derivative of the potential surface from which the model was derived. Such models have also been used to study the spread of diseases (Hooten et al. 2010*a*) and animal movement (Hooten et al. 2010*b*, Hanks et al. 2011), and are fundamentally linked to the movement of individuals and populations (Hooten et al. 2013). Extensions to the gradient-based function (Eq. 9) could include the local habitat variables themselves in addition to, or instead of, the differences. In the simpler case, Eq. 10 could be replaced with $x_{\mathbf{d},i,k} = x_k$ for each element k in $\mathbf{x}_{\mathbf{d},i}$.

The full models for both the homogeneous and gradient-based colonization, as applied to the myna, are presented in Appendix S2.1. Appendix S2.2 can be used as a glossary for the model symbols, and Appendix S2.3 provides the associated JAGS (Plummer 2003) code to fit the models.

As with most occupancy models, we assume that the detection probabilities are conditionally independent of each other and that there are no false positives (i.e., the possibility that a myna was reported as detected but, in truth, did not occupy the site of interest). Although these assumptions may not be valid in all situations, the framework we present can be generalized to accommodate these measurement discrepancies.

Data

The Southern African Bird Atlas Project, SABAP 2, is a large citizen science database of bird lists collected by volunteer bird-watchers from July 2007 to the present (*available online*).[9] Each bird list represents one survey of one site; non-detections are deduced by a species' absence from the list. For each survey, a volunteer spends a minimum time period of 2 h of intensive birding up to a maximum time period of 5 days conducting each survey and, in that time, all habitat types in the grid cell were expected to be visited.

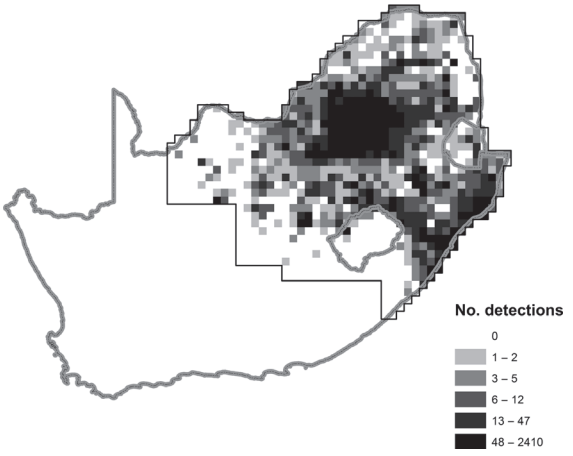

FIG. 2. The black outline in the map shows the sites included in our analyses. The gray boundaries delineate the separate countries of Lesotho and Swaziland. Each square is one quarter degree grid cell (QDGC) and is approximately 25 km × 25 km. The map shows the total number of detections of the Common Myna (*Acridotheres tristis*) at each site, for all years combined.

The sites of SABAP 2 are 5-min latitude by 5-min longitude grid cells, approximately 8 × 7.6 km each (Harebottle et al. 2007). South Africa is covered by 17444 of these sites. We aggregated the data into quarter degree grid cells (QDGC) to compare our model results against an earlier version of the bird atlas project, SABAP 1, which occurred mainly from 1987 to 1991. Each QDGC is 15-min latitude by 15-min longitude and is equal to nine of the smaller grid cells. A total of 1946 QDGC cover South Africa. We limited our analyses to the eastern half of South Africa plus Lesotho and Swaziland because the myna exists primarily in that study area (Fig. 2). Therefore, our analysis included 1068 of the QDGC sites. Each year, between 613 and 862 of these sites were surveyed at least once. Maps of the numbers of surveys, myna detections, and myna reporting rates by year are provided in Appendix S3.

The myna data are indexed by six time periods, one for each year of data starting in January 2008, and ending in December 2013. We fit the diffusion occupancy models to the data from January 2008–December 2012, with each year of data representing one time period. We held out the 2013 data to

[9] http://sabap2.adu.org.za/

compare the diffusion occupancy models' predictive performance against the estimated occupancies from a single-season occupancy model. The number of times that sites were surveyed was variable but the median number of surveys per site was five. To prevent detection probability coefficients from being dominated by a few well sampled sites, the number of surveys per site per year was limited to 15. Preliminary analyses showed that inference was not sensitive to this censoring.

In an exploratory analysis, we identified an initial set of covariates affecting myna occurrence by fitting single-season, nonspatial occupancy models for the 2008 data using the *unmarked* package in R (Fiske and Chandler 2011) and used AIC to select the best predictive model. The following site-specific covariates were considered: the logarithm of the human population density (LOG HUMAN; Balk et al. 2011); the proportion of the site that was pastureland (PASTURE; Ramankutty et al. 2010*b*); the proportion of the site that was cropland (CROP; Ramankutty et al. 2010*a*); the proportion of the site that was in a protected area such as a national park or game reserve (PA; Rouget et al. 2004); and the distance from the site's centroid to the center of the closer of Johannesburg or Durban (DIST). We included these variables because it has been proposed that the myna's distribution is most associated with human population density and habitat transformations and because its two source populations have historically been Johannesburg and Durban (Peacock et al. 2007). In addition to the site-specific covariates, the following survey-specific covariates were considered when building the detection probability function: the log of the number of hours spent intensively birding for the checklist (INTENSIVE); the log of the total numbers of hours spent on the checklist (TOTAL), which will include the intensive hours birding plus time spent passively birding; and the number of species found on the bird list (NSPP), scaled by its mean and standard deviation.

After the best nonspatial model for 2008 was selected, a spatial random effect was added to the occupancy component of the model. The spatial occupancy models were fit using the *stocc* package in R (Johnson 2013). We added an RSR (restricted spatial regression) random effect to account for residual spatial patterns while minimizing the confounding with the fixed effects of interest (Broms et al. 2014). RSR is a dimension-reduced version of an *intrinsic conditional autoregressive* (ICAR) model. Thus, our model for the initial time period was

$$\begin{aligned} \text{logit}(\boldsymbol{\psi}_1) &= \mathbf{X}_{\psi}\boldsymbol{\beta}_{\psi} + \mathbf{K}\boldsymbol{\alpha} \\ \boldsymbol{\alpha} &\sim \text{Normal}\left(\mathbf{0}, \sigma^2\left(\mathbf{K}'\mathbf{Q}\mathbf{K}\right)^{-1}\right) \end{aligned} \tag{11}$$

where $\mathbf{K}$ are eigenvectors associated with the Moran operator matrix (Hughes and Haran 2013) and $\mathbf{Q}$ is the ICAR precision matrix whose elements are −1 if sites are neighbors, 0 if sites are not neighbors, and is equal to the number of neighbors of each site along its diagonal. We restricted the random effect to include 250 eigenvectors following the recommendations of Broms (2013). Further details of the spatial component of this model may be found in Johnson et al. (2013) and Hanks et al. (2015). Heuristically, our $\mathbf{K}\alpha$ serves as an autocovariate relating a site to the occupancy status of its neighbors. If parameters became nonsignificant with the addition of the spatial covariate, they were dropped from the model. We then fit three different diffusion occupancy models containing neighborhood colonization probabilities that were specified as follows. Two models used the homogeneous neighborhood colonization, one allowed the neighbor colonization probabilities to vary among neighbors, and one assumed a constant neighborhood colonization in each direction. The other model used the gradient-based neighborhood colonization, as in Eq. 9, with human population density as the environmental covariate. For simplicity, we assumed constant persistence and long-distance dispersal probabilities across time and space. The covariates that influenced detection were selected with the 2008 single season model and then the same covariates were used for the observation process for subsequent years.

Relatively vague priors were specified for all parameters, as described in Appendix S2.1. We obtained three MCMC chains for 160 000 iterations with a burn-in of 10 000 iterations and a thinning rate of 20, resulting in a total of 22 500 samples for each model. The model fits required about 31 h.

Model selection

We compared models using out-of-sample validation with a logarithmic scoring rule to assess predictive performance (Gneiting and Raftery 2007, Hooten and Hobbs 2015). We predicted 2013 occupancies and detections using the posterior predictive distributions and compared these 2013 predictions against the true detections from 2013. In our calculations, we only compared sites that had at least one survey conducted in 2013. For each iteration s of the MCMC, we calculated the log-score as negative the logarithm of the integrated likelihood:

$$L_{\text{log}}^{(s)} = -\sum_{i=1}^{M}\sum_{j=1}^{J_i} \log[y_{ij}|\psi_i^{(s)}, p_{ij}^{(s)}] \tag{12}$$

$$= -\sum_{i=1}^{M}\sum_{j=1}^{J_i} y_{ij}\log(\psi_i^{(s)}p_{ij}^{(s)}) + (1-y_{ij})\log(1-\psi_i^{(s)}p_{ij}^{(s)}) \tag{13}$$

where M is the number of sites with surveys conducted in 2013 and J_i is the number of surveys associated with that site. The log score was calculated as the posterior

mean of $L_{\log}$. The lowest logarithmic score indicated the best predictive model.

Simulation study

We conducted a simulation study to investigate the convergence and inference characteristics for the dynamic components of our model. Four scenarios were tested: long-distance dispersal was either constant or a function of a covariate, and the neighborhood dispersal was either homogeneous or gradient-based.

For all scenarios, we assumed a grid of $30 \times 30 = 900$ sites, of which a random subset of 75% of the sites were surveyed. Following the approach of Yackulic et al. (2012), we simulated data using a constant detection probability of 0.5 and assumed four surveys per site. Occupancy probabilities for the first year were a function of the scaled x-coordinate of the data:

$$\text{logit}(\boldsymbol{\psi}_1) = -1.5 + 1.5\mathbf{x}. \quad (14)$$

These parameters led to occupancy probabilities ranging from 0.02 to 0.73, with a median probability of 0.18. The persistence probability, φ, was set at 0.90.

For half of the simulations, long-distance dispersal, γ, was set at 0.05; for the other half of the simulations, it was a function of the scaled x-coordinate:

$$\text{logit}(\boldsymbol{\gamma}) = -3 + 1\mathbf{x} \quad (15)$$

leading to a range of long-distance dispersal probabilities from 0.01 to 0.21, with a median of 0.05. For half of the simulations, neighborhood dispersal was homogeneous, with neighborhood colonization probabilities of (0.20, 0.20, 0.02, 0.20, 0.02, 0.02, 0.02, 0.02) for $k = 1, \ldots, 8$, respectively. These probabilities imply that a site was most likely colonized from the northwest direction, and its range is therefore expanding in the southeast direction. The other set of simulations had gradient-based neighborhood colonizations with

$$\text{logit}(\mathbf{d_i}) = -2 + 2\mathbf{x}_{\mathbf{d},i}. \quad (16)$$

The $\mathbf{x}_{d,\,i}$ were based on the scaled x-coordinate, and were calculated using Eq. 10. The gradient-based model had neighborhood colonization probabilities range from 0.005 to 0.79 with a median of 0.12.

For all scenarios, occupancy probabilities for the first year were fit using:

$$\text{logit}(\boldsymbol{\psi}_1) = \beta_0 + \boldsymbol{\eta} \quad (17)$$

where η is an RSR spatial random effect. This model specification was different from the data generation to mimic reality in that all environmental variables affecting occupancy may not be known or measurable. The code to generate the data and fit the models may be found in the Supplement.

To investigate the sensitivity under differing data scenarios, 10 simulations were performed for each scenario. Each model fit included three chains with 5000 iterations each, all thinned by five and with a burn-in of 500 samples, leaving a total of 2700 samples for approximating posterior quantities. The gradient-based model simulations required 27–30 min and the model with homogeneous neighborhood colonizations required 42–45 min on a 3.5 GHz Intel Core i7 desktop computer. Parameter estimates were obtained as the medians from the marginal posterior distributions. To determine model performance, relative biases were then calculated as

$$\text{Bias} = \frac{1}{S}\sum_{s=1}^{S}\frac{\hat{\theta}_s - \theta}{\theta} \quad (18)$$

where the averages are taken over the S simulations.

Results

Simulation study

The simulation study demonstrated that the model performed well for a variety of data sets. For all scenarios, the number of sites occupied each year was estimated accurately (Appendix S4). The persistence probability and detection probability estimates were also unbiased. The models were able to recover the directionality of the neighborhood colonizations, but the estimates of the long-distance dispersal and neighborhood colonization probabilities were variable with no distinct patterns in the biases, suggesting less precision in their estimates.

Myna results

The detection probabilities were positively correlated with the number of species reported, human population density, and the proportion of the site that was cropland. The positive correlation with number of species probably is related to observer skill level. Detection was negatively correlated with the proportion of the site that was part of a protected area and the distance from Johannesburg or Durban (Table 1). The myna occupancy probabilities of year 2008 were originally correlated with the distance from Johannesburg or Durban and the proportion of the site that was pasture. Once the RSR random effect was included, only the distance covariate affected occupancies. The probability of occupancy increased with proximity to the city centers (Table 1).

The dynamic models produced very similar estimates for the parameters that overlapped among them. The long-distance dispersal probability was estimated at 0.02, with a 95% credible interval of 0.002–0.06 for the homogeneous models and a 95%

Table 1. Parameter estimates from the best-fitting model for the Common Myna (*Acridotheres tristis*) data from Africa, which used gradient-based neighborhood colonizations as a function of human density.

Parameter	Median	SE	95% CI	
			Lower	Upper
Detection coefficients				
Intercept	−1.81	0.157	−2.14	−1.51
NSPP	0.45	0.017	0.42	0.49
PA	−0.60	0.020	−0.63	−0.56
DIST	0.26	0.013	0.23	0.28
CROP	0.50	0.116	0.25	0.72
LOG_HUMAN	−1.35	0.138	−1.63	−1.09
2008 occupancy coefficients				
Intercept	5.60	0.810	4.18	7.38
DIST	−1.84	0.278	−2.45	−1.35
2008 spatial parameter, σ	5.46	1.091	3.61	7.90
Persistence, φ	0.94	0.008	0.92	0.95
Long-distance dispersal, γ	0.02	0.014	0.002	0.05
Neighborhood colonization				
Intercept	−2.28	0.110	−2.51	−2.07
LOG_HUMAN	−0.38	0.206	−0.83	−0.01
Number of sites occupied				
Year 2008	582	20.1	544	623
Year 2009	628	16.8	597	663
Year 2010	670	16.1	640	703
Year 2011	707	16.7	675	740
Year 2012	748	18.6	712	785
Year 2013	780	22.5	735	824

Note: Abbreviations are NSPP, number of species on the bird list; PA, proportion of the site that was in a protected area such as a national park or game reserve; DIST, distance from the site's centroid to the center of the closer of Johannesburg or Durban; CROP, proportion of the site that was cropland; LOG HUMAN, logarithm of the human population density.

credible interval of 0.002–0.05 for the gradient-based model. The persistence probability was estimated at 0.94, with a 95% credible interval of 0.92–0.95 for all models.

In the homogeneous model with constant neighborhood colonization, the neighborhood colonization probability was 0.09, with a 95% credible interval of 0.08–0.11 (Appendix S5: Table S2).

In the homogeneous model with varying neighborhood colonizations, those probabilities ranged from 0.03 to 0.21 (Appendix S5: Table S1). Higher colonization probabilities related to neighbors 6 and 7 implies that a site is most likely to be colonized from its southern and southwestern neighbors. Therefore, the range of the myna is mostly expanding in the north and northeast directions according to this model.

The gradient-based model with neighborhood colonization as a function of human density estimated the neighbor colonization probabilities to range from 0.0005 to 0.12, with a median probability of 0.003. The negative coefficient associated with the human population suggests that the myna is dispersing away from the large cities into the less populated surrounding areas, possibly because the myna populations are already saturated in the more heavily populated sites: Fig. 3 was created from Eq. 9 and provides a visualization of the neighborhood colonizations and the potential routes along which the myna expands its range.

The gradient-based model had the best predictive performance, with a log-score equal to 2418.75. In contrast, the homogeneous model with homogeneous colonizations had a log-score of 2422.09 and the homogeneous model with varying colonizations had a log-score of 2424.72.

All models estimated an increase in the number of sites becoming occupied over time. For the homogeneous model with varying colonizations, the estimated number of sites occupied in 2008 was 581 and increased

FIG. 3. Gradient surface of neighborhood colonizations. The myna is likely to disperse to the darker areas. Gray sites are the sites of known occurrence in 2008.

to 785 by the end of 2013 (Appendix S5: Table S1), with a rate of spread faster in the beginning: 8.6% in year 2008, and ending at 4.2%. The number of new sites becoming occupied each year decreased from a high of 50 new sites from 2008 to 2009, to a low of 32 new sites from 2012 to 2013. The homogeneous model with constant colonization was very similar (Appendix S5: Table S2). For the gradient-based model, the estimated number of sites occupied in 2008 was 582 and increased to 780 by 2013 (Table 1, Fig. 4), suggesting rates of spread from 8.1% to 4.3% a year. The number of new sites similarly decreased from a high of 46 to a low of 32 by the end of the study period. The rate of spread had decreased, but remained greater than zero.

DISCUSSION

Range expansions are often a focus of mathematical ecologists, but usually in the context of continuous space and integrodifference equations (e.g., Skellam 1951, Van den Busch et al. 1992, Kot et al. 1996, Neubert et al. 2000, Shigesada and Kawasaki 2002). We merged a Bayesian occupancy model with a discrete form of diffusion model to learn how an invasive species spreads across a landscape, but for data collected on relatively small-scale areal units or patches. The colonization process was a function of how many neighbors of a site were occupied in the previous time period; a site was more likely to be colonized if more of its neighbors were formerly occupied and if it had better habitat than neighboring sites, but a site could also be colonized through long-distance dispersal if it did not have occupied neighbors. These explicit connections were intuitive and provided insight into the ecological processes. In particular, this model was sensible for the myna, a species whose range was believed to be expanding.

The occupancy model is flexible and is gaining familiarity with ecologists (Bailey et al. 2014), whereas the Bayesian hierarchical framework allows for latent states and added complexity through its conditional probabilities (Hooten et al. 2003, Latimer et al. 2006). This framework allowed for occupancy in year 1 to be a function of site-specific covariates and a spatial random effect. Because our data collection was initiated after the myna had already begun its spread in South Africa, it was important to recognize the relationships that had developed, and the spatial autocovariate, incorporated through the RSR random effect, acknowledged that there were unmeasured processes additionally affecting the myna's distribution.

The derivation of the colonization process from a diffusion model sets our model apart from other spatially explicit, dynamic occupancy models (Bled et al. 2011, 2013, Yackulic et al. 2012, Eaton et al. 2014, Sutherland et al. 2014). However, the different frameworks may be complementary, as they represent different underlying mechanisms. Adding an autocovariate to the temporal components of the model as in Bled et al. (2011, 2013), Yackulic et al. (2012), and Eaton et al. (2014) is computationally convenient but less mechanistic. This may or may not be desired, depending on the data and research questions. Sutherland et al. (2014) used count data collected on discrete patches and explicitly modeled the relationship among those patches from metapopulation theory. Our model relied on conventional spatiotemporal modeling concepts (e.g., Wikle and Hooten 2010, Cressie and Wikle 2011) and mathematical theory for the movement of animals (Turchin 1998).

One metric derived from our model was a potential surface of spread from the neighborhood colonizations (Fig. 3). As far as we are aware, previous studies using multi-season occupancy models have not included such gradient maps for the colonization or extinction processes. These dispersal gradient maps can inform managers about which sites are more likely to be colonized in the future, and hence where to focus containment resources. For the myna, the map showed the flow of colonizations and the spread of the myna northward into Zimbabwe, eastward into Mozambique and Swaziland, and westward into South Africa's interior. The myna's range expansion probably will continue along these routes in the near future. Indeed, there are incidence records of the myna in parts of Botswana, Zimbabwe, and Mozambique (Peacock et al. 2007).

Previous studies suggest that human population densities and land transformations are positively correlated with the myna's spread (Peacock et al. 2007, Hugo and Rensburg 2009). Although there was some evidence from the 2008 model that the proportion of pasture was negatively associated with myna occurrences and some evidence from the 2013 model that protected areas are negatively associated myna occurrences, these

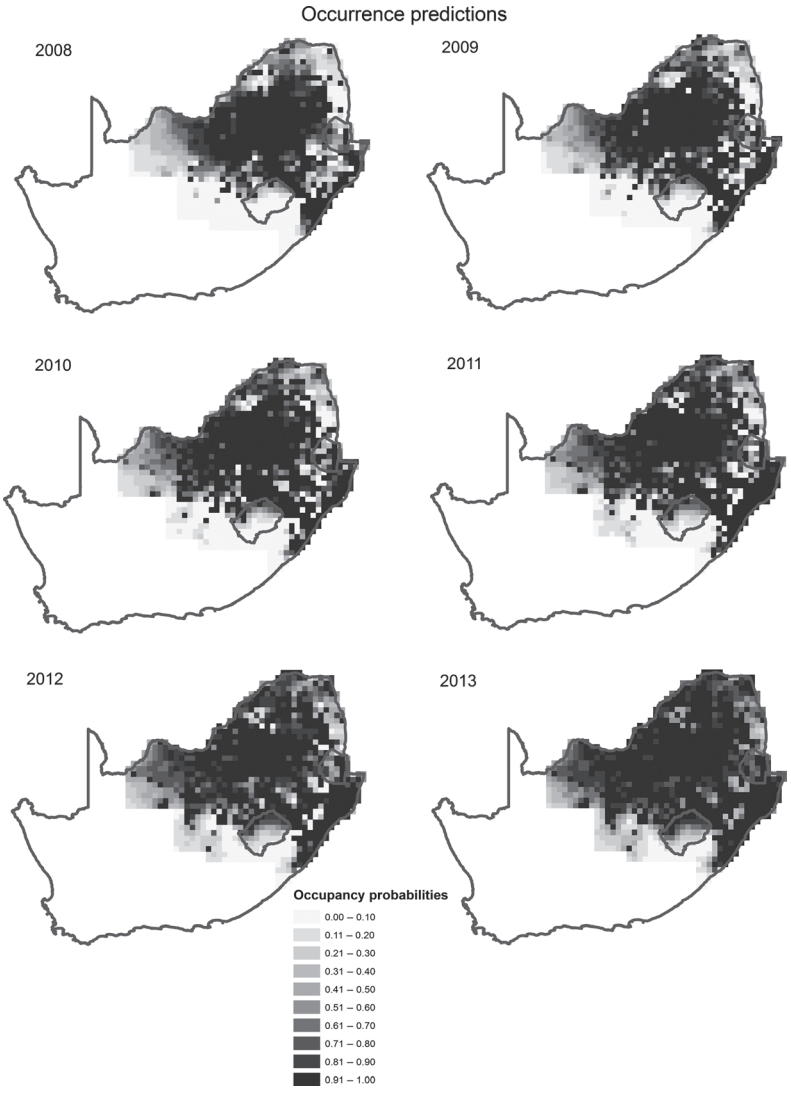

FIG. 4. The mean occurrence predictions (occupancy probabilities) for mynas for each site and year; these predictions are similar to the conditional occupancy probabilities estimated through a likelihood framework.

parameters became nonsignificant when spatial structure was added to the models. The detection parameters suggested that myna were more likely to be detected in more populated areas and croplands, and were less likely to be detected in protected areas. These results may be due to observers' expectations of where they might see myna, or it may be due to different abundance levels of myna among the landscapes. Therefore, our findings supported the relationship between mynas and human population density, but were less conclusive about how land transformations related to the myna distribution.

In the models, the long-distance dispersal probability was estimated to be 2% and the lower bound of its 95% credible interval was close to 0%. Therefore, most if not all of the myna's range expansion was through its neighborhood colonization. The gradient-based occupancy model fit the 2013 data slightly better than the homogeneous model, lending further support to the suggestion that human populations are driving the myna populations. However, the coefficient related to human population density was negative, so the myna expansion began in areas of high human population density, but then expanded away from densely populated areas.

Finally, the models suggest that the myna's range continues to expand at a rate of more than 4% a year. Because the myna population has not yet stabilized, resource managers should continue to be aware of the likelihood of myna expansions, and biologists need to be aware of the nonequilibrium resulting from the lack of stabilization when trying to determine occupancy–environment relationships (Yackulic et al. 2015).

Many other extensions of our model are possible. For example, the process component of the model could be adapted to accommodate other diffusion processes (Wikle and Hooten 2010), such as jump-diffusion (e.g., Li et al. 2014). Given that the method is based on a Lagrangian implementation of a partial differential equation, it is also possible to use optimal mathematical solution methods such as "homogenization" when fitting these models (e.g., Hooten et al. 2013). This would be especially useful for longer time series and larger spatial domains than those considered in our myna example. In the case of the myna, an invasive species whose range is expanding, we focused on the colonization process and included probabilities for extinction and for long-distance dispersal. For other species, the neighborhood colonization coefficients may vary temporally to reflect colonization patterns that change from year to year; the persistence and long-distance dispersal probabilities could be functions of spatial or temporal covariates; and the neighborhood colonization probabilities could be a function of more than one environmental gradient variable. Alternatively, the models could be extended to better understand extinction probabilities by explicitly allowing persistence to evolve dynamically, as we did with the neighborhood colonization.

Acknowledgments

The authors would like to thank all the volunteers who contributed to the Southern African Bird Atlas Project, and the reviewers for their suggestions on an earlier version of the manuscript. Any use of trade, firm, or product names is for descriptive purposes only and does not imply endorsement by the U.S. Government. R. Altwegg was supported by the National Research Foundation of South Africa (Grant 85802). The NRF accepts no liability for opinions, findings, and conclusions or recommendations expressed in this publication.

Literature Cited

Bailey, L. L., D. I. MacKenzie, and J. D. Nichols. 2014. Advances and applications of occupancy models. Methods in Ecology and Evolution 5:1269–1279.

Balk, D., U. Deichmann, G. Yetman, F. Pozzi, S. Hay, and A. Nelson. 2011. Global Rural–Urban Mapping Project, Version 1 (GRUMPv1): population count grid. http://sedac.ciesin.columbia.edu/data/dataset/grump-v1-population-count.

Bled, F., J. A. Royle, and E. Cam. 2011. Hierarchical modeling of an invasive spread: the Eurasian Collared-Dove (*Streptopelia decaocto*) in the United States. Ecological Applications 21:290–302.

Bled, F., J. D. Nichols, and R. Altwegg. 2013. Dynamic occupancy models for analyzing species' range dynamics across large geographic scales. Ecology and Evolution 3:4896–4909.

Broms, K. M. 2013. Using presence–absence data on areal units to model the ranges and range shifts of select South African bird species. Dissertation. University of Washington, Seattle, Washington, USA.

Broms, K. M., D. S. Johnson, R. Altwegg, and L. L. Conquest. 2014. Spatial occupancy models applied to atlas data show Southern Ground Hornbills strongly depend on protected areas. Ecological Applications 24:363–374.

Cressie, N., and C. K. Wikle. 2011. Statistics for spatio-temporal data. Wiley, New York, New York, USA.

Eaton, M. J., P. T. Hughes, J. E. Hines, and J. D. Nichols. 2014. Testing metapopulation concepts: effects of patch characteristics and neighborhood occupancy on the dynamics of an endangered lagomorph. Oikos 123:662–676.

Fiske, I., and R. Chandler. 2011. unmarked: an R package for fitting hierarchical models of wildlife occurrence and abundance. Journal of Statistical Software 43:1–23.

Gneiting, T., and A. E. Raftery. 2007. Strictly proper scoring rules, prediction, and estimation. Journal of the American Statistical Association 102:359–378.

Hanks, E. M., M. B. Hooten, D. S. Johnson, and J. T. Serling. 2011. Velocity-based movement modeling for individual and population level inference. PLoS ONE 6:e22795.

Hanks, E. M., E. Schliep, M. B. Hooten and J. A. Hoeting. 2015. Restricted spatial regression in practice: geostatistical models, confounding, and robustness under model misspecification. Environmetrics 26:243–254.

Harebottle, D. M., N. Smith, L. G. Underhill, and M. Brooks. 2007. Southern African Bird atlas: quick-start guide. http://sabap2.adu.org.za/docs/sabap2_instructions_v5.pdf.

Hooten, M. B., and N. T. Hobbs. 2015. A guide to Bayesian model selection for ecologists. Ecological Monographs 85:3–28.

Hooten, M. B., and C. K. Wikle. 2010. Statistical agent-based models for discrete spatio-temporal systems. Journal of the American Statistical Association 105:236–248.

Hooten, M. B., D. R. Larsen, and C. K. Wikle. 2003. Predicting the spatial distribution of ground flora on large domains using a hierarchical Bayesian model. Landscape Ecology 18:487–502.

Hooten, M. B., J. Anderson, and L. A. Waller. 2010a. Assessing North American influenza dynamics with a statistical SIRS model. Spatial and Spatio-Temporal Epidemiology 1:177–185.

Hooten, M. B., D. S. Johnson, E. M. Hanks, and J. H. Lowry. 2010b. Agent-based inference for animal movement and selection. Journal of Agricultural, Biological, and Environmental Statistics 15:523–538.

Hooten, M. B., M. J. Garlick, and J. A. Powell. 2013. Computationally efficient statistical differential equation modeling using homogenization. Journal of Agricultural, Biological, and Environmental Statistics 18:405–428.

Hughes, J., and M. Haran. 2013. Dimension reduction and alleviation of confounding for spatial generalized linear mixed models. Journal of the Royal Statistical Society B 75:139–159.

Hugo, S., and B. J. V. Rensburg. 2009. Alien and native birds in South Africa: patterns, processes and conservation. Biological Invasions 11:2291–2302.

Johnson, D. S. 2013. stocc: fit a spatial occupancy model via Gibbs sampling. R package, version 1.0–5. https://cran.r-project.org/web/packages/stocc/index.html.

Johnson, D. S., P. B. Conn, M. Hooten, J. Ray, and B. Pond. 2013. Spatial occupancy models for large data sets. Ecology 94:801–808.

Kéry, M., G. Guillera-Arroita, and J. J. Lahoz-Monfort. 2013. Analysing and mapping species range dynamics using occupancy models. Journal of Biogeography 40:1463–1474.

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

Latimer, A. M., S. Wu, A. E. Gelfand, and J. A. Jr Silander. 2006. Building statistical models to analyze species distributions. Ecological Applications 16:33–50.

Li, D., J. Cui, and G. Song. 2014. Asymptotic behaviour and extinction of delay Lotka-Volterra model with jump-diffusion. Journal of Applied Mathematics. 2014. doi:10.1155/2014/249504.

Lowe, S., M. Browne, S. Boudjelas and M. De Poorter. 2000. 100 of the world's worst invasive alien species: a selection from the Global Invasive Species Database. Species Survival Commission, World Conservation Union, Auckland, New Zealand.

MacKenzie, D. I., J. D. Nichols, J. E. Hines, M. G. Knutson, and A. B. Franklin. 2003. Estimating site occupancy, colonization, and local extinction when a species is detected imperfectly. Ecology 84:2200–2207.

Neubert, M. G., M. Kot, and M. A. Lewis. 2000. Invasion speeds in fluctuating environments. Proceedings of the Royal Society B 267:1603–1610.

Peacock, D. S., B. J. van Rensburg, and M. P. Robertson. 2007. The distribution and spread of the invasive alien common myna, *Acridotheres tristis* L. (Aves: Sturnidae), in southern Africa. South African Journal of Science 103:465–473.

Pejchar, L., and H. A. Mooney. 2009. Invasive species, ecosystem services and human well-being. Trends in Ecology and Evolution 24:497–504.

Plummer, M. 2003. JAGS: a program for analysis of Bayesian graphical models using Gibbs Sampling. Proceedings of the Third International Workshop on Distributed Statistical Computing, 20–22 March 2003, Vienna, Austria. https://www.r-project.org/conferences/DSC-2003/Proceedings/Plummer.pdf

Ramankutty, N., A. T. Evan, C. Monfreda, and J. A. Foley. 2010a. Global agricultural lands: croplands. 2000. NASA Socioeconomic Data and Applications Center (SEDAC). http://sedac.ciesin.columbia.edu/es/aglands.html.

Ramankutty, N., A. T. Evan, C. Monfreda, and J. A. Foley. 2010b. Global agricultural lands: pastures. 2000. NASA Socioeconomic Data and Applications Center (SEDAC). http://sedac.ciesin.columbia.edu/es/aglands.html.

Rouget, M., B. Reyers, Z. Jonas, P. Desmet, A. Driver, K. Maze, B. Egoh, and R. M. Cowling. 2004. South African National Biodiversity Assessment 2004: Technical Report. Volume 1: Terrestrial component. South African National Biodiversity Institute, Pretoria, South Africa.

Royle, J. A., and R. M. Dorazio. 2008. Hierarchical modeling and inference in ecology. 1st edition. Academic Press, New York, New York, USA.

Royle, J. A., and M. Kéry. 2007. A Bayesian state-space formulation of dynamic occupancy models. Ecology 88:1813–1823.

Shigesada, N., and K. Kawasaki. 2002. Invasion and the range expansion of species: effects of long-distance dispersal. Pages 350–373 *in* J. M. Bullock, R. E. Kenward, and R. S. Hails, eds. Dispersal ecology. Blackwell Science, Malden, Massachusetts, USA.

Skellam, J. 1951. Random dispersal in theoretical populations. Biometrika 38:196–218.

Sutherland, C. S., D. Eaton, and X. Lambin. 2014. A demographic, spatially explicit path occupancy model of metapopulation dynamics and persistence. Ecology 95:3149–3160.

Turchin, P. 1998. Quantitative analysis of movement: measuring and modeling population redistribution in plants and animals. Sinauer Associates, Sunderland, Massachusetts, USA.

Van den Busch, F., R. Hengeveld and J. A. J. Metz. 1992. Analysing the velocity of animal range expansion. Journal of Biogeography 19:135–150.

Wikle, C. K. and M. B. Hooten. 2010. A general science-based framework for dynamical spatio-temporal models. Test 19:417–451.

Wilcove, D. S., D. Rothstein, J. Dubow, A. Phillips and E. Losos. 1998. Quantifying threats to imperiled species in the United States. BioScience 48:607–615.

van Wilgen, B. W., R. M. Cowling, C. Marais, K. Esler, M. McConnachie and D. Sharp. 2012. Challenge in invasive alien plant control in South Africa. South African Journal of Science 108:11–13.

Yackulic, C. B., J. Reid, R. Davis, J. E. Hines, J. D. Nichols and E. Forsman. 2012. Neighborhood and habitat effects on vital rates: expansion of the Barred Owl in the Oregon Coast Ranges. Ecology 93:1953–1966.

Yackulic, C. B., J. D. Nichols, J. Reid and R. Der. 2015. To predict the niche, model colonization and extinction. Ecology 96:16–23.

Supporting Information

Additional supporting information may be found in the online version of this article at http://onlinelibrary.wiley.com/doi/10.1890/15-0416.1/suppinfo