Appendix 1: Supplemental Methods

The performance of presence-based and process-based species distribution models

Tim M. Szewczyk, Marek Petrik, Jenica M. Allen

Contents

This appendix contains supplemental methods pertaining to the virtual species generation, species distribution model specifics, and scenario particulars.

---

1 General model structure

1.1 Integral Projection Model overview

To generate fully-known true distributions for the virtual species, we used the general structure of an Integral Projection Model (IPM) to calculate the intrinsic growth rate λ in each cell of the gridded landscape, and adapted the regression-based structure of an IPM into an individual-level, simulation-based cellular automata (CA) model to produce spatiotemporally dynamic abundance distributions. The equations below apply to both models.

IPMs use the size distribution, z, of individuals at time t, along with a kernel $K(z', z)$, to predict the size distribution, z', at time $t + 1$:

$$n_{t+1}(z') = \int_{\Omega} K(z', z) n_t(z) dz \tag{1}$$

Here, Ω represents the range of possible sizes for the species or population. The kernel $K(z', z)$ is composed of a growth and survival component, $P(z, z')$, representing the fate of individuals from time t to $t + 1$, and a fecundity component, $F(z, z')$, representing new individuals added between time t and $t + 1$. In practice, the integral is approximated using a discretized transition matrix, and the intrinsic growth rate λ is calculated as the first eigenvalue of the transition matrix.

The P and F kernels are decomposed further into more mechanistic conditional probabilities and parameters, many of which are functions of the size distribution z. For example:

$$\begin{aligned} K(z', z) &= P(z', z) + F(z', z) \\ &= s(z)g(z'|z) + p_{flower}(z)f_{seeds}(z)p_{estab}f_{rcrSize}(z') \end{aligned} \tag{2}$$

where $s(z)$ is the survival probability of individuals based on size, $g(z'|z)$ is the probability density of size z' for an individual of size z, $p_{flower}(z)$ is the probability that an individual of size z produces flowers, $f_{seeds}(z)$ is the expected number of seeds produced by an individual of size z given that they flower, p_{estab} is the probability that a seed germinates and establishes as a new recruit, and $f_{rcrSize}(z')$ is the expected size distribution of new recruits at time $t + 1$. On a gridded landscape, the population in each cell could be

modelled independently using the above structure, with environmental effects incorporated by allowing the environment to influence parameter values (e.g., f_{seeds}).

1.2 Adding a seed bank

This basic IPM structure is very flexible, allowing for discrete life stages, reproduction-dependent mortality, and environmental covariates. To incorporate a seed bank where seeds that do not germinate between time t and $t+1$ may survive to $t+2$ or beyond, the fecundity kernel is altered, with seed bank B, such that:

$$n_{t+1}(z') = B_t s_{rcrB} p_{estab} f_{rcrSize}(z') + \int_\Omega [P(z',z) + F(z',z)] n_t(z) dz \tag{3}$$

$$F(z',z) = p_{flower}(z) f_{seeds}(z) s_{rcrDirect} p_{estab} f_{rcrSize}(z') \tag{4}$$

$$B_{t+1} = B_t s_{survB}(1 - s_{rcrB}) + p_{flower}(z) f_{seeds}(z)(1 - s_{rcrDirect}) s_{survB} n_t(z) dz \tag{5}$$

where B_t is the number of seeds in the seed bank at time t, s_{rcrB} is the probability a seed recruits from the seed bank, s_{survB} is the probability a seed survives in the seed bank from time t to $t+1$, and $s_{rcrDirect}$ is the probability a seed produced in year t germinates between year t and $t+1$. Two notes about the structure and definitions: 1) a seed added to the seed bank must fail to recruit in time t and must also survive from t to $t+1$, and 2) the probability of recruiting is best interpreted as the probability of germinating and is therefore separate from the probability of establishing. Both of these could be defined differently to combine each set of processes, though keeping them separate allows for seeds to perish.

1.3 Adding dispersal

The above equations assume isolated populations in each cell. However, for a typical plant species, dispersal occurs when seeds move from the cell where they are produced to a different cell. In an IPM with a seed bank, this will affect the fecundity kernel $F(z',z)$ and the seed bank B. Specifically, the number of seeds in the seed bank in cell i at time $t+1$ will be the number of seeds surviving in the seed bank B_i from time t to $t+1$, plus the number of seeds produced in cell i at time t that remain in cell i and do not recruit directly, plus the number of seeds entering cell i from cells $j = 1, ..., J$ as immigrants in time t that then fail to recruit directly:

$$\begin{aligned} B_{i,t+1} = \; & B_{i,t} s_{survSB}(1 - s_{rcrSB}) \; + \\ & \int_\Omega p_{flower,i}(z) f_{seeds,i}(z)(1 - p_{emig})(1 - s_{rcrDirect}) s_{survB} n_{i,t}(z) dz \; + \\ & \sum_{j=1}^{J} \int_\Omega [p_{flower,j}(z) f_{seeds,j}(z) p_{emig} n_{t,j}(z) dz] p_{SDD,ji}(1 - s_{rcrDirect}) s_{survB} \end{aligned} \tag{6}$$

$$\begin{aligned} n_{i,t+1}(z') = \; & B_{i,t} s_{rcrSB} p_{estab_i} f_{rcrSize}(z') \; + \\ & \int_\Omega [s_i(z) g_i(z'|z) + p_{flower_i}(z) f_{seeds_i}(z)(1 - p_{emig}) s_{rcrDirect} p_{estab_i} f_{rcrSize}(z')] n_{i,t}(z) dz \; + \\ & \sum_{j=1}^{J} \int_\Omega [p_{flower,j}(z) f_{seeds,j}(z) p_{emig} n_{t,j}(z) dz] p_{SDD,ji} s_{rcrDirect} p_{estab,i} f_{rcrSize}(z') \end{aligned} \tag{7}$$

for each cell i which is a target cell of each cell j of J cells, where the integral describes the seed production in each cell j and $p_{SDD,ji}$ is the probability that a seed dispersed from j lands in i. Note that, as above, seeds added to the seed bank must survive overwinter as well as fail to recruit directly.

2 Regression equations for the virtual species

We used the above IPM structure, including a seed bank and dispersal, as the basis for the virtual species. Thus, the population in each cell i has size distribution z, where z' is the size distribution the next year, and $\boldsymbol{z_i}$ is a matrix with columns for: 1, z, z^2, and z^3.

2.1 Survival

Annual survival, $\boldsymbol{s_i}$, was modelled for each individual in cell i as a binary outcome (0: mortality; 1: survival) following a Bernoulli distribution with probability $\boldsymbol{\psi_{si}}$ such that:

$$\boldsymbol{s_i} \sim Bern(\boldsymbol{\psi_{si}}) \tag{8}$$

$$logit(\boldsymbol{\psi_{si}}) = \boldsymbol{z_i}\beta_s + \boldsymbol{X_i}\theta_s \tag{9}$$

where β_s is a vector of covariates for size, $\boldsymbol{X_i}$ is a set of cell-level environmental covariates, and θ_s is a vector of responses to the environmental covariates.

2.2 Growth

The size distribution, $\boldsymbol{z_i'}$ of individuals in cell i for time $t+1$ was distributed normally about the vector of expected sizes, $\boldsymbol{\mu_{gi}}$ with standard deviation σ_g, such that

$$\boldsymbol{z_i'} \sim Norm(\boldsymbol{\mu_{gi}}, \sigma_g) \tag{10}$$

$$\boldsymbol{\mu_{gi}} = \boldsymbol{z_i}\beta_g + \boldsymbol{X_i}\theta_g \tag{11}$$

where β_g is a vector of covariates for size, $\boldsymbol{X_i}$ is a set of cell-level environmental covariates, and θ_g is a vector of responses to the environmental covariates.

2.3 Flowering

Individual flowering, $\boldsymbol{l_i}$, was modelled for each individual in cell i as a binary outcome (0: no flowers; 1: flowers) following a Bernoulli distribution with probability $\boldsymbol{\psi_{li}}$ such that:

$$\boldsymbol{l_i} \sim Bern(\boldsymbol{\psi_{li}}) \tag{12}$$

$$logit(\boldsymbol{\psi_{li}}) = \boldsymbol{z_i}\beta_l + \boldsymbol{X_i}\theta_l \tag{13}$$

where β_l is a vector of covariates for size, $\boldsymbol{X_i}$ is a set of cell-level environmental covariates, and θ_l is a vector of responses to the environmental covariates.

2.4 Seeds

The number of seeds produced, $\boldsymbol{d_i}$ by each flowering individual in cell i for time t was Poisson distributed about the vector of expected seed counts, $\boldsymbol{\mu_{di}}$, such that

$$\boldsymbol{d_i} \sim Poisson(\boldsymbol{\mu_{di}}) \tag{14}$$

$$log(\boldsymbol{\mu_{di}}) = \boldsymbol{z_i}\beta_d + \boldsymbol{X_i}\theta_d \tag{15}$$

where β_d is a vector of covariates for size, $\boldsymbol{X_i}$ is a set of cell-level environmental covariates, and θ_d is a vector of responses to the environmental covariates. In each cell i, the seeds produced stay in the cell, emigrate through short distance dispersal, or perish. Seeds that survive may either enter the seed bank or recruit directly.

2.5 Dispersal

The total number of immigrant seeds arriving in a cell, D_i is calculated as the sum of the seeds from nearby cells that disperse from cell j to cell i, such that:

$$D_i = \sum_{j=1}^{J} \sum_{n=1}^{N_j} \boldsymbol{d_j} p_{emig} p_{ji} \tag{16}$$

where J is the number of cells dispersing into i, N_j is the number of individuals in cell j, d_j is the vector of seed numbers produced by individuals in cell j, p_{emig} is the probability that seeds produced in cell j emigrate, and p_{ji} is the probability that a seed emigrating from cell j is dispersed to cell i.

2.6 Recruits

The number of recruits, $n_{rcr,i}$ in cell i in time t, may originate either from the seed bank, from seeds produced in cell i in year t, or from immigrant seeds arriving in year t, such that:

$$n_{rcr,i} = B_i p_{rcrB} p_{est} + \sum_{n=1}^{N_i} \boldsymbol{d_i} (1 - p_{emig}) p_{rcrDirect} p_{est} + D_i p_{rcrDirect} p_{est} \tag{17}$$

$$\boldsymbol{z'_{rcr,i}} \sim Norm(\mu_{rcr.z}, \sigma_{rcr.z}) \tag{18}$$

where B_i is the number of seeds in the seed bank, p_{rcrB} is the probability that a seed germinates from the seed bank, p_{est} is the probability that a new seedling establishes, $p_rcrDirect$ is the probability that a seed germinates in the year it was produced, $z'_{rcr,i}$ is the size distribution of new recruits in year $t+1$, $\mu_{rcr.z}$ is the mean recruit size, and $\sigma_{rcr.z}$ is the standard deviation of recruit size.

2.7 Seed Bank

Finally, the seed bank for year $t+1$ is calculated as the sum of the seeds remaining in the seed bank, the seeds produced in cell i and entering into the seed bank, and the immigrant seeds entering into the seed bank, such that:

$$B'_i = B_i (1 - p_{rcrB}) s_B + \sum_{n=1}^{N_i} \boldsymbol{d_i} (1 - p_{emig}) (1 - p_{rcrDirect}) s_B + D_i (1 - p_{rcrDirect}) s_B \tag{19}$$

where s_B is the probability that a seed survives in the seed bank to the next year.

---

3 Species Distribution Model details

This section of the appendix contains additional information regarding the structure of the species distribution models. The full R code is available from https://github.com/Sz-Tim/sdmMethodComp. Include more information on the CA_p model and some on the CA_i model.

---

4 Scenario details

This section of the appendix contains additional information regarding the data and modelling scenarios. The full R code is available from https://github.com/Sz-Tim/sdmMethodComp. Include the table showing the exact implementation of each scenario.