Mathematical Biosciences 268 (2015) 45–51

Contents lists available at ScienceDirect

Mathematical Biosciences

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

Equilibrium distributions and simulation methods for age structured populations

Fredrik Olsson*, Ola Hössjer

Department of Mathematics, Division of Mathematical Statistics, Stockholm University, Stockholm, Sweden

ARTICLE INFO

Article history:
Received 18 December 2014
Revised 3 August 2015
Accepted 5 August 2015
Available online 11 August 2015

Keywords:
Age structured population
Simulation method
Demographic variation
Genetic variation
Matrix analytic method

ABSTRACT

A simulation method is presented for the demographic and genetic variation of age structured haploid populations. First, we use matrix analytic methods to derive an equilibrium distribution for the age class sizes conditioned on the total population size. Knowledge of this distribution eliminates the need of a burn-in time in simulations. Next, we derive the distribution of the alleles at a polymorphic locus in various age classes given the allele frequencies in the total population and the age size composition. For the time dynamics, we start by simulating the dynamics for the total population. In order to generate the inheritance of the alleles, we derive their distribution conditionally on the simulated population sizes. This method enables a fast simulation procedure of multiple loci in linkage equilibrium.

1. Introduction

Simulation studies are important as a tool for checking the validity of various assumptions and approximations in population genetic models. Fast and accurate simulation techniques are therefore of interest in order to obtain reliable results. Age structured population models with deterministic growth have been of interest for a long time [4,13,14] and a good overview can be found e.g. in [12] with extended models that account both for demographic and environmental noise. In this paper, we present simulation techniques for an age structured population in a constant environment in which the age class sizes, as well as the total population size, fluctuate stochastically.

We present a discrete time model, where the age composition at each time step is described by means of matrix recursions. Using this model, we derive an approximate distribution for the age composition, given the total population size. By knowing this distribution, the need for a burn-in time is eliminated in simulations since a random draw from this distribution can be used as a starting point.

In population genetics, the genetic information at various loci is important for calculating and estimating e.g. inbreeding and effective population sizes, as reviewed by [2] and [15]. We present a method for simulation of alleles at independent loci given the trajectory of age class sizes. This method can be applied to models in which the age class sizes are either constant, or vary stochastically according to some demographic model.

The paper is organized as follows. In Section 2 we define the demographic population model. In the following section we derive an approximate conditional distribution for the age composition, given the total population size. The accuracy of these approximations is tested by means of simulations in Section 4. In Section 5, we present a method for simulation of allele frequencies at loci in linkage equilibrium. A discussion is found in Section 6, derivations in the appendices and a list of the most important notation is given in Table 1.

2. Demographic model

Consider a population divided into J age classes and let

$$\boldsymbol{N}_t = (N_{t0}, \ldots, N_{tJ-1})',$$

with $'$ referring to vector transposition, be the number of individuals in each age class at time t. Let Y_{tjh} be the number of offspring of an individual h in age class j at time t. For fixed t and j, all Y_{tjh} are independent and identically distributed random variables with mean b_j and variance σ_j^2 . The survival I_{tjh} of an individual h from time t to $t+1$ is Bernoulli distributed with probability s_j and independent of other individuals' survival. We also allow for a correlation ρ_j between the number of progeny Y_{tjh} and survival I_{tjh} of h. Let $N_{t+1,0j}$ denote the total number of newborns at time t of individuals in age class j. Then, the dynamics of the population

* Corresponding author. Tel.: +0046 8 16 45 61.
E-mail address: fredriko@math.su.se (F. Olsson).

http://dx.doi.org/10.1016/j.mbs.2015.08.003
0025-5564/

Table 1
List of notation used in the paper.

Notation	Definition
b_j	Mean number of offspring for an individual in age class j
l_j	Probability that an individual survives to age class j
s_j	Probability that an individual in age class j survives to age class $j+1$
ρ_j	Correlation between the number of progeny and survival for an individual in age class j
I_{tjh}	Survival indicator of individual h in age class j from time t to $t+1$
Y_{tjh}	Number of offspring of individual h in age class j at time t
$\boldsymbol{N}_t$	Vector containing number of individuals in all age classes at time t
N_{tj}	Number of individuals in age class j at time t
N_{tj0}	Number of newborns at time t of individuals in age class j
$\boldsymbol{G}_t$	Projection matrix of vital rates for $\boldsymbol{N}_t$
$\boldsymbol{g}$	Expected projection matrix
$\boldsymbol{\epsilon}_t$	Matrix of serially uncorrelated demographic noise for the $\boldsymbol{N}$-process
λ	Multiplicative growth rate and largest eigenvalue of $\boldsymbol{g}$
$\boldsymbol{u}$	Vector with components proportional to the center point of the equilibrium age distribution of $\boldsymbol{N}$
$\boldsymbol{v}$	Vector with components proportional to the reproductive values
$\tilde{N}_t$	Population size when age classes are weighted by $\boldsymbol{v}$
Z_{taj}	Number of individuals with allele a in age class j at time t
$\boldsymbol{Z}_{ta}$	Vector containing number of individuals with allele a in all age classes
$\tilde{Z}_{ta}$	Number of individuals with allele a when age classes are weighted by $\boldsymbol{v}$
p_{ta}	Age averaged allele frequency, with respect to $\boldsymbol{v}$, of allele a at time t

is given by

$$\begin{aligned} N_{t+1,j+1} &= \sum_{h=1}^{N_{tj}} I_{tjh}, \quad j = 0,\dots,J-2, \\ N_{t+1,0j} &= \sum_{h=1}^{N_{tj}} Y_{tjh}, \quad j = 0,\dots,J-1, \\ N_{t+1,0} &= \sum_{j=0}^{J-1} N_{t+1,0j}. \end{aligned} \tag{1}$$

This implies that the length of each age class is the same and it equals one unit of time. Following [17], the time dynamics of the population size can also be described using matrix population models (cf. [1]). Let

$$\boldsymbol{N}_{t+1} = \boldsymbol{G}_t \boldsymbol{N}_t = \boldsymbol{g}\boldsymbol{N}_t + \boldsymbol{\epsilon}_{t+1} \tag{2}$$

where $\boldsymbol{G}_t$ is a $J \times J$ projection matrix of vital rates, $\boldsymbol{g}$ is the expected projection or Leslie matrix [13] and $\boldsymbol{\epsilon}_{t+1}$ is a column vector with $E(\boldsymbol{\epsilon}_{t+1}|\boldsymbol{N}_t) = 0$ that represents serially uncorrelated demographic noise. Let $\lambda_0,\dots,\lambda_{J-1}$ be the complex-valued eigenvalues of $\boldsymbol{g}$ in descending order with respect to their moduli, and let $\boldsymbol{g} = \boldsymbol{Q}\boldsymbol{\Lambda}\boldsymbol{Q}^{-1}$ be its Jordan canonical form. The columns (rows) of the matrix $\boldsymbol{Q}$ ($\boldsymbol{Q}^{-1}$) are the right (left) eigenvectors of $\boldsymbol{g}$, and $\boldsymbol{\Lambda}$ is an upper triangular matrix with $\lambda_0,\dots,\lambda_{J-1}$ along the diagonal (see for instance [7]).

The largest eigenvalue $\lambda = \lambda_0$ of $\boldsymbol{g}$, which is real-valued, positive and unique according to Perron–Frobenius theorem, represents the multiplicative growth rate of the population. The right eigenvector $\boldsymbol{u}$ corresponding to λ consists of the stable age distribution and the elements of the left eigenvector $\boldsymbol{v}$ are proportional to the age specific reproductive values [6]. It is assumed that the elements of $\boldsymbol{u}$ and $\boldsymbol{v}$ are normalized so that $\sum_{j=0}^{J-1} u_j = \sum_{j=0}^{J-1} u_j v_j = 1$. The age specific reproductive values are of importance for age structured populations. For instance, if they are used as weights when calculating the variance effective population size it is possible to determine the long term genetic drift [5,11,17,20].

3. Distribution of the age composition

Suppose that the reproductively weighted population size at time t is $\tilde{N}_t = \boldsymbol{v}\boldsymbol{N}_t$. Here, we will derive an approximate age distribution for both the total population as well as for different alleles at a specific chromosomal locus. We show in Appendix A that recursion (2) can be rewritten as

$$\boldsymbol{N}_{t+1} - \tilde{N}_{t+1}\boldsymbol{u} = \boldsymbol{g}(\boldsymbol{N}_t - \tilde{N}_t\boldsymbol{u}) + \boldsymbol{\Pi}_2\boldsymbol{\epsilon}_{t+1}, \tag{3}$$

where $\boldsymbol{\Pi}_2 = \boldsymbol{Q}\boldsymbol{I}_2\boldsymbol{Q}^{-1}$ and $\boldsymbol{I}_2 = \mathrm{diag}(0,1,\dots,1)$ are $J \times J$ matrices. Iterating (3) with respect to t we can express the deviation from the stable age distribution $\tilde{N}_t\boldsymbol{u}$ at time t as

$$\boldsymbol{N}_t - \tilde{N}_t\boldsymbol{u} = \sum_{\tau=0}^{\infty} \boldsymbol{g}^\tau \boldsymbol{\Pi}_2 \boldsymbol{\epsilon}_{t-\tau}. \tag{4}$$

Following calculations in [3] and [17], the noise covariance matrix is

$$\mathrm{Cov}(\boldsymbol{\epsilon}_t|\boldsymbol{N}_{t-1}) \approx \tilde{N}_{t-1}\boldsymbol{\Sigma}, \tag{5}$$

where $\boldsymbol{\Sigma} = (\Sigma_{ij})$ has non-zero elements given by

$$\begin{aligned} \Sigma_{00} &= \sum_{j=0}^{J-1} u_j \sigma_j^2, \\ \Sigma_{j+1,j+1} &= u_j s_j (1-s_j), \quad j = 0\dots,J-2, \\ \Sigma_{0,j+1} &= \Sigma_{j+1,0} = u_j \sigma_j \sqrt{s_j(1-s_j)}\rho_j, \quad j = 0\dots,J-2. \end{aligned} \tag{6}$$

In formula (5), the number of individuals in each class j at time $t-1$ is approximated by $\tilde{N}_{t-1}u_j$, so that for instance the variance of the total reproductive success of all age j individuals is roughly $\tilde{N}_{t-1}u_j\sigma_j^2$. Since $\{\boldsymbol{\epsilon}_t\}$ are martingale differences, it follows from (4), (5) and the central limit theorem for martingales [8,10] that

$$\begin{aligned} \boldsymbol{N}_t - \tilde{N}_t\boldsymbol{u} | \{\tilde{N}_{t-\tau-1}\}_{\tau=0}^{\infty} &\approx N(\boldsymbol{0}, \sum_{\tau=0}^{\infty} \tilde{N}_{t-\tau-1}\boldsymbol{g}^\tau \boldsymbol{\Pi}_2 \boldsymbol{\Sigma} \boldsymbol{\Pi}_2' (\boldsymbol{g}^\tau)') \\ &\approx N(\boldsymbol{0}, \tilde{N}_{t-1}\boldsymbol{V}), \end{aligned} \tag{7}$$

is a good approximation if the sum does not converge too rapidly, so that many terms contribute. In the last step we assumed that $\tilde{N}_{t-\tau-1}/\tilde{N}_{t-1} \approx \lambda^{-\tau}$, so that the covariance matrix is proportional to

$$\boldsymbol{V} = \sum_{\tau=0}^{\infty} \lambda^{-\tau} \boldsymbol{g}^\tau \boldsymbol{\Pi}_2 \boldsymbol{\Sigma} \boldsymbol{\Pi}_2' (\boldsymbol{g}^\tau)', \tag{8}$$

and in order for the sum in (8) to converge, it is necessary that $|\lambda_1|^2 < \lambda = \lambda_0$, see [17]. Hence,

$$\boldsymbol{N}_t | \tilde{N}_t, \tilde{N}_{t-1} \approx N(\tilde{N}_t\boldsymbol{u}, \tilde{N}_{t-1}\boldsymbol{V}), \tag{9}$$

is the conditional distribution of the age composition given the weighted population size.

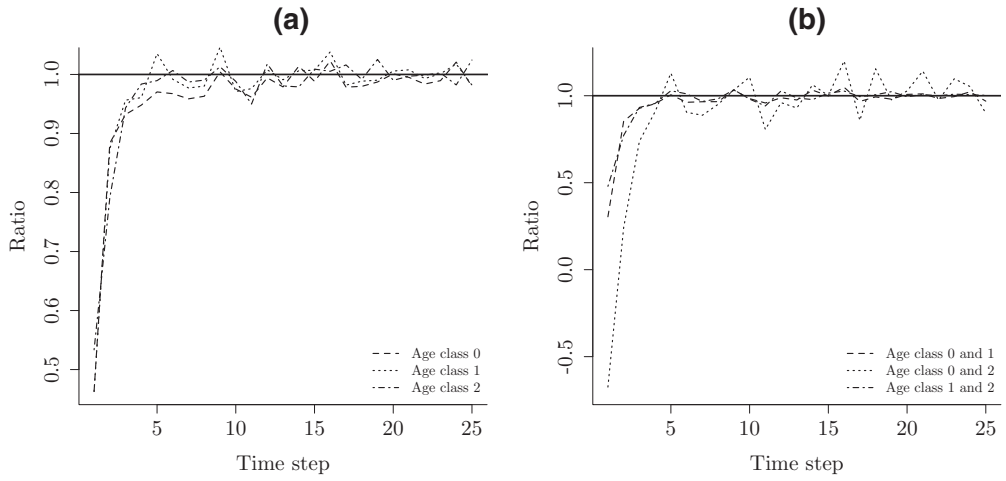

Fig. 1. Estimated variances (a) and covariances (b) from simulations divided by the corresponding element in the analytically derived covariance matrix (9), as a function of time. The estimates are based on 5000 simulated values for each age class.

Now, assume that the population is haploid, and consider a selectively neutral gene with A alleles at the locus of interest. Since the population is haploid, each individual carries a single copy of the gene. Let $\mathcal{Z}_{tja} \subset \{h = 1, \dots, N_{tj}\}$ refer to the individuals in age class j at time t with allele a, and put $Z_{tja} = |\mathcal{Z}_{tja}| = N_{tj} p_{tja}$, where p_{tja} is the frequency of a in age class j at time t. Also, let $\boldsymbol{Z}_{ta} = (Z_{ta0}, \dots, Z_{ta,J-1})'$ be a vector with the number of individuals having allele a in all age classes and let $\tilde{Z}_{ta} = \boldsymbol{v}\boldsymbol{Z}_{ta}$ be the weighted number of individuals with this allele. Hence,

$$p_{ta} = \frac{\tilde{Z}_{ta}}{\tilde{N}_t}$$

is the weighted frequency of allele a at time t. Then, the vector of total age class sizes can be expressed as

$$\boldsymbol{N}_t = \sum_{a=1}^{A} \boldsymbol{Z}_{ta}.$$

Suppose we know the age-weighted frequencies $p_{t1}, \dots, p_{tA}$ for all alleles at time t, but not how they are distributed over age classes. We can repeat the argument leading to (9) for each allele separately. In analogy with (4)–(8), since different alleles reproduce independently, we find that

$$\begin{pmatrix} \boldsymbol{Z}_{t1} - \tilde{Z}_{t1}\boldsymbol{u} \\ \vdots \\ \boldsymbol{Z}_{tA} - \tilde{Z}_{tA}\boldsymbol{u} \\ \boldsymbol{N}_t - \tilde{N}_t\boldsymbol{u} \end{pmatrix} \Big| \{\tilde{Z}_{t-1,a}\}_{a=1}^{A} \approx N\left(\begin{pmatrix} \boldsymbol{0} \\ \vdots \\ \boldsymbol{0} \\ \boldsymbol{0} \end{pmatrix}, \tilde{N}_{t-1}\boldsymbol{P}_1 \otimes \boldsymbol{V} \right), \tag{10}$$

where $\otimes$ is the tensor product of matrices and

$$\boldsymbol{P}_1 = \begin{pmatrix} p_{t1} & 0 & \dots & 0 & p_{t1} \\ 0 & p_{t2} & \ddots & \vdots & p_{t2} \\ \vdots & \ddots & \ddots & 0 & \vdots \\ 0 & \dots & 0 & p_{tA} & p_{tA} \\ p_{t1} & p_{t2} & \dots & p_{tA} & 1 \end{pmatrix}.$$

The approximate conditional allele frequency distribution given all age class sizes, is given by

$$\begin{pmatrix} \boldsymbol{Z}_{t1} \\ \vdots \\ \boldsymbol{Z}_{tA} \end{pmatrix} \Big| \boldsymbol{N}_t, \{\tilde{Z}_{ta}\}_{a=1}^{A}, \{\tilde{Z}_{t-1,a}\}_{a=1}^{A} \approx N\left(\begin{pmatrix} p_{t1}\boldsymbol{N}_t \\ \vdots \\ p_{tA}\boldsymbol{N}_t \end{pmatrix}, \tilde{N}_{t-1}\boldsymbol{P}_2 \otimes \boldsymbol{V} \right), \tag{11}$$

Table 2
Life table data for sparrows, where b_j is the mean number of progeny for an individual in age class j, $l_j = \prod_{i=0}^{j-1} s_i$ is the probability for an individual to survive to age class j and s_j the probability that an individual in age class j survives to age class $j+1$. Each age class represents 1 year.

	Sparrow	
Age class	l_j	b_j
0	1.000	0.000
1	0.167	3.018
2	0.083	3.202
3	0.048	3.416
4	0.012	3.602
5	0.006	3.842

where

$$\boldsymbol{P}_2 = \begin{pmatrix} p_{t1}(1-p_{t1}) & -p_{t1}p_{t2} & \dots & -p_{t1}p_{tA} \\ -p_{t2}p_{t1} & \ddots & & \vdots \\ \vdots & & \ddots & -p_{t,A-1}p_{tA} \\ -p_{tA}p_{t1} & \dots & -p_{tA}p_{t,A-1} & p_{tA}(1-p_{tA}) \end{pmatrix}.$$

In order to derive the mean vector and covariance matrix in (11) we used (10) and formulas, given in Appendix B, for the conditional distribution $\boldsymbol{X}|\boldsymbol{Y}$ of two jointly multivariate normal vectors $\boldsymbol{X}$ and $\boldsymbol{Y}$.

4. Simulation of demographic equilibrium

In order to illustrate the performance of the proposed method we compare our approximate age composition distribution (9) with simulated values. Using the methods in Section 2, we simulate a population with demographic parameters according to Table 2. We let $\tilde{N}_0 = 1000$ be the initial size and $\boldsymbol{u}$ the age composition and simulate the population for 25 time steps $t = 0, 1, \dots, 25$. For each time step t, we calculate the reproductively weighted population size $\tilde{N}_t$ and keep the age composition only if $\tilde{N}_t$ is rounded to 1000. We repeat the simulation until we have 5000 sampled age distributions for each time step.

In Fig. 1, we show how the estimated variances and covariances, divided by the corresponding values of the covariance matrix (9), vary

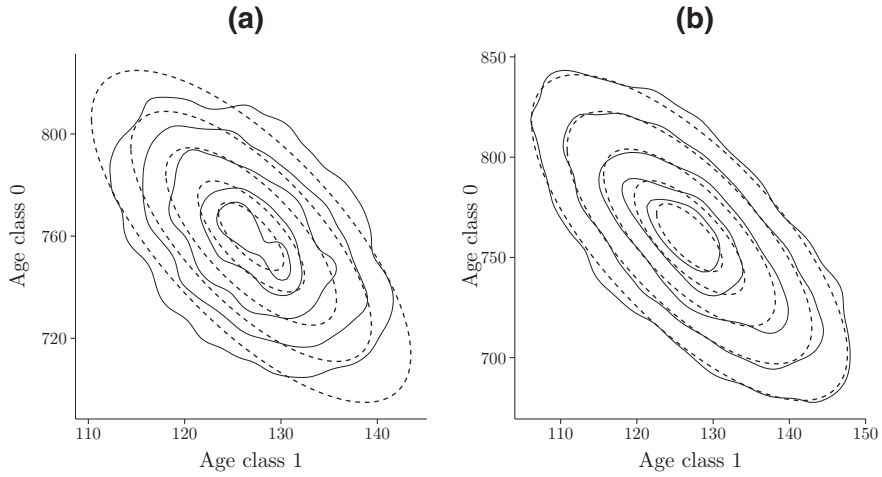

Fig. 2. Estimated quantiles from simulations (solid lines) and theoretical quantiles according to (9) (dashed lines) for the bivariate age class size distribution (N_{t0}, N_{t1}) of age classes 0 and 1. The estimates are based on 5000 simulated values for each age class and are sampled after $t = 1$ time step in (a) and after $t = 20$ time steps in (b). The bivariate quantiles from simulations are computed according to function kde2d of the package MASS [19] in R [18].

over time. We see that both variances and covariances approach the true value after only a few time steps. Only values from the first three age classes are shown, however, the variances and covariances for the other age classes follow a similar pattern.

For age classes 0 and 1, we have estimated the 10% , 25% , 50% , 75% and 90% quantiles from the simulations, as shown in Fig. 2. We compare the estimated quantiles of the bivariate distribution (N_{t0}, N_{t1}) after one and 20 time steps with the corresponding quantiles from the bivariate normal distribution in (9). As expected, we obtain a good fit if a longer burn-in time is allowed for.

5. Demographic and multilocus genetic simulation

It is straightforward to generate demographic and genetic data $\{N_{tj}, Z_{tja};\ j = 0, \dots, J-1,\ a = 1, \dots, A\}$ jointly for time $t = 0, \dots, T$ at one specific locus, according to the model of Section 2, keeping track of which genes that have the different alleles. However, we will present a different approach, by first generating all N_{tj} and then conditionally on them the allele counts Z_{tja}. As an intermediate step we also need the total offspring numbers N_{tj0} of all age classes j at all time points t. The simplest option is to generate all N_{tj0} simultaneously with the N_{tj} variables, as described in Section 2. But in order to allow also for prechosen scenarios, where N_{tj} but not N_{tj0} are prespecified, one may obtain the offspring numbers N_{tj0} conditionally on $\{N_{tj}\}$. For instance, if all

$$Y_{tjh} \sim \mathrm{Po}(b_j) \tag{12}$$

are Poisson distributed, addition of independent Poisson distributed random variables implies $N_{tj0} \sim \mathrm{Po}(N_{tj} b_j)$. Given the total number of offspring $N_{t+1,0}$, children choose age of parents multinomially

$$(N_{t+1,00}, \dots, N_{t+1,J-1,0}) | N_{t+1,0} \sim \mathrm{Mult}(N_{t+1,0}; Q_{t0}, \dots, Q_{t,J-1}) \tag{13}$$

for $t = 0, \dots, T$, with probabilities $Q_{tj} = N_{tj} b_j / \sum_{i=0}^{J-1} N_{ti} b_i$. These probabilities simplify to $Q_j = u_j b_j / (u_0 \lambda)$ for a population in exact demographic equilibrium $N_{tj} = N_t u_j$. For other offspring distributions we either use (13) as an approximation, or simulate from the exact conditional distribution of $\{N_{t+1,0j}\}_{j=0}^{J-1}$ by first generating $\{N_{t+1,0j}\}_{j=0}^{J-1}$ from the unconditional distribution of Y_{tjh} and then accept draws for which the lower equation of (1) holds.

The major advantage of generating $\{N_{tj}\}$ before $\{Z_{tja}\}$ is that this approach allows us to generate allele frequencies independently for loci in linkage equilibrium in one single population, with the same demographic history $\{N_{tj}\}$ at all loci. The price to be paid is a more complicated algorithm for generating allele frequencies. In order to allocate alleles to all age classes at all time points we start at $t = 0$, assuming that the weighted allele frequencies $p_{01}, \dots, p_{0A}$ are known, for instance drawn from allele frequency spectrum data. Then we generate $\{Z_{0ja}\}_{j=0,a=1}^{J-1,A}$ according to (11) and derive a recursive scheme for the allele frequency change in all age classes from time t to $t+1$, conditionally on $\{N_{tj}, N_{t+1,j}, N_{t+1,j0}, \{Z_{tja}\}_{a=1}^{A}\}_{j=0}^{J-1}$, for $t = 0, \dots, T-1$. To this end, we need to find the conditional distribution of survival

$$Z_{t+1,j+1,a} = \sum_{h \in \mathcal{Z}_{tja}} I_{tjh} \tag{14}$$

and reproduction

$$Z_{t+1,0,a} = \sum_{j=0}^{J-1} Z_{t+1,j0a} \tag{15}$$

where

$$Z_{t+1,j0a} = \sum_{h \in \mathcal{Z}_{tja}} Y_{tjh} \tag{16}$$

is the number of offspring of individuals in age class $j = 0, \dots, J-1$ at time $t = 0$ that have allele $a = 1, \dots, A$.

Starting with survival from time t to $t+1$, the conditional distribution

$$Z_{t+1,j+1,1}, \dots, Z_{t+1,j+1,A} | N_{tj}, N_{t+1,j+1}, Z_{tj1}, \dots, Z_{t,j,A} \sim \mathrm{MultHyp}\left(N_{tj}, N_{t+1,j+1}; \frac{Z_{tj1}}{N_{tj}}, \dots, \frac{Z_{tjA}}{N_{tj}}\right) \tag{17}$$

follows a multivariate hypergeometric distribution for $j = 0, \dots, J-2$, since survival of individuals in each age class j are independent and Bernoulli distributed random variables with the same probability s_j.

For the conditional distribution of reproduction, we need to generate newborns at time $t+1$ in each age class j. We first assume that $\rho_j = 0$, and in order to obtain an explicit distribution we need to make further assumptions on the distribution of Y_{tjh}. If Y_{tjh} are Poisson distributed (12), reproduction follows a Wright–Fisher model within each age class, with

$$(Y_{tj1}, \dots, Y_{tjN_{tj}}) | N_{tj}, N_{t+1,j0} \sim \mathrm{Mult}\left(N_{t+1,j0}; \frac{1}{N_{tj}}, \dots, \frac{1}{N_{tj}}\right). \tag{18}$$

Together with (16) this implies

$$(Z_{t+1,0j1},\dots,Z_{t+1,0jA})|N_{tj},N_{t+1,j0},Z_{tj1},\dots,Z_{tjA} \sim \text{Mult}\left(N_{t+1,j0};\frac{Z_{tj1}}{N_{tj}},\dots\frac{Z_{tjA}}{N_{tj}}\right). \quad (19)$$

Alternatively, in order to allow for overdispersion, we may assume that all offspring numbers

$$Y_{tjh} \sim \text{NegBin}(m_j,q_j)$$

have a negative binomial distribution, with expected value

$$E(Y_{tjh}) = \frac{m_j(1-q_j)}{q_j} = b_j$$

and variance

$$\text{Var}(Y_{tjh}) = \frac{m_j(1-q_j)}{q_j^2} = \sigma_j^2$$

in age class j. Since $Y_{tjh} \sim \text{Po}(\Lambda_{tjh})$ has a mixed Poisson distribution with a gamma distributed mean $\Lambda_{tjh} \sim \Gamma(m_j,(1-q_j)/q_j)$, it follows that the offspring numbers of age class j and time t have a Dirichlet multinomial distribution

$$\begin{aligned}(Y_{tj1},\dots,Y_{tjN_{tj}})|N_{t+1,j0},\boldsymbol{P}_{tj} &\sim \text{Mult}(N_{t+1,j0},\boldsymbol{P}_{tj}),\\ \boldsymbol{P}_{tj}|N_{tj} &\sim \text{Dir}(m_j,\dots,m_j),\end{aligned} \quad (20)$$

where $\boldsymbol{P}_{tj} = (P_{tj1},\dots,P_{tjN_{tj}})$, see for instance [9] and references therein. By the marginalization property of the Dirichlet distribution, (16) implies

$$\begin{aligned}(Z_{t+1,j01},\dots Z_{t+1,j0A})|N_{t+1,j0},\boldsymbol{P}_{tj} &\sim \text{Mult}(N_{t+1,j0},\overline{\boldsymbol{P}}_{tj}),\\ \overline{\boldsymbol{P}}_{tj}|N_{tj},Z_{tj1},\dots,Z_{tjA} &\sim \text{Dir}(Z_{tj1}m_j,\dots,Z_{tjA}m_j),\end{aligned} \quad (21)$$

with $\bar{\boldsymbol{P}}_{tj} = (\bar{P}_{tj1},\dots,\bar{P}_{tjA})$, and $\bar{P}_{tja} = \sum_{h\in\mathcal{Z}_{tja}} P_{tjh}$. The Poisson case (19) corresponds to $q_j \to 1$ and $m_j \to \infty$ while b_j is kept fixed. Then, $\sigma_j^2 = b_j/q_j$ converges to b_j and $\bar{P}_{tja}|N_{tj},Z_{tja}$ converges in probability to Z_{tja}/N_{tj}, the non-random probabilities of the multinomial distribution in (19).

Suppose $q_j = q$ (and hence also the amount of overdispersion) is the same in all age classes j, whereas m_j (and hence also the expected number of offspring b_j) varies. Then the total offspring numbers within each age class can be generated from a distribution

$$\begin{aligned}(N_{t+1,00},\dots,N_{t+1,J-1,0})|N_{t+1,0} &\sim \text{Mult}(N_{t+1,0},\boldsymbol{P}_t)\\ \boldsymbol{P}_t &\sim \text{Dir}(N_{t1}m_1,\dots,N_{tA}m_A).\end{aligned}$$

that generalizes (13).

When $\rho_j \neq 0$ things are much more complicated. Formula (17), describing survival, will be unchanged but reproduction will be conditioned on survival, and a simulation method is presented in Appendix C.

To generate multiple loci, we use the same N_{tj} and N_{tj0} at all loci. If they are in linkage equilibrium, we may either repeat the above procedure and generate all Z_{tj} and Z_{tj0} independently at each locus. Alternatively, in order to mimic a diploid situation more accurately, we employ the *same* Y_{tjh} and I_{tjh} at all loci, and then pick the sets $\mathcal{Z}_{tja}$ randomly and independently between loci, according to the allele frequencies at time t for each age class and locus combination.

6. Discussion

In this paper, we present a method to simulate demographics and DNA for an age structured population at loci in linkage equilibrium. By specifying the initial population size and aged averaged allele frequencies we first derive novel formulas for the initial approximate age distribution of the total population as well as how the alleles distribute over age classes independently at different loci. Given that the initial age distribution is known, it is straightforward to generate future generations of the population according to the method described in Section 5. In Fig. 3, we give a schematic overview of the simulation process.

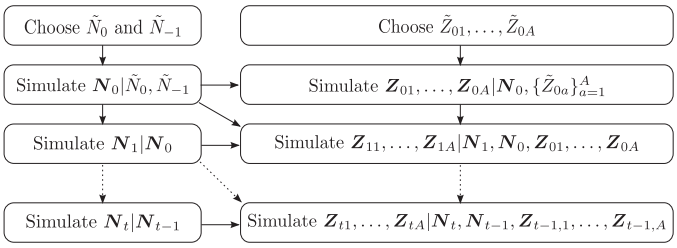

Fig. 3. Schematic overview of how the demography of a population and its genetic variation is simulated over time. The left part shows how the demography is generated and the right part describes simulation of genetic data at one locus. Only the right part is repeated in order to generate data at multiple loci.

An advantage of knowing the distribution of the age composition is that the need for a burn-in time is eliminated in the simulations. For populations in which the population size fluctuates, it enables us to draw the initial distribution of the age composition and ensures that each simulation starts with the same total population size.

For age structured populations with constant age class sizes, it is easy to apply a burn-in time to ensure that the distribution of alleles has reached its stationary distribution. However, with a burn-in time it is harder to choose an exact allele frequency to start with. With the proposed method, the initial allele frequency can be specified for each allele at a given locus, and the method can be used regardless of the scenario by which the population was generated, as long as the assumptions (2) and (5) of a time invariant expected projection matrix and demographic noise are reasonable.

The method to generate DNA for loci in linkage equilibrium gives a more realistic and efficient simulation procedure compared to if age class sizes and number of offspring for each individual were simulated at each locus. This method has recently been used to simulate data in order to illustrate the performance of a multilocus estimator of the variance effective population size [16].

It is also possible to extend the demographic simulation method to a model where all reproduction and survival parameters $b_j = b_j(\boldsymbol{N})$, $\sigma_j^2 = \sigma_j^2(\boldsymbol{N})$, $s_j = s_j(\boldsymbol{N})$ and $\rho_j = \rho_j(\boldsymbol{N})$ depend on the current population size vector $\boldsymbol{N}_t = \boldsymbol{N}$. This incorporates populations with a finite carrying capacity. Recursion (2) then generalizes to

$$\begin{aligned}\boldsymbol{N}_{t+1} &= \boldsymbol{f}(\boldsymbol{N}_t) + \boldsymbol{\epsilon}_{t+1}\\ &\approx \boldsymbol{f}(\boldsymbol{n}) + \boldsymbol{g}(\boldsymbol{N}_t - \boldsymbol{n}) + \boldsymbol{\epsilon}_{t+1}\end{aligned} \quad (22)$$

if $\boldsymbol{n} = (n_0,\dots,n_{J-1})'$ is a fix point $\boldsymbol{f}(\boldsymbol{n}) = \boldsymbol{n}$, and $\boldsymbol{g} = \boldsymbol{D}\boldsymbol{f}(\boldsymbol{n})$ the non-negative derivative matrix at the fix point. If the fix point is stable, and if $\boldsymbol{g}$ is irreducible and aperiodic, $\boldsymbol{g}$ has a unique real-valued largest eigenvalue $\lambda = \lambda_0 < 1$. It is then possible to apply (9) with $\tilde{N}_t = \boldsymbol{v}(\boldsymbol{N}_t - \boldsymbol{n})$ rather than $\tilde{N}_t$, so that

$$\boldsymbol{N}_t|\tilde{N}_t \approx N(\boldsymbol{n} + \tilde{N}_t\boldsymbol{u}, n\boldsymbol{V}),$$

where $\boldsymbol{v}$ and $\boldsymbol{u}$ are the left and right eigenvectors of $\boldsymbol{g}$ corresponding to λ, $n = \sum_{j=0}^{J-1} n_j$, $\boldsymbol{V}$ is given by (7), but $\boldsymbol{\Sigma}$ is defined differently, with n_j/n replacing u_j everywhere. In Appendix D, we give an example of a population for which the carrying capacity affects the offspring distribution.

On the other hand, it is not as straightforward to generalize the allele frequency distribution (11) when the carrying capacity is finite. The reason is that different alleles do not change independently, as in (10). Finding this allele frequency distribution is an interesting topic for further research.

Acknowledgments

Financial support from the Swedish Research Council, contracts nr. 621-2008-4946 and 621-2013-4633, and the Gustafsson Foundation for Research in Natural Sciences and Medicine to Ola Hössjer is acknowledged. We also wish to thank a reviewer for helpful comments on the manuscript.

Appendix A. Derivation of (3)

Starting with $\boldsymbol{N}_{t+1} - \tilde{N}_{t+1}\boldsymbol{u}$, we apply recursion (2) and have that

$$\begin{aligned}\boldsymbol{N}_{t+1} - \tilde{N}_{t+1}\boldsymbol{u} &= \boldsymbol{g}\boldsymbol{N}_t + \boldsymbol{\epsilon}_{t+1} - \tilde{N}_{t+1}\boldsymbol{u}\\ &= \boldsymbol{g}\boldsymbol{N}_t - \tilde{N}_t\lambda\boldsymbol{u} + \boldsymbol{\epsilon}_{t+1} - (\tilde{N}_{t+1} - \lambda\tilde{N}_t)\boldsymbol{u}\\ &= \boldsymbol{g}(\boldsymbol{N}_t - \tilde{N}_t\boldsymbol{u}) + \boldsymbol{\epsilon}_{t+1} - (\tilde{N}_{t+1} - \lambda\tilde{N}_t)\boldsymbol{u}\\ &= \boldsymbol{g}(\boldsymbol{N}_t - \tilde{N}_t\boldsymbol{u}) + \boldsymbol{\Pi}_1\boldsymbol{\epsilon}_{t+1} + \boldsymbol{\Pi}_2\boldsymbol{\epsilon}_{t+1} - (\tilde{N}_{t+1} - \lambda\tilde{N}_t)\boldsymbol{u}\\ &= \boldsymbol{g}(\boldsymbol{N}_t - \tilde{N}_t\boldsymbol{u}) + \boldsymbol{\Pi}_2\boldsymbol{\epsilon}_{t+1},\end{aligned}$$

where $\boldsymbol{\Pi}_1 = \boldsymbol{Q}\boldsymbol{I}_1\boldsymbol{Q}^{-1}$ and $\boldsymbol{I}_1 = \mathrm{diag}(1, 0, \ldots, 0)$ are $J \times J$ matrices. In the third step we used that $\boldsymbol{g}\boldsymbol{u} = \lambda\boldsymbol{u}$ and in the last step that

$$\begin{aligned}0 &= \boldsymbol{v}(\boldsymbol{N}_{t+1} - \tilde{N}_{t+1}\boldsymbol{u})\\ &= \lambda\boldsymbol{v}(\boldsymbol{N}_t - \tilde{N}_t\boldsymbol{u}) + \boldsymbol{v}\boldsymbol{\Pi}_1\boldsymbol{\epsilon}_{t+1} + \boldsymbol{v}\boldsymbol{\Pi}_2\boldsymbol{\epsilon}_{t+1} - \boldsymbol{v}(\tilde{N}_{t+1} - \lambda\tilde{N}_t)\boldsymbol{u}\\ &= \boldsymbol{v}\big[\boldsymbol{\Pi}_1\boldsymbol{\epsilon}_{t+1} - (\tilde{N}_{t+1} - \lambda\tilde{N}_t)\boldsymbol{u}\big].\end{aligned}$$

By definition of the matrix $\boldsymbol{\Pi}_1$, $\boldsymbol{\Pi}_1\boldsymbol{\epsilon}_{t+1} - (\tilde{N}_{t+1} - \lambda\tilde{N}_t)\boldsymbol{u} = c\boldsymbol{u}$ for some constant c. The normalization of $\boldsymbol{u}$ and $\boldsymbol{v}$ implies that $\boldsymbol{v}[\boldsymbol{\Pi}_1\boldsymbol{\epsilon}_{t+1} - (\tilde{N}_{t+1} - \lambda\tilde{N}_t)\boldsymbol{u}] = c\boldsymbol{v}\boldsymbol{u} = c$. Therefore $\boldsymbol{v}[\boldsymbol{\Pi}_1\boldsymbol{\epsilon}_{t+1} - (\tilde{N}_{t+1} - \lambda\tilde{N}_t)\boldsymbol{u}] = 0$ leads to $c = 0$ and hence also $\boldsymbol{\Pi}_1\boldsymbol{\epsilon}_{t+1} - (\tilde{N}_{t+1} - \lambda\tilde{N}_t)\boldsymbol{u} = \boldsymbol{0}$.

Appendix B. Mean vector and covariance matrix in (11)

Let $\boldsymbol{X}$ and $\boldsymbol{Y}$ be two jointly multivariate normal vectors such that

$$\begin{pmatrix}\boldsymbol{X}\\ \boldsymbol{Y}\end{pmatrix} \sim N\left(\begin{pmatrix}\boldsymbol{\mu}_x\\ \boldsymbol{\mu}_y\end{pmatrix}, \begin{pmatrix}\boldsymbol{\Sigma}_{xx}\boldsymbol{\Sigma}_{xy}\\ \boldsymbol{\Sigma}_{yx}\boldsymbol{\Sigma}_{yy}\end{pmatrix}\right). \tag{B.1}$$

The conditional distribution $\boldsymbol{X}|\boldsymbol{Y} = \boldsymbol{y}$ is also multivariate normal with mean vector $\boldsymbol{\mu}_{x|y} = \boldsymbol{\mu}_x + \boldsymbol{\Sigma}_{xy}\boldsymbol{\Sigma}_{yy}^{-1}(\boldsymbol{y} - \boldsymbol{\mu}_y)$ and covariance matrix $\boldsymbol{\Sigma}_{x|y} = \boldsymbol{\Sigma}_{xx} - \boldsymbol{\Sigma}_{xy}\boldsymbol{\Sigma}_{yy}^{-1}\boldsymbol{\Sigma}_{yx}$.

In order to derive (11) from (10), we apply (B.1) with

$$\begin{aligned}\boldsymbol{X} &= (\boldsymbol{Z}_{t1}, \ldots, \boldsymbol{Z}_{tA})'|\{\tilde{Z}_t\}_{a=1}^A, \{\tilde{Z}_{t-1}\}_{a=1}^A,\\ \boldsymbol{Y} &= \boldsymbol{N}_t|\tilde{N}_t, \tilde{N}_{t-1},\\ \boldsymbol{y} &= \boldsymbol{N}_t,\\ \boldsymbol{\mu}_x &= (\tilde{Z}_{t1}\boldsymbol{u}', \ldots, \tilde{Z}_{tA}\boldsymbol{u}')',\\ \boldsymbol{\mu}_y &= \tilde{N}_t\boldsymbol{u},\\ \boldsymbol{\Sigma}_{xx} &= \tilde{N}_{t-1}\mathrm{diag}(p_{t1}, \ldots, p_{tA}) \otimes \boldsymbol{V},\\ \boldsymbol{\Sigma}_{yy} &= \tilde{N}_{t-1}\boldsymbol{V},\\ \boldsymbol{\Sigma}_{yx} &= \boldsymbol{\Sigma}_{xy}' = \tilde{N}_{t-1}(p_{t1}, \ldots, p_{tA}) \otimes \boldsymbol{V}.\end{aligned}$$

We also use that, for any square matrices $\boldsymbol{A}$ and $\boldsymbol{B}$, $(\boldsymbol{A} \otimes \boldsymbol{V})(\boldsymbol{B} \otimes \boldsymbol{V}) = (\boldsymbol{A}\boldsymbol{B} \otimes \boldsymbol{V})$ and if $\boldsymbol{A}$ is also invertible, then $(\boldsymbol{A} \otimes \boldsymbol{V})^{-1} = \boldsymbol{A}^{-1} \otimes \boldsymbol{V}^{-1}$. Note also that we made the approximation $p_{t-1,a} = p_{ta}$ in (10) and (11), for $a = 1, \ldots, A$. □

Appendix C. Simulation of populations with correlation between reproduction and survival

Assume $Y_{tjh} \sim \mathrm{NegBin}(m_j, q_j)$ has a negative binomial distribution and that survival conditionally on offspring numbers follows a logistic regression model

$$P(I_{tjh} = x|Y_{tjh} = y) = \frac{(e^{\beta_{0j}+\beta_{1j}y})^{\{x=1\}}}{1 + e^{\beta_{0j}+\beta_{1j}y}}$$

for $x = 0, 1$. There is a one-to-one correspondence between the regression parameters β_{0j} and β_{1j} for age class j and s_j, ρ_j, the survival probability and correlation coefficient. Note in particular that $\beta_{1j} = 0$ is equivalent to $\rho_j = 0$.

When $\rho_j \neq 0$, we can still generate survival variables I_{tjh} according to (17). But then we must generate offspring numbers conditionally on survival as

$$\begin{aligned}P(Y_{tjh} = y|I_{tjh} = x) &\propto P(I_{tjh} = x|Y_{tjh} = y)P(Y_{tjh} = y)\\ &= \frac{exp(\beta_{0j}+\beta_{1j}y)^{\{x=1\}}}{1+exp(\beta_{0j}+\beta_{1j}y)}\binom{m_j+y-1}{y}(1-q_j)^{m_j}q_j^y\\ &=: f_j(y|x),\end{aligned} \tag{C.1}$$

and condition on the total number of offspring of individuals in age class j as well, so that

$$\begin{aligned}P\Bigg(Y_{tj1} = y_1, \ldots, Y_{tjN_{tj}} = y_{N_{tj}}|\sum_{h=1}^{N_{tj}} Y_{tjh}\\ = N_{t+1,j0}, I_{tj1} = x_1, \ldots, I_{tjN_{tj}} = x_{N_{tj}}\Bigg) = C\prod_{h=1}^{N_{tj}} f_j(y_h|x_h),\end{aligned} \tag{C.2}$$

where C is a normalizing constant assuring that the conditional probabilities in (C.2) sum to 1. One possibility is to simulate from another conditional distribution, where the constraint $\sum_{h=1}^{N_{tj}} Y_{tjh} = N_{t+1,j,0}$ is removed, so that all $Y_{tj1}, \ldots, Y_{tjN_{tj}}$ are conditionally independent with distributions $f_j(\cdot\,|I_{tjh})$.

Then, in the end, only draws with $\sum_{h=1}^{N_{tj}} Y_{tjh} = N_{t+1,j,0}$ are accepted. Finally, the allele frequencies are updated in all age classes according to (14)–(16).

Appendix D. Demography example with linear decrease in expected value of the offspring distribution

Here, we give an example of a population with a carrying capacity K and how it affects the offspring distribution. Suppose that the Poisson assumption (12) holds, but let the expected number of offspring of a parent in age class j

$$b_j = b_j(\boldsymbol{N}_t) = b_j'\left[\max\left(0, 1 - \frac{\sum_{i=0}^{J-1} N_{ti}}{K}\right)\right],$$

and the corresponding variance $\sigma_j^2(\boldsymbol{N}_t) = b_j(\boldsymbol{N}_t)$, depend on the current population size, where b_j' is a constant. Then, if survival s_j and correlation ρ_j between survival and reproduction for any age class j do not depend on the population size, we can apply recursion (22) with

$$\boldsymbol{f}(\boldsymbol{N}_t) = \begin{pmatrix}f_0(\boldsymbol{N}_t)\\ f_1(\boldsymbol{N}_t)\\ \vdots\\ f_{J-1}(\boldsymbol{N}_t)\end{pmatrix} = \begin{pmatrix}\sum_{j=0}^{J-1} b_j(\boldsymbol{N}_t)N_{tj}\\ s_0N_{t0}\\ \vdots\\ s_{J-2}N_{t,J-2}\end{pmatrix},$$

so that

$$\boldsymbol{n} = n_0\begin{pmatrix}l_0\\ l_1\\ \vdots\\ l_{J-1}\end{pmatrix}$$

is the stable fix point, with

$$n_0 = \frac{K\big[1 - (\sum_{j=0}^{J-1} b_j'l_j)^{-1}\big]}{\sum_{j=0}^{J-1} l_j}.$$

References

[1] H. Caswell, Matrix Population Models: Construction, Analysis, and Interpretation, Sinauer Associates Inc, Sunderland, MA, 2001.
[2] B. Charlesworth, Effective population size and patterns of molecular evolution and variation, Nat. Rev. Genet. 10 (3) (2009) 195–205.
[3] S. Engen, R. Lande, B. Saether, Effective size of a fluctuating age-structured population, Genetics 170 (2) (2005) 941–954.
[4] L. Euler, A general investigation into the mortality and multiplication of the human species, Theor. Popul. Biol. 1 (3) (1970) 307–314 (Originally published in 1760).
[5] J. Felsenstein, Inbreeding and variance effective numbers in populations with overlapping generations, Genetics 68 (4) (1971) 581–597.
[6] R. Fisher, The Genetical Theory of Natural Selection, second ed., Dover Publications, Inc., New York, 1958.
[7] G. Grimmett, D. Stirzaker, Probability and Random Processes, Oxford University Press, USA, 2001.
[8] P. Hall, C.C. Heyde, Martingale Limit Theory and Its Applications, Academic, New York, 1980.
[9] T. Huillet, M. Möhle, Population genetics models with skewed fertilities: a forward and backward analysis, Stochastic Models 27 (3) (2011) 521–554.
[10] J. Jacod, A.N. Shiryaev, Limit Theorems for Stochastic Processes, vol. 288, Springer-Verlag, Berlin, 1987.
[11] P. Jorde, N. Ryman, Temporal allele frequency change and estimation of effective size in populations with overlapping generations, Genetics 139 (2) (1995) 1077–1090.
[12] R. Lande, S. Engen, B. Sæther, Stochastic Population Dynamics in Ecology and Conservation, Oxford University Press, USA, 2003.
[13] P. Leslie, On the use of matrices in certain population mathematics, Biometrika 33 (3) (1945) 183–212.
[14] A. Lotka, Elements of Physical Biology, Williams & Watkins, Baltimore, 1924 (Reprinted and revised as Elements of Mathematical Biology, 1956, Dover, New York).
[15] G. Luikart, N. Ryman, D. Tallmon, M. Schwartz, F. Allendorf, Estimation of census and effective population sizes: the increasing usefulness of DNA-based approaches, Conserv. Genet. 11 (2) (2010) 355–373.
[16] F. Olsson, O. Hössjer, Estimation of the variance effective population size in age structured populations, Theor. Popul. Biol. 101 (2015) 9–23.
[17] F. Olsson, O. Hössjer, L. Laikre, N. Ryman, Characteristics of the variance effective population size over time using an age structured model with variable size, Theor. Popul. Biol. 90 (2013) 91–103.
[18] R Core Team, R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria, 2014.
[19] W.N. Venables, B.D. Ripley, Modern Applied Statistics with S, fourth ed., Springer, New York, 2002 ISBN 0-387-95457-0.
[20] R.S. Waples, M. Yokota, Temporal estimates of effective population size in species with overlapping generations, Genetics 175 (1) (2007) 219–233.