Inbreeding, Effective Population Sizes and Genetic Differentiation
- A Mathematical Analysis of Structured Populations

Fredrik Olsson
Inbreeding, Effective Population Sizes and Genetic Differentiation
A Mathematical Analysis of Structured Populations

Fredrik Olsson
Abstract

This thesis consists of four papers on various aspects of inbreeding, effective population sizes and genetic differentiation in structured populations, that is, populations that consist of a number of subpopulations. Three of the papers concern age structured populations, where in the first paper we concentrate on calculating the variance effective population size ($N_eV$) and how $N_eV$ depends on the time between measurements and the weighting scheme of age classes. In the third paper we develop an estimation procedure of $N_eV$ which uses age specific demographic parameters to obtain approximately unbiased estimates. A simulation method for age structured populations is presented in the fourth paper. It is applicable to models with multiallelic loci in linkage equilibrium.

In the second paper, we develop a framework for analysis of effective population sizes and genetic differentiation in geographically subdivided populations with a general migration scheme. Predictions of gene identities and gene diversities of the population are presented, which are used to find expressions for effective population sizes ($N_e$) and the coefficient of gene differentiation ($G_{ST}$). We argue that not only the asymptotic values of $N_e$ and $G_{ST}$ are important, but also their temporal dynamic patterns.

The models presented in this thesis are important for understanding how different age decomposition, migration and reproduction scenarios of a structured population affect quantities, such as various types of effective sizes and genetic differentiation between subpopulations.
List of Papers

The following papers, referred to in the text by their Roman numerals, are included in this thesis.


**Contribution:** F. Olsson contributed with the writing whereas the modeling was done jointly.


**Contribution:** F. Olsson contributed with implementation of the methods and a major part of the examples. The writing was done jointly with O. Hössjer as leading author.


**Contribution:** F. Olsson contributed with the writing whereas the modeling was done jointly.


**Contribution:** F. Olsson contributed with the writing and the original idea. The modeling was done jointly.

---

Reprints were made with permission from the publishers.
Related articles that are not included in the thesis but have been published during the course of the PhD studies:

DOI: [10.1016/j.tpb.2015.03.006](https://doi.org/10.1016/j.tpb.2015.03.006)
## Contents

**Abstract**

**List of Papers**

### I Introduction

1 **Theoretical population genetics**

2 **Homogeneous populations**
   2.1 Reproduction
   2.2 Inbreeding, genetic drift and coalescence
   2.3 Effective population size
      - 2.3.1 Inbreeding effective population size
      - 2.3.2 Variance effective population size
      - 2.3.3 Eigenvalue effective population size
      - 2.3.4 Coalescent effective population size

3 **Structured populations**
   3.1 Spatially structured populations
   3.2 Age-structured populations
      - 3.2.1 Summary of paper I
      - 3.2.2 Summary of paper III
      - 3.2.3 Summary of paper IV

4 **Outlook**

**Sammanfattning**

**References**

**Tack**

### II Papers
Part I

Introduction
1. Theoretical population genetics

In the mid 19th century, Gregor Mendel conducted experiments on peas to investigate rules of inheritance of traits. By crossbreeding peas with different traits he discovered some of the fundamental laws of inheritance. Because of these discoveries, Mendel is often referred to as the father of genetics. Since then, genetic research has evolved and is now subdivided into many research fields.

Theoretical population genetics is a branch of genetics in which models of populations are studied. The foundations were developed mainly by Sewall Wright, R. A. Fisher and J. B. S. Haldane in the beginning of the 20th century (Crow & Kimura 1970). The main interest is to study alleles in a population over time and in particular how distributions of the alleles change. An allele is a variant of a part of the chromosome, for instance a single base pair or a whole gene, found at a specific position in the genome of an organism, referred to as a genetic locus. A locus is called biallelic if there are two alleles present in the population and multiallelic if there are more than two variants.

There are many different processes that can affect the transmission process of alleles in the population, e.g. mutation, selection, genetic drift and gene flow/migration. Over the years, many models have been developed to capture the effect these and other processes have on the genetic composition of the population. The papers included in this thesis mainly concern genetic drift, although it is possible to include mutations in the model as well in paper II. The model species are haploid, i.e. they only have one copy of their chromosomes in the cell nucleus whereas a diploid organism has two copies. It is common to consider a haploid population of $2N$ individuals in order to mimic a diploid population of size $N$.

There are many reasons for studying theoretical population genetics and parameters in the models are used e.g. for understanding microevolutionary dynamics (Ewens 2004), in conservation biology (Franklin 1980, Allendorf & Ryman 2002, Jamieson & Allendorf 2012) and in artificial breeding programs (Lande & Barrowclough 1987).
The models in the papers of this thesis concern populations with some type of structure. However, we will first introduce vital concepts such as inbreeding, effective population size and genetic drift for homogeneous populations. Then, these concepts will be generalized to spatially structured populations together with a summary of paper II, and lastly to age structured populations along with a summary of the other three papers.
2. Homogeneous populations

Consider a biallelic locus in a haploid population with constant size \( 2N \). Let \( A \) and \( a \) be the two alleles present at that locus and let \( p_t \) denote the frequency of the \( A \)-allele at time \( t \). We assume that both alleles are selectively neutral, so that an individual’s fitness does not depend upon its allele. Time is counted discretely and individuals only survive one time step, i.e. the population has discrete or non-overlapping generations.

2.1 Reproduction

The genetic processes in the population, e.g. inbreeding, genetic drift and coalescence will generally depend on the population’s reproduction model. In order to formalize reproduction at time \( t \), let \( Y_{t1}, \ldots, Y_{t,2N} \) be the number of offspring of the \( 2N \) genes, conditioned that \( \sum_{i=1}^{2N} Y_{ti} = 2N \) to keep the population size constant. The so called Cannings model of a homogeneous population [Cannings, 1974] treat all \( Y_{ti} \) symmetrically as exchangeable random variables, i.e. the joint distribution of \( \{Y_{ti}\}_{i=1}^{2N} \) does not depend on the order of the indices. One model for reproduction is of special interest, when all offspring choose their parents randomly with replacement and independently of each other. Then, the number of offspring for a gene

\[
Y_{t1} \sim \text{Bin}(2N, \frac{1}{2N})
\]

follows a binomial distribution. This model is called the Wright-Fisher model and serves as an important reference model in population genetics. In Figure 2.1 the Wright-Fisher model is illustrated with \( 2N = 10 \) genes during 8 generations.

By varying the reproduction scenario, an endless number of models can be created that deviate from the Wright-Fisher model. For example in the breeder model a number \( 2N_e \leq 2N \) of the genes in the parental generation are chosen randomly without replacement as breeders and the offspring choose their parents randomly with replacement among the breeders.
Figure 2.1: A realization of 8 generations of the Wright-Fisher process for $2N = 10$ genes. The genes are sorted in each generation so that it is easier to follow the ancestral lineages.
2.2 Inbreeding, genetic drift and coalescence

Breeding of individuals that are closely related is called inbreeding. In a population of haploid individuals, inbreeding can be quantified by means of the gene identity. Suppose two genes are picked randomly, with or without replacement, from the population at time $t$ where a gene is a part of a chromosome containing two or more alleles. The gene identity, $F_t$, at time $t$ is the probability that these genes have the same allele. By saying that two genes have the same allele we either mean that they are identical by descent (IBD), i.e. that they are descendants from the same ancestor, or that they are identical by state (IBS) which mean that they have the same variant. The probability that the two alleles are different is called gene diversity and is denoted by $H_t = 1 - F_t$. Strictly speaking, only IBD-sharing corresponds to inbreeding, although the recursions presented below are valid for IBS-sharing as well.

We will generally be interested in predictions of the gene identity and the gene diversity. Let $t = 0$ denote present time and $t > 0$ the future. Then, write

$$f_t = E_0(F_t),$$

as the expected gene identity at time $t$ conditioned on the present. Similarly, let

$$h_t = E_0(H_t),$$

be the expected gene diversity at time $t$ conditioned on the present.

The expected gene diversity can be described by means of a recursion. First, let $p$ be the probability that two randomly chosen distinct genes at time $t$, descend from the same parent in generation $t - 1$. For a population following the Wright-Fisher model this probability $p = 1/(2N)$, whereas $p = 1/(2N_e)$ if it follows the breeder model. The probability that two genes, chosen without replacement at time $t$, are different follows the recursion

$$h_t = (1 - p)h_{t-1}. \tag{2.2}$$

For a homogeneous population, the recursion (2.2) also holds if genes are chosen with replacement. Noticing that $f_t = 1 - h_t$, the expected gene identities can be expressed by the recursion

$$f_t = p + (1 - p)f_{t-1}.$$

The probability $p$ above is called the coalescence probability and for a general Cannings model it equals

$$p = \frac{E[Y_{t1}(Y_{t1} - 1)]}{2N - 1}. \tag{2.3}$$
Figure 2.2: The ancestral history of five genes randomly chosen from the extant genes in Figure 2.2 with \( \tau_5 = 2, \tau_4 = 1, \tau_3 = 3 \) and \( \tau_2 = 1 \), where \( \tau_i \) is the number of generations with \( k \) ancestral lineages. The time to the most recent common ancestor is \( \sum_{k=2}^{5} \tau_k = 7 \).

Suppose we sample \( k = 5 \) of the genes in the present generation in Figure 2.1 and track their ancestors backwards in time. Then their genealogical history forms a tree as shown in Figure 2.2. If we measure time in units of \( 2N \) and let \( N \to \infty \), the amount of time \( T_k = \tau_k / (2N) \) with \( k \) lineages will have an exponential distribution with mean \( 2/(k(k-1)) \). The resulting continuous time Markov process, which at each time point records the number of ancestral lineages, is called the Kingman coalescent and the tree in Figure 2.3 is called a coalescent tree [Kingman 1982, Durrett 2008].

We have assumed that the alleles are selectively neutral, hence the expected number of type A-alleles in the next generation will equal the number of alleles today. However, due to the stochastic nature of the process, the number of type A-alleles will vary over time and eventually the population will be fixated for one of the two alleles. Genetic drift describes these stochastic fluctuations of the allele frequency \( p_t \) over time.
Figure 2.3: The coalescent tree for the five chosen genes with $T_5 = 0.4$, $T_4 = 0.2$, $T_3 = 0.6$ and $T_2 = 0.2$, where $T_i = \tau_k/(2N)$ is the amount of time with $k$ ancestral lineages. The time to the most recent common ancestor is $\sum_{k=2}^{5} T_k = 1.4$. 
The genetic drift in a population depends on the reproduction scenario. For the Wright-Fisher model, the conditional distribution of the allele frequency follows a binomial distribution and for the conditional variance we have that

\[ \text{Var}(p_{t+1}|p_t) = \frac{p_t(1-p_t)}{2N}. \] (2.4)

For the breeders model with \(2N_e\) breeders we have the corresponding conditional variance

\[ \text{Var}(p_{t+1}|p_t) = \frac{p_t(1-p_t)}{2N_e}. \] (2.5)

### 2.3 Effective population size

We have previously seen that two populations of the same size \(N\) behave differently depending on their reproductive model. The population size alone is therefore not enough for predicting e.g. inbreeding or genetic drift. Another important concept in population genetics is the effective population size \((N_e)\). It is the size of an idealized population, i.e. a population that follows the Wright-Fisher model, which has the same expected change of some genetic characteristic as the studied population. The effective population size can be used to compare different population models and analyze how various assumptions affect the population. It is used in practice e.g. for determining desirable sizes of endangered species for their long and short term conservation.

The concept of the effective population size was first introduced by Wright (1931), and have since that evolved into many different versions depending on the genetic characteristic under consideration.

#### 2.3.1 Inbreeding effective population size

Wright (1931) defined the inbreeding effective population size \((N_{eI})\), which is the size of a Wright-Fisher population that has the same decrease of gene diversity over one time step as the studied population. Combining (2.1), (2.2) and (2.3) we see that the gene diversity

\[ h_{t+1} = \left(1 - \frac{1}{2N}\right) h_t \] (2.6)
is reduced by a factor $1/(2N)$ each time step for a Wright-Fisher model. By solving (2.6) for $N$, we obtain

$$N_{eI} = \frac{1}{2 \left(1 - \frac{h_{t+1}}{h_t}\right)},$$

as the inbreeding effective population size where $h_t$ is the gene diversity for the studied population.

For the breeders model, the coalescent probability of two genes is

$$p = \frac{1}{2N_e},$$

and from (2.2) we find that the inbreeding effective population size is

$$N_{eI} = \frac{1}{2 \left(1 - \left(1 - \frac{1}{2N_e}\right)\right)} = N_e.$$

### 2.3.2 Variance effective population size

The variance effective population size ($N_{eV}$) was introduced by [Crow (1954)](Crow1954). It is defined as the size of a Wright-Fisher population that has equal variance of allele frequency change as the studied population. Let $\text{Var}(p_{t+1} | p_t)$ be the variance of allele frequency change for the model under consideration. Equating $\text{Var}(p_{t+1} | p_t)$ with the variance (2.4) under the Wright-Fisher model and solving for $N$ we obtain

$$N_{eV} = \frac{p_t(1 - p_t)}{2 \text{Var}(p_{t+1} | p_t)},$$

and for the breeders model it follows from (2.5) that

$$N_{eV} = N_e.$$

### 2.3.3 Eigenvalue effective population size

Starting with the gene diversity $h_0$ where $t = 0$ denotes present time, we see that (2.6) can be rewritten as

$$h_t^{1/t} = h_0^{1/t} \left(1 - \frac{1}{2N}\right).$$

The long term multiplicative rate at which the gene diversity decreases is obtained from the limit

$$\lim_{t \to \infty} h_t^{1/t} = 1 - \frac{1}{2N}. \quad (2.7)$$
The eigenvalue effective population size

\[
N_{eE} = \frac{1}{2(1 - \lim_{t \to \infty} h_t^{1/t})}
\]  

is defined by solving \(2.7\) for \(N\). In the next section, we will see that the long term rate at which the gene diversity decreases in structured populations can be described by the largest eigenvalue of a certain matrix, which motivates the name of \(N_{eE}\).

2.3.4 Coalescent effective population size

Another version of the effective population size is the coalescent effective population size \(N_{eC}\) (Nordborg & Krone, 2002; Sjödin et al., 2005). If the genealogy follows the Kingman coalescent asymptotically when time is scaled by \(2N_e\), then

\[N_{eC} = N_e\]

is the coalescent effective population size.

This effective size shares the property of \(N_{eE}\) as being an asymptotic single value which summarizes the population. On the other hand, it is easy to see that we could extend the definitions of \(N_{eI}\) and \(N_{eV}\) based on gene diversity change or genetic drift over \(t\) generations. For the population scenarios above, this dependence is of limited interest since both the rate at which the gene diversity decreases and the genetic drift are constant over time. However, for structured populations, time dependent versions of \(N_{eI}\) and \(N_{eV}\) are extensions that provide more information than one single effective size.
Almost all real populations are subdivided into subgroups in which some key features are shared. It could be fishes in a system of lakes, a species with two sexes or an organism with multiple age classes where each lake, sex or age class can be considered as a subgroup. Such structure in a population complicates all of the concepts in the previous chapter, and ignoring it can lead to seriously biased results.

3.1 Spatially structured populations

Consider a haploid population divided into $s$ homogeneous subpopulations of constant size, where subpopulation $i$ consists of $2N_i$ genes. It is no longer enough with a single number $F_t$ to describe the gene identity in the population since it can differ between the subpopulations. Let $F_{tti}$ be the gene identity in subpopulation $i$ at time $t$, with $E_0(F_{tti}) = f_{tti}$ its expected value conditionally on the present. Similarly, let $F_{tij}$ be the gene identity between subpopulation $i$ and $j$, i.e. the probability that two randomly chosen genes from $i$ and $j$ are identical, with expected value $E_0(F_{tij}) = f_{tij}$. Gene diversities $H_{tij} = 1 - F_{tij}$, and the corresponding expectations $h_{tij} = 1 - f_{tij}$ are defined analogously. Let $\mathbf{h}_t = (h_{tij}, 1 \leq i, j \leq s)'$ be a column vector of size $s^2$ with all gene diversities at time $t$, where $'$ denotes the transpose of a vector. In order to generalize recursion (2.2) to a spatially structured population, the coalescent probability for two genes from subpopulation $i$ and $j$ needs to be specified. These probabilities depend on the migration scheme in the population and the local population sizes.

In order to describe migration between subpopulations, matrix migration models are commonly used (Caswell, 2001). Let $M_{iji}$ be the expected fraction of offspring of a gene in subpopulation $j$ at time $t - 1$ that lives in subpopulation $i$ at time $t$. The matrix $\mathbf{M} = \{M_{iji}\}_{i,j=1}^s$ is called the forward migration matrix and we assume that its elements are independent of time. Looking backwards in time, let $\mathbf{B}$ be the backward migration matrix where element $B_{tij}$ is the probability that a gene in subpopulation $i$ at time $t$ originates from subpopulation $j$ at time $t - 1$. The elements of $\mathbf{B}$ can be found
from those of $M$ through

$$B_{ij} = \frac{N_j M_{ji}}{\sum_{k=1}^s N_k M_{ki}}.$$ 

The local effective population size $N_{ei}$ of subpopulation $i$ corresponds to the inbreeding effective size it would have if the subpopulation was an isolated homogeneous population.

If two randomly chosen genes are drawn with replacement then the recursion (2.2) can be generalized to

$$h_t = A h_{t-1}$$

for a certain matrix $A$ which describes the coalescence probabilities and depends on $B$ and the local effective and census population sizes, see for instance Malécot (1951) and Hössjer & Ryman (2014). If $B$ is irreducible and aperiodic, the largest eigenvalue $\lambda = \lambda_A$ of $A$, describes the long term rate at which gene identities approaches 1. It can be shown that (3.1) leads to

$$\lim_{t \to \infty} \frac{h_i^t}{t} = \lambda_A$$

for any $i = 1, \ldots, s$, and in view of (2.8) this gives the eigenvalue effective population size

$$N_{eE} = \frac{1}{2(1 - \lambda)}. 

(3.2)$$

If the two genes are drawn without replacement, a recursion similar to (3.1) holds with a slightly different matrix $D$ instead of $A$ with $\lambda_D = \lambda_A = \lambda$.

If $B$ is irreducible and aperiodic, it follows from theory of discrete time Markov chains that $B$ has a stationary distribution $\pi = (\pi_1, \ldots, \pi_s)'$. If we assume that $N_i = N_{ei}$ and $\sum_{i=1}^s N_i \to \infty$ such that $N_i / \sum_{j=1}^s N_j \to \kappa_i$ then, if $\sum_{i=1}^s N_i$ is large,

$$N_{eC} = \frac{\sum_{i=1}^s N_i}{\sum_{i=1}^s \frac{\pi_i}{\kappa_i}} 

(3.3)$$

is the coalescent effective population size (Notohara 1990, Nordborg & Krone 2002, Durrett 2008). It can be shown that (3.2) and (3.3) are asymptotically equivalent for large populations (Hössjer 2014).

One of the most studied matrix migration model is the island model (Wright 1949), sketched for the case $s = 5$ in Figure 3.1a. In this model, the subpopulation sizes $N_i = N_{ei} = N$ are equal for all $i$, and migration is described by $M_{ii} = 1 - m$ and $M_{ij} = m/(s-1)$ for $i \neq j$ where $m$ is the overall migration
rate. Consider an island model with an infinite number of subpopulations and a biallelic locus. Then,

\[ F_{ST} = \frac{\text{Var}(p)}{\bar{p}(1 - \bar{p})} \]

is a measure of genetic differentiation between the subpopulations of a structured population introduced by Wright (1943; 1949), where \( \text{Var}(p) \) is the variance of the allele frequency in subpopulations and \( \bar{p} \) their average allele frequency. This measure of differentiation can also be expressed as

\[ F_{ST} = \frac{F_T - F_S}{1 - F_T} \]

where \( F_S \) is the average gene identity within the subpopulations and \( F_T \) the average gene identity in the whole population. Nei (1973; 1975) generalized this concept and introduced

\[ G_{ST} = \frac{F_T - F_S}{1 - F_T} \]

for multiallelic loci.

Other well studied matrix migration models include the stepping stone models, whose circular version is shown in Figure 3.1b. Most of the well studied matrix migration models posses some kind of symmetry which make the analysis easier. Assume for example that, at time \( t \), the gene identities of all subpopulations in an island model are equal and also that the gene identities between all pairs of different subpopulations are equal. Then, a 2-dimensional vector \( h_t = (h_{11}, h_{12})' \) consisting of gene diversities within and between subpopulations is enough to describe the gene identities of the whole population (Nei 1975; Li 1976). In the second paper, the aim was to develop a general framework for more complex migration schemes, such as the ones in Figure 3.1c-d.

3.1.1 Summary of paper II

The Swedish wolf population is highly inbred and questions about the conservation genetic situation of this population motivated the second paper. There we present a framework for modeling migration of a general class of subdivided populations in order to develop theoretical means for assessing effective population sizes and genetic differentiation. It lead to a generalization of the recursion (3.1) including mutations as well as letting the demographics, census population sizes and effective population sizes vary in
Figure 3.1: Illustration of the island model (a), the circular stepping stone model (b), a model with a more complex migration scheme (c) and an age structured population (d). Migration occurs according to the arrows. There is also migration within the subpopulations, except for (d), where migration only occurs according to the arrows. All four populations (a)-(d) have $s = 5$ subpopulations, represented as circles, with sizes proportional to their local census sizes.
an arbitrary way. This setup allows for many types of population scenarios including global and local bottlenecks, subpopulation extinction and recolonization, cyclic changes and exponential growth.

Using the time profile of the predicted gene identities, the gene differentiation and inbreeding population size can be calculated for each time step using a general weighting scheme for the subpopulations. In the paper we argue that an asymptotic effective population size such as the eigenvalue effective population size is not enough. It should be accompanied with e.g. the time dynamics $N_{eI}([0, t])$ of the inbreeding effective population size over time intervals $[0, t]$ of various lengths.

For migration models with symmetries, such as the island model, we introduce a method for reducing the state space of the model which is important when the number of subpopulations is large. In paper A, which is not included in the thesis, this general framework has been extended to accommodate e.g. diploid organisms with separate sexes and various reproduction scenarios, such as selfing and mating.

### 3.2 Age-structured populations

Consider a population of constant size $2N$ that differs from the Wright-Fisher model in that only one gene dies and the other $2N - 1$ genes survive at each time step. In order to keep the population size constant, one new gene must be born each time step. If that gene chooses its parent randomly from the surviving genes we obtain the Moran model \cite{Moran1958}. This modification of the Wright-Fisher model introduces overlapping generations. Moran \cite{Moran1958} showed that asymptotically, the gene diversity decreases at a rate $1 - 2/(2N)^2$ per time step. Let $T$ be the expected age of a parent of a newborn, also called the generation time. For the Moran model, $T = 2N$, and

$$
\left(1 - \frac{2}{(2N)^2}\right)^{2N} \approx 1 - \frac{2}{2N}
$$

is the rate at which the gene diversity decreases per generation, with an approximation that is accurate for large values of $N$. Hence, the inbreeding effective population size per generation

$$
N_{eI} \approx \frac{N}{2}
$$

is approximately half of the census population size.

In the Moran model, the generation time increases with the population size, which is an unrealistic feature of the population. Another way of modeling
an age structured population is to split the population into \( n \) homogeneous age classes where age class \( i \) consists of \( 2N_i \) haploid individuals at time \( t \). Let \( Y_{th} \) be the independent and identically distributed random variables for \( h = 1, \ldots, 2N_i \) describing the number of progeny gene \( h \) in age class \( i \) has at time \( t \) with expected value \( E(Y_{th}) = b_i \) and variance \( \text{Var}(Y_{th}) = \sigma_i^2 \). Survival from \( t \) to \( t + 1 \) of gene \( h \) in age class \( i \) can be described by the indicator variable \( I_{th} \) with \( E(I_{th}) = s_i \). Let \( N_t = (N_{t0}, \ldots, N_{tn-1})' \) be a column vector with the number of genes in each age class. Then, the recursion

\[
N_{t+1} = G_t N_t
\]
describes the population vector over time, in terms of the projection matrix

\[
G_t = \begin{pmatrix}
\frac{1}{2N_0} \sum_{h=1}^{2N_0} Y_{t0h} & \cdots & \frac{1}{2N_{t,n-1}} \sum_{h=1}^{2N_{t,n-1}} Y_{t,n-1,h} \\
\frac{1}{2N_0} \sum_{h=1}^{2N_0} I_{t0h} & 0 & \cdots & 0 \\
0 & \ddots & \ddots & \ddots \\
0 & \cdots & \frac{1}{2N_{t,n-2}} \sum_{h=1}^{2N_{t,n-2}} I_{t,n-2,h} & 0 \\
\end{pmatrix}
\]

Let

\[
E(G_t) = g = \begin{pmatrix}
b_0 & b_1 & \cdots & b_{n-1} \\
s_0 & 0 & \cdots & 0 \\
0 & \ddots & \ddots & \ddots \\
0 & \cdots & s_{n-2} & 0 \\
\end{pmatrix}
\]

be the expected projection matrix, the so called Leslie matrix [Leslie 1945]. It is analogous to the forward migration matrix \( M \) for a geographically subdivided population with survival rates \( s_i \) and birth rates \( b_i \) as migration rates. The largest eigenvalue \( \lambda \) of \( g \), which is positive and unique according to Perron-Frobenius Theorem, is the multiplicative growth rate of the population with \( \lambda < 1, \lambda = 1 \) and \( \lambda > 1 \) for a population whose expected size decreases, is constant or increases. The right eigenvector, \( u = (u_0, \ldots, u_{n-1})' \), corresponding to \( \lambda \), normalized such that \( \sum_{i=0}^{n-1} u_i = 1 \), is the stable age distribution of the population. The left eigenvector \( v = (v_0, \ldots, v_{n-1}) \) corresponding to \( \lambda \) is proportional to the age specific reproductive values [Fisher 1958]. If \( v \) is normalized such that \( \sum_{i=0}^{n-1} v_i u_i = 1 \), then the average number of descendants to a gene in age class \( i \) after \( \tau \) time steps is \( \lambda^\tau v_i \) when \( \tau \) is large.

Expressions for \( N_{eV} \) and \( N_{eI} \) in age structured populations have previously been derived by [Felsenstein 1971] and [Hill 1972 1979] for populations with
constant age class sizes. Engen et al. (2005) derived an expression for \( N_{eV} \) for populations with a fluctuating population size using diffusion processes.

3.2.1 Summary of paper I

In the first paper we consider a population of haploid individuals with overlapping generations where the age classes sizes vary stochastically. The main goal was to derive an expression for the variance effective population size \( (N_{eV}) \) over time as a function of weights assigned to each of the age classes.

We use matrix population models (cf. Caswell, 2001) to model the time dynamics of the population as well as a specific allele of selectively neutral gene at a biallelic locus. Let \( \mathbf{c} = (c_0, \ldots, c_{n-1}) \) be a row vector of age specific arbitrary weights, normalized so that \( \mathbf{c}^T \mathbf{1} = 1 \). We define the variance effective population size as

\[
N_{eV}^c(\tau) \approx \frac{\tau p_t^c (1 - p_t^c)}{TE \left[ (p_{t+\tau}^c - p_t^c)^2 | N_t^c, p_t^c \right]}
\]

where \( p_{ti} \) is the frequency of the specified allele in age class \( i \) at time \( t \), \( p_t^c = \sum_{i=0}^{n-1} p_{ti} N_t c_i / N_t^c \) and \( N_t^c = \sum_{i=0}^{n-1} N_{ti} c_i \) are the weighted allele frequency and population size over all age classes at time \( t \), and \( \tau \) is the time between the measurements.

Previous expressions for \( N_{eV} \) in age structured populations all use the reproductive values \( \mathbf{v} \) as weights. This weighting scheme eliminates initial fluctuations of \( N_{eV} \) due to age structure and predicts the long term genetic drift for a population where the population size is expected to be constant.

In the paper we examine the influence of different weighting schemes on \( N_{eV} \) for different time intervals and present a lower bound on the number of generations that are needed between the measurements in order to obtain a certain relative discrepancy between \( N_{eV}^c(\tau) \) and \( N_{eV}^v(\tau) \).

This paper inspired further investigation on the estimation procedure of \( N_{eV} \) in age structured populations.

3.2.2 Summary of paper III

A common method for estimating \( N_{eV} \) is the temporal method (Krimbas & Tsakas, 1971). By comparing alleles from two samples from the population that are separated by a time \( \tau \), one can estimate the genetic drift in the population. Neglecting the fact that a population has overlapping generations can result in highly biased results. In paper III we extend the method of
Jorde & Ryman (1995), which uses the demographic parameters to calculate a correction factor in order to obtain unbiased estimates of $N_eV$.

The method utilizes theory from paper I and II in order to calculate a correction factor based on the expected genetic drift for the chosen weighting scheme. The correction factor is defined as a function of the age specific offspring and survival distribution and it is possible to include correlation between survival and reproductive success. Knowing these parameters, it is possible to calculate the ratio $N_eV/N$ using methods from paper I. Combining this information with the estimate of $N_eV$ gives a joint estimator of the variance effective population size and the census population size.

We also derive confidence intervals for the estimates as well as an optimal way of weighting age classes and loci. In paper I we found that using the reproductive values as weights, the variance effective population size was constant over time. Using these weights when estimating $N_eV$ with the method above, the correction constant equals 1. Hence, we do not need to specify the variance of the offspring distribution or potential correlation between survival and reproductive success in order to obtain unbiased estimates. The method still requires knowledge of the expected number of offspring and survival probability in each age class though in order to calculate the weights.

3.2.3 Summary of paper IV

Simulations are a powerful tool in order to assess the performance of different methods. A fast and reliable simulation procedure is essential when carrying out these performance tests. In the fourth paper, we present such a simulation procedure for age structured populations that is used in the simulations of paper III.

A schematic overview of the simulation process is shown in Figure 3.2 and the idea behind the method is as follows. We start by choosing a population size $\tilde{N}_0 = \sum_{i=0}^{n-1} N_0 i v_i = N_0^v$ for the present time $t = 0$. This is not the actual population size but the size where age classes are weighted according to their reproductive values $v_i$. Given the initial population size, we use a normal approximation to simulate the age class sizes at time $t = 0$. These sizes can in turn be used to simulate the age class sizes in the following time steps.

We use a similar approach for simulation of the number of haploid individuals $Z_{tia}$ with allele $a$ and the allele frequencies

$$p_{tia} = \frac{Z_{tia}}{N_{tii}}$$
of age class $i$ at time $t$ at a locus with $A$ alleles $a = 1, \ldots, A$. However, in the simulation we also condition on the age class composition of the total population. This simulation method eliminates the need of a burn in period which makes it easy to repeat under identical initial conditions.

The sums $\tilde{N}_t = \sum_{i=0}^{n-1} N_{ti} v_i$ and $\tilde{Z}_{ta} = \sum_{i=0}^{n-1} Z_{ti} v_i$ denote the total number of individuals at time $t$, weighted with their age specific reproductive values $v_i$, in the population and with allele $a$ respectively, whereas $N_t = (N_{t0}, \ldots, N_{tn-1})'$ and $Z_{ta} = (Z_{t0a}, \ldots, Z_{tn-1,a})'$ denote the age composition of the total population and the individuals with allele $a$.

**Figure 3.2:** 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.
4. Outlook

We have studied quantities that can be expressed in terms of pairs of alleles, such as gene identity and gene diversity. A number of other quantities of interest are more general functions of the allele frequency spectrum (Durrett, 2008, Ch. 2). This includes probabilities of identity by descent or state configurations of more than two genes, the total number of alleles or the number of rare alleles. In conservation biology, it is well known that, during a population bottleneck, many of the rare alleles are lost (Maruyama & Fuerst, 1985; Cornuet & Luikart, 1996; Luikart et al., 1998; Garza & Williamson, 2001). Hence, it would be of interest to include the allele frequency spectrum into the models and study the effect of various population scenarios.

In the first and second paper, we only consider genetic information from a single locus. By extending the models to allow for multiple loci, expressions for linkage disequilibrium measures of effective sizes (Hill, 1981; Waples & England, 2011; Waples et al., 2014) could be developed. It would also be of interest to generalize the estimation procedure in paper III to allow for loci in linkage disequilibrium.

Although many real world species are dioecious, the models in this thesis all concern monoecious species. In paper A the methods from paper II have been generalized to dioecious species. These methods could also be used to generalize the correction constant in paper III in order to develop an approximately unbiased estimator of variance effective size for structured dioecious populations.

In this thesis we have focused on how migration affects genetic drift. Wakeley & Sargsyan (2009) included mutations when defining the coalescence effective size $N_{ec}$. It would also be of interest to allow for e.g. mutation and selection for the various other types of effective population size.

When the work with this thesis started, the practical motivation concerned harvesting strategies of a moose population. Over the time, other motivating examples have been added, such as conservation strategies of the highly inbred Swedish wolf population, with various migration scenarios of neighboring populations and management of Baltic Sea species metapopulations. In order to study these problems and to make the methods more
accessible to conservation biologists and other researchers, we are in the process of implementing some of the methods from paper II and A in an R-based software called Genetic Exploration of Structured Populations (GESP).
Sammanfattning

Då nära besläktade individer får avkommor kallas det inavel. Att beskriva hur inavelsgraden i en strukturerad population förändras utgör en central del av populationsgenetik och bevarandebiologi. Genom matematisk modellering kan man undersöka hur olika egenskaper hos populationen påverkar dess genetiska sammansättning. Eftersom alla sådana modeller är förenklingar av verkligheten är det viktigt att undersöka effekten av dessa förenklingar. Jämförelser med mer avancerade modeller kan exempelvis ge en ökad förståelse för hur olika strukturer i en population påverkar dess genetiska sammansättning. Denna avhandling, som består av fyra artiklar, behandlar olika aspekter av inavel, effektiva populationsstorlekar och genetisk differentiering i strukturerade populationer.

Ett viktigt begrepp inom populationsgenetiken är relaterat till hur inavelsgraden förändras över tid, den så kallade effektiva populationsstorleken ($N_e$). Det finns flera varianter av effektiva storlekar och i den första artikeln presenterar vi en metod för att beräkna varianseffektiva populationsstorleken ($N_{eV}$) i en åldersstrukturerad population. Vi studerar hur $N_{eV}$ beror på hur åldersklasserna viktas samt effekten av längden på det tidsintervall man är intresserad av. Dessa metoder används senare i den tredje artikeln då vi presenterar en metod för att skatta $N_{eV}$ med hjälp av ett urval från populationen. Denna skattningsmetod ger approximativt väntevärdesriktiga skattningar om vi kan tillföra demografisk information om populationen.


I avhandlingens andra artikel presenterar vi metoder för att analysera effektiva populationsstorlekar och genetisk differentiering i spatialt strukturerade populationer med generella migrationsmodeller. Vi utvecklar metoder för prediktion av inavel i populationen med hjälp av rekursionsformler. Dessa prediktioner kan i sin tur användas för att exempelvis beräkna olika effektiva storlekar inom subpopulationerna och totalpopulationen samt populationens genetiska differentiering.
References


Jag är otroligt tacksam för att ha fått möjligheten att grotta ner mig i ett ämne som jag tycker är både roligt och intressant och ni är många som har förgyllt min tid som doktorand:

Först och främst vill jag tacka min handledare Ola Hössjer. Ditt tålamod och din förmåga att förklara komplicerade saker på ett förståeligt vis har varit ovärderlig. Jag är väldigt glad att ha fått jobba med dig.

Nils Ryman och Linda Laikre för ett gott samarbete, ni har verkligen gett modellerna en mening.

Keith Humphreys för tiden på Karolinska Institutet, även om vår modell till-slut bara passade någon utomjordisk varelse.


Jag vill även tacka min familj som alltid har stöttat mig och mina vänner som ibland har tvingat ut mig i den "verkliga" världen. Slutligen, tack Frida för ditt stöd, din kärlek och all humor.
Part II

Papers